

POSITION PAPER

Opening a conversation on responsible environmental data science in the age of large language models

Ruth Y. Oliver¹ , Melissa Chapman², Nathan Emery³, Lauren Gillespie⁴, Natasha Gownaris⁵, Sophia Leiker¹, Anna C. Nisi⁶, David Ayers⁷, Ian Breckheimer⁸, Hannah Blondin⁹, Ava Hoffman¹⁰, Camille M.L.S. Pagniello¹¹ , Megan Raisle and Naupaka Zimmerman¹² 

¹Bren School of Environmental Science and Management, University of California Santa Barbara, Santa Barbara, CA, USA

²National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA, USA

³Center for Innovative Teaching, Research, and Learning, University of California Santa Barbara, Santa Barbara, CA, USA

⁴Department of Computer Science, Stanford University, Palo Alto, CA, USA

⁵Department of Environmental Studies, Gettysburg College, Gettysburg, PA, USA

⁶Department of Biology, Center for Ecosystem Sentinels, University of Washington, Seattle, WA, USA

⁷Wildlife, Fish and Conservation Biology Department, University of California Davis, Davis, CA, USA

⁸Rocky Mountain Biological Laboratory, Crested Butte, CO, USA

⁹Cooperative Institute for Marine and Atmospheric Studies (CIMAS), University of Miami, Miami, FL, USA

¹⁰Data Science Lab, Fred Hutchinson Cancer Center, Seattle, WA, USA

¹¹Hawai'i Institute of Marine Biology, University of Hawai'i at Ma'noa, Kaneohe, HI, USA

¹²Department of Biology, University of San Francisco, San Francisco, CA, USA

Corresponding author: Ruth Y. Oliver; Email: rutholiver@bren.ucsb.edu

Received: 05 May 2023; Revised: 15 January 2024; Accepted: 28 February 2024

Keywords: bias; ChatGPT; data ethics; generative AI; pedagogy

Abstract

The general public and scientific community alike are abuzz over the release of ChatGPT and GPT-4. Among many concerns being raised about the emergence and widespread use of tools based on large language models (LLMs) is the potential for them to propagate biases and inequities. We hope to open a conversation within the environmental data science community to encourage the circumspect and responsible use of LLMs. Here, we pose a series of questions aimed at fostering discussion and initiating a larger dialogue. To improve literacy on these tools, we provide background information on the LLMs that underpin tools like ChatGPT. We identify key areas in research and teaching in environmental data science where these tools may be applied, and discuss limitations to their use and points of concern. We also discuss ethical considerations surrounding the use of LLMs to ensure that as environmental data scientists, researchers, and instructors, we can make well-considered and informed choices about engagement with these tools. Our goal is to spark forward-looking discussion and research on how as a community we can responsibly integrate generative AI technologies into our work.

Impact Statement

With the recent release of ChatGPT and similar tools based on large language models, there is considerable enthusiasm and substantial concern over how these tools should be used. We pose a series of questions aimed at unpacking important considerations in the responsible use of large language models within environmental data science.

1. Introduction

Following its public release in late 2022, ChatGPT rapidly captured the world's attention. Its simple text chat interface allows users to interact with a powerful artificial intelligence model capable of generating shockingly human-like responses. As with any new technology that seems to break the bounds of what was previously thought possible, the release of ChatGPT provoked sizable reactions across academia, industry, and the general public. Some have lauded its potential while others have raised alarm bells over its potential misuse (ChatGPT is a Data Privacy Nightmare, and We Ought to be Concerned, 2023; Getahun, 2023; Marcus, 2023; Pause Giant AI Experiments, 2023; Rillig, 2023). Particularly concerning is the tendency of large language models (LLMs), like ChatGPT, to propagate stereotypes and social biases and provide false or misleading information, all while engendering unmerited trust due to their “human-like” qualities (Bender et al., 2021; Weidinger et al., 2021). In the clamor to understand the societal implications of ChatGPT and similar tools, there are growing calls for the scientific community to critically examine the potential ramifications of reliance on LLMs (Bommasani et al., 2022; van Dis et al., 2023).

Although artificial intelligence is not a novel addition to environmental data science, LLMs have not been widely adopted within the community. Therefore, environmental scientists may have cursory or inaccurate knowledge about the benefits, drawbacks, and limitations of these models or may have difficulty deciding whether and when to use these models in their work. While traditional artificial intelligence synthesizes data, LLMs are a form of generative artificial intelligence in which new content, such as text or images, is created based on existing data (e.g., Schmidt et al., 2021). The uptake of tools like ChatGPT is poised to rapidly increase; thus, there is an urgent need for the environmental data science community to acquire the literacy necessary to make informed choices regarding their use. LLMs share many of the issues recognized in other artificial intelligence and machine learning methods, including bias propagation and homogenization (McGovern et al., 2022). However, the ability to effortlessly generate text that can synthesize concepts, create new content, and even design analysis workflows raises a new set of questions about the responsible use of LLMs.

2. How do LLMs work?

LLMs are a family of machine learning models designed to intake text prompts and generate contextual text outputs. They are extremely large deep neural networks with upwards of hundreds of billions of parameters, also known as weights. LLMs are trained in a “self-supervised” manner, sometimes referred to as “autoregressive,” using an architecture called a transformer that relies on a surprisingly simple process called attention (Vaswani et al., 2017). These models are fed text with some words omitted and are trained to essentially “impute” the missing text. Training text comes from an incredibly large corpora of text often scraped from the internet (i.e., Reddit, Wikipedia, and the Internet Archive). A training example might be the sentence “I’m going to walk the.....” The model would predict the end of the sentence with the goal of maximizing the likelihood that the characters it predicts are the characters omitted in the original text. Assuming the original sentence was “I’m going to walk the dog,” the model’s parameters would be updated to maximize the likelihood that the three characters it predicted were indeed “dog.” By repeating this omission-based training process billions of times, models like ChatGPT “learn” to predict responses to a wide range of inputs ranging from patents to historical summaries (Brown et al., 2020).

In principle, this approach can be used on any text-based data format (e.g., DNA sequences) (Gankin et al., 2023), but human prose text is the most common. Of specific interest to the EDS community, these models have shown surprising capability in translating between computer code and natural language or between multiple coding languages (Merow et al., 2023).

The primary difference between previous versions of these models, such as GPT-2, and the current generation of GPT models—which as of writing was GPT-4—is that newer models have vastly more parameters and are trained on substantially more data. Nearly everything about the training paradigm, pipeline, and validation is the same as previous iterations, with the caveat that the ChatGPT implementations of these models are fine-tuned on human feedback data, referred to as reinforcement learning from

human feedback (RLHF), to better reflect human dialog rather than longform text. RLHF enables LLMs to align text with the complex values of users and reject questions that are inappropriate or outside of the scope of the model's knowledge, and seems to improve LLM's ability to carry on a dialog (Bai et al., 2022). Fine-tuning with RLHF has ostensibly improved the dialog performance of GPT models and driven the rapid adoption of ChatGPT, and indeed much of the attention granted to LLMs hinges on outputs that are lengthy, varied, flexible, and coherent, as selected for by the RLHF process (Clark et al., 2021).

