11-13-2023

Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam

Peter Mooney
National University of Ireland, Maynooth

Wencong Cui
Florida International University

Boyuan Guan
Florida International University

Levente Juhasz
Florida International University, ljuhasz@fiu.edu

Recommended Citation

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. 2023. Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geo- graphic Information Systems (GIS) Exam. In 6th ACM SIGSPATIAL Inter- national Workshop on AI for Geographic Knowledge Discovery (GeoAI '23), November 13, 2023, Hamburg, Germany. ACM, New York, NY, USA. https://doi.org/10.1145/3615886.3627745

Towards Understanding the Geospatial Skills of ChatGPT

Taking a Geographic Information Systems (GIS) Exam

Peter Mooney*
peter.mooney@mu.ie
Department of Computer Science, Maynooth University
Maynooth, Co. Kildare, Ireland

Wencong Cui
wecui@fiu.edu
GIS Center, Florida International University
Miami, FL, USA

Boyuan Guan
bguan@fiu.edu
GIS Center, Florida International University
Miami, FL, USA

Levente Juhász*
ljuhasz@fiu.edu
GIS Center, Florida International University
Miami, FL, USA

ABSTRACT

This paper examines the performance of ChatGPT, a large language model (LLM), in a geographic information systems (GIS) exam. As LLMs like ChatGPT become increasingly prevalent in various domains, including education, it is important to understand their capabilities and limitations in specialized subject areas such as GIS. Human learning of spatial concepts significantly differs from LLM training methodologies. Therefore, this study aims to assess ChatGPT's performance and ability to grasp geospatial concepts by challenging it with a real GIS exam. By analyzing ChatGPT's responses and evaluating its understanding of GIS principles, we gain insights into the potential applications and challenges of LLMs in spatially-oriented fields. We conduct our evaluation with two models, GPT-3.5 and GPT-4, to understand whether general improvements of an LLM translate to improvements in answering questions related to the spatial domain. We find that both GPT variants can pass a balanced, introductory GIS exam, scoring 63.3% (GPT-3.5) and 88.3% (GPT-4), which correspond to grades D and B+ respectively in standard US letter grading scale. In addition, we also identify specific questions and topics where the LLMs struggle to grasp spatial concepts, highlighting the challenges in teaching such topics to these models. Finally, we assess ChatGPT's performance in specific aspects of GIS, including spatial analysis, basic concepts of mapping, and data management. This granular analysis provides further insights into the strengths and weaknesses of ChatGPT's GIS literacy. This research contributes to the ongoing dialogue on the integration of AI models in education and can provide guidance for educators, researchers, and practitioners seeking to leverage LLMs in GIS. By focusing on specific questions or concepts that pose difficulties for the LLM, this study addresses the nuances of teaching spatial concepts to AI models and offers potential avenues for improvement in spatial literacy within future iterations of LLMs.

CCS CONCEPTS

• **Social and professional topics**; • **Human-centered computing** → **Interaction paradigms**; **Natural language interfaces**; • **Applied computing** → **Education**;

KEYWORDS

GIS, education, ChatGPT, Large Language Models, Generative AI, geospatial, foundation model

ACM Reference Format:
Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. 2023. Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam. In *6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3615886.3627745

1 INTRODUCTION AND MOTIVATION

ChatGPT, from OpenAI[1], needs little introduction at this point in time. ChatGPT is a publicly-available chatbot interface for the GPT family of large language model (LLM) artificial intelligence (AI) systems that generates human-like text in response to user text input. When presented with a query, ChatGPT will automatically generate a response, which is based on a massive corpus of data sources, often without further input from the user. ChatGPT (Generative Pretrained Transformer) was developed with a technique called Reinforcement Learning from Human Feedback (RLHF) [21] to train the language model, enabling it to be very conversational. ChatGPT is able to answer follow up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests due to this conversational format [23]. It integrates various abilities of natural language processing, including question answering, storytelling, logic reasoning, code debugging, machine translation, and so on. Foundation models, such as ChatGPT, are *pre-trained* and then *adapted* via fine-tuning learning strategies[8] and are subsequently deployed on a wide range of knowledge domains. This mitigates the need for task-specific training data[36, 42]. One must temper expectations for these models around domain-specific knowledge - in our case GIS and geospatial understanding. The *multimodal* nature (images, text, vector and raster data, semantic information) of GIS "hinders a straightforward application of existing (foundation)

*These authors contributed equally to this research.

GeoAI '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0348-5/23/11.
https://doi.org/10.1145/3615886.3627745

[1] https://openai.com/blog/chatgpt

models" across different geographic tasks [36]. Here in this work we are particularly interested in testing the ability of ChatGPT when it comes to GIS and spatial literacy. Spatial literacy, as argued by many authors, is as important as mathematical literacy (numeracy) and classic literacy—the ability to read and write [4]. King [28] suggests that "spatial literacy is clearly a highly important skill for students of geography, earth and environmental sciences to master". Spatial ability is a cognitive factor that has been linked to high performance in science and mathematics [34]. Spatial literacy is now a component of many professions and careers and interest in it from the research community is driven by a desire to understand the role that spatial literacy plays in implementation of geospatial technologies such as geographic information systems (GIS) [27]. One is said to be *spatially literate* if they have developed appropriate levels of knowledge and skills that enable them to think, act and reason about the world in spatial ways [32].

1.1 ChatGPT and Examinations

Already, at the time of writing in Summer 2023, researchers have published various accounts of experiments where ChatGPT was tasked with passing a particular examination. There is a fascinating range of examinations recorded. We provide just a flavour here. In Fijačko et al. [18], the authors tested the accuracy of ChatGPT's answers to the American Heart Association (AHA) Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) exams. ChatGPT did not reach the passing threshold for any of the exams. The work by Gilson et al. [20] aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination exams. The LLM achieved the equivalent of a passing score for a third-year medical student. In a short meta review, Newton [40] concluded that ChatGPT "fails to meet the passing grade on almost every MCQ exam that it is tested against, and performs significantly worse than the average human student". The concern for Newton is that despite this somewhat modest performance the use of ChatGPT and other LLM poses difficulties for the integrity of MCQ-based assessments, particularly those administered online. In other examination scenarios, Strong et al. [45] investigated if ChatGPT was capable of consistently meeting the passing threshold on free-response, case-based clinical reasoning assessments. ChatGPT did pass overall which was "an unremarkable result". However, the authors cautioned that for the LLM "to achieve a passing performance in nearly half of the cases analyzed demonstrates the need to revise clinical reasoning assessments and incorporate AI-related topics into medical curricula and practice". Finally, Alberts et al. [1] considered a nuclear medicine board examination and the authors concluded that ChatGPT would be unlikely to pass the exam in a real examination scenario. However, "this could change in the future with better training for the model". Bhayana et al. [5] reports that despite no radiology-specific pretraining, ChatGPT *nearly* passed a radiology board–style examination without images where it struggled on questions with "higher-order thinking, calculation, and classification". Deshpande and Szefer [13] found that ChatGPT was capable of doing very well in introductory computer engineering assessments its inability (currently) to analyze an image for its contents limits the types of assignments and assessments it can be provided with.

