Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Fine-tuning large language models for interdisciplinary environmental challenges

Yuanxin Zhang [a,b], Sijie Lin [a,c], Yaxin Xiong [a,b], Nan Li [d,*], Lijin Zhong [c], Longzhen Ding [a,b], Qing Hu [a,b,**]

[a] State Key Laboratory of Soil Pollution Control and Safety, Southern University of Science and Technology, Shenzhen, 518055, China
[b] School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China
[c] Beijing Huanding Environmental Smart Data Institute, Beijing, 100083, China
[d] School of Environment, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 7 April 2025
Received in revised form
27 July 2025
Accepted 27 July 2025

Keywords:
Environmental science
Artificial intelligence
Large language model
Fine-tuning
Instruction dataset

ABSTRACT

Large language models (LLMs) are revolutionizing specialized fields by enabling advanced reasoning and data synthesis. Environmental science, however, poses unique hurdles due to its interdisciplinary scope, specialized jargon, and heterogeneous data from climate dynamics to ecosystem management. Despite progress in subdomains like hydrology and climate modeling, no integrated framework exists to generate high-quality, domain-specific training data or evaluate LLM performance across the discipline. Here we introduce a unified pipeline to address this gap. It comprises EnvInstruct, a multi-agent system for prompt generation; ChatEnv, a balanced 100-million-token instruction dataset spanning five core themes (climate change, ecosystems, water resources, soil management, and renewable energy); and EnvBench, a 4998-item benchmark assessing analysis, reasoning, calculation, and description tasks. Applying this pipeline, we fine-tune an 8-billion-parameter model, EnvGPT, which achieves 92.06 ± 1.85 % accuracy on the independent EnviroExam benchmark—surpassing the parameter-matched LLaMA-3.1–8B baseline by ~8 percentage points and rivaling the closed-source GPT-4o-mini and the 9-fold larger Qwen2.5–72B. On EnvBench, EnvGPT earns top LLM-assigned scores for relevance (4.87 ± 0.11), factuality (4.70 ± 0.15), completeness (4.38 ± 0.19), and style (4.85 ± 0.10), outperforming baselines in every category. This study reveals how targeted supervised fine-tuning on curated domain data can propel compact LLMs to state-of-the-art levels, bridging gaps in environmental applications. By openly releasing EnvGPT, ChatEnv, and EnvBench, our work establishes a reproducible foundation for accelerating LLM adoption in environmental research, policy, and practice, with potential extensions to multimodal and real-time tools.

1. Introduction

Environmental science is a complex and interdisciplinary field that involves the study of interactions between, impacts on, and management of the environment and natural resources [1,2], to uncover complex relationships among the components of the Earth's interconnected physical and biological systems [3,4]. This field encompasses various disciplines, such as ecology, meteorology, geology, and hydrology [5–7]. Integrating methods from these disciplines enables a deeper exploration of environmental challenges [8,9] and supports prevention, mitigation, and resolution strategies through diverse perspectives, tools, and theories [10,11]. Recent advances in large language models (LLMs) [12,13] offer new opportunities to accelerate such interdisciplinary work by providing novel tools for tackling complex environmental problems [14]. Although LLMs have begun to reshape domains such as medicine and law [15–18], their impact on environmental science remains limited, as no dedicated framework has yet combined domain-specific data generation and rigorous evaluation [19,20]. Two obstacles dominate: specialized terminology that is absent from general training corpora [21,22] and the growing

[*] Corresponding author.
[**] Corresponding author. State Key Laboratory of Soil Pollution Control and Safety, Southern University of Science and Technology, Shenzhen, 518055, China.
E-mail addresses: li-nan@tsinghua.edu.cn (N. Li), huq@sustech.edu.cn (Q. Hu).

https://doi.org/10.1016/j.ese.2025.100608
2666-4984/

complexity and diversity of environmental data that LLMs must absorb and reason over. Addressing these issues requires creating high-quality domain datasets before model construction and ensuring output accuracy throughout development and deployment [23].

Recent domain-specific LLMs have shown promise in adapting large models to environmental research. WaterGPT augments a 7B-parameter backbone with 1.1 billion hydrology tokens and multimodal tools, and has achieved 83 % accuracy on the bespoke EvalWater benchmark [24]. ClimateGPT continually pre-trains 7–70B models on a 4.2 billion-token climate corpus and has performed competitively on ClimaBench [25]. OceanGPT couples a 67,000-document oceanography corpus with an automated instruction pipeline and the OceanBench evaluation set [26]. However, each system is confined to a single subdiscipline and employs an idiosyncratic data workflow, which hinders cross-domain comparison and reuse. To overcome this fragmentation, we introduce EnvGPT, an 8B-parameter model that unifies climate, hydrology, ecology, soil science, and renewable energy knowledge through a low-rank adaptation (LoRA) fine-tuning process. This model is built upon a single reproducible pipeline—comprising EnvInstruct, ChatEnv, and the 4998-item EnvBench—which enables consistent training, evaluation, and extension at a modest computational cost.

In this study, we assemble a cross-domain corpus comprising open-access articles from environmental science journals. Through the EnvInstruct procedure, this material is distilled into 112,000 balanced instruction–response pairs (ChatEnv) that span the five targeted subfields. An 8B-parameter base model is then adapted with parameter-efficient LoRA updates to create EnvGPT. We quantify performance using the 4998-item EnvBench, whose automated metrics capture factual accuracy, numerical precision, and multistep reasoning. We corroborate the results with a panel-based LLM expert evaluation that rates contextual relevance and practical utility for authentic environmental tasks. The findings highlight the crucial role of EnvInstruct in generating high-quality, domain-specific instruction datasets and stress the importance of supervised fine-tuning (SFT) in enhancing the performance of LLMs for specialized scientific applications.

2. Methodology

2.1. Methodological framework

We developed a framework for constructing LLMs in environmental science, which we used to build EnvGPT. This framework comprises four key components: EnvCorpus, an environmental science corpus; EnvInstruct, an instruction dataset-generation framework; an SFT process for training EnvGPT; and EnvBench, a benchmark for LLMs in environmental science (Fig. 1).

The core of the EnvGPT construction framework is EnvInstruct, which provides a high-quality SFT instruction dataset from EnvCorpus. We collected approximately 500 million tokens from open-access articles in environmental science journals to ensure the diversity and representativeness of the corpus. This collection spans five major areas of environmental science, ranging from climate change to water resource management. In constructing the instruction dataset, future scientific and technological expectations were integrated with cross-domain data, providing sufficient breadth and depth to support the model's SFT.

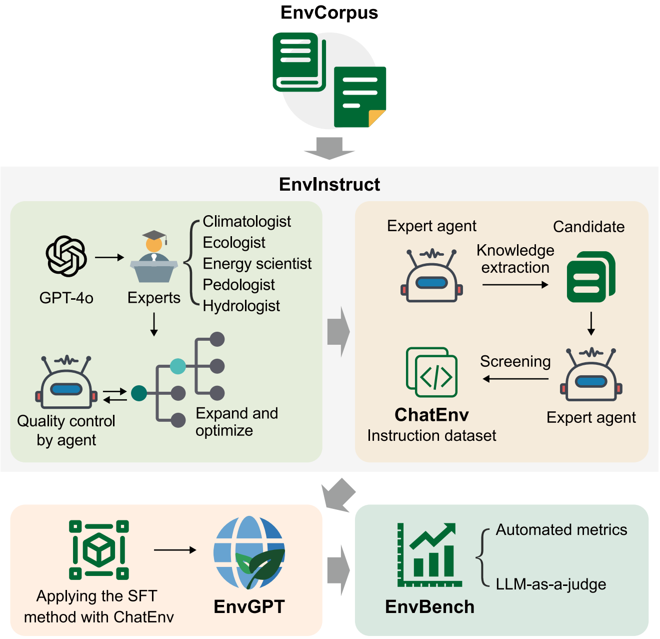

Fig. 1. Methodology framework of the EnvGPT. The framework outlines the development of EnvGPT, which is structured around four key components: EnvCorpus, EnvInstruct, ChatEnv, and EnvBench. SFT: supervised fine-tuning; LLM: large language model.