3. What are the current concerns with LLMs?

Although outputs generated by LLMs can sound human-like, coherence is often incorrectly conflated with understanding. It is critical to bear in mind that the process by which LLMs generate output is not a reliable source of facts (Bender et al., 2021; Weidinger et al., 2021; World Economic Forum [Internet], 2023). Just as human-generated prose or code can contain factual or formatting errors or reflect certain world views, LLM-generated text can also contain mistakes or recapitulate the dominant views. In general, the reliability of the outputs of ChatGPT varies substantially across domains; in particular, it tends to underperform on science-related prompts (Shen et al., 2023). LLMs may be prone to demographic, cultural, linguistic, ideological, and political biases, (Bolukbasi et al., 2016; Caliskan et al., 2017; Buolamwini and Gebru, 2018; Bender et al., 2021; Kirk et al., 2021; Ferrara, 2023a). Biases in the form of systematic misrepresentations, attribution errors, or factual inaccuracies can arise from decisions made in the development and implementation of LLMs, including the selection of training data, decisions about algorithm architecture, details of product design, and policies that control model behavior (Ferrara, 2023a). Also troubling is the fact that models can, with ostensibly equal confidence, create false but believable "facts," known as hallucinations (Weidinger et al., 2021; Liu et al., 2023).

One important yet underappreciated aspect of LLM performance is the curation of the training data corpus—the steps used to prepare cleaned training samples from raw text. This process often includes the removal of data considered harmful, irrelevant, or incorrect. The decisions about what falls into these categories are often subjective, and consequently the outcome of such decisions may vary depending on the data science workers building the dataset or the LLM's target audience (Chung, 2019; Miceli et al., 2022). While intended to remove objectionable content such as violence or hate speech, data curation can have unintended consequences and create temporal homogenization of viewpoints (Bender et al., 2021). Further, confirmation biases may arise from training data that was unintentionally curated to reflect individuals' viewpoints (Bolukbasi et al., 2016; Caliskan et al., 2017; Ferrara, 2023a). Unlike humans, who naturally censor new and emerging social taboos from their work, LLMs may cast a long shadow of outdated norms. Further, due to different amounts of training data available for different languages, ChatGPT performs more effectively with (e.g.) European languages that have lots of online training content available than with languages that lack this body of data (Jiao et al., 2023).

To return to the earlier example, when asked to predict what comes after the phrase "I'm going to walk the...", it is reasonable to think that most people would guess "dog," as this is a very common phrase in American English. However, it is certainly not true that the sentence *must* conclude with the word "dog." For example, perhaps the person writing has a pet cat that they take on daily walks! This example may feel forced or trivial, but when you begin to consider the sum total of text on the internet, there are many viewpoints, opinions, and people that are not represented and innumerable subtle representation biases in this large body of text. The probability of text generation is commensurate with its availability on the internet, and when that text is disproportionately generated by a small slice of the world's population, then that text will disproportionately reflect their world views. For example, Wikipedia contributors are primarily white, cisgender men, leading to subtle but important systemic biases in topic contribution and tone (Shaw and Hargittai, 2018; MIT Technology Review [Internet], 2023). When the text an LLM is trained on is the product of a limited worldview, the model output propagates a specific knowledge base and worldview.

An especially insidious way that this emerges is in errors of omission. For example, when asked who discovered the global warming potential of carbon dioxide, ChatGPT responded by making reference to

Box 1. Propagating bias through omission. When asked about the early recognition of carbon dioxide as a greenhouse gas, ChatGPT perpetuates a common erasure of contributions by Eunice Newton Foote. Conversation with ChatGPT 3.5 on May 2, 2023.

Prompt: who discovered the global warming potential of CO₂?

Response:

The concept of the global warming potential (GWP) of carbon dioxide (CO₂) was first introduced by scientists in the 1970s, particularly by J. S. Sawyer and Wallace Broecker. They recognized that different greenhouse gases have different warming potentials based on their ability to absorb and trap heat in the atmosphere.

However, the specific calculation of CO₂'s GWP as a factor of its ability to trap heat relative to other greenhouse gases, such as methane and nitrous oxide, was developed in the 1990s by the Intergovernmental Panel on Climate Change (IPCC) as part of their efforts to understand and quantify the impacts of human-caused climate change.

publications from the 1970s by John Sawyer and Wallace Broecker (Sawyer, 1972; Broecker, 1975) as well as the Intergovernmental Panel on Climate Change (Box 1). However, this response ignores the contributions of Eunice Newton Foote who correctly theorized the connection between increased carbon dioxide and planetary warming based on experimentation over a hundred years before (Huddleston, 2023). The erasure of her story from history is perpetuated when we use models trained on the consequences of this erasure.

Troublingly, the training data used in current LLMs are so voluminous that documenting and understanding what these models are based on is nearly impossible (Bender et al., 2021). Further, because of the size and complexities of model architecture and training corpus, even small decisions can have large reverberations in outcomes through the Butterfly Effect (Ferrara, 2023b). In addition, data curation and training processes are usually opaque (ZDNET, 2023), since many models are trained by large corporations that are not bound to open science data practices or peer-review processes and as such rarely follow best practices laid out by the AI ethics research community (Mitchell et al., 2019; Pushkarna et al., 2022). Financial barriers further limit the number of organizations with the capacity to train and host such models. This has left only a select few organizations such as Google, Meta, and OpenAI / Microsoft as competitive in this space. Open-source examples of LLMs, such as the BLOOM project, are even rarer. Open-source models generally have lower performance (Chen, 2023), potentially due to differences in the all-important data curation step, which is usually opaque to all but the modelers. However, the model weights for Meta's LLaMA (Large Language Model Meta AI) were leaked shortly after the model's February 2023 release, which has allowed several open-source and local implementations of LLMs to emerge (e.g., Alpaca [Stanford], Koala [Berkeley], and Vicuna [multiple institutions]). Despite the general trend toward the use of closed-source models, there are also increasing pushes for open LLM research, such as the AI Alliance, which could help bring more transparency to model development and deployment.

4. How might LLMs affect our research practices?

Given the capabilities of LLMs like ChatGPT, we may need to rethink how our community engages with our work and one another. We explore a hypothetical example of how ChatGPT's core functionalities (text generation, translation, and data analysis/visualization) could be leveraged in the development and execution of an environmental data science project (Figure 1; Supplementary material). In this case, a researcher is inspired by a recent publication that identified pervasive biases in biodiversity records in the United States of America due to historical residential redlining practices. The researcher is interested in