1.2 Our contribution

To the best of our knowledge at the time of writing, no study has been reported in the literature to test the ability of ChatGPT to tackle exams or assessments in GIS. Our paper contributes to this current gap. As no standardised or widely used spatial literacy or GIS exam exists, we have used our own experience as teachers in this domain to create an exam for ChatGPT (see Section 3). We are careful to point out that GIS Certification exists in many countries [15, 22] with professional certification often being portfolio based, competency based, or curriculum based. In some cases, certification may include elements of all three approaches. Whilst we describe some limitations around the study setup and environment (see section 5.1) it is our belief that this work will make an important first contribution to the understanding around the spatial literacy and geospatial skills of ChatGPT. This research has two specific research questions and these are outlined as follows:

- RQ1: Undertake an assessment of ChatGPT's performance and-geospatial skills by configuring it to take a real GIS exam. Exam questions were adapted from the instructor resources of a popular textbook for introductory GIS courses called GIS Fundamentals [6]
- RQ2: Quantitatively assess ChatGPT's performance in this exam for specific aspects of GIS including spatial analysis, basic concepts of mapping, and data management. This analysis will provide us with the opportunity to gain insights into the potential applications and challenges of LLMs in spatially-oriented fields. It can also provide further insights into the strengths and weaknesses of ChatGPT's GIS literacy.

Overall, we find that both models pass our exam with letter grades D and B+ respectively using standard US letter grading scheme. GPT-3.5 performed similarly to random guessing of the answers whereas GPT-4's answers are significantly different from guessing. However, as somewhat expected, GPT-4 outperformed GPT-3.5 in all topics. For complex questions that required computation, both models answered incorrectly. Both models passed our examination despite LLMs struggling with complex geospatial semantics tasks such as geographic question answering [35, 37] because they are unable to carry out implicit spatial reasoning in a way that is "grounded in the real world" [36]. We were only able to provide the models with text-based questions which is not an accurate reflection of real-world spatial literacy or GIS assessments and examinations. As multimodal content (images, diagrams, vector layers) cannot be processed and understood by ChatGPT *yet*, in the context of a spatial literacy or GIS examination, instructors in this domain are somewhat shielded from the potentially negative impacts of LLMs on academic integrity. However, future enhancements of ChatGPT and other publicly available chatbots will include these abilities [48] and instructors will need to be prepared to adapt and engage with these changes. It will be necessary that students are educated on the use and limitations of ChatGPT and its potential impact on academic integrity [33].

The remainder of the paper is structured as follows. Section 2 provides an overview of some of the most relevant literature to this work. The experimental setup is described in section 3 with the topics, GIS skills and competencies tested in the exam also included in this section. The results of our experimentation are described in

section 4. Our paper closes in section 5 where we discuss some of the main results of the work along with the limitations of the current study (in section 5.1). Some ideas for future work are presented in section 5.2. For reference, a small section of the questions selected from Bolstad and Manson [6] are shown in Appendix section A at the end of the paper.

2 RELATED WORKS

Peer-reviewed published academic literature related to ChatGPT is, at the time of writing, beginning to appear. As ChatGPT was only publicly released in November 2022 we are in the early stages of the research process around the impact of this publicly available chatbot. Many academic and industry-based studies have been published and made available on pre-print servers and this constitutes a significant body of early research work. The literature on foundation models (FMs), LLMs, Deep Learning, and Natural Language Processing (NLP) is much more mature. Zhao et al. [50] have produced survey of the resources for developing LLMs and a discussion of issues for future directions. Their survey provides an up-to-date review of the literature on LLMs. This section of our paper gives a brief overview of relevant literature to the core work of the paper. Other literature is discussed and referenced appropriately through the remainder of the paper. Released to the public in November 2022, ChatGPT sent the conversation around AI and its role in society beyond "a tipping point" [38]. This was the first time ever that a very sophisticated LLM had become accessible to the general public and it was exceptionally easy to use. There was broad amazement at the capabilities of ChatGPT and as remarked by Mbakwe et al. [38] the public media predicted that the public emergence of ChatGPT would "change our mind about how we work, how we think, and what human creativity really is". Thorp [47] called it a "cultural sensation" and entertainment value aside Thorp [47] warns that "there are serious implications for generative AI programs like ChatGPT in science and academia".

In a work by Mbakwe et al. [38] the authors commented on how ChatGPT passed the United States Medical Licensing Examinations (USMLE). Upon reflection, this examination rewards "memorizing the components of a system rather than analyzing how it works, how it fails, how it was created, how it is maintained". While ChatGPT's success at passing this exam is noteworthy, the authors emphasized that the success of a LLM in passing this exam "demonstrates some of the shortcomings in how we train and evaluate medical students". Kung et al. [31] suggested that ChatGPT's success in USMLE "may potentially assist human learners in a medical education setting, as a prelude to future integration into clinical decision-making". Milano et al. [39] argue that "excitement about ChatGPT and other LLM tools foreshadows the huge political issue of who owns and sets the standards for education in the age of AI" and that it may be required that future LLMs could be specifically developed for usage within educational settings. This would go a long way to ensuring that they are "more transparent with regards their human and environmental costs". In Tang and Kejriwal [46] the authors report on the findings from their pilot study of selectively evaluating the cognitive abilities (decision making and spatial reasoning) of ChatGPT and DALL-E 2[2]. Both tests require text as input, but the authors hypothesized that spatial reasoning is more directly tested through the production of visual output, while decision-making is better tested through contextualized conversation-style text output and is hence more appropriate for a large language model like ChatGPT. The spatial reasoning was limited to tests on DALL-E. An example of a test would be for DALL-E to produce an image of *a person standing right in front of the Eiffel Tower* or *a person standing 5 miles from the Eiffel Tower*. The decision making tests for ChatGPT did not involve any geographical or spatial component. The authors concluded that it was easier to provide more quantitative estimates for DALL-E 2's spatial reasoning. In an editorial comment by Chang and Kidman [12] the authors emphasize that there are many things that ChatGPT cannot do in geographical and environmental education. The authors provide examples including "innovative pedagogies, especially for fieldwork, and learning beyond textual, visual and auditory modes, such as using Geographic Information Systems (GIS)".