2.2. EnvCorpus construction

A well-curated corpus is essential for training domain-specific language models [27]. To construct EnvCorpus, we collected open-access research articles published in leading environmental science journals over the past five years (Supplementary Table S1). The five-year horizon maximized coverage of the latest analytical methods, data resources, disciplinary knowledge, and policy developments—dimensions that are typically underrepresented in general-purpose LLM corpora yet are critical for up-to-date environmental research.

Each article was mapped to one of five research themes—climate change and atmospheric science (CCAS), ecosystems and biodiversity conservation (EBC), water resources and aquatic environment (WRAE), soil and land-use management (SLUM), and renewable energy and environmental management (REEM). The rationale for this thematic structure and a concise description of each theme are provided separately (Supplementary Table S2). These themes capture the current research frontiers of environmental science while remaining mutually independent, ensuring balanced topical coverage across physical, biological, and engineering perspectives. Open-access licensing permits the redistribution of full texts, thereby guaranteeing legal reusability for benchmarking and future community extensions of EnvCorpus.

PDF documents were transformed into plain text using PyPDF, and subsequently tokenized employing GPT2TokenizerFast [28]. Page numbers, hyperlinks, nontext symbols, and other artifacts were removed through a rule-based filter. The resulting corpus comprised approximately 350 million tokens distributed across the five themes. Although a recency-focused, open-access strategy inevitably omits some seminal paywalled studies, foundational concepts remain accessible from review and perspective articles in the same journals. Future releases will broaden temporal and licensing coverage, as permissions allow, enabling a systematic assessment of the influence of historical literature on model generality.

2.3. Instruction dataset generation for environmental science

Instruction fine-tuning is an effective method for enhancing the capacity of a given LLM within a specific domain [29]. The instruction dataset guides the output of the LLM to incorporate domain-specific knowledge [18,30,31]. To build an LLM with environmental science knowledge on top of a base model and meet the complex and diverse needs of researchers, we introduced EnvInstruct, an instruction dataset generation framework for the environmental science domain based on multiple expert agents. EnvInstruct aims to generate instruction sets for the SFT of LLMs in a parallel and efficient manner via multi-agent collaboration.

GPT-4, the latest and most advanced flagship model developed by OpenAI [32], has led to transformative changes in several scientific fields [33–35]. EnvInstruct utilizes GPT-4 as part of a multi-agent system, rather than as an independent expert, to generate instructions from EnvCorpus under strict system-level guidance. EnvInstruct generates an environmental science instruction dataset through multi-agent collaboration. Different GPT-4o instances are assigned as expert agents in various subfields of environmental science, such as climatology, ecology, energy science, soil science, and hydrology, with EnvCorpus input into each agent. Each expert agent uses the input EnvCorpus to generate the corresponding initial instruction dataset based on its expertise. The dataset generation process is guided by a system-level instruction framework to ensure consistency in content and a focus on key environmental science issues. After the initial dataset is constructed, another expert agent refines and enriches it by adjusting details and filling gaps, improving its quality and aligning it with disciplinary standards. To ensure accuracy and relevance, the dataset undergoes rigorous quality control, with other expert agents expanding it to cover all key environmental science themes while maintaining data integrity and diversity. Through continuous optimization and verification, a high-quality, domain-specific dataset for environmental science instruction is produced to support the development of large-scale professional language models, such as EnvGPT (Fig. 1). Further details on the system-level instructions and agent responsibilities are available (Supplementary Text S1), along with relevant code examples to aid in replication and extension (Supplementary Text S2).

2.4. Environmental science benchmarking

Benchmarking enables the effective comparison of different LLMs, thereby assessing their performance within specific domains [36]. Currently, no representative benchmark exists for LLMs in the field of environmental science [21,22]. To bridge this gap, we designed EnvBench, a benchmark for evaluating the performance of LLMs in environmental science. EnvBench also includes the five themes of environmental science and follows the same process as EnvInstruct to generate data from the raw corpus. EnvBench comprises 4998 samples covering various task types, including analysis, question–answer, and description. To facilitate detailed analyses by researchers, we annotated each EnvBench sample with specific task types and thematic labels, enabling performance assessments by both question category and environmental science theme (Supplementary Fig. S1). We employed GPT-4o to classify each benchmark item by theme, allowing for an LLM performance evaluation within each theme, for which the system-level instructions and code are provided separately (Supplementary Text S3).

To measure the incremental benefit of our environmental science instruction set under comparable resource budgets, we benchmarked EnvGPT against four models of similar size: LLaMA-2-7B [37], LLaMA-3.1–8B [38], Vicuna-1.5–7B [39], and GPT-4o-mini [40]. These baselines encompass two successive open-source models: a general-purpose, fine-tuned model and a compact, closed-source model. To contextualize these findings against a substantially larger open-source alternative, we also included Qwen 2.5–72B as an upper-bound reference [41]. The diversity in parameter scale, licensing, and prior fine-tuning regimes provides a balanced test bed for quantifying the performance gains afforded by EnvInstruct and our unified pipeline across different model classes and sizes (Supplementary Table S3).

In terms of evaluation methods, three approaches were employed: automated evaluation metrics, the LLM-as-a-judge method, and LLM scoring, for a comprehensive assessment of the performance of EnvGPT [42]. First, for automated evaluation, we employed two widely used metrics: bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE). Widely applied in natural language processing, these metrics reflect the similarity between generated and reference texts, thereby providing an objective measure of text generation quality. BLEU focuses on matching accuracy between texts, making it suitable for evaluating lexical and syntactic precision in generated texts [43]. ROUGE, however, is a measure of coverage and coherence that uses ROUGE-1, ROUGE-2, and ROUGE-L to provide a comprehensive assessment across multiple levels [44]. Second, with the LLM-as-a-judge method, an LLM is used to evaluate generated responses, thereby enhancing the objectivity and consistency of the assessment [45]. We selected GPT-4o as an evaluation model because of its high performance in generative tasks and notable comprehension abilities, enabling detailed assessments of semantic accuracy, fluency, and task completion in generated texts. The evaluation code and system-level instructions for this method, which enable GPT-4o to classify outcomes as wins, ties, or losses, are provided separately (Supplementary Text S4). Compared with traditional automated metrics, the LLM-as-a-judge method offers a more flexible evaluation approach, particularly in cases requiring subjective assessment, thus enhancing adaptability and reliability. In addition, we implemented LLM scoring, an evaluation method in which an LLM applies predefined rubrics to assess responses from a domain-specific perspective [22]. For this purpose, we set the GPT-4o to score responses across four key dimensions: accuracy of facts, real-world applicability, logical reasoning and structure, and clarity and expression, with detailed evaluation criteria and scoring instructions provided separately (Supplementary Text S5). The specific evaluation criteria are detailed in a rubric (Supplementary Table S4). This choice ensures a scalable, consistent, and domain-specific evaluation method, which is particularly important for large-scale assessments [46]. By integrating LLM scoring with automated and LLM-as-a-judge methods, we can achieve a scalable, consistent, and robust multimethod evaluation, offering deeper insights into the model's performance on environmental science tasks.