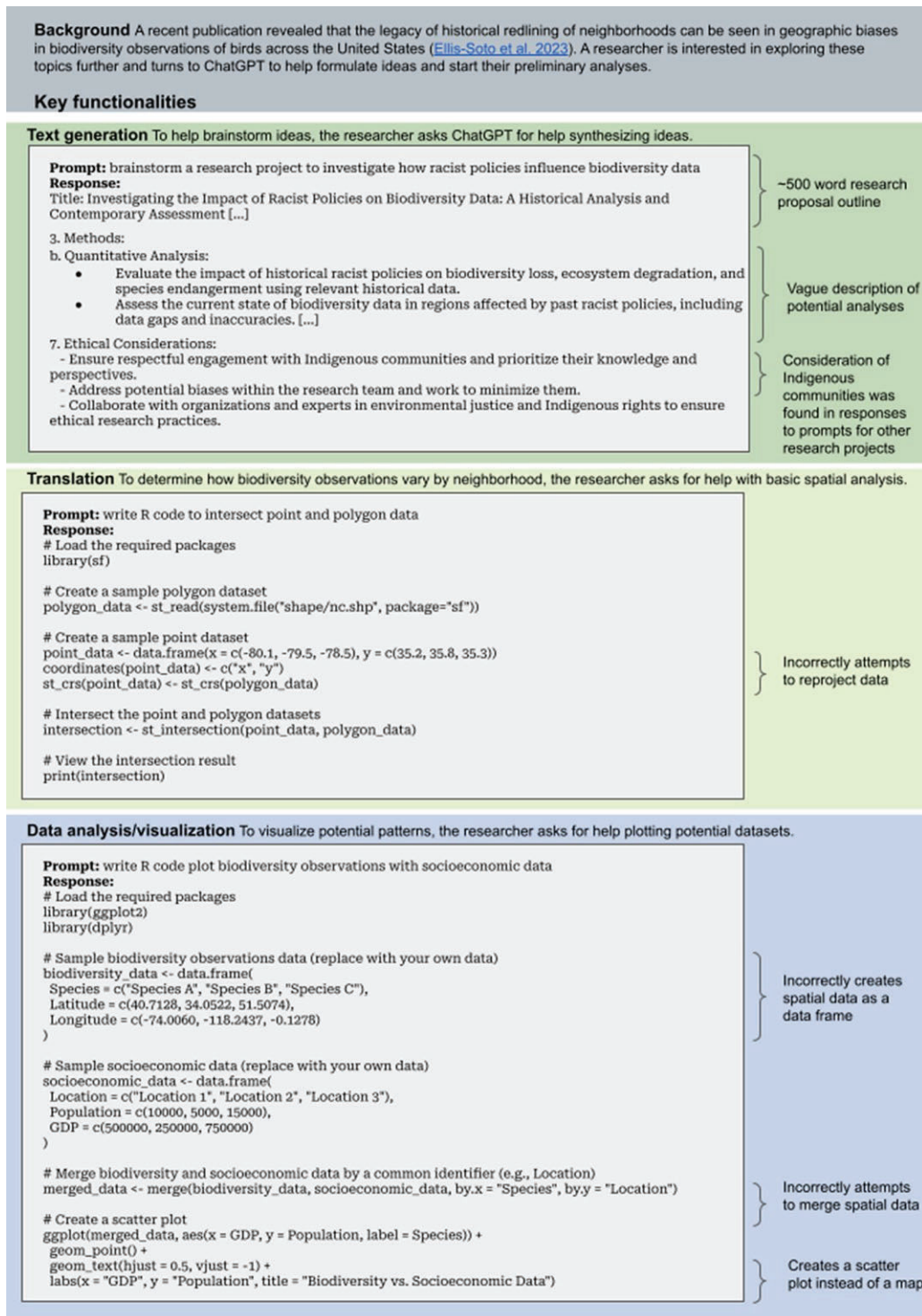


Figure 1. Collaborating with ChatGPT on an environmental data science project. In this hypothetical example, we explore how ChatGPT's (3.5) core functionalities may be used for environmental data science. We have annotated to highlight notable components of responses. Responses have been truncated for brevity. A full transcript can be found in the Supplementary material.

further exploring how racist policies impact biodiversity data collection and turns to ChatGPT to help brainstorm research questions and approaches as well as generate code to perform relevant analyses.

In response to a prompt to brainstorm a research project on the topic, ChatGPT quickly generated a lengthy research project proposal. While much of the text appears to be a useful starting place for a grant or dissertation proposal, many of the details are quite vague and seem to recapitulate obvious statements. For example, under a section entitled “Quantitative Analysis” the output suggests the user to “evaluate” and “assess” the impact of racist policies on biodiversity data collection, closely mirroring the original prompt without suggesting specific methodologies. ChatGPT is unable to provide more specific instructions on how to break down a complex research question into an actionable analytical framework without further prompting. Therefore, identifying the appropriate methods to use still requires expert knowledge. Interestingly, the output includes recommendations to meaningfully engage with Indigenous communities and center their knowledge and perspectives, which remains rare in much of environmental data science.

In response to prompts to provide code for potentially relevant analyses, ChatGPT returned code that handled simple tasks well, but it struggled with more complex and ambiguous prompting. For example, the outputted code correctly handled example datasets as spatial objects. However, minor issues arose when it suggested reassigning a dataset’s coordinate reference system instead of using a transformation. This is a relatively small error that would be straightforward for an experienced user to catch, but less obvious to a more novice user of spatial data. Additionally, the response to a subsequent, more ambiguous prompt for assistance in data visualization was largely incorrect. Despite creating example datasets to visualize as spatial objects, the suggested code for joining these datasets to one another does not treat them as such, in this case it suggests a standard merge instead of a spatial merge. This incoherence demonstrates a fundamental aspect and limitation of ChatGPT’s output generation process – it is simply predicting the most likely set of text, and therefore does not build a logically cohesive response.

While this is merely one example of the ways that ChatGPT may be used (or misused), in [Table 1](#) we provide a more general overview of the potential applications, benefits, and pitfalls of using LLMs in EDS research. One of the most powerful uses of AI will be the automation of routine tasks, which would ostensibly allow researchers to spend more time on interpretation and deep thinking. LLMs may also aid interpretation by identifying patterns and relationships in complex data and have the potential to improve reproducibility and consistency within EDS by lowering the barriers to writing code when reproducing other’s results. ChatGPT has been shown capable of creating code for ecological tasks. It performs best when generating shortcode blocks based on highly specific prompts, although almost all suggested code requires debugging (Merow et al., 2023). These tools also have the ability to translate code to text, which could facilitate consistent methodological descriptions across published literature and help students and other researchers interpret and use a new codebase. Further, ChatGPT may be able to translate between coding languages. As a highly interdisciplinary field with many communities of practice, more seamless translation between coding languages could help bolster collaboration. Consequently, LLMs could facilitate collaboration across disciplines, from biologists to managers to statisticians.

Despite these potential benefits, LLMs do not reduce the need for human-based validation and assessment. Using LLMs to perform data analysis can create situations where EDS practitioners who may have limited understanding of the underlying techniques might apply them inappropriately when recommended by ChatGPT, as users tend to overly trust AI-generated content (Perry et al., 2022). These considerations will be particularly important as we train the next generation of environmental data scientists, who will begin their learning in the time of LLMs. Since LLMs produce output based on their inputs and training data, their ability to construct novel approaches to scientific or data analysis problems is necessarily limited. Additionally, writing code is a skill that not only enables the interpretation of complex environmental data but also allows scientists to think creatively and critically about their data. Removing the scientist from methodological development could be detrimental to both the development of the scientist and the correct interpretation of nuances in the data and/or analysis. It is incumbent upon users to consider potential collateral damages inherent with a more “efficient” approach and likely best to think of AI tools as supporting, rather than replacing, existing duties.

Table 1. General overview of the key functionalities, applications, and potential benefits and pitfalls of ChatGPT

Key functionalities	Application	Potential benefits	Potential pitfalls
Text generation	• Editing, outlining, and/or generating draft text • Synthesizing concepts for learning, teaching, or brainstorming • Developing analogies	• Automating rote tasks (e.g., reducing word count) • Providing an overview of well-studied and/or wide-ranging topics • Outlining key concepts that can guide further reading • Providing opportunities for a focus on process and revision of written pieces • Revision for brevity and clarity	• Creating false citations • Presenting misleading or inaccurate conclusions • Reducing student engagement with the scientific literature, methodology, and process • Decreasing time spent on deeply understanding key concepts • Homogenizing and/or biasing knowledge
Translation (text to code, code to text, or code to code)	• Creating analysis workflows • Creating example workflows • Interpreting, refactoring, or annotating code • Translating between coding languages	• Automating rote tasks (e.g., making code more efficient) • Lowering barriers to learning how to code • Reducing time to identify errors in code and providing real-time feedback • Improve annotation • Improving reproducibility and standardization across coding languages	• Using others' code without proper attribution or in violation of copyright • Reducing need for skill building in critical thinking and perseverance • Widening opportunity gaps if only financially privileged can access premium services (e.g., the latest, largest LLMs) • Providing proprietary or sensitive code to the companies hosting the model
Data analysis and visualization	• Data summary statistics • Simple visualizations when provided data • Guidance on best approaches given data	• Improving the interpretability of figures • Creating more accessible graphics (e.g., colorblind friendly) with less time/expertise needed • Improving the understanding of, and application of appropriate statistical methods	• Misleading information and broken code due to model being trained on outdated packages • Suggesting inappropriate statistical analyses due to incomplete understanding of data and assumptions • Developing less thoughtful visualizations or visualizations that do not align with the broader goals of analysis • Reducing development of novel packages for visualization

Note. This summary is based on considerations with environmental data science, although they may apply more broadly.