As with the wider debate [16] in education around ChatGPT, Chang and Kidman [12] call for a "framework to consider how best we can use AI tools like ChatGPT to support good and meaningful geographical and environmental education". Educators can use technologies like ChatGPT to create more engaging and personalised learning experiences for students in all disciplines [20, 43] and not just within geography. By considering the opportunities and challenges posed by generative AI Chang and Kidman [12] concludes that "we must ensure that our efforts in geographical and environmental education do not degenerate".

3 EXPERIMENTAL OVERVIEW

Here we provide a brief overview of our experimental setup and the development of an examination to test ChatGPT. We describe the topics as well as the GIS skills and competencies tested in section 3.2.1 onward.

3.1 Exam Implementation

In the absence of a standardised or widely used GIS examination, we derived exam questions from a popular GIS textbook for introductory courses, GIS Fundamentals [6]. In addition to hundreds of study questions with solutions and explanations provided in the book, Bolstad and Manson supplied us with potential exam questions for each chapter that are released only as instructor resources. Crucially, since these are not available on the public facing web, we believe that there is little chance that they were included in the training set of GPT models. A balanced set of 60 questions were selected manually to simulate a real exam that covers most topics students of an introductory GIS course are expected to be familiar with. Both senior authors of this paper, Mooney and Juhász, teach spatial databases and introductory GIS courses at University Graduate level in Europe and the United States. We drew upon these experiences to extract a balanced and fair set of questions for five topics outlined as follows:

- Fundamental concepts of mapping and GIS (see Section 3.2.1)
- Data sources and tabular data (see Section 3.2.2)
- Spatial Analysis (see Section 3.2.3)
- Spatial statistics and interpolation (see Section 3.2.4)

[2]https://openai.com/dall-e-2

- Applied GIS workflows (see Section 3.2.5)

The workflow is shown in Figure 1. Answers to exam questions were collected from the ChatGPT models and stored in a CSV file for easier manipulation and analysis. At the time of writing, the rollout of GPT-4 was still not complete, and access to the API (application programming interface) could only be gained through a waitlist. For this reason, we opted to assess these models through their conversational, web interface, ChatGPT. The API's through which these models would be accessible allow to manipulate only a small number of parameters (e.g. model temperature, token length) that are not relevant to our case study. In addition, if students of an introductory GIS class were to use ChatGPT in an exam, it is almost certain that they would use a web-based interface as opposed to interacting with these models through an API. The process took place on April 19-20, 2023 using the version available on the web those days. Questions were manually pasted to the conversation, and ChatGPT was instructed to select the correct answer(s). Its responses were recorded and added to the exam dataset.

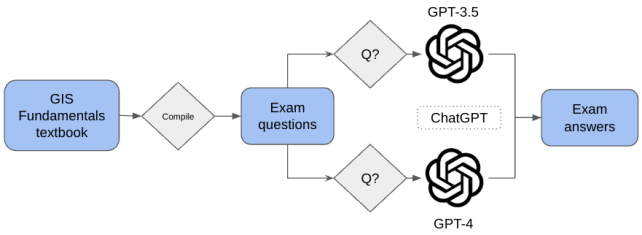

Figure 1: Experimental setup framework

All analysis was conducted using R statistics 4.1.2. Chi-square tests were used to determine the association between exam results and topics as well as question types. The exam dataset including ChatGPT's answers as well as the analysis steps are available from an open repository in Juhasz et al. [24]. The exam consists of 38 true/false, 17 simple choice and 5 multiple choice questions. Each question is worth one point. Table 1 summarizes the topics covered in our exam with corresponding book chapters from Bolstad and Manson [6] and the total number of questions for this topic. The following sections describe the topics as well as the GIS skills and competencies that were tested by the exam. Furthermore, a one representative question from each category is provided in Appendix A.2-A.6.

3.2 Examination Topics

Specific chapters from Bolstad and Manson [6] were grouped together into examination topics. For easier reproducibility of the work, chapter numbers and titles are listed in Table 1 and in the sections below.

3.2.1 ***Fundamental concepts of mapping and GIS.*** This topic consists of 11 simple choice, 2 multiple choice and 14 True/False questions and covers the content of the following chapters in Bolstad and Manson [6]:

(2) Data Models
(3) Geodesy, Datums, Map Projections, and Coordinate Systems
(4) Maps, Data Entry, Editing, and Output
(5) Global Navigation Satellite Systems (GNSS) and Coordinate Surveying
(6) Aerial and Satellite Images

This category is designed to test students' fundamental understanding of digital mapping and GIS. Specific topics include basic GIS data models, coordinate systems and projections, map scale, most common data collection methods, such as digitizing, GNSS, remote sensing and LiDAR (Light detection and ranging).

3.2.2 ***Data sources and tabular data.*** This topic consists of 1 simple choice, 3 multiple choice and 1 True/False questions and covers the content of the following chapters in Bolstad and Manson [6]:

(7) Digital Data
(8) Attribute Data and Tables

This category tests familiarity with digital spatial data sources (global and US specific) and their quality, and concepts in storing & handing attribute (tabular) data in GIS and relational database management systems.

3.2.3 ***Spatial analysis.*** This topic consists of 1 simple choice and 9 True/False questions and covers the content of the following chapters in Bolstad and Manson [6]:

(9) Basic Spatial Analysis
(10) Topics in Raster Analysis
(11) Terrain Analysis Objectives

This category tests understanding and familiarity with spatial analysis concepts and tools using both vector and raster data. Specific topics include spatial scope and spatial operations (geoprocessing), network analysis, map algebra, neighborhood functions as well as common methods of terrain analysis.

3.2.4 ***Spatial statistics and interpolation.*** This topic consists of 2 simple choice and 14 True/False questions and covers the content of Chapter 12: Spatial Estimation: Interpolation, Prediction, and Core Area Delineation in Bolstad and Manson [7]. Specific topics include spatial sampling design, most common interpolation methods, spatial prediction (regression, kriging) and core area mapping (kernel functions, hull methods).

