Open Access

Advances in Environmental and Engineering Research

LIDSEN Publishing Inc.

Review

Applications of Generative Large Language Models in Environmental Science: A Systematic Review

Masoume M. Raeissi *, Rob Knapen

Environmental Sciences Group, Wageningen University & Research, Droevendaalsesteeg 3 6708PB, The Netherlands; E-Mails: masoume.raeissi@wur.nl; rob.knapen@wur.nl

* **Correspondence:** Masoume M. Raeissi; E-Mail: masoume.raeissi@wur.nl

Academic Editor: José L. Segovia-Juárez

Special Issue: Artificial Intelligence in Environmental Research

Adv Environ Eng Res
2025, volume 6, issue 3
doi:10.21926/aeer.2503028

Received: April 07, 2025
Accepted: July 17, 2025
Published: August 12, 2025

Abstract

Environmental science addresses critical global challenges, including climate change, biodiversity loss, and sustainability. These complex topics generate a vast amount of both structured and unstructured data, from remote sensing output to policy documents, guidelines, and scientific literature. Effectively processing and utilizing this information is essential for advancing research and supporting decision-making. Large Language Models (LLMs) have shown remarkable capabilities in natural language understanding and generation across various domains. They offer a promising solution for extracting insights and synthesising knowledge and present potential benefits for environmental science research and engineering. This study presents a systematic review of 51 peer-reviewed articles on the use of LLMs within environmental science. We analyze their applications, usage types, and the challenges or limitations identified by researchers. Key trends show that knowledge extraction is the most common application, with climate science being the dominant domain. Our findings map the current landscape, highlight research gaps, and outline open problems that need to be addressed in future work. This review serves as a resource for researchers applying LLMs in environmental contexts, supporting more effective and informed decision making.

Keywords
Large language models; environmental science; generative artificial intelligence

1. Introduction

The increasing urgency of environmental challenges, from climate change and biodiversity loss to sustainable resource management, demands innovative solutions. However, the large amount and complexity of environmental data, such as scientific papers, policy documents, citizen science reports, and remote sensing data, create significant challenges for researchers and policy makers. For example, aligning national biodiversity targets with global goals remains complex due to the diversity of data [1]. Traditional retrieval methods for finding information about environmental events are inefficient [2], and addressing climate misinformation requires scalable AI tools that can operate at a large scale [3]. Climate change communication differs widely across cultures and ideological contexts, influencing public perception and policy responses [4].

Given these conditions, there is a growing need for tools that can make sense of vast, fragmented, and constantly evolving environmental knowledge. Generative Large Language Models (LLMs), such as GPT [5], PaLM [6], and LLaMA [7], offer a new paradigm for addressing this need by enabling scalable analysis, synthesis, and interaction with textual environmental data. Their ability to extract insights, generate summaries, and interact via natural language makes them highly relevant for supporting environmental decision-making, public engagement, and interdisciplinary research workflows.

Environmental science, by nature, spans disciplines and requires integrating information from diverse sources—scientific literature, legal regulations, policy documents, sensor data annotations, and public discourse. LLMs are uniquely positioned to handle this diversity, not only by retrieving or summarizing content, but also by generating new hypotheses, detecting misinformation, or translating technical findings into accessible language. This makes them a potentially transformative tool in bridging gaps between science, policy, and society.

Despite their promise, the integration of LLMs into environmental science remains limited and underexplored. This gap raises critical questions about their practical applications, limitations, and reliability in addressing environmental problems. Our work seeks to fill this gap by systematically reviewing existing studies that apply generative LLMs in environmental contexts.

The goal is to illustrate how LLMs are being used in various environmental challenges, offering practitioners a clearer understanding of their potential use cases. For each study, we analyse the specific environmental science topic, how LLMs were applied in the study, the limitation identified in the implementation, and the type of evaluation that was carried out.

Upon reviewing existing surveys on this topic, we found that our work stands out in several ways. Other surveys either concentrate on specific environmental science challenges or do not focus on LLMs. The study most similar to ours [8] utilized LLMs to extract the latest literature in environmental science, specifically focusing on studies published between 2021 and 2023, using the Web of Science database. In contrast, our study explicitly examines research that utilizes LLMs to address environmental science challenges. We conducted a comprehensive search across both

Scopus and Web of Science, spanning the period from 2018 to 2025. To summarise, the contributions of this paper are as follows:

- We present a systematic review focused specifically on the use of generative LLMs in environmental science, covering 51 peer-reviewed studies.
- We propose a structured categorisation of how LLMs have been applied, including use cases, evaluation methods, and common limitations across environmental domains.
- We provide a critical assessment of the strengths, weaknesses, opportunities, and threats of applying LLMs in environmental science, offering a practical reference point for future research, interdisciplinary collaboration, and responsible development.

This paper is organised as follows. Section 2 introduces the background and foundational concepts of LLMs. Section 3 details the research methodology used in this review. In Section 4, we present the results, structured around key research questions. Finally, in Section 5, we discuss our findings, and in Section 6, we conclude the article and outline future research directions.

2. Background: Large Language Models

Large Language Models are a class of artificial intelligence systems designed to understand and generate human language. Their development has been driven by advances in *Natural Language Processing (NLP)*, a subfield of artificial intelligence that focuses on enabling machines to process and interpret human language. NLP techniques such as statistical modelling, Recurrent Neural Networks (RNNs), and word embeddings laid the groundwork for the evolution of LLMs. However, these earlier models faced challenges in capturing long-range dependencies and efficiently processing large-scale text corpora.

A significant breakthrough in NLP came with the introduction of the *attention mechanism*, which allowed models to dynamically focus on the relevant parts of the input sequence during processing [9]. This innovation led to the development of the *Transformer* architecture, which replaced sequential computation in recurrent models with a fully parallelisable structure. Unlike RNNs, Transformers compute word relationships in a single step, drastically improving training efficiency and enabling models to process vast amounts of text tractably.

In 2018, BERT was introduced and quickly became widely adopted [10]. Although the original transformer has both encoder and decoder blocks, BERT is an encoder-only model. Around the same time, other types of LLMs, such as GPT (Generative Pre-trained Transformers), emerged, enabling task-solving through prompting–providing input text with instructions to a LLM to guide its response toward generating relevant, coherent, and contextually appropriate output. Unlike BERT, GPT models are decoder-only, leveraging their generative capabilities to create new content effectively.

The scalability of Transformers led to the realisation that model performance improves significantly with increased data, computational power, and model parameters. This principle is encapsulated in the scaling laws for LLMs [11], which demonstrated that larger models trained on more data with sufficient computation yield superior language understanding and generation capabilities. The ability to scale LLMs has resulted in models that achieve state-of-the-art performance on various NLP benchmarks. Table 1 shows the key concepts in LLMs.

Table 1 An overview of the key concepts related to Large Language Models.