Just as the outputs of LLMs are based on our scientific knowledge base, its biases are also our own. As environmental data scientists, we are responsible for ensuring that we do not propagate and amplify embedded biases in the outputs of LLMs. However, it is not just the outputs that should be used responsibly. We also have a responsibility to design studies that, to the extent possible, contribute unbiased information back into the training corpus that will be used for future LLM training. Recognizing and completely eliminating bias can be challenging; thus, one potential use of LLMs could involve flagging elements of our prose or data analyses that introduce biases, with the caveat that these tools still may overly homogenize what is considered biased (Chung, 2019; Miceli et al., 2022). Ultimately, scientists may one day outsource the coding itself, but will still need to be trained in how to prompt AI tools appropriately (Zamfirescu-Pereira et al., 2023), how to assess the validity of their outputs (Passi and Vorvoreanu, n.d.; *Zombies in the Loop?* 2023), and to consider the societal implications and applications of these outputs (Tomašev et al., 2020; Krügel et al., 2023).

Looking forward, generative AI could accelerate the broader application and integration of insights from the fields of applied science and environmental justice, helping to link scientific findings with ethical considerations and management decisions across local, regional, and global scales. When deciding whether to use LLMs as a tool, environmental data scientists should reflect on what biases LLMs might introduce in the specific context of their work, and when biases or factual errors exist, how they will identify them. Researchers must have enough comfort with their field's approaches that they can anticipate areas where LLMs will fall short (e.g., if a commonly used package has been replaced since the LLM training data were updated), and researchers should also be aware of any policies that might restrict their use of LLMs in proposing projects (e.g., grant writing), analyzing data (e.g., data confidentiality), and in communicating their findings (e.g., article drafting) (Box 2).

5. How might LLMs affect our teaching practices?

The widespread availability of LLMs will shape how we train the next generation of environmental data scientists, though the advantages and disadvantages of using systems like ChatGPT in the classroom are still being explored (Baidoo-Anu and Owusu Ansah, 2023; Meyer et al., 2023; Milano et al., 2023). Students are already using ChatGPT in primary school (Khan, 2023) and higher education settings (Cu and Hochman, 2023) with many institutions struggling to keep pace. At one extreme, the New York City Department of Education initially banned ChatGPT on all devices and networks within the city's school system (NBC News, 2023). Alternately, some school districts have opted to partner with Khan Academy and OpenAI to introduce automated tutoring with GPT-4, with tens of thousands of students using these AI tutors to date (Khan, 2023). The decision on whether or not LLMs are more helpful or harmful for students' education remains to be seen, with many institutions opting for a "wait and see" approach on whether to adopt or to ban these tools in the classroom.

Proponents argue that these resources should remain free and easily accessible (Mills et al., 2023), as they could help to reduce barriers to learning, improve learner confidence, and support diversity and inclusion in environmental data science (Samuel et al., 2020). LLMs also have the potential to help instructors assist students in large enrollment courses (Popenici and Kerr, 2017) and to foster a self-directed learning environment (Wilcox, 1996). As novices, students greatly benefit from practicing with feedback (Keuning et al., 2019), and for data science skills like coding, a responsive LLM that understands code could provide personalized practice for students struggling to learn coding languages. Additionally, by using an LLM for coding, students may feel more empowered to apply their knowledge and skills to new environmental problems and attain agency over their own learning, an established equitable teaching practice (Madkins et al., 2020). However, learning how to code without LLMs helps students to build troubleshooting skills and perseverance (Calışkan, 2023); thus, there are concerns that ChatGPT would reduce the need for students to develop these skills.

When approaching how we teach EDS in the world of LLMs, it may be helpful to consider an asset-based approach (Alim et al., 2020). This entails teaching practices and instructional materials that promote a variety of approaches to knowing and doing. While LLMs can suggest new approaches, a risk with using

Box 2. Discussion prompts for responsibly using LLMs in EDS research. These prompts were developed by the team of authors as a starting place for researchers reflecting on the use of LLMs.

In research communities

Responsibly and effectively using LLMs in research requires a thoughtful consideration of the limitations of these tools in the specific context of the research question posed. On their own, within their lab, and among their professional networks, environmental data scientists should consider the prompts below.

► What are some biases in the type of data that I use that LLMs might not be aware of? Are there caveats to these data due to how they were collected? Are there gaps in these data due to historical underrepresentation of some geographic locations and/or groups of people? Is there a way to make LLMs aware of these limitations and, if not, will this lack of awareness overly bias the LLM's responses?

To responsibly use LLMs, environmental data scientists must have sufficient knowledge of the biases and caveats of their specific datasets and 1) prime LLMs to acknowledge these biases and/or caveats in analyses, 2) interpret the outputs of LLMs in light of these biases and/or caveats. For example, in analyses of spatial data on biodiversity, LLMs are unlikely to be aware of the impact of systemic racism on biodiversity observations (Ellis-Soto et al., 2023).

► Have there been any technical changes that have occurred in the analytical approach that I use that an LLM would not be aware of? Are there new packages, or packages that no longer exist? Are there best practices that have been developed since the LLM training data was collected?

Researchers must know enough about their field to understand the technical limitations of LLMs due to a lag in the time period covered by the training data. For example, LLMs might provide code that relies on a package that has since been deprecated. There may also be changes in the language used to discuss certain phenomena, and environmental data scientists must be aware of these changes to best practices so as to not perpetuate problematic language. For example, biology education researchers no longer use the term “achievement gaps,” as this term places blame on the student, but instead now use “opportunity gaps.”

► How will I identify errors created by LLMs? Will I use code generated by LLMs, even if I do not personally understand what each line of code does? If I use LLMs to help with literature review, how will I ensure that none of the information or resources that it produces are “hallucinated” or the novel ideas it generates are not plagiarized?

Environmental data scientists must be able to critically assess the outputs provided by LLMs. These models cannot be held accountable for any errors produced in the coding or writing process, and the responsibility still ultimately lies with scientists to ensure the accuracy of their work. LLMs may produce code or rely on models that are inappropriate for the researcher's data or may make up or “hallucinate” scientific references or statistical packages, particularly in the context of newer or less commonly used types of analyses for which LLMs have less training data. As they are trained on information available online, text taken directly from LLMs could also lead to accidental plagiarism. As a result, research outputs should not include code or text taken directly from LLMs without editing and verification (Buriak et al., 2023).

► What policies do I need to be aware of regarding generative AI? If I am reviewing a paper or proposal, would uploading the text to a company's LLM for guidance go against confidentiality agreements? Is the data that I am using confidential or sensitive in any way and, if so, is it ethical to upload these data into an LLM platform? If I use an LLM to help write articles, what is the benchmark that I will use to disclose this use of generative AI?