3.2.5 ***Applied GIS workflow.*** This topic consists of 2 simple choice questions and builds on top of the content of Chapter 13: Spatial Models and Modeling in Bolstad and Manson [6]. More specifically, these questions present an analytic objective (e.g. *Identify flat building sites, outside of the floodplain, within 1/4 mile of a road*) and a list of available data layers and asks students to select analysis steps that result in solving the analytic objective. To answer these questions correctly, students need to demonstrate a strong understanding of GIS concepts, including data models, data manipulation and GIS data operations, and be able to connect these concepts and tools. An example is provided as an appendix in Appendix A.6.

4 EXPERIMENTAL RESULTS

Results were manually recorded for both models for every exam question. For replication purposes, the generated answers and question scorings are provided in Juhasz et al. [24]. Table 2 shows the

Table 1: Select of exam questions from Bolstad and Manson [6]

Topic	Corresponding chapters	Number of questions
Fundamental concepts of mapping and GIS	2, 3, 4, 5, 6	27
Data sources and tabular data	7, 8	5
Spatial analysis	9, 10, 11	10
Spatial statistics and interpolation	12	16
Applied GIS workflow	13	2

Table 2: Performance of the two models on the exam dataset based on two evaluation methods (n=60)

	GPT-3.5 n (%)	GPT-4 n (%)
Correct	38 (63.3)	53 (88.3)
Incorrect	22 (36.7)	7 (11.7)
Letter grade	D	B+

performance of GPT-3.5 and GPT-4 in the exam. Both models would pass the exam with letter grades D and B+ respectively using standard US letter grading scheme. GPT-3.5 performed similarly to random chance, however, GPT-4's answers are significantly different from guessing ($p<0.0001$). GPT-4 performed better (88.3%) than GPT-3.5 (63.3%). In addition to the 38 questions answered correctly by GPT-3.5, it accumulated 15 extra correct answers. This could indicate that the general improvements of the GPT-4 family of models, claimed by OpenAI[3], translated to the spatial domain as well. To account for partially correct answers in multiple choice questions, we repeated the same process using point scores calculated using a scheme to reward partial knowledge and penalize guessing [11]. This resulted only in minor improvements for both GPT-3.5 (+3.5%) and GPT-4 (+1.4%). We present the remainder of the results considering only correct and incorrect answers for simplicity.

Results by specific GIS topic areas described in Sections 3.2.1-3.2.5 are shown in Figure 2. GPT-4 outperformed GPT-3.5 in all topics. The highest and lowest scoring categories, not counting the Applied GIS workflow category that was assessed with only two questions, are consistent between the models, which implies that their strengths and weaknesses remain consistent. GPT-3.5's answers were not consistent across topics, ($p=0.04$), however, the association between topic and answers becomes non-significant in GPT-4. This may suggest that GPT-4 picked up a general knowledge that allows it to perform more consistently across a wider spectrum of GIS topics. Figure 3 shows results by question type. The two models behave similarly when assessed by question type. There is a significant association between question type and answers for GPT-3.5 ($p<0.01$) which becomes non-significant for GPT-4. This suggests that model improvements allow GPT-4 to perform more consistently regardless of the question type. We note that even though GPT-3.5 achieved zero entirely correct answers out of five multiple choice questions, it received partial credit in four questions (2.08 points) which would increase its performance to 41.6% for multiple choice questions.

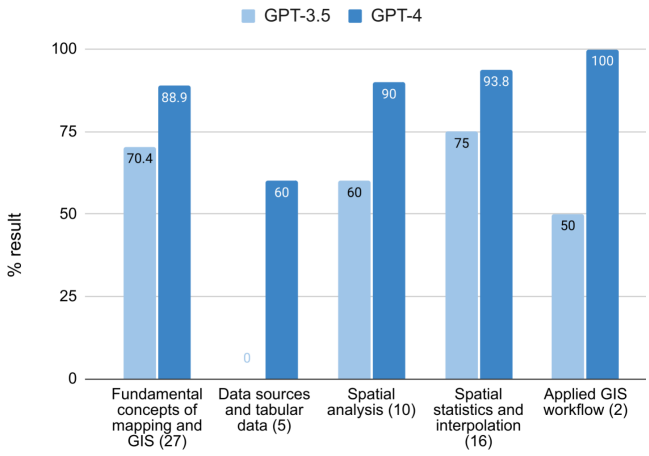

Figure 2: Performance of the two models in specific GIS topic areas. While GPT-3.5's answers to questions in the Data sources and tabular data topic were never entirely correct, it would have achieved 26.6% if awarded fractional points for partially correct answers. The number of questions in a topic is provided in parentheses after the topic label.

4.1 ChatGPT: Exam performance

Assessing the strengths of LLMs in a GIS context is particularly difficult and challenging in the absence of a widely accepted examination or other mechanisms to measure proficiency. As multi-modal questions cannot be currently tested with ChatGPT (see Section 5.1), we are limited to assessment of questions or problems which can be expressed as text. Where LLMs, such as ChatGPT, can really excel within GIS is in the ability to generate programming code effectively and efficiently. This is something many students in GIS courses can really struggle with [14]. ChatGPT, as Borji [9] suggests, is "a proficient coder, but falls short of being a top-notch software engineer". It will generate very good boilerplate and template code which students in GIS courses could find very useful. Stokel-Walker and Van Noorden [44] suggests it offers many facilities for learning and improving coding skills and can be an "excellent debugging assistant". While LLMs like ChatGPT gather knowledge to perform simple mathematical calculations, these models are designed to resemble human speech and not to compute

[3]https://openai.com/gpt-4

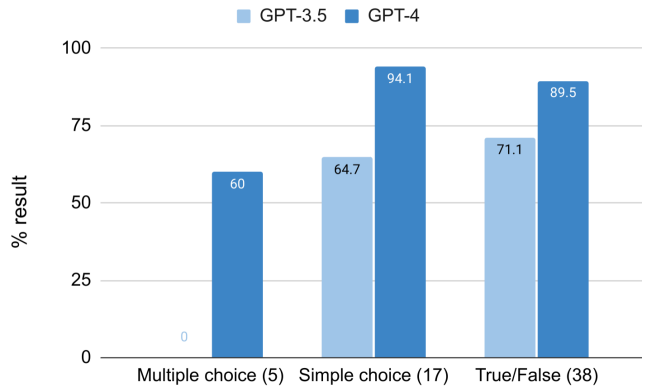

Figure 3: Performance of the two models across question types. While GPT-3.5's answers to multiple choice questions were never entirely correct, it would have achieved 41.5% if awarded fractional points for partially correct answers. The number of questions per type is provided in parentheses in the type label.

complex mathematical expressions. It was shown that ChatGPT's ability to solve math problems decreases with complexity [19]. Our exam assessed student ability to perform surface distance calculations assuming a spherical Earth model. It also required the student to convert latitudes and longitudes between decimal degrees and degree-minute-second format. Surprisingly, GPT-4 answered three out of four of these questions correctly, while GPT-3.5 achieved two correct results. The first question below was answered by both models correctly.