Concept	Description
Transformer Architecture	A deep learning model structure that enables parallel processing of text, improving efficiency in understanding and generating language.
Tokenization	The process of breaking text into smaller units (words, subwords, or characters) to be processed by the model
Embedding	Each token is converted into an embedding vector.
Self-Attention Mechanism	A key feature of transformers that allows the model to weigh the importance of different words in a sentence to improve contextual understanding
Pretraining	The initial phase where an LLM is trained on vast amounts of text data
Fine-Tuning	The process of adapting a pretrained model to specific tasks or domains using labeled data to improve performance
Reinforcement Learning with Human Feedback (RLHF)	A training approach where human reviewers help guide the model's behavior by ranking outputs, improving alignment with human expectations
Prompt	Crafting inputs strategically to guide LLM responses

The following section outlines the methodology used to identify, select, and analyse relevant studies, ensuring a comprehensive and structured evaluation of existing research.

3. Methodology

This systematic review aims to answer four key research questions concerning the use of large language models (LLMs) in environmental science:

- What were the application domains, that is, the specific environmental science topics, for which LLM was used?
- What was the use case of LLMs, that is, how LLMs were utilized, along with the model type?
- What evaluation metrics were used to measure performance?
- What were the challenges and limitations addressed?

We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [12] to ensure methodological rigor in conducting and reporting this review. In line with these guidelines, we clearly defined our inclusion/exclusion criteria, sources of information, and selection procedures. The selection process is visually summarized in a PRISMA-compliant flowchart (Figure 1).

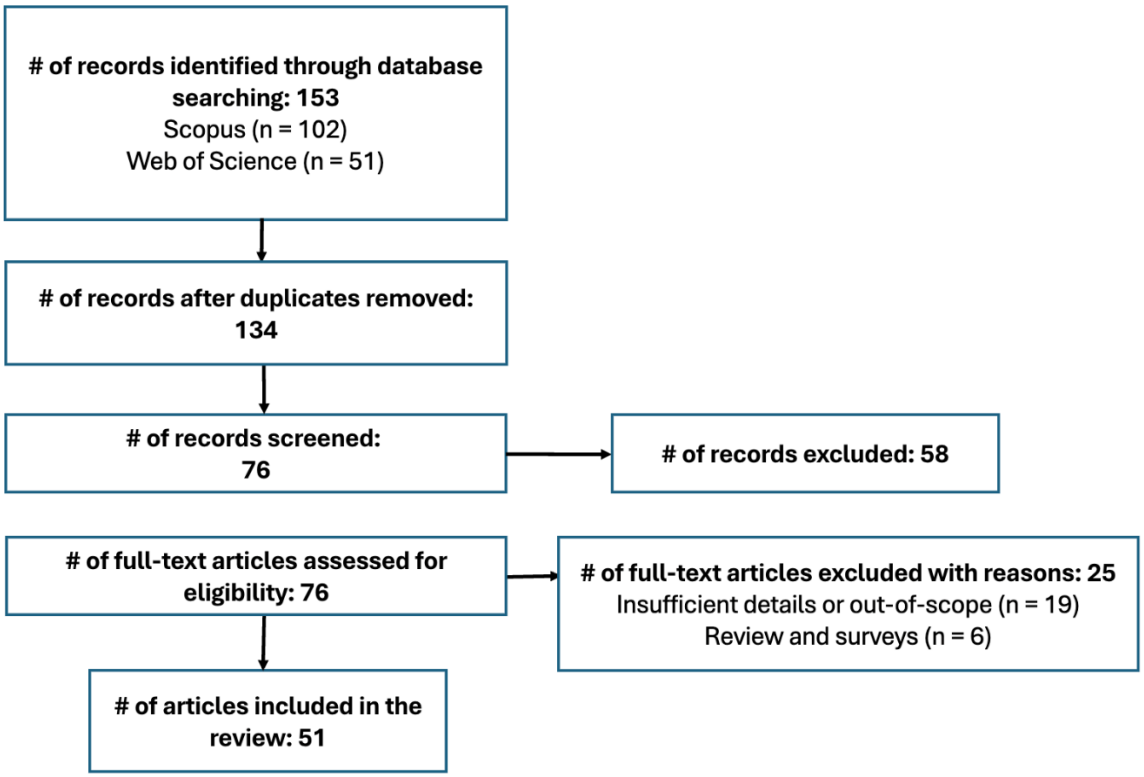

Figure 1 Diagram of the article selection process for evaluation.

Since our analysis focuses on the application domains and use cases of LLMs, we did not consider survey reviews in our evaluation. In Section 1, we clarified how our work differs from these other surveys.

To answer our research questions, we collected relevant articles from Scopus and Web of Science using the official APIs of each database. We applied identical search keywords across both databases, adjusting only for formatting differences. To narrow the scope, we focused only on generative large language models such as GPT, T5, LLaMA, PaLM, and more. For environmental science, our search included keywords related to environmental science, policy, communication, climate-related topics, and biodiversity as core areas. The term environmental science enables us to explore various subdomains and challenges, including energy, air, and water pollution, as well as other related issues. For each database, the title, abstract, and keywords were searched in February 2025.

3.1 Inclusion & Exclusion Criteria

For inclusion, we considered studies focusing on LLM applications in environmental science, published between 2018 and 2025. This time frame captures the introduction and adoption of generative LLMs. Only articles written in English were included. For exclusion, we did not consider editorials, conference reviews, or book chapters.

Since our analysis focuses on the application domains and use cases of LLMs, we did not consider survey reviews in our evaluation.

Figure 1 outlines the screening and selection process. Initially, 153 papers were retrieved (102 from Scopus and 51 from Web of Science). After removing duplicates and ineligible studies, 77

papers remained. Ineligibility was determined through a quick abstract scan, excluding papers where "LLM" did not refer to Large Language Models or where the term "environmental" was unrelated to environmental science topics.

We used ASReview [13] to accelerate the screening process before selecting papers for final inclusion. ASReview is an open source tool that uses active learning to prioritise relevant studies efficiently.

Figure 2 illustrates the distribution of publications over the years, highlighting a significant increase in 2024 and a rapid growth forecast for the future.

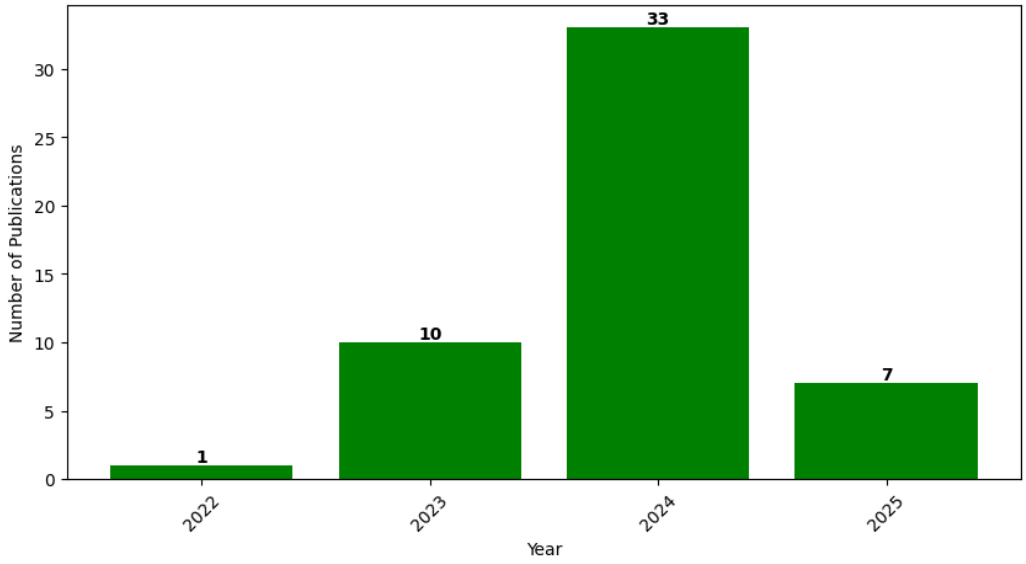

Figure 2 Number of publications per year.