If environmental data scientists choose to use LLMs in their research and professional service, they should be fully aware of the current policies on use of these models (A Quick Guide of Using GAI for Scientific Research, 2023). At the proposal stage, environmental data scientists are fully accountable for any proposed work that they receive funding for. This landscape is changing rapidly, and agencies

that do not currently explicitly prohibit use of LLMs in their proposals may do so in the near future. LLMs should not be used to help review papers or proposals, as reviewers are expected to treat these items as confidential. Additionally, some types of data (e.g., healthcare data; (Varghese and Chapiro, 2023)) may also have expectations of confidentiality that preclude use of LLMs for analysis. Researchers should receive consent before using LLMs to analyze data related to individuals or communities. When submitting manuscripts, authors should ensure that they are aware of journal policies on use of LLMs and, when LLMs are not prohibited, should be fully transparent about how and where these models were used as a tool and of any limitations that these tools may have created in their work. As with granting agencies, these policies currently differ across journals and may be subject to change.

LLMs in this context is that (as mentioned earlier) training data can be biased (Wich et al., 2020), which may lead to a more privileged set of students' needs being met. Furthermore, such tools can be expensive and potentially further economic inequity. Even non-profit systems such as Khan Academy's GPT-4-powered AI tutor are so expensive to run that individual students wishing to access the system are required to donate \$20 to help offset the cost (Khan, 2023).

We urge a transparent and cautionary approach for instructors using LLMs in classroom settings (Eager and Brunton, 2023; Meyer et al., 2023). In deciding how and when to use LLMs in the classroom, instructors should consider the technical background of their students and especially ensure that their students understand the limitations of LLMs. They should also consider their learning goals for the course and use these goals to decide which assessments will allow use of LLMs. Once they have made these decisions, instructors should be explicit with their students about their expectations regarding the use of LLMs and should provide some basic training on the use of these tools (e.g., specificity when prompting for a code example) so that students have equal access to them (Box 3).

6. How will LLMs impact how we engage with one another?

The issues raised above pose larger questions of how we engage with each other as a community of practice. As LLMs are increasingly used to support analysis, interpretation, and decision-making, whom do we hold accountable when things go awry? How far upstream do we hold individuals responsible? What is our responsibility in shaping how AI uses the information that we generate? Given that these technologies do not credit the sources of the knowledge they repurpose, there is a potential threat to open science principles and practices. For example, if EDS practitioners decide to withhold their data or insights in order to preclude their use in training LLMs, then the use and reuse of those data and results would likely be restricted for use by human researchers as well.

In terms of accountability, the January 2023 policy for *Nature* family journals states that "... no LLM tool will be accepted as a credited author on a research paper. That is because any attribution of authorship carries with it accountability for the work, and AI tools cannot take such responsibility" (No authors, 2023). Presumably this type of approach also applies to code generated for the purposes of scientific analysis. LLMs may produce outputs based on inputs that were not intended to be re-used, compared to resources like StackOverflow, where contributions are explicitly volunteered for reuse. More broadly, citation and attribution is an important way of understanding the origin and credibility of statements in research. The inability of LLMs like ChatGPT (and AI-based coding assistants like GitHub's CoPilot) to properly cite their sources suggests a need for more specialized tools that can provide proper attribution. Attribution tools are important for scientists not only for assessing the veracity of provided outputs, but also to maintain legal compliance with copy-left licenses like GPL and CC-BY-SA. To that note, some LLMs like Google's Bard or Ought's Elicit can already properly cite sources (web URLs and scientific articles, respectively), although these tools are still in early development.

Box 3. Discussion prompts for responsible use of LLMs in EDS teaching. These prompts were developed by the team of authors as a starting place for researchers reflecting on the use of LLMs.

In Teaching Communities

Educators must consider how they will treat the use of LLMs in their classes. The prompts below provide a starting point for reflection for individual educators as well as in dialogue with educators at learning centers, workshops, and conferences.

► How can I ensure that my students have the necessary *conceptual* background to responsibly use LLMs? Are they aware of the potential biases and shortcomings of LLMs as they relate to EDS? What are specific examples that I can use that will resonate with them?

EDS students are increasingly interested in issues of environmental justice, but are unlikely to understand the ramifications of LLM use in this context. Educators must ensure that students are aware of the costs and biases associated with LLM development and use. They should provide students with conceptually and culturally relevant examples and understand that these examples may not be common enough in an LLM's training corpus for the LLM to generate helpful responses.

► How can I ensure that my students have the necessary *technical* background to effectively use LLMs in their work? Do they have the technical language necessary to prompt an LLM to help with coding and other analytical tasks? Are they able to identify and troubleshoot issues in code?

LLMs do have potential to benefit environmental data scientists, from novices to experienced users. LLMs are most effective when prompts are specific and small (Merow et al., 2023). Students without the necessary technical background may not know how to properly prompt LLMs, and if their prompts fail, they may not understand why. Additionally, this technical background may be uneven among students, in which case LLMs may further exacerbate opportunity gaps. Even when prompted correctly, LLMs often produce flawed code that may only get the user 80-90% of the way to functional code (Merow et al., 2023). Debugging code is a difficult technical skill and it may be more difficult for new learners to start with code that is 80% correct than to start with a blank script.

► What do I hope my students take from this class? Is it important that they know which type of analyses to run, how to run these analyses, or how to interpret the output? What type of coursework have they done prior to taking this class?

This prompt motivates educators to think from a perspective of backwards design, whereby educators design assessments specifically to match previously outlined learning objectives (Vanderbilt University, 2023). These learning objectives will differ depending on the course. For example, one of the goals with a 100-level data science course is likely to introduce students to commonly used packages, functions, and practices. The use of LLMs would have different implications for this type of course than for an upper-level course focused on more advanced and specific learning objectives.

► How will I make my expectations about the use of LLMs clear to students? What are my “cut-offs” for cheating, and what will the repercussions be? How do I want students to report when they’ve used LLMs? If I allow use of LLMs, do all of my students have equal access to these tools?

Some educators may choose to prohibit all use of LLMs in their courses, as has been the stance of several academic journals. A complete ban on the use of these tools is likely to be unrealistic and detrimental to students, who will increasingly come in contact with LLMs in their daily lives. When educators do choose to allow use of LLMs, they should be clear about their policies from the start of class and, where possible, work together to develop policies that are consistent across courses (within departments, schools, institutions). This collection of Syllabi Policies for AI Generative Tools provides a useful overview of educator policies for the use of LLMs in their courses.

Another potential path forward is the integration of LLMs with more mechanistic or symbolic systems like Wolfram Alpha. OpenAI, the developer of ChatGPT, has recently added the ability for ChatGPT to interact with “plugins,” including one for Wolfram Alpha, that query other services and return results, but these plugins are still in early development. These plugins have already started to lead to new applications of LLMs in novel areas, like task automation with auto-GPT, but these tools have the potential to fundamentally alter how we communicate as scientists if everything from calendar invites to emails to even Slack messages may be generated by ChatGPT rather than by our colleagues. While these integrated LLM tools are nascent, their profound implications warrant proactive discussion around community norms concerning communication in an age where any piece of text we read in any context could theoretically have been generated by an AI model.

7. What are the costs of LLMs?

While we have yet to fully grasp the opportunities enabled by ChatGPT and other LLMs in the environmental domain, if these tools deliver on the promise to rapidly synthesize concepts and generate complex code, they stand to transform society’s capacity to address some of our most pressing and complex planetary crises by allowing more people to engage on these interdisciplinary problems (Biswas, 2023). While the power of these tools may at times seem a panacea, it is critical to bear in mind that they exist within the physical and social realities of the world. In using these tools, we must consider the ethical implications of the costs of their development, which reflect and are likely to exacerbate existing class and racial injustices.