Q: *A degree of longitude spans approximately 110,574 meters at the Equator. How many meters are spanned by a second of longitude at the Equator?*

A: **a)** 30.7(correct); **b)** 22.2; **c)** 123

The second, slightly more complex question was answered correctly by GPT-4, but incorrectly by GPT-3.5:

Q: *Three seconds of longitude span approximately 16.1 meters at a latitude of 80 degrees. How many meters are spanned by three minutes of longitude at 80 degrees?*

A: **a)** 30.7; **b)** 966.0(correct); **c)** 1024.6

Perhaps the most complex question that required computation was answered incorrectly by both models. This highlights the behavior of LLMs concerning computations. The question is shown below and GPT-3.5's response through the web-based ChatGPT interface is provided in Figure 4.

Q: *Assume an Earth radius of 6,278 kilometers. What is the ground surface distance, expressed in meters, of an arc that spans 14.2 seconds of arc? Please enter your answer to the nearest 0.1 meter*

A: **a)** 864.4; **b)** 492.5; **c)** 432.2(correct)

Figure 4 reveals that while the model was aware of the correct formula to calculate the length of an arc, however, it made multiple mistakes. For example, it failed to recognize the need for unit conversion (arc seconds to degrees), which is necessary before calculating the fraction of the circumference of the circle as appears

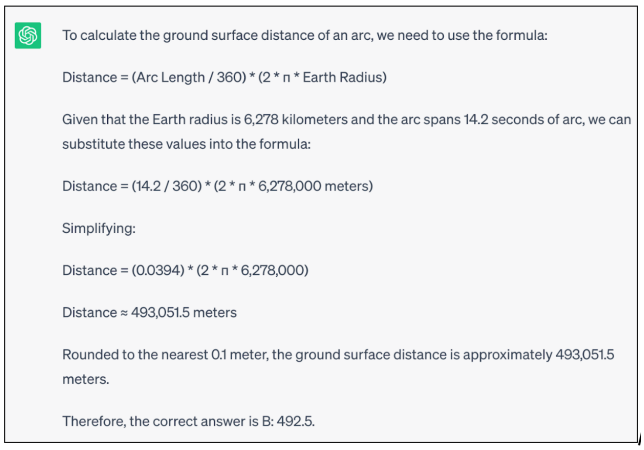

Figure 4: GPT-3.5's incorrect response

in the formula. The model also hallucinated [2] the result when computing its own formula $(0.0394) * (2 * \pi * 6278000)$, and came up with the distance 493,051.5 meters. The result, according to its own formula should have been 1554165.992. However, neither of these answers are correct as the ground surface distance of a 14.2 arc second central angle is 432.2 meters. This highlights limitations of LLMs in numerical calculations. Mai et al. [36] suggests that developing foundation models for vector data is the next major step towards foundation models for GeoAI applications. It is a very challenging problem to encode or decode different kinds of vector data.

5 CONCLUSIONS AND FUTURE WORK

In this paper we have reported on a study designed to assess ChatGPT's performance around geospatial skills and GIS by challenging it with a real-world GIS exam. Through a quantitative and qualitative analysis of ChatGPT's responses, we have shown that there are valuable insights into the potential applications and challenges of LLMs in spatially-oriented fields of inquiry and application. Within Section 1 we proposed two research questions. In achieving RQ1 we undertook an assessment of ChatGPT's performance and geospatial skills by supplying it with questions from a real GIS exam and then assess it's performance in this exam for specific aspects of GIS including spatial analysis, basic concepts of mapping, and data management. Both ChatGPT models achieved a passing grade in the examination. This is impressive but not remarkable. The three categories with the most answers correct from both models were Fundamental concepts of mapping and GIS (19 correct from 27), Spatial statistics and interpolation (12 correct from 16) and Spatial analysis (6 correct from 10). GPT-4's 53 correct answers included all 38 questions that GPT-3.5 answered correctly as well as 15 additional questions. The exam contained only a limited number of questions and unfortunately this does not allow us to reliably and robustly identify specific topics of strength in the models. However, a few interesting and impactful observations emerged from our analysis. It appears that both GPT variants are particularly strong at answering simple questions about basic GIS data models, including

scoring 100% (7 out of 7) for questions about basic GIS data models including data types of attributes. Questions include, for example:

Q: *What type of attribute is human population (the number of people) in a U.S. county data layer?*
A: **a)** interval/ratio (correct); **b)** nominal; **c)** ordinal

Q: *Is the following statement true or false? Vector data models are often better for describing discrete themes such as counties, voting areas, or zip codes?*
A: **a)** True (correct); **b)** False

The Applied GIS workflow category (section 3.2.5) assessed the models' ability to conduct an applied GIS analysis task. The questions provided a description of available data layers and the desired outcome. For these types of questions students are prompted to choose a sequence of steps that results in the desired outcome. In a real exam, to answer these correctly, students need to demonstrate a strong understanding of GIS concepts and a working knowledge of analysis tools. A sample question is provided in Appendix A.6. The exam contained two questions of this kind. GPT-4 answered both of them correctly, while GPT-3.5 achieved one correct answer. While the small number of questions in this category does not allow to draw definitive conclusions, these results suggest that GPT models may be able to understand complex GIS questions as well as create analysis flows. To further strengthen this point, we instructed the models to answer a third question of this kind that was not part of the exam dataset. GPT-4 answered correctly, while GPT-3.5 gave an incorrect answer. This suggests a promising direction to use the newer generation of LLMs to assists human analysts.