To provide an independent and robust assessment of model performance, we employed the EnviroExam benchmark, which consists of university-level multiple-choice examination questions across various environmental science disciplines [47]. Furthermore, to comprehensively evaluate the quality of model-generated responses on EnvBench using LLM scoring, we adopted a structured scoring framework from recent research to assess relevance, factuality, completeness, and style [48] (Supplementary Table S5). This combined approach enables objective benchmarking and in-depth evaluation of model responses, thus ensuring a thorough understanding of their capabilities within environmental science tasks.

To evaluate the real-world applicability of our model, we used the environmental large language model evaluation (ELLE) dataset—a structured approach specifically designed to test LLMs in solving real-world environmental science problems (https://elle.ceeai.net/). The dataset was developed by numerous experts in environmental science and contains 1130 question–answer pairs across 16 environmental themes [49]. We classified the ELLE dataset into three difficulty levels: easy, medium, and hard. The tasks were further categorized by type: calculation, reasoning, knowledge, transdisciplinary, and integration. In the evaluation, we used expert ratings across different difficulty levels and task types to assess the model's real-world applicability. By using the ELLE dataset, we can demonstrate the performance of different models on tasks of varying complexity and themes within the field of environmental science.

Hallucinations in LLMs are a recognized challenge that must be mitigated [50]. Prior research has shown that careful tuning of generation parameters, such as temperature and nucleus sampling (top-*p*), can substantially reduce spurious outputs. Very high-temperature settings tend to increase the risk of hallucinations in applications with open-ended prompts, whereas moderate top *p* values remove unlikely tokens and reduce incoherent content [51,52]. To explore mitigation strategies, we tested the base model LLaMA 3.1 8B under various parameter configurations and examined the representative responses (Supplementary Table S6). Based on these results, we selected moderate values of 0.6 for temperature and 0.8 for top *p* for all subsequent evaluations. This configuration provides a suitable trade-off between output diversity and factual reliability.

2.5. Base model and fine-tuning

EnvGPT is an LLM fine-tuned via LLaMA-3.1–8B, which is specifically designed for tasks in the field of environmental science. LLaMA-3.1–8B, a current-generation instruction-optimized model, offers significant advantages over its predecessor [38]. First, during the pretraining phase, approximately 15 trillion tokens from publicly available datasets spanning diverse semantic domains were utilized, thereby equipping the model with a robust knowledge foundation. Additionally, the fine-tuning phase included the use of publicly available instruction datasets and over 25 million synthetically generated examples, further increasing the model's ability to understand and execute complex task instructions [53]. Second, the eight-billion parameter size balances performance and resource consumption, enabling the model to capture rich contextual patterns and domain-specific features while ensuring high training and inference efficiency levels, rendering it ideal for building domain-specific models. Furthermore, the model excels in knowledge representation and generation, efficiently managing highly technical tasks, particularly those involving academic content and specialized texts [54]. Additionally, LLaMA-3.1–8B demonstrates high generalization capabilities across various benchmarks, adapting flexibly to diverse task scenarios—an essential feature for meeting the diverse demands of environmental science tasks.

During EnvGPT training, we employed the LoRA method to fine-tune the base model. LoRA is an efficient parameter-tuning method that aims to replace full-parameter updates with low-rank matrices, thereby significantly reducing computational and storage requirements while maintaining high performance for specific tasks [55]. In contrast to traditional full-parameter fine-tuning methods, the LoRA method maintains the weights of the pretrained model at a constant value, and only the added low-rank matrices are optimized. This approach significantly reduces the number of parameters to be updated and decreases memory usage during fine-tuning, making it highly practical in resource-limited environments [56].

The selection of the LoRA method was based on several key factors. First, its low-rank decomposition mechanism significantly reduces computational costs during fine-tuning, enabling efficient multidomain tuning of large models. Second, owing to its modular design, the fine-tuning process of LoRA can be seamlessly integrated with pretrained knowledge of the base model, which is a crucial feature for managing complex and diverse environmental science tasks. Finally, LoRA exhibits notable transferability, enabling rapid adaptation to instruction sets and task requirements across various domains. Its exceptional flexibility is particularly notable in handling cross-domain data within environmental science. The LoRA method offers a practical solution for efficiently fine-tuning EnvGPT, enabling the model to achieve exceptional performance and broad adaptability in specific environmental science tasks.

3. Results and discussion

3.1. Details of EnvGPT

The EnvGPT model was trained using the LoRA method to efficiently fine-tune the base model (i.e., LLaMA-3.1–8B) to meet the complex task requirements of the environmental science domain. During training, hyperparameters were optimized to balance model performance and computational resource usage (Supplementary Table S7). The training configuration included four NVIDIA RTX 4090 GPUs, a batch size of eight, and a total training time of approximately three days. The key training hyperparameters included the learning rate, LoRA rank, alpha value, and dropout rate. These settings work synergistically to ensure robust convergence and excellent adaptability for environmental science tasks.

The loss curve during training provides a visual representation of model convergence and performance changes (Supplementary Fig. S2). The training loss curve showed a steady downward trend, indicating a stable learning process during fine-tuning for environmental science tasks. In contrast, the validation loss curve fluctuated during its descent, reflecting the model's adjustments and adaptations to the complex domain data. The model file is now available on the Hugging Face platform for researchers and developers to further validate and apply. Hugging Face is a well-known machine learning model-sharing platform that is widely used for storing and distributing various pretrained models, with convenient support for model loading, inference, and fine-tuning processes [57]. We have provided detailed usage examples and loading code on the model page to help users start quickly, including how to load EnvGPT via Hugging Face's transformer library. The model file and more information can be accessed via the following link: https://huggingface.co/SustcZhangYX/EnvGPT.

3.2. ChatEnv instruction dataset

In this study, we developed the ChatEnv instruction dataset for the environmental science domain to address the domain's diverse task requirements. ChatEnv is based on raw samples from the environmental science corpus and is generated using an instruction generation framework that covers five core domains, comprising a total of 112,946 samples. The dataset contains over 100 million tokens and is available on the Hugging Face platform for validation and application by researchers and developers (https://huggingface.co/datasets/SustcZhangYX/ChatEnv).

The distribution of original samples and generated instruction samples across five themes highlights the ability of the EnvInstruct framework to produce a balanced instruction set (Fig. 2). The framework ensures an even distribution of instructions across domains, underscoring its ability to generate high-quality,

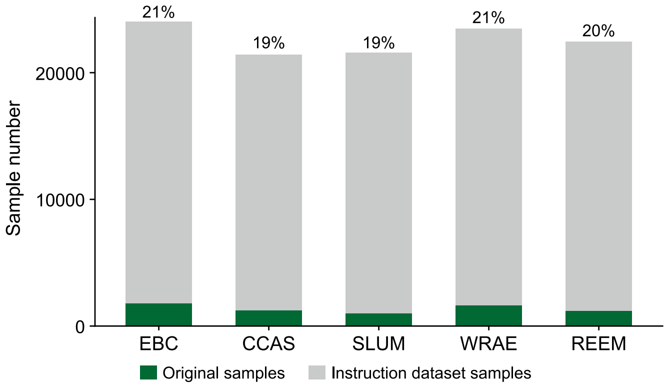

Fig. 2. Comparison of the original samples and the generated instruction dataset across the five themes in ChatEnv. Original samples (green) and generated instruction dataset samples (grey) are shown for each theme, with percentages indicating the balanced distribution achieved by EnvInstruct. The five themes are: renewable energy and environmental management (REEM), water resources and aquatic environment (WRAE), soil and land-use management (SLUM), climate change and atmospheric science (CCAS), and ecosystems and biodiversity conservation (EBC).

domain-specific content. The CCAS theme, with 1207 original samples, generated 21,437 instructions, whereas the EBC, with 1764 original samples, expanded to 24,018 instructions. The WRAE theme, starting with 1613 samples, produced 23,487 instructions. Respectively, SLUM and REEM, with 1006 and 1190 original samples, generated 21,563 and 22,441 instructions. Despite differences in initial sample sizes, EnvInstruct has balanced instruction generation across themes. This distribution ensures comprehensive coverage of environmental science domains, fostering interdisciplinary integration and supporting diverse environmental tasks. This approach reflects the adaptability and scalability of the EnvInstruct framework, which successfully expanded the dataset across multiple environmental domains while maintaining a fair distribution of content. By maintaining a balanced instruction dataset, this dataset can be utilized for SFT LLMs, enabling them to acquire knowledge of environmental science across multiple domains. The proportional distribution of instructions across themes is as follows: EBC (21 %), WRAE (21 %), REEM (20 %), SLUM (19 %), and CCAS (19 %). The balanced distribution reflects EnvInstruct's capacity to generate instructions evenly across environmental science domains, ensuring robust support for diverse research and applications.