For example, substantial human labor is needed to counteract the dissemination of the darkest aspects of the internet within LLMs (Rome, 2023; Time, 2023). Because ChatGPT primarily relies on text scraped from the internet, OpenAI leverages RLHF and a tool developed to detect violent, sexist, or racist language and preclude its inclusion in the training corpus. To learn to detect this language, the tool required many examples of violence, sexual abuse, and hate speech that had been coded as such by human workers. OpenAI outsourced labor for this task to workers in Kenya, paying <\$2 per hour for the identification of harmful text samples. Some workers were made physically ill by these images and descriptions of violence and abuse (Ellis-Soto et al., 2023). Unfortunately, this outsourcing is a common practice for training AI models that need large sets of labeled data, presenting a social and ethical cost largely unseen by end users (Miceli and Posada, 2021). As an alternative censoring approach, rule-based, or “constitutional,” AI might help reduce reliance on large amounts of human-classified training data, but would also require subjective decisions about what to include in these “constitutions” (Anthropic, 2023).

Developing, training, and running ChatGPT has an environmental cost as well. The training of GPT-3 led to an estimated 552 metric tons of CO₂ equivalent emissions, roughly the same as generated by three round-trip jet plane flights between San Francisco and New York. ChatGPT has an estimated daily carbon footprint at the scale of 23 kg CO₂e (Patterson et al., 2022; Ludvigsen, 2023). These costs will likely grow with the use of ChatGPT, but are still currently orders of magnitude lower than the carbon emissions attributed to airborne travel to conferences (Patterson et al., 2021). Nevertheless, more transparent and rigorous accounting of efficiency gains versus energy usage is needed, especially for LLMs hosted by private companies that currently have no mandatory emissions reporting requirements (Bender et al., 2021).

In addition to the human and environmental costs of developing, training, and maintaining ChatGPT, its use, and implementation in environmental data science raises several considerations around policy dynamics and imbalances. Leveraging ChatGPT for addressing our most pressing planetary crises risks handing over power historically held by government agencies and resource users to the companies that develop these tools (Kalluri, 2020). Disparities in underlying data and decisions about model design and training stand to create not only biased content but also biased research agendas and biased researcher

ideologies. For example, it is not difficult to imagine a case where the political and financial interests of big tech companies are misaligned with global environmental stewardship.

Finally, access to LLMs will likely be inequitable. Currently, ChatGPT is free to use (assuming access to a computing device and the internet) but OpenAI also sells a paid tier subscription (20 USD/month) that provides priority access, faster response times, and access to GPT-4 (vs. 3.5 in the free version). For larger queries to the OpenAI application programming interface (API), OpenAI charges per-token (currently 3 USD for every 1,000 tokens for GPT-4 and 15 USD for every 1,000 tokens for GPT-3.5) (<https://openai.com/pricing>). While prices are low per-token for now, with ever-larger datasets and more complicated queries and data manipulation steps, costs can quickly add up and introduce disparities between the volume of data that researchers can afford to manipulate with privately-held LLMs. Furthermore, it is unclear how the costs of using these tools will evolve in the future and whether or not access to higher-performing tools will mirror current inequities in access to proprietary computational software (Ghose and Welcenbach, 2018). On top of many other ethical issues with these products, access to these tools might be the most immediate way that this technology reproduces or exacerbates existing equity issues in the field of EDS.

8. Where do we go from here?

While the future of LLMs is not yet clear, what is certain is that these technologies are here to stay. There are substantial social, ethical, and environmental issues with the creation and maintenance of these tools, and their adoption will have potentially profound implications for the practice of science. Thus, our community needs to not only proactively consider the implications of their use for our own research and training tasks, but to also consider how we might leverage our uniquely human perspectives to guide the development of these tools and their use more broadly. Tools like ChatGPT can assist EDS practitioners in the development of ideas and translating them into code. However, creating meaning from the output still requires expert knowledge. While LLMs may save time on certain tasks, their use runs the risk of propagating biases embedded in their training and therefore requires caution. As a community, we see the need for greater dialogue about the issues raised by the emerging use of LLMs. We offer a series of discussion prompts for engaging with community members within a range of contexts: within labs, classrooms, departments, and conferences (Box 2, 3).

Taken together, these changes pose a fundamental question about the nature of our roles as environmental data scientists going forward. Ultimately, LLMs are built off of human ideas, and the continued relevance of these tools depends on the sustained generation and distribution of novel ideas. Without this, scientific insight will stagnate and scientific thinking will become increasingly homogenized. Even while we as scientists continue to do what models cannot yet accomplish—to innovate and to think critically about the challenging environmental issues of our time—we need to comprehensively reassess the skills that are needed to succeed in these roles in the age of LLMs.

Acknowledgments. We would like to thank the Steering Committee and participants of the Environmental Data Science Summit hosted by the National Center for Ecological Analysis and Synthesis, with support from NSF RCN grant #2021151.

Author contribution. All co-authors contributed to the conceptualization of the work. R.O., M.C., N.E., L.G., N.G., S.L., and A.N. contributed to the writing of the original draft. All co-authors reviewed the work.

Competing interest. The authors declare none.

Data availability statement. This article did not incorporate any data.

Funding statement. The concept for this article originated at the meeting hosted with support from NSF RCN grant #2021151.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/eds.2024.12>.

References

- A Quick Guide of Using GAI for Scientific Research [Internet] (2023) MIDAS [cited 2023 Oct 9]. Available at <https://midas.umich.edu/generative-ai-user-guide/> (accessed 28 February 2024).
- Alim HS, Paris D and Wong CP (2020) Culturally sustaining pedagogy: A critical framework for centering communities. In *Handbook of the Cultural Foundations of Learning*. New York, NY: Routledge.
- Anthropic [Internet] (2023) Measuring Progress on Scalable Oversight for Large Language Models [cited 2023 May 1]. Available at <https://www.anthropic.com/index/measuring-progress-on-scalable-oversight-for-large-language-models> (accessed 28 February 2024).
- Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B and Kaplan J (2022) Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback [Internet]. [cited 2023 Oct 23]. Available at <http://arxiv.org/abs/2204.05862> (accessed 28 February 2024).
- Baidoo-Anu D and Owusu Ansah L (2023) Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning [Internet]. Rochester, NY [cited 2023 May 1]. Available at <https://papers.ssrn.com/abstract=4337484> (accessed 28 February 2024).
- Bender EM, Gebru T, McMillan-Major A and Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* [Internet] [cited 2023 Aug 16]. Virtual Event Canada: ACM, pp. 610–623. doi:10.1145/3442188.3445922.
- Biswas SS (2023) Potential use of chat GPT in global warming. *Annals of Biomedical Engineering* 51, 1126–1127 [Internet]. [cited 2023 May 1]. doi:10.1007/s10439-023-03171-8.
- Bolukbasi T, Chang KW, Zou JY, Saligrama V and Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc [cited 2023 Oct 2]. Available at https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html (accessed 28 February 2024).
- Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Wei Koh P, Krass M, Krishna R, Kudipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li XC, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Sung Park J, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K and Liang P (2022) On the Opportunities and Risks of Foundation Models [Internet]. arXiv; 2022 [cited 2023 May 5]. Available at <http://arxiv.org/abs/2108.07258> (accessed 28 February 2024).
- Broecker WS (1975) Climatic change: Are we on the brink of a pronounced global warming? *Science* 189(4201), 460–463.
- Brown LA, Meier C, Morris H, Pastor-Guzman J, Bai G, Lerebourg C, Gobron N, Lanconelli C, Clerici M and Dash J (2020) Evaluation of global leaf area index and fraction of absorbed photosynthetically active radiation products over North America using Copernicus ground based observations for validation data. *Remote Sensing of Environment* 247, 111935.
- Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* [Internet] [cited 2023 Oct 2]. PMLR, pp. 77–91. Available at <https://proceedings.mlr.press/v81/buolamwini18a.html> (accessed 28 February 2024).
- Buriak JM, Akinwande D, Artzi N, Brinker CJ, Burrows C, Chan WCW, Chen C, Chen X, Chhowalla M, Chi L, Chueh W, Crudden CM, Di Carlo D, Glotzer SC, Hersam MC, Ho D, Hu TY, Huang J, Javey A, Kamat PV, Kim I-D, Kotov NA, Lee TA, Lee YH, Li Y, Liz-Marzán LM, Mulvaney P, Narang P, Nordlander P, Oklu R, Parak WJ, Rogach AL, Salanne M, Samori P, Schaak RE, Schanze KS, Sekitani T, Skrabalak S, Sood AK, Voets IK, Wang S, Wang S, Wee ATS and Ye J (2023) Best practices for using AI when writing scientific manuscripts. *ACS Nano* 17(5), 4091–4093.
- Calışkan E (2023) The effects of robotics programming on secondary school students' problem-solving skills. *World Journal on Educational Technology: Current Issues* 12(4), 217–230 [Internet]. [cited 2023 May 1]. Available at <https://un-pub.eu/ojs/index.php/wjet/article/view/5143> (accessed 28 February 2024).
- Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334), 183–186.
- ChatGPT is a Data Privacy Nightmare, and We Ought to be Concerned Ars Technica [Internet]. [cited 2023 May 1] (2023). Available at <https://arstechnica.com/information-technology/2023/02/chatgpt-is-a-data-privacy-nightmare-and-you-ought-to-be-concerned/> (accessed 28 February 2024).
- Chen E (2023) Human Evaluation of Large Language Models: How Good is Hugging Face's BLOOM? [Internet]. [cited May 1]. Available at <https://www.surgehq.ai/blog/how-good-is-hugging-faces-bloom-a-real-world-human-evaluation-of-language-models> (accessed 28 February 2024).