3.2 Data Analysis Methods

For our final screening, we excluded the survey or review papers as they do not fit our analysis. To answer the research questions, we applied different strategies for each question. For example, for the first question, identifying environmental science topics, we used a combination of index keywords, author-provided keywords, and terms found in the abstracts. From these sources, we extracted all available keywords and manually selected those relevant to environmental science.

However, identifying LLM use cases required a more comprehensive approach, as relying solely on the provided keywords was insufficient. In this case, we also extracted keywords from the abstract and performed a manual review of the article to confirm relevance. Some studies included LLMs as part of a larger system; in those instances, we considered the broader context when classifying the use case (e.g., [14-17]).

Similarly, for research questions related to evaluation and challenges, we examined the papers in greater detail, including both the abstract and full text. To conduct the analysis, we grouped related terms mentioned by the authors into categories, as detailed in Section 4.

4. Results

This section presents the results of the analysis addressing the four research questions outlined in Section 3. For each research question, we grouped similar articles into 4 to 5 main categories. In

some cases, a paper fit into multiple categories, which we accounted for accordingly. Table 2 provides an overview of the categorisation, showing which papers fall under each category for each research question.

Table 2 Classification of papers by topic, usecase, evaluation, and limitation categories.

Paper	Topic	Usecase	Evaluation	Limitation
[18]	Climate	Factual evaluation	Comparison with base models, Human feedback	Ethical concerns, Technical issue
[19]	Environmental science, Sustainability	Knowledge extraction	Human feedback	Unreliable output
[20]	Marine ecosystem, Pollution	Predictive modelling	Human feedback	Technical issue, Unreliable output
[21]	Climate	Factual evaluation, Evaluation framework	Human feedback	Ethical concerns, Interpretability issues, Technical issue
[22]	Air pollution	Knowledge extraction, Predictive modelling	Human feedback	Technical issue, Unreliable output
[23]	Climate	Factual evaluation	Automatic evaluation	Ethical concerns
[16]	Biodiversity, Environmental science	Knowledge extraction	Automatic evaluation	Ethical concerns, Unreliable output
[1]	Biodiversity	Decision support	Automatic evaluation, Human feedback. Train/test evaluation	Ethical concerns, Interpretability issues, Technical issue
[24]	Climate, Sustainability	Knowledge extraction	Human feedback	Technical issue
[25]	Forestry	Knowledge extraction, Agent-based modelling	Comparison with base models, Human feedback	None (limit)
[26]	Biodiversity	Knowledge extraction	Automatic evaluation	Data limitations, Unreliable output
[27]	Environmental science	Question-answering	Human feedback	Ethical concerns
[28]	Climate	Factual evaluation	Human feedback	Technical issue
[29]	Climate, Crowdsourcing	Knowledge extraction	Human feedback	Technical issue
[30]	Climate	Predictive modelling	Comparison with base models	Technical issue, Unreliable output
[31]	Climate	Question-answering	Human feedback	Technical issue

[32]	Carbon footprint, Climate	Question-answering	Human feedback	Data limitations, Ethical concerns
[33]	Sustainability, Nuclear energy, Renewable energy	Factual evaluation	Automatic evaluation, Human feedback	Data limitations
[34]	Carbon footprint, Climate	Decision support	Automatic evaluation	Unreliable output
[35]	Climate	Predictive modelling	Train/test evaluation	None (limit)
[36]	Sustainability	Predictive modelling	Comparison with base models, Train/test evaluation	None (limit)
[37]	Climate	Knowledge extraction	Comparison with base models, Train/test evaluation	Data limitations, Ethical concerns, Technical issue, Unreliable output
[38]	Climate, Agricultural sector	Question-answering	Comparison with base models	Data limitations, Ethical concerns, Technical issue
[39]	Climate	Factual evaluation	Comparison with base models, Human feedback	Ethical concerns
[17]	Carbon footprint, Climate	Decision support, Question-answering	None (eval)	Technical issue
[40]	Biodiversity	Knowledge extraction	Automatic evaluation	Technical issue, Unreliable output
[41]	Climate, Environmental science	Question-answering	Automatic evaluation, Comparison with base models, Human feedback. Train/test evaluation	Data limitations
[42]	Climate	Question-answering	Comparison with base models, Human feedback	Data limitations, Unreliable output
[43]	Biodiversity, Marine ecosystem	Question-answering	Automatic evaluation, Comparison with base models	Data limitations, Unreliable output
[44]	Biodiversity	Knowledge extraction	Human feedback	Ethical concerns
[15]	Forest Smoke detection, Wildfire containment	Predictive modelling	Automatic evaluation, Comparison with base models	Technical issue
[45]	Environmental science	Question-answering	Human feedback	Data limitations
[46]	Sustainability, Water management	Decision support, Interactive Question-answering	Automatic evaluation	Data limitations, Technical issue, Unreliable output

[47]	Environmental science, Ocean science	Knowledge extraction, Multi-agent collaboration	Human feedback	Ethical concerns, Unreliable output
[48]	Climate	Factual evaluation	Automatic evaluation, Human feedback	Unreliable output
[49]	Carbon footprint, Climate	Decision support, Knowledge extraction	Comparison with base models	Interpretability issues, Unreliable output
[50]	Climate, Sustainability	Question-answering, Conversational agent	Automatic evaluation, Comparison with base models	Technical issue, Unreliable output
[51]	Climate	Factual evaluation	Automatic evaluation, Human feedback	Unreliable output
[52]	Climate	Knowledge extraction	Comparison with base models, Human feedback	Interpretability issues, Technical issue, Unreliable output
[53]	Climate	Question-answering	Human feedback	Unreliable output
[54]	Environmental science	Knowledge extraction	Comparison with base models	Unreliable output
[55]	Climate	Question-answering	Human feedback	Ethical concerns, Unreliable output
[56]	Climate, Sustainability	Factual evaluation	Automatic evaluation, Human feedback. Train/test evaluation	Unreliable output
[14]	Environmental science, Marine ecosystem	Predictive modelling	Automatic evaluation, Comparison with base models	Technical issue
[57]	Climate, Sustainability	Knowledge extraction	Comparison with base models	Ethical concerns, Unreliable output
[58]	Climate	Question-answering	Human feedback	Ethical concerns, Unreliable output
[2]	Climate, Environmental events	Decision support, Knowledge extraction	Automatic evaluation	Data limitations, Ethical concerns, Interpretability issues
[3]	Climate	Factual evaluation	Human feedback	Technical issue, Unreliable output
[59]	Climate	Factual evaluation	Automatic evaluation	Technical issue, Unreliable output
[60]	Sustainability	Knowledge extraction	Automatic evaluation, Human feedback	Unreliable output
[4]	Climate	Knowledge extraction	Human feedback	Data limitations, Unreliable output

4.1 Application Areas of LLMs in Environmental Science

This research question aims to provide an overview of the most common application areas within environmental science where LLMs have been utilized. In some articles, the authors referred only to the general term *environmental science*, which we accepted as a valid topic. Figure 3 and Figure 4 show the most frequently addressed application topics. Among them, *climate* has the highest appearance (i.e., about 60%) in our review, highlighting its critical importance within environmental science research.