The development of ChatEnv has established a sound foundation for creating LLMs tailored to the field of environmental science. As the first instruction dataset specifically designed for environmental science, ChatEnv offers a rich and diverse collection of task instructions spanning multiple key areas of the domain. The data distribution of ChatEnv effectively mirrors current research hotspots and trends in environmental science. The diverse instruction set generated by EnvInstruct covers the five core themes of environmental science, ensuring the richness and breadth of the ChatEnv dataset while guiding future researchers in expanding and refining model application scenarios to meet real-world needs. By offering training data closely tied to real-world applications, ChatEnv enables LLMs to address complex tasks effectively in environmental science, enhancing their understanding of environmental issues and the value of practical applications.

3.3. *Performance evaluation via automated metrics*

For automated metric evaluation, EnvBench was utilized as the benchmark dataset (https://huggingface.co/datasets/SustcZhangYX/EnvBench), encompassing tasks across the five core themes of environmental science. EnvBench provides extensive coverage and notable domain adaptability, offering a robust foundation for comprehensive model performance evaluation. The performance of EnvGPT and the four baseline models (LLaMA-2-7B, LLaMA-3.1–8B, GPT-4o-mini, and Vicuna-1.5–7B) was evaluated using the ROUGE and BLEU metrics across the five major themes, revealing notable performance differences and trends (Fig. 3) [58]. Overall, EnvGPT consistently outperformed the baseline models across all the themes and evaluation metrics, demonstrating its superior ability to solve environmental science tasks. The comparison with Vicuna-1.5–7B revealed a marked decline in performance when fine-tuned on non-environmental science data, underscoring the critical importance of domain-specific training for optimal performance. Notably, the performance differences among the baseline models revealed the complex effects of model architecture, parameter scaling, and domain specialization on their performance. In detailed evaluations, LLaMA-3.1–8B generally outperformed LLaMA-2-7B, especially in terms of ROUGE-1 and ROUGE-L scores. This suggests that increasing the parameter size (from 7B to 8B) and refining the instruction alignment processes notably enhanced the text generation capabilities of the model [59]. The notable lead of EnvGPT and GPT-4o-mini in terms of the ROUGE and BLEU scores highlights their strengths in semantic understanding and generation consistency. Additionally, these results demonstrate the differing evaluation focuses of the two metrics: ROUGE emphasizes content coverage and similarity, whereas BLEU prioritizes grammatical and phrasing consistency.

Specifically, regarding the ROUGE metrics, EnvGPT and GPT-4o-mini presented notable advantages, particularly in terms of the ROUGE-1 and ROUGE-L scores. This can be attributed to the advanced instruction alignment and text-generation optimization capabilities of GPT-4o-mini and EnvGPT. For example, in the CCAS and EBC categories, EnvGPT achieved ROUGE-1 and ROUGE-L scores of 51.93 and 38.48 (Supplementary Fig. S3a) and 51.19 and 33.42 (Supplementary Fig. S3b), respectively, with GPT-4o-mini closely trailing. This suggests that both models possess superior capabilities in terms of word coverage and text structural consistency, allowing them to generate content that is more precisely aligned with domain-specific task requirements. In contrast, LLaMA-3.1–8B, LLaMA-2-7B, and Vicuna-1.5–7B achieved lower ROUGE scores, likely because of the absence of instruction fine-tuning in the environmental science domain and relatively limited task adaptability. With respect to the BLEU metric, EnvGPT

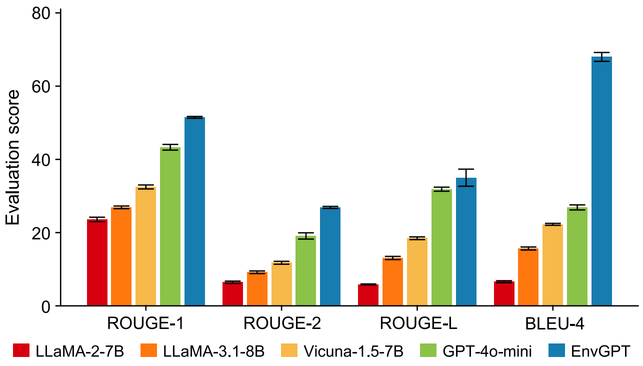

Fig. 3. Performance comparison of the EnvGPT and baseline models across multiple evaluation metrics. Bar chart showing the evaluation scores of EnvGPT, GPT-4o-mini, Vicuna-1.5–7B, LLaMA-2-7B, and LLaMA-3.1–8B on the basis of the ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) and BLEU metrics. ROUGE: recall-oriented understudy for gisting evaluation; BLEU: bilingual evaluation understudy.

provided an even greater advantage. BLEU emphasizes precise text matching, which is especially critical for environmental science tasks that demand the accurate generation of technical terms and complex semantics. For example, in the REEM, SLUM, and WRAE categories, EnvGPT achieved BLEU scores of 69.33, 68.42, and 69.06, respectively (Supplementary Figs. S3c, d, e), notably surpassing those of the other models. This notable lead can be attributed to the high-quality instruction coverage of the ChatEnv dataset, which allows EnvGPT to generate more precise and technically robust content. Furthermore, the BLEU results highlight the linguistic fluency and attention to detail of EnvGPT in managing complex tasks, which are particularly crucial for highly specialized tasks in environmental science.

EnvGPT consistently outperformed the baseline models, demonstrating exceptional accuracy and consistency in language generation, with significantly higher BLEU scores than those of the GPT-4o-mini, Vicuna-1.5–7B, and LLaMA series models. In terms of the ROUGE metrics, EnvGPT and GPT-4o-mini achieved high mean scores, underscoring the importance of domain-specific datasets and advanced instruction alignment techniques in managing complex domain tasks. The remarkable advantage of EnvGPT lies in its integration of domain-specific datasets during pretraining and the application of instruction fine-tuning, enabling greater consistency and accuracy and thus establishing a new benchmark for performance in environmental science tasks. EnvGPT targets efficient domain-specific performance, outperforming GPT-4o-mini despite its smaller size, making it ideal for resource-limited environmental applications. Additionally, compared with the LLaMA series, Vicuna-1.5–7B demonstrated improvements in language generation performance due to SFT. However, the SFT process utilized non-environmental science datasets, leading to a considerable gap in performance compared with EnvGPT. This comprehensive analysis revealed that the differences between the models were influenced not only by parameter size but also by dataset specialization and instruction optimization strategies, offering valuable insights for the development of future LLMs in environmental science.