- Chung AW (2019) How Automated Tools Discriminate Against Black Language – MIT Center for Civic Media [Internet] [cited 2023 May 5]. Available at <https://civic.mit.edu/index.html%3Fp=2402.html> (accessed 28 February 2024).
- Clark E, August T, Serrano S, Haduong N, Gururangan S and Smith NA (2021) All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [Internet] [cited 2023 May 1]. Association for Computational Linguistics, pp. 7282–7296. Available at <https://aclanthology.org/2021.acl-long.565> (accessed 28 February 2024).
- Cu MA and Hochman S (2023) Scores of Stanford students used ChatGPT on final exams [Internet] [cited 2023 May 1]. Available at <https://stanforddaily.com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests/> (accessed 28 February 2024).
- Eager B and Brunton R (2023) Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice* [Internet] 20(5), 2. Available at <https://ro.uow.edu.au/jutlp/vol20/iss5/02> (accessed 28 February 2024).
- Ellis-Soto D, Chapman M and Locke DH (2023) Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nature Human Behaviour* 7, 1–9.
- Ferrara E (2023a) Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models [Internet]. arXiv; 2023 [cited 2023 Oct 2]. Available at <http://arxiv.org/abs/2304.03738> (accessed 28 February 2024).
- Ferrara E (2023b) The Butterfly Effect in Artificial Intelligence Systems: Implications for AI Bias and Fairness [Internet]. arXiv; 2023 [cited 2023 Sep 26]. Available at <http://arxiv.org/abs/2307.05842> (accessed 28 February 2024).
- Future of Life Institute (2023) Pause Giant AI Experiments: An Open Letter [Internet]. [cited 2023 May 1]. Available at <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (accessed 28 February 2024).
- Gankin D, Karollus A, Grosshauser M, Klemon K, Hingerl J and Gagneur J (2023) Species-aware DNA language modeling [Internet]. bioRxiv; [cited 2023 May 1]. p. 2023.01.26.525670. doi:10.1101/2023.01.26.525670v1.
- Getahun H (2023) Insider. ChatGPT could be used for good, but like many other AI models, it's rife with racist and discriminatory bias [cited May 1]. Available at <https://www.insider.com/chatgpt-is-like-many-other-ai-models-rife-with-bias-2023-1> (accessed 28 February 2024).
- Ghose R and Welchenbach T (2018) Power to the people": Contesting urban poverty and power inequities through open GIS. *Canadian Geographies / Géographies Canadiennes* 62(1), 67–80.
- Huddleston A (2023) Happy 200th birthday to Eunice Foote, hidden climate science pioneer | NOAA Climate.gov [Internet]. [cited May 2]. Available at <http://www.climate.gov/news-features/features/happy-200th-birthday-eunice-foote-hidden-climate-science-pioneer> (accessed 28 February 2024).
- Jiao W, Wang W, Huang J tse, Wang X, Shi S and Tu Z (2023) Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine [Internet]. arXiv; 2023 [cited 2024 Jan 4]. Available at <http://arxiv.org/abs/2301.08745> (accessed 28 February 2024).
- Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583(7815), 169.
- Keuning H, Jeuring J and Heeren B (2019) A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education* 19(1), 1–43.
- Khan S (2023) Khan Academy Blog. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access [cited 2023 May 5] Available at <https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/> (accessed 28 February 2024).
- Kirk HR, Jun Y, Volpin F, Iqbal H, Benussi E, Dreyer F, Shtedritski A and Asano YM (2021) Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc, pp. 2611–2624. [cited 2023 Oct 2]. Available at <https://proceedings.neurips.cc/paper/2021/hash/1531beb762df4029513ebf9295e0d34f-Abstract.html> (accessed 28 February 2024).
- Krügel S, Ostermaier A and Uhl M (2023) ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports* 13 (1), 4569.
- Liu NF, Zhang T and Liang P (2023) Evaluating Verifiability in Generative Search Engines [Internet]. arXiv; 2023 [cited 2023 May 5]. Available at <http://arxiv.org/abs/2304.09848> (accessed 28 February 2024).
- Ludvigsen KGA (2023) Medium. The Carbon Footprint of ChatGPT [cited 2023 May 5]. Available at <https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d> (accessed 28 February 2024).
- Madkins TC, Howard NR and Freed N (2020) Engaging equity pedagogies in computer science learning environments. *Journal of Computer Science Integration* 3(2), 1.
- Marcus G (2023) Scientific American. AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous [cited 2023 May 1]. Available at <https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/> (accessed 28 February 2024).
- McGovern A, Ebert-Uphoff I, Gagne DJ and Bostrom A (2022) Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science* 1, e6.
- Merow C, Serra-Diaz JM, Enquist BJ and Wilson AM (2023) AI chatbots can boost scientific coding. *Nature Ecology & Evolution* 7(7), 960–962.
- Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G and Moore JH (2023) ChatGPT and large language models in academia: Opportunities and challenges. *Biodata Mining* 16(1), 20.