In achieving RQ2 we gathered a number of interesting observations. Since LLMs were designed to resemble human speech, and not to perform numerical computations we found that asking the models to perform distance calculations proved to be challenging (see Section 4.1). This presents an opportunity for educators concerned about students using ChatGPT to cheat in their exams. For example, instead of asking to recite or select a correct formula in a simple choice question, instructing students to perform computations would likely be effective against the use of LLMs in an exam. However, not all questions can be transformed into calculations, and foundation models might also get better at performing numerical calculations in time. When creating the examination from the questions provided in Bolstad and Manson [6] we realised how much our own student examinations in spatial databases and spatial analysis used images and diagrams as part of the question text and answers. All questions in our exam for this work were supplied to ChatGPT web interface as text. Some studies, such as Deshpande and Szefer [13] have reported that the web-based ChatGPT performed better than the OpenAI API, for examinations because the web-based version retained context about the questions. Just as we cannot supply diagrams or figures as part of question matter to ChatGPT any examination it is asked to undertake cannot ask for figures, diagrams, sketches, and so on to be provided wholly or partially as answers. This severely limits the types of questions and associated answer templates that can be used in examinations. Finally, we observed that it would be very insightful to provide this same examination to human students in undergraduate and graduate courses in order to allow a comparison between ChatGPT and human participants in the examination. To undertake an experiment such as this is currently beyond the scope of our work and resource allocation. However, with appropriate planning this is an achievable task for future consideration.

5.1 Limitations of this study

There are a number of potential limitations to our study which are outlined below as follows. Some of these issues are not limitations in the strictest sense but rather a description of the study environment and the assumptions made on our behalf.

- **Question types:** In the absence, to our knowledge, of a standardised GIS examination or assessment used at University or College level we extracted questions from the instructor resources in the well known text book on GIS by Bolstad and Manson [6]. We choose a representative sample of questions in our opinion. However, we were constrained in that we could not utilise the large number of questions which contained diagrams, flowcharts, or maps within these resources since currently, mainstream foundation models lack capabilities to understand these inputs [5]. As a result, these questions were not included in our exam. An example of a question that could not be asked is provided in Appendix A.1. All of our questions are simple, multiple choice, or true/false questions with text only used for the question description.
- **Prompting:** We did not undertake any prompting of ChatGPT [49] to clarify its answers or update the answers generated. While this is not necessarily a limitation, in a classroom exam settings students undertaking an exam could be provided with prompts from the teacher or indeed ask for clarification about a specific question. Therefore, we have not tried to replicate this aspect of the exam environment.
- **Fine-tuning and training:** Fine-tuning [4] and other ways to provide additional GIS context to the models would likely, we believe, increase the models' performance in our exam. We did not attempt these approaches as they did not correspond to normally observed exam behaviour. It is unlikely that students of an introductory GIS course will have the skills and experience to provide additional training materials and context to LLMs. It is also unlikely that students, with the appropriate skills, would have enough time to carry out training of the models during moderated examination.
- **Interface:** All questions were posed to ChatGPT using the web interface, as explained in section 3.1. We did not use any of the available APIs[5] as we believe that most students, in an exam-based scenario, would use the web interface to ChatGPT for ease-of-use and time saving.
- **Model improvement:** As time passes and there is additional model development of ChatGPT and other LLMs it is very likely that their ability to score higher in these types of spatial literacy assessment will grow. The paper by Brown et al. [10], on arXiv, from OpenAI, outlines the scale of the training datasets used in the development of ChatGPT. Our paper provides an important milestone in the tracking or

[4] https://platform.openai.com/docs/guides/fine-tuning
[5] https://openai.com/blog/introducing-chatgpt-and-whisper-apis

observation of the evolution of spatial literacy and GIS proficiency within publicly available LLMs. As development of the foundation models will continue, the OpenAI team conclude that "despite many limitations and weaknesses, very large language models may be an important ingredient in the development of adaptable, general language systems" in the future.

- **Multimodal nature of GIS:** Foundation models, such as ChatGPT, are *pre-trained* and then *adapted* via fine-tuning learning strategies[8] and are subsequently deployed on a wide range of knowledge domains [10]. This mitigates the need for task-specific training data[36, 42]. We must then subsequently temper our expectations for these models around domain-specific knowledge (in our case spatial literacy and GIS). The *multimodal* nature (images, text, vector and raster data, semantic information) of GIS "hinders a straightforward application of existing models" across different geographic tasks [36], however, there are examples of utilizing multimodal models in participatory mapping [25].

5.2 Future work

There are many interesting directions for future work on the topic of assessment of ChatGPT and other LLMs' spatial literacy and GIS proficiency. The use of LLMS in education is a promising area of research that offers many opportunities to enhance the learning experience for students and support the work of teachers [26]. While more focussed on the impact of prompting in question asking for ChatGPT, Kocoń et al. [29] state that "it is still an open question what would happen if ChatGPT was finetuned using the datasets for specific tasks". No finetuning or prompting was used in our experimentation and this followed closely the experimental setup for examples outlined in section 1 and section 2. Examination of geospatial skills and GIS proficiency requires a multimodal approach and this cannot be extensively tested with the current LLMs available. Mai et al. [36] suggests that the major challenge in developing a foundational model for GeoAI is the challenge of the "multimodality of geospatial tasks". As argued by Koh et al. [30], many real-world tasks have additional metadata (e.g., spatial location coordinates, environmental information) which may provide additional structure for generalization of models across different geographic regions. As LLMs, such as ChatGPT, improve their spatial literacy the impact for GIS education, teaching and learning, and so on will need to be assessed. We believe that it is a little early to know exactly what this impact will be. However, it is critical that these discussions starts now. The potential impact of LLMs, ChatGPT, chatbots, AI, and so on in education permeates into every educational discipline or subject [43]. The debates and conversations around how tools such as ChatGPT can or cannot be used by students and teachers will have commonalities across subject domains [41] but will also have domain-specific characteristics. Educators must begin thinking about what ChatGPT can and cannot do from that perspective within their own subject domain. We do believe, as shown in this paper, that Geography and GIS, contain special and unique concepts and ideas needed for spatial literacty and GIS proficiency. How we, as an educational community, are examining or measuring this proficiency should be discussed, as a matter of urgency, within the Geography GIS community. ChatGPT brings both new opportunities and new complexity. The ubiquity of this technology, and the likely increased availability of similar and more powerful tools in the future, means that educators need to be aware how to use it, the associated dangers, and how to encourage safe use [17]. We have attempted to probe its capacities and limitations, understand what we are seeing, and then suggest a path forward. Teachers, meanwhile, at this time have a responsibility to train students to use the technology properly [3].