3.4. Expert evaluation via LLM-as-a-judge

The task generation quality of EnvGPT and the four baseline models (LLaMA-2-7B, LLaMA-3.1–8B, GPT-4o-mini, and Vicuna-1.5–7B) was evaluated using the LLM-as-a-judge method with the EnvBench benchmark dataset, which shows the detailed distributions of the win, tie, and loss rates (Fig. 4). Overall, EnvGPT attained a notable advantage, particularly over the Vicuna-1.5–7B and LLaMA series models, with win rates exceeding 60 %. This demonstrates the effectiveness of domain-specific fine-tuning in enhancing content quality and domain adaptability. However, compared with GPT-4o-mini, the advantage of EnvGPT decreased, with its win rate decreasing to 50.37 % and its tie rate increasing to 31.88 %, reflecting the high competitiveness of GPT-4o-mini in general instruction tasks.

Specifically, EnvGPT achieved a win rate of 61.56 % and a loss rate of 35.55 % compared with LLaMA-3.1–8B. This result indicates that despite the larger parameter size and optimized instruction alignment of LLaMA-3.1–8B, the absence of domain-specific fine-tuning restricts its adaptability and accuracy in environmental science tasks. This performance gap can be attributed to the ChatEnv dataset, which enables EnvGPT to better comprehend and generate complex content in the environmental science domain. Similarly, EnvGPT achieved a win rate of 60.67 % and a loss rate of 24.86 % compared with LLaMA-2-7B, further demonstrating that parameter scaling alone offers limited performance improvements without domain-specific fine-tuning. Compared with Vicuna-1.5–7B, EnvGPT achieved the highest win rate, at 63.08 %. This suggests that fine-tuning with non-environmental science data, such as that performed with Vicuna-1.5–7B, may diminish the model's ability to retain domain-specific knowledge in environmental science. Compared with GPT-4o-mini, EnvGPT achieved a win rate of 50.37 %, with the tie rate increasing to 31.88 %. This suggests that although EnvGPT significantly outperformed GPT-4o-mini in terms of the BLEU and ROUGE metrics (Fig. 3), their performance levels were more comparable in subjective evaluations. This phenomenon may be attributed to the broader fine-tuning of general instruction data in GPT-4o-mini, which enables it to maintain a certain degree of text fluency and acceptability despite its limited domain adaptability. Nevertheless, EnvGPT maintained its lead by exhibiting superior performance in generating domain-specific terminology and structured information in environmental science. This comparison highlights the critical role of domain-specific fine-tuning while revealing the limitations of general language models that lack domain-specific training.

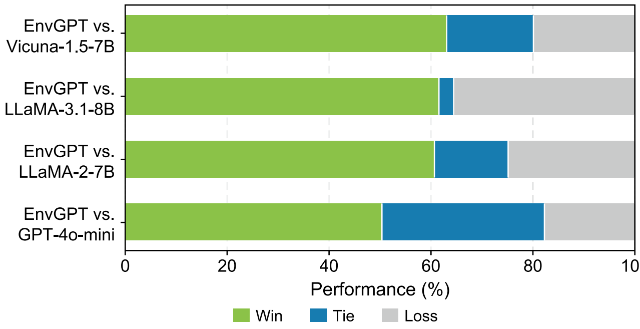

Fig. 4. Comparative evaluation of EnvGPT against four baseline models via the large language model-as-a-judge approach. The chart shows the win, tie, and loss rates of EnvGPT in comparison with those of LLaMA-2-7B, LLaMA-3.1–8B, GPT-4o-mini, and Vicuna-1.5–7B.

The significance of this evaluation lies not only in validating the exceptional performance of EnvGPT in environmental science tasks but also in emphasizing the pivotal role of the domain-specific instruction dataset ChatEnv. Supported by high-quality data, EnvGPT outperforms both general models and baseline models that lack domain-specific fine-tuning in specialized fields, thereby providing a valuable reference for the development of more efficient and accurate domain-specific LLMs in the future. Furthermore, these results suggest that although general models (e.g., GPT-4o-mini) exhibit certain strengths in task generalization, targeted fine-tuning remains a pivotal strategy for enhancing performance in specialized domains.

3.5. Independent domain benchmarking and rubric-based evaluation

To complement the automated metrics and the LLM-as-a-judge analysis, we performed two additional evaluations focusing on factual accuracy and answer quality from distinct perspectives. The first evaluation utilized the EnviroExam benchmark, which comprises university-level multiple-choice questions covering the core disciplines of environmental science [47]. EnvGPT achieved an average accuracy of 92.06 ± 1.85 % (95 % confidence interval) (Fig. 5). This result surpassed the performance of the parameter-matched baselines of LLaMA-3.1–8B (84.06 ± 3.19 %) and Vicuna-1.5–7B (63.77 ± 5.09 %). Paired t-tests confirmed these differences as highly significant ($p < 5 \times 10^{-6}$). The accuracy of EnvGPT was statistically indistinguishable from that of GPT-4o-mini (91.11 ± 2.74 %, $p = 0.53$) and slightly less than the considerably

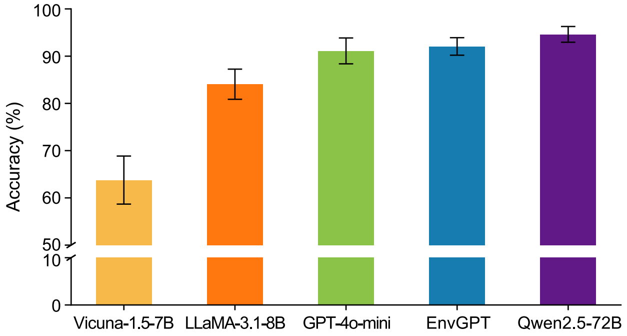

Fig. 5. EnviroExam benchmark accuracy for Vicuna-1.5–7B, LLaMA-3.1–8B, GPT-4o-mini, EnvGPT, and Qwen2.5–72B. Columns display the mean correct response rate for each model; error bars indicate the corresponding 95 % confidence intervals.

larger value of Qwen2.5–72B (94.61 ± 1.67 %, $p = 0.046$). These findings indicate that EnvGPT achieves performance comparable to that of a significantly larger open-source model, despite using fewer parameters.

The second evaluation assessed free-form answers from EnvBench using LLM scoring, for which GPT-4o assigned ratings (0–5 scale) for relevance, factuality, completeness, and style using a detailed rubric (Supplementary Table S5) [48]. EnvGPT obtained the highest or joint-highest scores across all dimensions, showing particularly notable advantages in terms of completeness (4.38 ± 0.19) and factuality (4.70 ± 0.15) (Fig. 6). Compared with the LLaMA-3.1–8B baseline, EnvGPT improved scores by +0.52 in relevance, +0.56 in factuality, +0.87 in completeness, and +0.40 in style, clearly demonstrating the effectiveness of the EnvInstruct dataset. GPT-4o-mini performed comparably in relevance and style but was inferior to EnvGPT in completeness and cross-domain factual depth. In contrast, Vicuna-1.5–7B consistently showed limited performance across all criteria because of its lack of environmental content in the fine-tuning corpus.