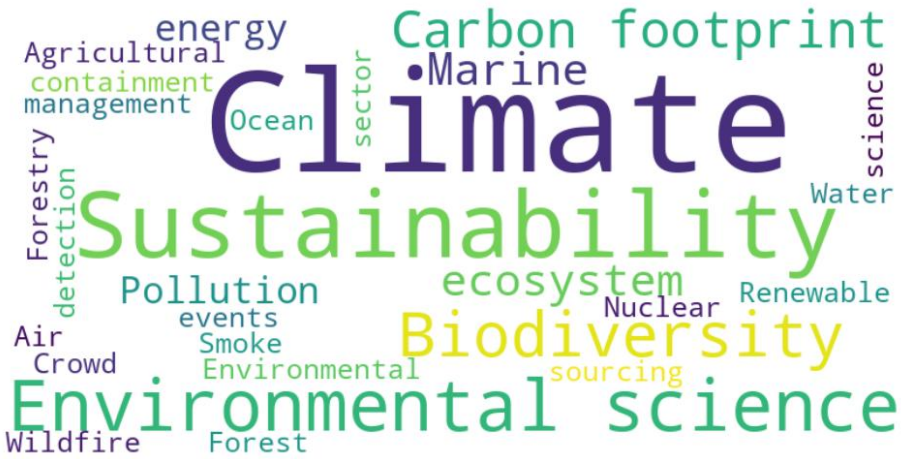

Figure 3 Wordcloud of application topics.

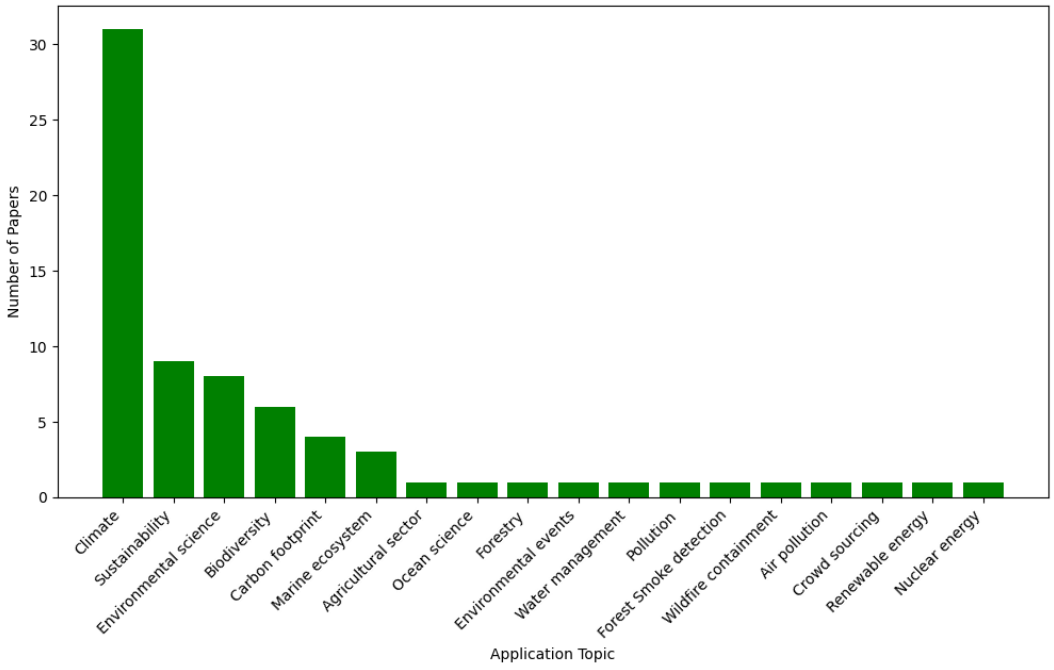

Figure 4 Number of papers per application topics.

4.2 Usecase of LLMs

In this section, we analyse how LLMs were used in environmental science research. We identified five main categories of application: knowledge extraction, question answering, factual evaluation, predictive modeling, and decision support.

The most common use case was knowledge extraction, where LLMs were used to summarise scientific literature, extract insights from policy documents, and retrieve specific information from environmental datasets [19, 24, 29]. These models helped reduce the manual workload required to process large volumes of unstructured and structured text.

Question answering was another prevalent use case, often implemented via chatbots or conversational agents. These systems enabled both experts and non-specialists to interact with environmental information through natural language queries. Ferreri-Pic [27] developed an LLM-based interface for public engagement with environmental guidelines by generating future scenarios. Giudici et al. [50] created a conversational system that guided users through domestic sustainability. Similarly, Labonnote et al. [53] developed a chatbot designed to enhance user accessibility and awareness of climate-related topics.

Another notable application was factual evaluation. In factual evaluation, LLMs were applied to verify the accuracy of environmental statements, detect misinformation, or perform sentiment analysis on climate-related discourse. These applications are particularly relevant for monitoring public narratives or validating the trustworthiness of sources. For example, Christodoulou et al. [23] leveraged LLMs to detect hate speech and stance in climate-related social media posts. Corral et al. [48] utilized LLMs to detect and categorize accusations of hypocrisy in climate discourse. Aremu et al. [18] applied LLMs to assess their reliability in responding accurately to misinformed prompts.

LLMs were also used in predictive modeling tasks, where they contributed to classification, forecasting, or hybrid modeling approaches. Balcacer et al. [20] integrated LLMs with environmental sensors to predict pollution levels in marine ecosystems. Grasso et al. [30] used LLMs to forecast climate indicators by analyzing past event reports and scientific literature. Wei et al. [15] incorporated LLMs in wildfire containment systems, combining visual context and language-based prompts to improve smoke detection.

Ultimately, several studies have utilized LLMs to support decision-making processes. In these cases, LLMs were tasked with synthesising policy recommendations, suggesting mitigation strategies, or identifying emerging environmental trends from complex datasets. For example, DeSantis et al. [1] applied LLMs to align biodiversity goals with global policy frameworks, while Li et al. [34] and Ghinassi et al. [49] used them to inform decisions on carbon footprint management. Ochieng et al. [17] integrated LLMs as post-prediction recommender tools, and Arslan et al. [46] demonstrated how LLMs can facilitate interactive support in sustainability and water management scenarios.

While knowledge extraction dominated the reviewed literature, the diversity of other use cases highlights the adaptability of LLMs in environmental research. We chose to distinguish between knowledge extraction and question answering to reflect their different technical structures and user-facing goals, one being passive, the other interactive and end-user oriented. Figure 5 illustrates the distribution of LLM use cases identified in our review.