REFERENCES

[1] Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. 2023. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging* 50, 6 (2023), 1549–1552.

[2] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[3] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484* (2023).

[4] Sarah Witham Bednarz and Karen Kemp. 2011. Understanding and nurturing spatial literacy. *Procedia - Social and Behavioral Sciences* 21 (2011), 18–23. https://doi.org/10.1016/j.sbspro.2011.07.004 International Conference: Spatial Thinking and Geographic Information Sciences 2011.

[5] Rajesh Bhayana, Satheesh Krishna, and Robert R Bleakney. 2023. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. *Radiology* (2023), 230582.

[6] Paul Bolstad and Steven Manson. 2022. *GIS Fundamentals: A first text on Geographic Information Systems* (7th ed.). Eider Press.

[7] Paul Bolstad and Steven Manson. 2022. *GIS fundamentals: A first text on geographic information systems* (7th ed.). Vol. 620. Eider Press White Bear Lake, MN.

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]

[9] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL]

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[11] Martin Bush. [n. d.]. A Multiple Choice Test that Rewards Partial Knowledge. 25, 2 ([n. d.]), 157–163. https://doi.org/10.1080/03098770120050828 Publisher: Routledge _eprint: https://doi.org/10.1080/03098770120050828.

[12] Chew-Hung Chang and Gillian Kidman. 2023. The rise of generative artificial intelligence (AI) language models-challenges and opportunities for geographical and environmental education. *International Research in Geographical and Environmental Education* (2023), 1–5.

[13] Sanjay Deshpande and Jakub Szefer. 2023. Analyzing ChatGPT's Aptitude in an Introductory Computer Engineering Course. arXiv:2304.06122 [cs.CY]

[14] Thomas R. Etherington. 2016. Teaching introductory GIS programming to geographers using an open source Python approach. *Journal of Geography in Higher Education* 40, 1 (2016), 117–130. https://doi.org/10.1080/03098265.2015.1086981

[15] Todd D Fagin and Thomas A Wikle. 2011. The instructor element of GIS instruction at US colleges and universities. *Transactions in GIS* 15, 1 (2011), 1–15.

[16] Mohammadreza Farrokhnia, Seyyed Kazem Banihashem, Omid Noroozi, and Arjen Wals. 2023. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International* 0, 0 (2023), 1–15. https://doi.org/10.1080/14703297.2023.2195846

[17] Mohammadreza Farrokhnia, Seyyed Kazem Banihashem, Omid Noroozi, and Arjen Wals. 2023. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International* (2023), 1–15.

[18] Nino Fijačko, Lucija Gosak, Gregor Štiglic, Christopher T Picard, and Matthew John Douma. 2023. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation* 185 (2023).

[19] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. [n. d.]. Mathematical Capabilities of ChatGPT. https://doi.org/10.48550/arXiv.2301.13867 arXiv:2301.13867 [cs]

[20] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* 9, 1 (2023), e45312.

[21] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).

[22] Jiří Horák. 2015. The role of certification in GIS&T education. *Procedia-Social and Behavioral Sciences* 174 (2015), 1356–1363.

[23] Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745 [cs.CL]

[24] Levente Juhasz, Wencong Cui, Peter Mooney, and Boyuan Guan. 2023. Replication data: Towards understanding the spatial literacy of ChatGPT. https://doi.org/10.17605/OSF.IO/RU6MF

[25] Levente Juhász, Peter Mooney, Hartwig H. Hochmair, and Boyuan Guan. 2023. ChatGPT as a mapping assistant: A novel method to enrich maps with generative AI and content derived from street-level photographs. In *Spatial Data Science Symposium 2023 Short Paper Proceedings*. UC Santa Barbara: Center for Spatial Studies. https://doi.org/10.25436/E2ZW27

[26] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[27] Minsung Kim and Robert Bednarz. 2013. Effects of a GIS Course on Self-Assessment of Spatial Habits of Mind (SHOM). *Journal of Geography* 112, 4 (2013), 165–177. https://doi.org/10.1080/00221341.2012.684356

[28] Helen King. 2006. Understanding spatial literacy: cognitive and curriculum perspectives. *Planet* 17, 1 (2006), 26–28. https://doi.org/10.11120/plan.2006.00170026

[29] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* (2023), 101861. https://doi.org/10.1016/j.inffus.2023.101861

[30] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[31] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* 2, 2 (02 2023), 1–12. https://doi.org/10.1371/journal.pdig.0000198

[32] Diarmaid Lane, Raymond Lynch, and Oliver McGarr. 2019. Problematizing spatial literacy within the school curriculum. *International Journal of Technology and Design Education* 29, 4 (2019), 685–700.

[33] Chung Kwan Lo. 2023. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences* 13, 4 (2023). https://doi.org/10.3390/educsci13040410

[34] Thomas R Lord and Joy L Rupert. 1995. Visual-spatial aptitude in elementary education majors in science and math tracks. *Journal of Elementary Science Education* 7, 2 (1995), 47–58.

[35] Gengchen Mai, Yingjie Hu, Song Gao, Ling Cai, Bruno Martins, Johannes Scholz, Jing Gao, and Krzysztof Janowicz. 2022. Symbolic and subsymbolic GeoAI: Geospatial knowledge graphs and spatially explicit machine learning. *Transactions in GIS* 26, 8 (2022), 3118–3124. https://doi.org/10.1111/tgis.13012 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.13012

[36] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2023. On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence. arXiv:2304.06798 [cs.AI]

[37] Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. 2021. Geographic question answering: challenges, uniqueness, classification, and future directions. *AGILE: GIScience series* 2 (2021), 8.

[38] Amarachi B. Mbakwe, Ismini Lourentzou, Leo Anthony Celi, Oren J. Mechanic, and Alon Dagan. 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digital Health* 2, 2 (02 2023), 1–3. https://doi.org/10.1371/journal.pdig.0000205

[39] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* (2023), 1–2.

[40] Philip Mark Newton. 2023. ChatGPT performance on MCQ-based exams. (2023). https://doi.org/sytu3

[41] Junaid Qadir. 2023. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1–9.

[42] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv:2302.06476 [cs.CL]

[43] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6, 1 (2023).

[44] Chris Stokel-Walker and Richard Van Noorden. 2023. What ChatGPT and generative AI mean for science. *Nature* 614, 7947 (2023), 214–216.