Taken together, the EnviroExam scores and LLM scoring results consistently indicate that SFT on domain-specific data critically enhances model performance. Despite having order-of-magnitude fewer parameters than Qwen 2.5–72B, EnvGPT matched or surpassed its performance in multiple knowledge and quality metrics. This confirms that targeted SFT with high-quality instructions could effectively compensate for the reduced model size. The curated ChatEnv dataset provides high-quality instructions, and EnvBench serves as a transparent, diverse task platform that reveals model strengths and weaknesses in detail. Collectively, these resources provide a reproducible framework for constructing and rigorously evaluating future environmental science LLMs. They

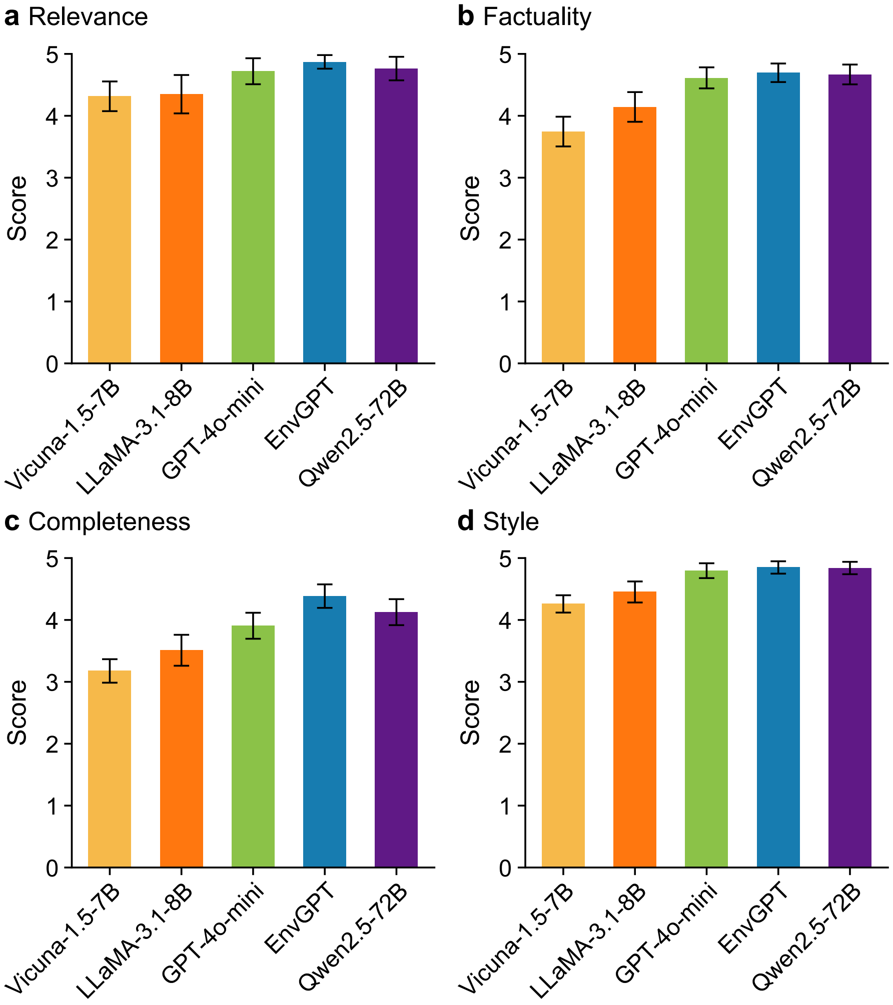

Fig. 6. Large language model scoring evaluation of EnvBench responses across four quality dimensions. Panels show the mean scores assigned by GPT-4o (0–5 scale) for relevance (**a**), factuality (**b**), completeness (**c**), and style (**d**). The error bars denote 95 % confidence intervals.

highlight the broader principle that carefully designed, domain-specific data are essential for successfully adapting LLMs to specialized scientific applications.

3.6. Evaluating the real-world applicability of LLMs

To assess the real-world applicability of LLMs in environmental science, we employed LLM scoring to evaluate the performance of LLaMA-2-7B, LLaMA-3.1–8B, Vicuna-1.5–7B, GPT-4o-mini, and EnvGPT across four key dimensions: factual accuracy, real-world applicability, logical reasoning and structure, and clarity and expression. The evaluation used the ELLE dataset. This dataset, developed by environmental science experts, was specifically designed to evaluate LLM performance on real-world environmental tasks. Model performance examples across different task types and difficulty levels, along with LLM scoring details, are provided separately (Supplementary Table S8–S10).

The five models were compared across the four evaluation dimensions using LLM scoring (Fig. 7). The box plots highlight performance differences, with EnvGPT consistently outperforming baseline models, particularly in terms of factual accuracy and real-world applicability ($p < 0.05$). LLaMA-3.1–8B was better suited for complex environmental science tasks than LLaMA-2-7B ($p < 0.05$). This improvement resulted from the larger model size and its ability to utilize a broader range of training data, thereby improving generalizability and task performance. In contrast, Vicuna-1.5–7B scored lower than LLaMA-3.1–8B, GPT-4o-mini, and EnvGPT across all four dimensions ($p < 0.05$). This was due to the fine-tuning of domain-specific data of Vicuna-1.5–7B, which lacks environmental science–specific data, limiting its effectiveness in addressing environmental science tasks. This underscores the importance of fine-tuning domain-specific data to optimize LLMs for specialized tasks. In contrast, GPT-4o-mini performed like EnvGPT in terms of factual accuracy and clarity ($p > 0.05$) but lagged behind EnvGPT in real-world applicability ($p < 0.05$). This demonstrates that domain-specific SFT is still a viable method for improving LLM performance.

The performance of the LLMs was assessed across different task difficulties and types (Fig. 8). In terms of real-world applicability, the five models were compared across three difficulty levels (easy, medium, and hard) (Fig. 8a). Clearly, EnvGPT outperformed all the baseline models at every difficulty level. In the easy and medium categories, the EnvGPT scores were significantly higher than those of the other models, at 83.50 ± 7.63 (easy), 88.04 ± 5.90 (medium), and 84.66 ± 9.24 (hard). These results highlight the model's advantages in addressing real-world environmental tasks. In