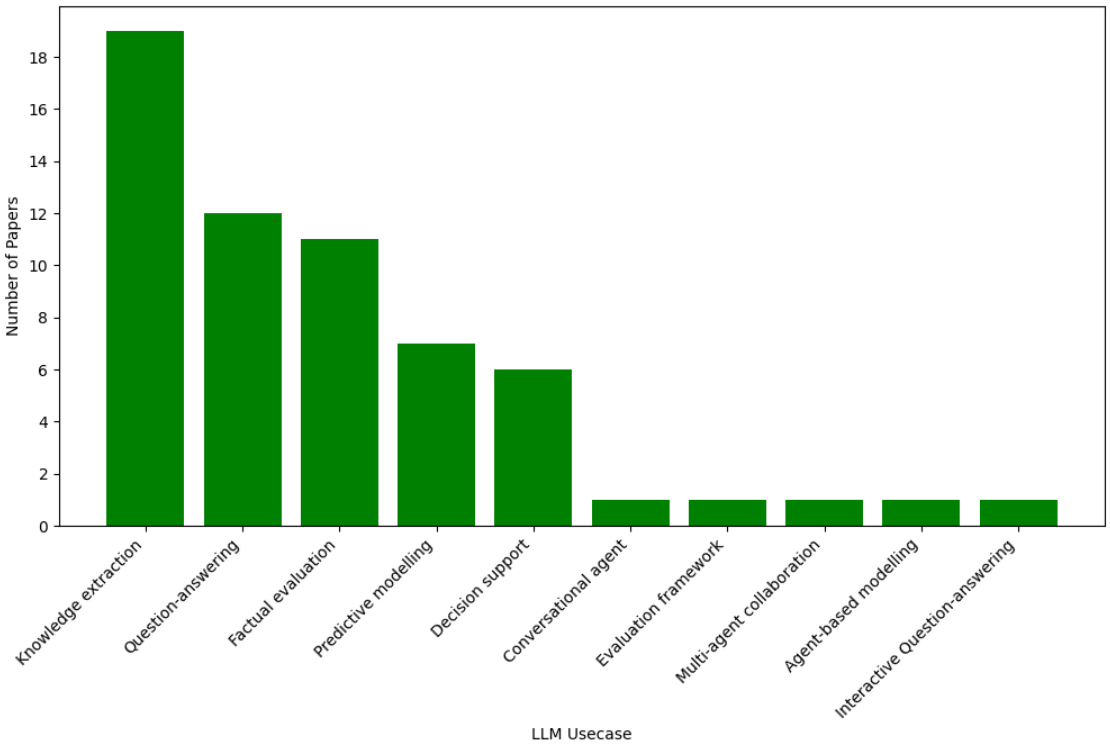

Figure 5 Number of papers per use case.

4.3 Evaluation Types

For each article, we analysed the evaluation process and associated metrics. We classified the types of evaluation into four main categories: human feedback, train/test evaluation, automatic evaluation, and comparison with baseline models. Some papers used more than one method, while others did not include a clear evaluation strategy for the LLM component.

Human feedback was the most frequently used evaluation method. In these studies, researchers relied on expert or non-expert reviewers to assess the outputs of LLMs, often judging their usefulness, accuracy, or relevance. This approach was prevalent in tasks such as summarization or question answering, where subjective quality matters [18, 47, 52].

Train/test evaluation was used in a smaller set of papers, particularly where LLMs were involved in classification or prediction tasks. These studies split datasets into training and testing subsets to evaluate performance in a more controlled and reproducible way. For example, Li et al. [35] employed standard train-test splits to assess predictive accuracy in climate-related models. Lin et al. [36] followed a similar approach in modeling sustainability outcomes, and Sun et al. [41] combined train/test evaluation with human review to validate both model performance and user satisfaction.

Automatic evaluation was another standard method, particularly in tasks like summarization or sentiment analysis where standardized metrics such as BLEU, ROUGE, or accuracy scores could be applied. This approach also included cases where another LLM or script was used to evaluate output quality. For example, Elliott et al. [26] trained a logistic regression model to predict correctness of LLMs' answers. Corral et al. [48] applied automatic scoring to sentiment classification outputs, and Giudici et al. [50] combined scripted evaluation with human judgment in their analysis of chatbot performance.

Some studies opted to compare LLMs against baseline models or previously established methods. This category included both qualitative and quantitative comparisons, aiming to highlight where LLMs offered improvements (or not) over traditional techniques. For example, Grasso et al. [30] benchmarked LLM-driven forecasts against existing predictive models for climate variables. Bi et al. [47] conducted comparisons between LLM-based agents and rule-based systems in ocean science tasks. Zhou et al. [60] tested the output of LLMs against older NLP models to assess gains in interpretability and accuracy.

Figure 6 presents the number of papers in each evaluation category. One notable exception was the study by Ochieng et al. [17], which did not include any evaluation specific to the LLM component. While the study evaluated a predictive system overall, it lacked metrics or validation strategies for the language model used in generating recommendations, a gap that reflects the broader challenge of evaluating LLMs in multi-component pipelines.

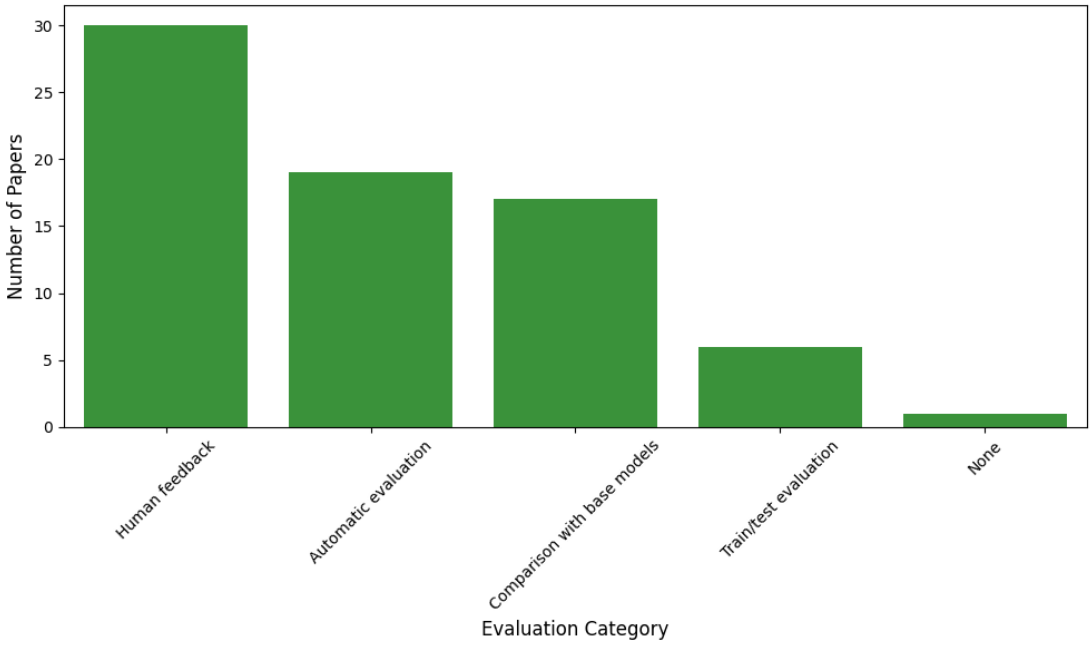

Figure 6 Number of papers per evaluation category.

As seen in Figure 6, most studies relied on some form of human evaluation, which can introduce subjectivity. This trend may be because many use cases fall under knowledge extraction and question-answering, where human judgment plays a significant role. Table 2 provides a more detailed view of the evaluation approaches.