- Miceli M and Posada J (2021) Wisdom for the Crowd: Discursive Power in Annotation Instructions for Computer Vision [Internet]. arXiv 2021; [cited 2023 May 2]. Available at <http://arxiv.org/abs/2105.10990> (accessed 28 February 2024).
- Miceli M, Posada J and Yang T (2022) Studying up machine learning data: Why talk about bias when we mean power? In *Proceedings of the ACM on Human-Computer Interaction*.
- Milano S, McGrane JA and Leonelli S (2023) Large language models challenge the future of higher education. *Nature Machine Intelligence* 5(4), 333–334.
- Mills A, Bali M and Eaton L (2023) How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process. *Journal of Applied Learning and Teaching* 6(1), 16–30.
- MIT Technology Review [Internet] (2023) Computational Linguistics Reveals How Wikipedia Articles Are Biased Against Women. Available at <https://www.technologyreview.com/2015/02/02/169470/computational-linguistics-reveals-how-wikipedia-articles-are-biased-against-women/> (accessed 28 February 2024).
- Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID and Gebru T (2019) Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* [Internet]. New York, NY: Association for Computing Machinery, pp. 220–229 [cited 2023 May 5]. doi:10.1145/3287560.3287596.
- (2023) Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* 613(7945), 612.
- NBC News [Internet] (2023) ChatGPT banned from New York City public schools' devices and networks [cited 2023 May 1]. Available at <https://www.nbcnews.com/tech/tech-news/new-york-city-public-schools-ban-chatgpt-devices-networks-rcna64446> (accessed 28 February 2024).
- Passi S and Vorvoreanu M (2022) Overreliance on AI Literature Review. Microsoft Research.
- Patterson D, Gonzalez J, Hölzle U, Le Q, Liang C, Munguia LM, Rothchild D, So D, Texier M and Dean J (2022) The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.
- Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, So D, Texier M and Dean J (2021) Carbon Emissions and Large Neural Network Training [Internet]. arXiv; 2021 [cited 2023 May 5]. Available at <http://arxiv.org/abs/2104.10350> (accessed 28 February 2024).
- Perry N, Srivastava M, Kumar D and Boneh D (2022) Do users write more insecure code with AI assistants? [Internet]. arXiv; 2022 [cited 2023 Oct 23]. Available at <http://arxiv.org/abs/2211.03622> (accessed 28 February 2024).
- Popenici SAD, Kerr S (2017) Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning* 12(1), 22.
- Pushkarna M, Zaldivar A and Kjartansson O (2022) Data cards: Purposeful and transparent dataset documentation for responsible AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* [Internet]; [cited 2023 May 5]. New York, NY: Association for Computing Machinery, pp. 1776–1826. doi: 10.1145/3531146.3533231.
- Rillig MC (2023) How can AI-powered language models (like ChatGPT) be useful for researchers in the environmental sciences? [Internet]. The Ecological Mind: ecology, environment, research. [cited 2023 May 1]. Available at <https://matthiasrillig.substack.com/p/how-can-ai-powered-language-models> (accessed 28 February 2024).
- Rome TK (2023) ChatGPT bot tricked into giving bomb-making instructions, say developers. [cited 2023 May 5]; Available at <https://www.thetimes.co.uk/article/chatgpt-bot-tricked-into-giving-bomb-making-instructions-say-developers-rvktrxb5> (accessed 28 February 2024).
- Samuel Y, George J and Samuel J (2020) Beyond STEM, How Can Women Engage Big Data, Analytics, Robotics and Artificial Intelligence? An Exploratory Analysis of Confidence and Educational Factors in the Emerging Technology Waves Influencing the Role of, and Impact Upon, Women [Internet]. arXiv; 2020 [cited 2023 May 1]. Available at <http://arxiv.org/abs/2003.11746> (accessed 28 February 2024).
- Sawyer JS (1972) Man-made carbon dioxide and the “greenhouse” effect. *Nature* 239(5366), 23–26.
- Schmidt V, Luccioni AS, Teng M, Zhang T, Reynaud A, Raghupathi S, Cosne G, Juraver A, Vardanyan V, Hernandez-Garcia A and Bengio Y (2021) ClimateGAN: Raising Climate Change Awareness by Generating Images of Floods [Internet]. arXiv; 2021 [cited 2023 Aug 16]. Available at <http://arxiv.org/abs/2110.02871> (accessed 28 February 2024).
- Shaw A and Hargittai E (2018) The pipeline of online participation inequalities: The case of Wikipedia editing. *The Journal of Communication* 68(1), 143–168.
- Shen X, Chen Z, Backes M and Zhang Y (2023) In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT [Internet]. arXiv; 2023 [cited 2023 Sep 27]. Available at <http://arxiv.org/abs/2304.08979> (accessed 28 February 2024).
- Time [Internet] (2023) Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer [cited 2023 May 1]. Available at <https://time.com/6247678/openai-chatgpt-kenya-workers/> (accessed 28 February 2024).
- Tomašev N, Cornebise J, Hutter F, Mohamed S, Picciariello A, Connelly B, Belgrave DCM, Ezer D, van der Haert FC, Mugisha F, Abila G, Arai H, Almiraat H, Proskurnia J, Snyder K, Otake-Matsuura M, Othman M, Glasmachers T, de Wever W, Teh YW, Khan ME, De Winne R, Schaul Tand Clopath C (2020) AI for social good: Unlocking the opportunity for positive impact. *Nature Communications* 11(1), 2468.
- van Dis EAM, Bollen J, Zuidema W, van Rooij R and Bockting CL (2023) ChatGPT: Five priorities for research. *Nature* 614 (7947), 224–226.
- Vanderbilt University [Internet] (2023) Understanding by Design [cited 2023 Oct 9]. Available at <https://cft.vanderbilt.edu/guides-sub-pages/understanding-by-design/> (accessed 28 February 2024).

- Varghese J and Chapiro J (2023) ChatGPT: The transformative influence of generative AI on science and healthcare. *Journal of Hepatology* [Internet] [cited 2023 Oct 9]. Available at <https://www.sciencedirect.com/science/article/pii/S0168827823050390> (accessed 28 February 2024).
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is All You Need [Internet]. arXiv; 2017 [cited 2023 May 1]. Available at <http://arxiv.org/abs/1706.03762> (accessed 28 February 2024).
- Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z, Brown S, Hawkins W, Stepleton T, Biles C, Birhane A, Haas J, Rimell L, Hendricks LA, Isaac W, Legassick S, Irving G and Gabriel I (2021) Ethical and social risks of harm from Language Models [Internet]. arXiv; 2021 [cited 2023 Sep 27]. Available at <http://arxiv.org/abs/2112.04359> (accessed 28 February 2024).
- Wich M, Bauer J and Groh G (2020) Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* [Internet]. Association for Computational Linguistics [cited 2023 May 1], pp. 54–64. Available at <https://aclanthology.org/2020.alw-1.7> (accessed 28 February 2024).
- Wilcox S (1996) Fostering self-directed learning in the university setting. *Studies in Higher Education* 21(2), 165–176.
- World Economic Forum [Internet] (2023) Here's why ChatGPT raises issues of trust [cited 2023 Sep 25]. Available at <https://www.weforum.org/agenda/2023/02/why-chatgpt-raises-issues-of-trust-ai-science/> (accessed 28 February 2024).
- Zamfirescu-Pereira JD, Wong R, Hartmann B and Yang Q (2023) Why Johnny can't prompt: How non-ai experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI[CMT15] Conference on Human Factors in Computing Systems* [Internet]. [cited 2023 May 1]. doi:10.1145/3544548.3581388.
- ZDNET [Internet] (2023) With GPT-4, OpenAI opts for secrecy versus disclosure [cited 2023 May 5]. Available at <https://www.zdnet.com/article/with-gpt-4-openai-opts-for-secrecy-versus-disclosure/> (accessed 28 February 2024).
- Zombies in the Loop? (2023) Humans Trust Untrustworthy AI-Advisors for Ethical Decisions | SpringerLink [Internet]. [cited 2023 May 1]. doi:10.1007/s13347-022-00511-9.

Cite this article: Oliver RY, Chapman M, Emery N, Gillespie L, Gownaris N, Leiker S, Nisi AC, Ayers D, Breckheimer I, Blondin H, Hoffman A, Pagnello CMLS., Raisle M and Zimmerman N (2024). Opening a conversation on responsible environmental data science in the age of large language models. *Environmental Data Science*, 3: e14. doi:10.1017/eds.2024.12