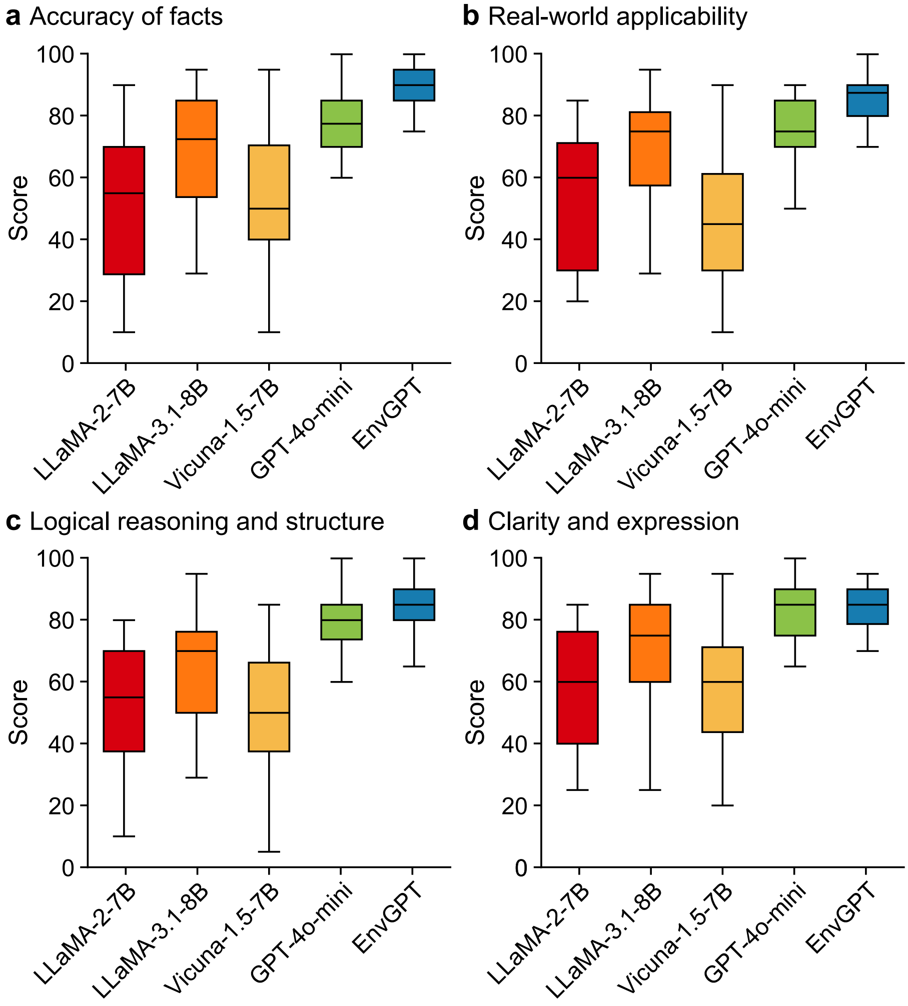

Fig. 7. Comparison of the performance of five large language models (LLMs) across four evaluation dimensions via LLM scoring. Box plots display the distribution of scores for five models (LLaMA-2-7B, LLaMA-3.1–8B, Vicuna-1.5–7B, GPT-4o-mini, and EnvGPT) across four key dimensions: accuracy of facts (**a**), real-world applicability (**b**), logical reasoning and structure (**c**), and clarity and expression (**d**).

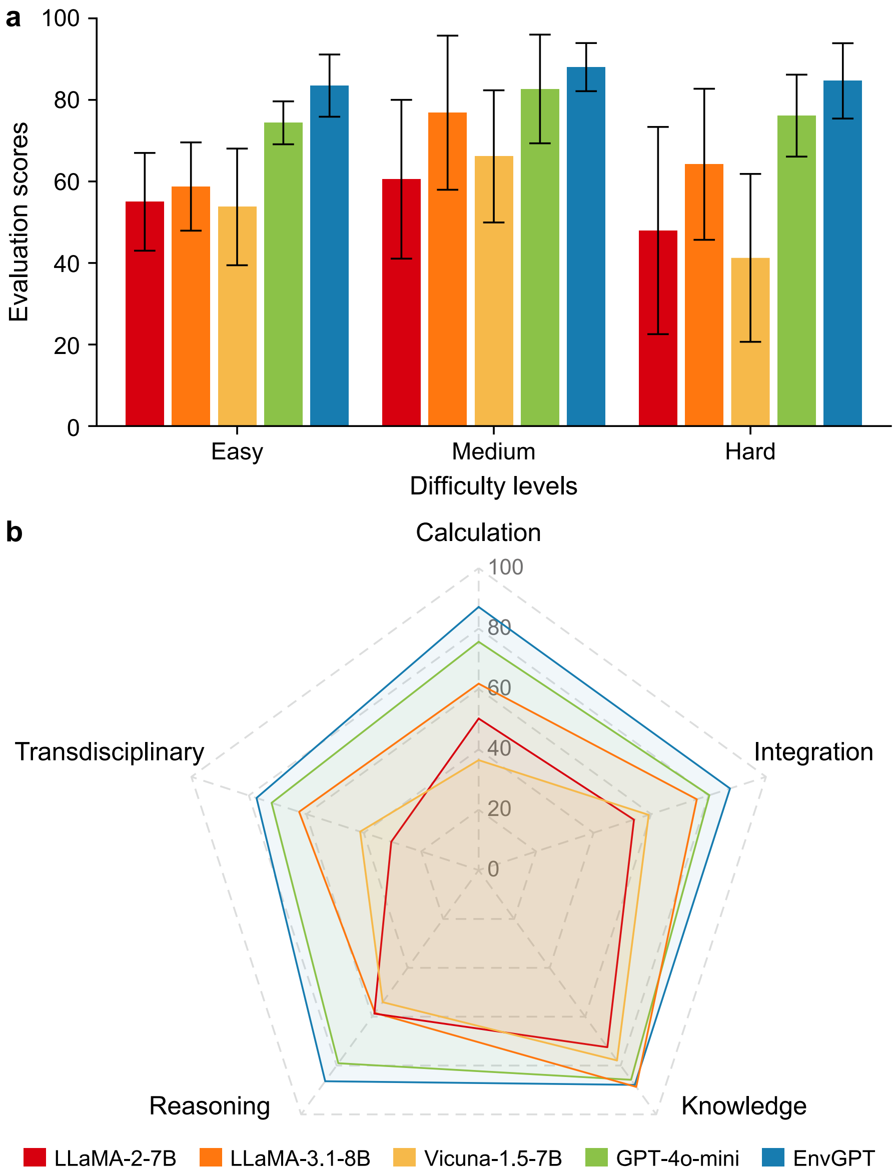

Fig. 8. Performance of large language models across task difficulties and types in real-world environmental science applications. **a**, Evaluation scores for real-world applicability across three difficulty levels (easy, medium, and hard) for LLaMA-2-7B, LLaMA-3.1–8B, Vicuna-1.5–7B, GPT-4o-mini, and EnvGPT. **b**, Evaluation scores across different task types, including calculation, integration, knowledge, reasoning, and transdisciplinary tasks, for LLaMA-2-7B, LLaMA-3.1–8B, Vicuna-1.5–7B, GPT-4o-mini, and EnvGPT.

contrast, GPT-4o-mini also performed well, particularly in the medium category, with a score of 82.68 ± 13.33. Although Vicuna-1.5–7B was fine-tuned on domain-specific data, its scores remained lower than those of LLaMA-3.1–8B, GPT-4o-mini, and EnvGPT, particularly in the hard category. For example, its score in the hard category was only 41.25 ± 20.59—much lower than the scores of the other LLMs. This is because the fine-tuning of Vicuna-1.5–7B on non-environmental science data limited its ability to address environmental science problems effectively. Similarly, LLaMA-3.1–8B, with a larger training dataset and improved architecture, outperformed LLaMA-2-7B, especially in the medium and hard categories, with scores of 76.86 ± 18.90 (medium) and 64.20 ± 18.52 (hard), whereas LLaMA-2-7B scored 60.54 ± 19.47 (medium) and 47.95 ± 25.41 (hard). The performance of these models was also broken down by different task types, including calculation, integration, knowledge, reasoning, and transdisciplinary tasks (Fig. 8b). EnvGPT scored 87.17 in calculation, 87.50 in integration, 87.92 in knowledge, 86.44 in reasoning, and 77.33 in transdisciplinary tasks. In contrast, Vicuna-1.5–7B showed a clear performance gap, especially in the calculation and transdisciplinary tasks, with scores of 36.43 and 41.25, respectively. LLaMA-3.1–8B outperformed LLaMA-2-7B in these dimensions, scoring 61.75 in calculation and 75.94 in integration, compared with LLaMA-2-7B's scores of 50.21 and 54.06, respectively. Overall, LLaMA-3.1–8B still lagged behind EnvGPT, particularly in reasoning tasks, in which EnvGPT excelled with a score of 86.44, whereas LLaMA-3.1–8B scored 58.44. The GPT-4o-mini performed excellently in terms of knowledge and reasoning, with scores of 85.83 and 79.06, respectively. However, it lagged behind EnvGPT in transdisciplinary tasks (GPT-4o-mini: 72.08, EnvGPT: 77.33) and in integrating domain-specific knowledge across all dimensions.

The results highlight EnvGPT's robust performance across a range of environmental science tasks, demonstrating superior real-world applicability. The model excels in integrating knowledge from multiple disciplines—a crucial capability for addressing complex, real-world environmental issues. While models such as LLaMA-3.1–8B and GPT-4o-mini showed strong overall performance, EnvGPT consistently outperformed them, particularly in integration and interdisciplinary tasks. These findings underscore the significant impact of domain-specific SFT in optimizing LLMs for specialized environmental science challenges.