4.4 Challenges & Limitations Identified

In this section, we provide an overview of the main challenges and limitations reported by authors when applying LLMs within environmental science contexts. Although the terminology used to describe these issues varied across studies, we grouped the identified challenges into five key categories: data limitations, technical matters, interpretability challenges, ethical concerns, and unreliable outputs.

Data limitations were widely discussed, with several authors noting that existing models often struggle due to unbalanced datasets, a lack of domain-specific training data, and limited access to

high-quality, up-to-date environmental information [26, 33, 41]. These challenges often impact the relevance and accuracy of LLM outputs in environmental applications, where specialized information is critical.

Technical issues were also commonly reported. These included the high computational and energy costs associated with deploying LLMs or integrating them with existing data pipelines, the complexity of prompt engineering, difficulties in incorporating human feedback, and limitations due to short context windows or low inference efficiency, e.g., [30, 40, 59]. Such barriers were seen as significant obstacles to practical deployment.

Interpretability presented another area of concern. Several studies pointed to the "black-box" nature of LLMs, which makes it difficult to trace how outputs are generated or explain model decisions. This lack of transparency was particularly problematic in cases where LLMs were used to support decision-making or policy-related tasks that require a clear rationale, e.g. [2, 21, 49, 52].

Ethical concerns were also raised throughout the reviewed literature. Issues such as fairness, bias, copyright, and privacy were mentioned, particularly in studies that involved sensitive content or public-facing applications. While these concerns are common in AI in general, they are essential in environmental domains where trust and public accountability are critical, e.g. [37, 55, 58].

Finally, unreliable output was among the most frequently cited limitations. This included hallucinated facts, confabulation, factual inaccuracies, inconsistencies, and generalisation, e.g. [22, 42, 60]. Authors emphasized the need for more robust evaluation methods and safeguards, particularly in high-stakes environmental scenarios where accuracy and reliability are crucial.

Figure 7 shows the distribution of papers across these categories. Some papers, [25, 35, 36], did not mention any specific limitations or challenges related to the use and implementation of LLMs. As shown in Figure 7, unreliable outputs and technical difficulties were the most frequently cited issues. While many of these concerns, such as hallucinations, bias, and scalability, are not specific to environmental science, a few challenges are more domain-relevant. In particular, issues related to training data, such as accessibility, data imbalance, and quality, were especially prominent in this field and can be considered by researchers in the environmental science field.

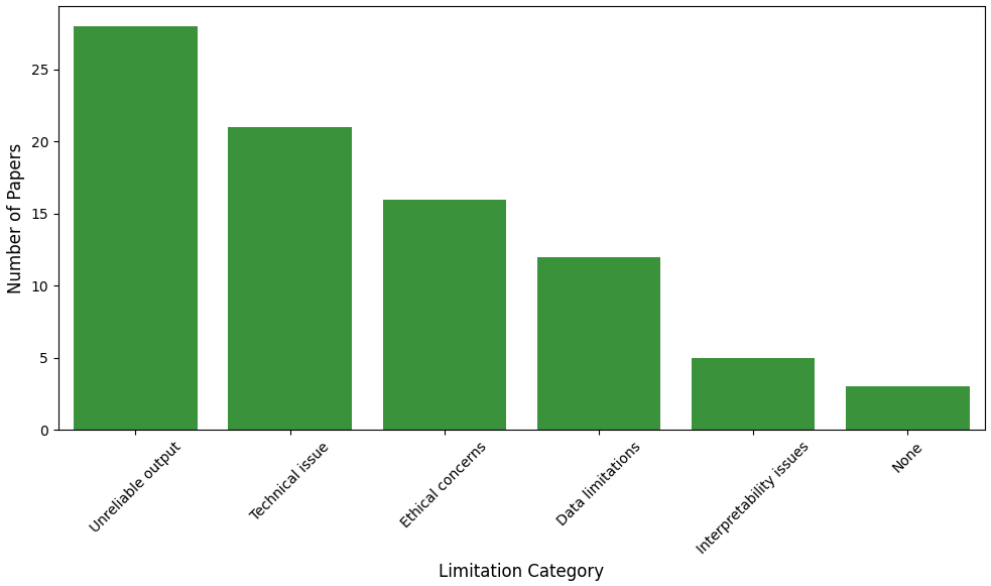

Figure 7 Number of papers per limitation category.

5. Discussion

The analysis of the reviewed papers reveals several emerging trends. The climate domain stands out as the most frequently addressed topic, indicating a strong research interest in using LLMs to support climate-related studies. The most common use case observed was knowledge extraction, such as summarisation and information retrieval, demonstrating the value of LLMs in reducing the manual effort required to process vast and rapidly growing volumes of environmental data. From the analysis, we also identify several open challenges and future research directions that could help advance the field:

- **Human-in-the-loop integration:** A key challenge lies in effectively incorporating human expertise throughout the development lifecycle, particularly in data selection, feedback, and evaluation stages. Future work can focus on standard platforms to facilitate this integration in a structured and reusable way.
- **Data limitations:** Issues such as lack of domain-specific training data, unbalanced datasets, and poor data quality were frequently mentioned. Addressing these limitations requires a greater adoption of FAIR principles (Findable, Accessible, Interoperable, and Reusable) and open-source frameworks, which can enhance data accessibility and model reproducibility. Although these practices are not new, they have yet to be widely adopted in the context of LLMs within environmental science.
- **Evaluation frameworks:** Evaluation is considered to be a very challenging part of using LLMs. While human evaluation is regarded as the most accurate and reliable method, it can also introduce subjectivity and bias. Developing standardized, domain-aware evaluation frameworks that balance human judgment with automated metrics is an essential direction for future work.
- **Interdisciplinary collaboration:** A significant gap exists in collaboration between AI researchers and environmental science experts. Many challenges, especially those that are domain-specific, could be addressed more effectively through co-development approaches.

While several studies explored LLMs for knowledge extraction and decision support, we found limited evidence of their use in processing real-time environmental data streams. As LLMs become more integrated into decision-making systems, leveraging them for dynamic, time-sensitive applications — such as responding to sensor data — could be a valuable direction for future research. Moreover, as LLMs evolve from language understanding to problem-solving, new opportunities arise. Multimodal LLMs, capable of processing text, images, and other data types, could support more comprehensive environmental modeling. Furthermore, agent-based LLM architectures, where individual agents specialize in distinct subfields of environmental science, can tackle complex and interdisciplinary challenges. These agents could collaborate as part of a larger problem-solving system, leveraging previously developed models and knowledge extraction tools to enhance their capabilities. This vision highlights the importance of open-source and FAIR data, which facilitate reuse and integration across various applications.

While LLMs show potential in various environmental applications, it is crucial to critically assess their validity through a SWOT (Strengths, Weaknesses, Opportunities, and Threats) lens. Strengths include their ability to process vast unstructured datasets (e.g., policy, literature, citizen reports) and generate human-readable outputs, which can support knowledge discovery and communication. They also lower the barrier for non-experts to access complex environmental

information. Weaknesses, however, include the lack of transparency, dependency on non-domain-specific training data, and challenges in evaluating output reliability. These limitations raise concerns about the direct use of LLMs in high-stakes decision-making without human oversight. On the opportunity side, LLMs could serve as intelligent assistants for interdisciplinary research, help automate the synthesis of environmental assessments, and support scalable environmental communication tools. However, threats include the risk of misinformation (e.g., hallucinations), model biases, and environmental costs associated with training and running large-scale models. The risk of overreliance on these systems without proper validation could hinder their safe deployment in sensitive policy or scientific workflows.