[45] Eric Strong, Alicia DiGiammarino, Yingjie Weng, Preetha Basaviah, Poonam Hosamani, Andre Kumar, Andrew Nevins, John Kugler, Jason Hom, and Jonathan H Chen. 2023. Performance of ChatGPT on free-response, clinical reasoning exams. *medRxiv* (2023). https://doi.org/10.1101/2023.03.24.23287731

[46] Zhisheng Tang and Mayank Kejriwal. 2023. A Pilot Evaluation of ChatGPT and DALL-E 2 on Decision Making and Spatial Reasoning. *arXiv preprint arXiv:2302.09068* (2023).

[47] H. Holden Thorp. 2023. ChatGPT is fun, but not an author. *Science* 379, 6630 (2023), 313–313. https://doi.org/10.1126/science.adg7879 arXiv:https://www.science.org/doi/pdf/10.1126/science.adg7879

[48] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.

[49] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[50] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

A EXAMPLE QUESTIONS

A.1 Non-suitable graphical question

Q: Based on Figure 5, match the map letter to the most appropriate type for these data collected by county

A: Feature - a); Contour - b); Dot density - d); Coropleth -c)

A.2 Fundamental concepts of mapping and GIS

Q: When long/lat earth coordinates are plotted on a Cartesian plane, shape distortion is greatest in what direction?

A: **a)** north-south; **b)** east-west (correct); **c)** at oblique angles; **c)** they're equal in all directions

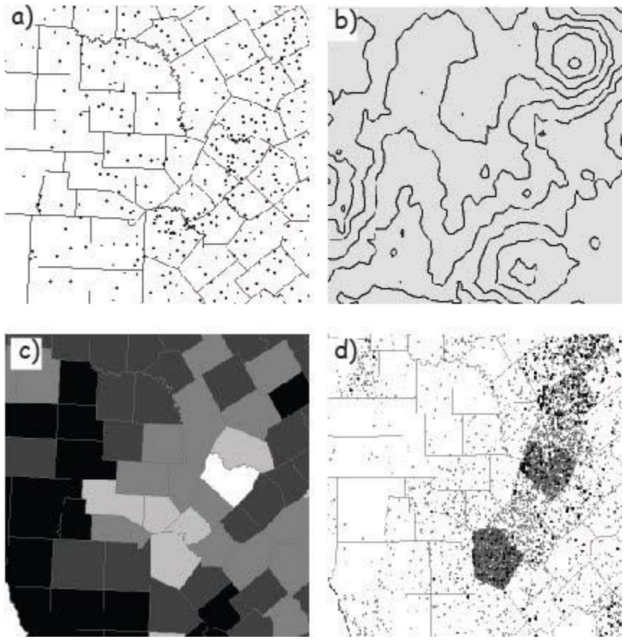

Figure 5: Multi-modal questions could not be tested due to ChatGPT's current limitation on understanding modalities other than text.

A.3 Data sources and tabular data

Q: Is the following statement true or false? A client in a DBMS is the person using the system.
A: **a)** True **b)** False (correct)

A.4 Spatial analysis

Q: Is the following statement true or false? Proximity functions may be applied to both raster and vector data
A: **a)** True (correct); **b)** False

A.5 Spatial statistics and interpolation

Q: Is the following statement true or false? Moran's I is a measure of spatial autocorrelation
A: **a)** True (correct); **b)** False

A.6 Applied GIS analysis flow

Q: You are asked to do a spatial analysis that may include some of the following data layers:

- BOUND - Study area boundary (vector polygon);
- FEMA - 100-year floodplain map (vector poly);
- CENSUS population block data (vector poly);
- ROAD and rail data (vector line);
- NASS landcover data (raster, 25 m res.)
- DEM - USGS 10m DEM (raster, 10 m);
- WETLAND - USFWS data (vector polys)

Select the sequence of steps that comes closest to describing how to complete your task: **Identify flat building sites, outside of the floodplain, within 1/4 mile of a road.** Note that intermediate layers in the answers are written in ALL CAPS, and –> means output.

A: - The correct answer is **a)**

a)
(1) Select all wetlands from WETLAND, dissolve, calculate area, and select those > 10 hectares –> 10WTL
(2) Select city by high population density from CENSUS, reclass, dissolve, buffer at 200m –> NEARCT
(3) Buffer wetlands at 1.6 km (give 9sq km area), without dissolving output across separate wetlands –> WTBUFF
(4) Reclassify NASS to Corn/noncorn –> CRN_RC
(5) Select 10WTL by location, against NEARCT –> CTWET
(6) Intersect CTWET with CRN_RC, once for each individual wetland –> CRN_WET
(7) Summarize area for CRN_WET, select those that have greater than 50% area in corn nearby –> FINAL LAYER

b)
(1) Select city NLCD, dissolve, buffer at 200m–>CTBUFF
(2) Select all wetlands from WETLAND, calculate area, and select those > 10 hectares –> 10WTL
(3) Buffer wetlands, dissolve across wetlands –> WTBUFF
(4) Reclassify NLCD to crop/noncrop –> CRN_RC
(5) Select 10WTL by location, against CTBUF –> CTWET
(6) Intersect CTWET with CRN_RC –> CRN_WET
(7) Summarize area for CRN_WET, select those that have greater than 50% area in crop nearby –> FINAL LAYER

c)
(1) Select all wetlands from WETLAND, dissolve, calculate area, and select those > 10 hectares –> 10WTL
(2) buffer each wetland in 10WTL, without dissolving output across separate wetlands –> WTBUFF
(3) select city by high population density from CENSUS, reclass, dissolve –> NEARCT
(4) select 10WTL by location, against NEARCT –> CTWET
(5) intersect CTWET with NASS summarize area for CT_WET, select those that have greater than 50% area in corn nearby –> FINAL LAYER

d)
(1) Select city NLCD, dissolve, buffer 200m –> CTBUFF
(2) select all wetlands from WETLAND, calculate area, and select those > 10 hectares –> 10WTL
(3) buffer each wetland, dissolving output across separate wetlands –> WTBUFF
(4) reclassify NLCD to crop/noncrop –> CRN_RC
(5) select 10WTL by location, against CTBUF –> CTWET
(6) intersect CTWET with CRN_RC –> CRN_WET
(7) summarize area for CRN_WET, select those that have greater than 50% area in crop nearby –> FINAL LAYER