6. Conclusions

This review focuses on the application of LLMs in addressing challenges within environmental science. By analyzing 51 papers sourced from Scopus and Web of Science databases, we identified main application topic trends, use cases, and research gaps. The *climate* domain emerged as the most frequently targeted area, while *knowledge extraction* was the most common use case of LLMs. In addition to mapping these trends, we investigated the evaluation methods used and discussed the significant challenges and limitations reported across studies. Finally, we outlined several directions for future research, emphasizing the need for improved evaluation frameworks, facilitating human-in-the-loop, and the adoption of open and FAIR data practices.

Author Contributions

MMR: Conceptualization, Investigation, Formal analysis, Paper screening, Writing, Editing. RK: Investigation, Paper screening, Writing, Editing.

Competing Interests

The authors have declared that no competing interests exist.

AI-Assisted Technologies Statement

All factual contents, ideas and decisions were created and verified by the authors. The AI tool, ChatGPT, served solely as a writing assistant.

References

1. DeSantis N, Supples C, Phillips L, Pigot J, Ervin J, Wade T. Leveraging AI for enhanced alignment of national biodiversity targets with the global biodiversity goals. Nat Based Solutions. 2025; 7: 100198.
2. Tian Y, Li W, Hu L, Chen X, Brook M, Brubaker M, et al. Advancing large language models for spatiotemporal and semantic association mining of similar environmental events. Trans GIS. 2025; 29: e13282.
3. Zanartu F, Otmakhova Y, Cook J, Frermann L. Generative debunking of climate misinformation. arXiv. 2024. doi: 10.48550/arXiv.2407.05599.

4. Zhou H, Hobson D, Ruths D, Piper A. Large scale narrative messaging around climate change: A cross-cultural comparison. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
5. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners [Internet]. OpenAI blog; 2019. Available from: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.
6. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling language modeling with pathways. J Mach Learn Res. 2023; 24: 1-113.
7. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv. 2023. doi: 10.48550/arXiv.2302.13971.
8. Ji X, Wu X, Deng R, Yang Y, Wang A, Zhu Y. Utilizing large language models for identifying future research opportunities in environmental science. J Environ Manag. 2025; 373: 123667.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 December 4-9; Long Beach, CA, USA. Aachen, Germany: CEUR-WS Team.
10. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers); 2019 June 2-7; Minneapolis, MN, USA. Stroudsburg, PA: Association for Computational Linguistics.
11. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv. 2020. doi: 10.48550/arXiv.2001.08361.
12. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Int J Surg. 2010; 8: 336-341.
13. Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021; 3: 125-133.
14. Raine S, Marchant R, Kusy B, Maire F, Fischer T. Image labels are all you need for coarse seagrass segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2024 January 3-8; Waikoloa, HI, USA. Piscataway Township: IEEE.
15. Wei T, Kulkarni P. Enhancing the binary classification of wildfire smoke through vision-language models. Proceedings of the 2024 Conference on AI, Science, Engineering, and Technology (AIxSET); 2024 September 30; Laguna Hills, CA, USA. Piscataway Township: IEEE.
16. Dasgupta D, Mondal A, Chakrabarti PP. Can synthetic plant images from generative models facilitate rare species identification and classification? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 June 17-18; Seattle, WA, USA. Piscataway Township: IEEE.
17. Ochieng B, Onyango F, Kuria P, Wanjiru M, Maake B, Awuor M. AI-Driven carbon emissions tracking and mitigation model. Proceedings of the 2024 IST-Africa Conference (IST-Africa); 2024 May 20-24; Dublin, Ireland. Piscataway Township: IEEE.
18. Aremu T, Akinwehinmi O, Nwagu C, Ahmed SI, Orji R, Amo PA, et al. On the reliability of Large Language Models to misinformed and demographically informed prompts. AI Mag. 2025; 46: e12208.

19. Ashby C, Weir D, Fussey P. Understanding public views on electric vehicle charging: A thematic analysis. Transp Res Interdiscip Perspect. 2025; 29: 101325.
20. Balcacer A, Hannon B, Kumar Y, Huang K, Sarnoski J, Liu S, et al. Mechanics of a Drone-based System for Algal Bloom Detection Utilizing Deep Learning and LLMs. Proceedings of the 2023 IEEE MIT Undergraduate Research Technology Conference (URTC); 2023 October 6-8; Cambridge, MA, USA. Piscataway Township: IEEE.
21. Bulian J, Schäfer MS, Amini A, Lam H, Ciaramita M, Gaiarin B, et al. Assessing large language models on climate information. Proceedings of the 41st International Conference on Machine Learning; 2024 July 21-27; Vienna, Austria.
22. Chen A, Du J, Rodriguez A, Rodriguez R, Higgins J, Podmore R, et al. Viability of applying large language models to indoor climate sensor and health data for scientific discovery. Proceedings of the 2024 IEEE Global Humanitarian Technology Conference (GHTC); 2024 October 23-26; Radnor, PA, USA. Piscataway Township: IEEE.
23. Christodoulou C. Nlpdame at climate Activism 2024: Mistral sequence classification with PEFT for hate speech, targets and stance event detection. Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024); 2024 March 21-22; St. Julians, Malta. Stroudsburg, PA: Association for Computational Linguistics.
24. Dimmelmeier A, Doll H, Schierholz M, Kormanyos E, Fehr M, Ma B, et al. Informing climate risk analysis using textual information-A research agenda. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
25. Du S, Tang S, Wang W, Li X, Guo R. Tree-GPT: Modular large language model expert system for forest remote sensing image understanding and interactive analysis. arXiv. 2023. doi: 10.48550/arXiv.2310.04698.
26. Elliott MJ, Fortes JA. Toward reliable biodiversity information extraction from large language models. Proceedings of the 2024 IEEE 20th International Conference on e-Science (e-Science); 2024 September 16-20; Osaka, Japan. Piscataway Township: IEEE.
27. Ferrer i Picó J, Catta-Preta M, Trejo Omeñaca A, Vidal M, Monguet i Fierro JM. The time machine: Future scenario generation through generative AI tools. Future Internet. 2025; 17: 48.
28. Fore M, Singh S, Lee C, Pandey A, Anastasopoulos A, Stamoulis D. Unlearning climate misinformation in large language models. arXiv. 2024. doi: 10.48550/arXiv.2405.19563.
29. Gimpel H, Laubacher R, Meindl O, Wöhl M, Dombetzki L. Advancing content synthesis in macro-task crowdsourcing facilitation leveraging natural language processing. Group Decis Negot. 2024; 33: 1301-1322.
30. Grasso F, Locci S. Assessing generative language models in classification tasks: Performance and self-evaluation capabilities in the environmental and climate change domain. In: International Conference on Applications of Natural Language to Information Systems. Cham, Switzerland: Springer; 2024. pp. 302-313.
31. Gursesli MC, Taveekitworachai P, Abdullah F, Dewantoro MF, Lanata A, Guazzini A, et al. The chronicles of ChatGPT: Generating and evaluating visual novel narratives on climate change through ChatGPT. In: International Conference on Interactive Digital Storytelling. Cham, Switzerland: Springer Nature; 2023. pp. 181-194.

32. Han T, Cong RG, Yu B, Tang B, Wei YM. Integrating local knowledge with ChatGPT-like large-scale language models for enhanced societal comprehension of carbon neutrality. Energy AI. 2024; 18: 100440.
33. Kwon OH, Vu K, Bhargava N, Radaideh MI, Cooper J, Joynt V, et al. Sentiment analysis of the United States public support of nuclear power on social media using large language models. Renew Sustain Energy Rev. 2024; 200: 114570.
34. Li Z, Tang P, Wang X, Liu X, Mou P. PCF-RWKV: Large language model for product carbon footprint estimation. Sustainability. 2025; 17: 1321.
35. Li B, Fu E, Yang S, Lin J, Zhang W, Zhang J, et al. Measuring China's Policy Stringency on Climate Change for 1954–2022. Sci Data. 2025; 12: 188.
36. Lin LH, Ting FK, Chang TJ, Wu JW, Tsai RT. Gpt4esg: Streamlining environment, society, and governance analysis with custom ai models. Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB); 2024 April 19-21; Taipei, Taiwan. Piscataway Township: IEEE.
37. Li N, Zahra S, Brito M, Flynn C, Görnerup O, Worou K, et al. Using LLMs to build a database of climate extreme impacts. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
38. Nguyen V, Karimi S, Hallgren W, Harkin A, Prakash M. My climate advisor: An application of NLP in climate adaptation for agriculture. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
39. Nisbett N, Spaiser V. How convincing are AI-generated moral arguments for climate action? Front Clim. 2023; 5: 1193350.
40. Scheepens D, Millard J, Farrell M, Newbold T. Large language models help facilitate the automated synthesis of information on potential pest controllers. Methods Ecol Evol. 2024; 15: 1261-1273.
41. Sun T, Carr J, Kazakov D. A hybrid question answering model with ontological integration for environmental information. Proceedings of the DAO-XAI 2024: Workshop on Data meets Applied Ontologies in Explainable AI; 2024 October 19; Santiago de Compostela, Spain. Aachen, Germany: CEUR-WS Team.
42. Vaghefi SA, Stammbach D, Muccione V, Bingler J, Ni J, Kraus M, et al. ChatClimate: Grounding conversational AI in climate science. Commun Earth Environ. 2023; 4: 480.
43. Wang X, Zhang M, Liu H, Ma X, Liu Y, Chen Y. ChatBBNJ: A question–answering system for acquiring knowledge on biodiversity beyond national jurisdiction. Front Mar Sci. 2024; 11: 1368356.
44. Weaver WN, Ruhfel BR, Lough KJ, Smith SA. Herbarium specimen label transcription reimagined with large language models: Capabilities, productivity, and risks. Am J Bot. 2023; 110: e16256.
45. Zhu JJ, Yang M, Jiang J, Bai Y, Chen D, Ren ZJ. Enabling GPTs for expert-level environmental engineering question answering. Environl Sci Technol Lett. 2024; 11: 1327-1333.
46. Arslan M, Munawar S, Riaz Z. Sustainable urban water decisions using generative artificial intelligence. Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA); 2024 December 11-12; Manama, Bahrain. Piscataway Township: IEEE.

47. Bi Z, Zhang N, Xue Y, Ou Y, Ji D, Zheng G, et al. OceanGPT: A large language model for ocean science tasks. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 August 11-16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
48. Corral PG, Green A, Meyer H, Stoll A, Yan X, Reuver M. A few hypocrites: Few-shot learning and subtype definitions for detecting hypocrisy accusations in online climate change debates. arXiv. 2024. doi: 10.48550/arXiv.2409.16807.
49. Ghinassi I, Catalano L, Colella T. Efficient aspect-based summarization of climate change reports with small language models. arXiv. 2024. doi: 10.48550/arXiv.2411.14272.
50. Giudici M, Abbo GA, Belotti O, Braccini A, Dubini F, Izzo RA, et al. Assessing llms responses in the field of domestic sustainability: An exploratory study. Proceedings of the Third International Conference on Digital Data Processing (DDP); 2023 November 27-29; Luton, United Kingdom. Piscataway Township: IEEE.
51. Jin Z, Lalwani A, Vaidhya T, Shen X, Ding Y, Lyu Z, et al. Logical fallacy detection. arXiv. 2022. doi: 10.48550/arXiv.2202.13758.
52. Joe ET, Koneru SD, Kirchhoff CJ. Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights. arXiv. 2024. doi: 10.48550/arXiv.2407.12826.
53. Labonnote N, Caetano L, Kind R. Empowering Citizens for Climate Adaptation in Norway: Leveraging (AI-Driven) Emerging Technologies. Proceedings of the 9th International Conference on Smart and Sustainable Technologies (SpliTech); 2024 June 25-28; Bol and Split, Croatia. Piscataway Township: IEEE.
54. Mishra L, Dhibi S, Kim Y, Ramis CB, Gupta S, Dolfi M, et al. Statements: Universal information extraction from tables with large language models for ESG KPIs. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.
55. Nguyen H, Nguyen V, López-Fierro S, Ludovise S, Santagata R. Simulating climate change discussion with large language models: Considerations for science communication at scale. Proceedings of the eleventh ACM conference on learning@ scale; 2024 July 18-20; Atlanta GA USA. New York, NY: Association for Computing Machinery, Inc.
56. Ni J, Bingler J, Colesanti-Senni C, Kraus M, Gostlow G, Schimanski T, et al. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2023 December 6-10; Singapore. Stroudsburg, PA: Association for Computational Linguistics.
57. Mullappilly SS, Shaker A, Thawakar O, Cholakkal H, Anwer RM, Khan S, et al. Arabic Mini-ClimateGPT: A climate change and sustainability tailored Arabic LLM. Proceedings of the Findings of the Association for Computational Linguistics: EMNLP; 2023 December 6-10; Singapore. Stroudsburg, PA: Association for Computational Linguistics.
58. Theophilou E, Koyutürk C, Yavari M, Bursic S, Donabauer G, Telari A, et al. Learning to prompt in the classroom to understand AI limits: A pilot study. In: International conference of the Italian association for artificial intelligence. Cham, Switzerland: Springer Nature; 2023. pp. 481-496.
59. Zhang H, Zhu Z, Zhang Z, Devasier J, Li C. Granular analysis of social media users' truthfulness stances toward climate change factual claims. Proceedings of the 1st Workshop on Natural

Language Processing Meets Climate Change (ClimateNLP 2024); 2024 August 16; Bangkok, Thailand. Stroudsburg, PA: Association for Computational Linguistics.

60. Zhou Y, Gu X, Ding J, Chen S, Perzylo A. Accessing the Capabilities of KGs and LLMs in Mapping Indicators within Sustainability Reporting Standards. Proceedings of the Workshop on Natural Language Processing for Knowledge Graph Creation (NLP4KGC) at International Conference on Semantic Systems (SEMANTICS); 2024 September 17; Amsterdam, Netherlands. Aachen, Germany: CEUR-WS Team.