3.7. Limitations

SFT on ChatEnv significantly enhances EnvGPT; however, several limitations should be acknowledged. Temporal drift is inevitable since an SFT model captures knowledge only at the time of training. Frequent retraining is thus necessary to incorporate new datasets, regulations, and discoveries, resulting in increased computational and environmental costs. In contrast, retrieval-augmented systems, such as ChatClimate and GeoGPT, update an external knowledge base instead of model parameters, providing a more efficient maintenance approach [60,61].

Another limitation concerns coverage. Although ChatEnv covers five major subfields, any instruction set is inherently limited in scope. Consequently, long-tail topics or interdisciplinary queries may not be fully represented. Retrieval pipelines can complement SFT by incorporating current domain-specific evidence at inference time, thus addressing limitations in the static fine-tuning corpus.

Finally, balancing the model scale and resource demand remains crucial. Compared with the significantly larger models, the eight-billion-parameter EnvGPT reduces the accuracy gap. However, larger architectures still retain a slight absolute advantage over benchmarks that demand extensive knowledge.

4. Conclusions

This study proposed a reproducible and extensible workflow for developing LLMs that serve the needs of environmental science. The workflow began with EnvInstruct, a multi-agent procedure that extracts high-quality prompts from the primary literature and assembles ChatEnv, a 100 million token instruction set balanced across climate and atmospheric science, ecosystems and biodiversity, water resources and aquatic environments, soil and land-use management, and renewable energy and environmental management. SFT of an eight billion parameter backbone on ChatEnv yielded EnvGPT, whereas the companion benchmark EnvBench provided 4998 task instances with explicit theme and task-type labels for rigorous evaluation.

Comprehensive testing demonstrated the value of this domain-specific fine-tuning strategy. EnvGPT outperformed the parameter-matched baselines in BLEU and ROUGE, outperformed the majority of the comparisons in LLM-as-a-judge tests, and achieved the highest rubric-based scores for relevance, factuality, completeness, and style on EnvBench. For the independent EnviroExam benchmark, EnvGPT attained 92 % accuracy, matching or exceeding closed-source GPT-4o-mini and approaching the performance of the much larger Qwen 2.5–72B model. These results confirm that carefully engineered, domain-specific SFT can narrow or even close the gap to models that rely primarily on scale.

Nevertheless, fine-tuning alone cannot address every limitation. Knowledge becomes outdated as the underlying literature evolves, and niche or cross-disciplinary questions may fall outside the scope of any finite instruction set. Future work will therefore explore hybrid approaches that couple an SFT-adapted backbone with retrieval-augmented generation, tool calling, or knowledge-graph integration, as demonstrated in recent systems, such as ChatClimate and GeoGPT. Additional optimization avenues include reinforcement or active-learning loops in which domain experts curate challenging queries, as well as the incorporation of geospatial and other multimodal data sources to extend reasoning beyond text.

To support these directions, ChatEnv and EnvBench will be updated on a rolling basis to reflect newly published research, revised regulations, and emerging data modalities. EnvGPT itself will be maintained through periodic resource-efficient refresh cycles, ensuring that the model remains an accurate, current, and versatile assistant for environmental science research and practice.

CRediT authorship contribution statement

Yuanxin Zhang: Writing – original draft, Investigation, Formal analysis, Data curation. **Sijie Lin:** Validation, Software, Methodology, Conceptualization. **Yaxin Xiong:** Visualization, Investigation, Data curation. **Nan Li:** Writing – review & editing, Supervision, Conceptualization. **Lijin Zhong:** Visualization, Resources, Methodology, Data curation. **Longzhen Ding:** Visualization, Investigation, Formal analysis. **Qing Hu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the National Key Research and Development Program of China (2024YFC3711800) and the High-level University Special Fund (G03050K001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2025.100608.

References

[1] S. Sauvé, S. Bernard, P. Sloan, Environmental sciences, sustainable development and circular economy: alternative concepts for trans-disciplinary research, Environ. Dev. 17 (2016) 48–56, https://doi.org/10.1016/j.envdev.2015.09.002.

[2] H. Deng, S. Liu, M. Hu, Y. Shen, Y. Qin, W. Pan, et al., Promoting the integration of ESG and eco-environmental sciences, Sheng Tai Xue Bao 44 (2024) 8774–8783, https://doi.org/10.20103/j.stxb.202404080749.

[3] F. Su, P. Li, X. He, V. Elumalai, Set pair analysis in Earth and environmental sciences: development, challenges, and future prospects, Expo. Health 12 (2020) 343–354, https://doi.org/10.1007/s12403-020-00368-3.

[4] T. Gundersen, D. Alinejad, T. Branch, B. Duffy, K. Hewlett, C. Holst, et al., A new dark age? Truth, trust, and environmental science, Annu. Rev. Environ. Resour. 47 (2022) 5–29, https://doi.org/10.1146/annurev-environ-120920-015909.

[5] B.B. Cael, K. Bisson, E. Boss, S. Dutkiewicz, S. Henson, Global climate-change trends detected in indicators of ocean ecology, Nature 619 (2023) 551–554, https://doi.org/10.1038/s41586-023-06321-z.

[6] C. Deser, F. Lehner, K.B. Rodgers, T. Ault, T.L. Delworth, P.N. DiNezio, et al., Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Change 10 (2020) 277–286, https://doi.org/10.1038/s41558-020-0731-2.

[7] L. Lai, Y. Liu, Y. Zhang, Z. Cao, Y. Yin, X. Chen, et al., Long-term spatiotemporal mapping in lacustrine environment by remote sensing:Review with case study, challenges, and future directions, Water Res. 267 (2024), https://doi.org/10.1016/j.watres.2024.122457.

[8] L. Li, M. Fu, Y. Zhu, H. Kang, H. Wen, The current situation and trend of land ecological security evaluation from the perspective of global change, Ecol. Indic. 167 (2024), https://doi.org/10.1016/j.ecolind.2024.112608.

[9] N. Pidgeon, B. Fischhoff, The role of social and decision sciences in communicating uncertain climate risks, Nat. Clim. Change 1 (2011) 35–41, https://doi.org/10.1038/nclimate1080.

[10] J.-J. Zhu, Z.J. Ren, The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining, Resour. Conserv. Recycl. 190 (2023) 106876, https://doi.org/10.1016/j.resconrec.2023.106876.

[11] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, et al., Machine learning: new ideas and tools in environmental science and engineering, Environ. Sci. Technol. 55 (2021) 12741–12754, https://doi.org/10.1021/acs.est.1c01339.

[12] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, et al., A brief overview of ChatGPT: the history, status quo and potential future development, IEEE-CAA J. Autom. Sin. 10 (2023) 1122–1136, https://doi.org/10.1109/JAS.2023.123618.

[13] J.-B. Mouret, Large language models help computer programs to evolve, Nature 625 (2024) 452–453, https://doi.org/10.1038/d41586-023-03998-0.

[14] M.C. Rillig, A. Kasirzadeh, AI personal assistants and sustainability: risks and opportunities, Environ. Sci. Technol. 58 (2024) 7237–7239, https://doi.org/10.1021/acs.est.4c03300.

[15] J. Barile, A. Margolis, G. Cason, R. Kim, S. Kalash, A. Tchaconas, et al., Diagnostic accuracy of a large language model in pediatric case studies, JAMA Pediatr. 178 (2024) 313–315, https://doi.org/10.1001/jamapediatrics.2023.5750.

[16] J. Lai, W. Gan, J. Wu, Z. Qi, P.S. Yu, Large language models in law: a survey, AI Open 5 (2024) 181–196, https://doi.org/10.1016/j.aiopen.2024.09.002.

[17] M. Leippold, Thus spoke GPT-3: interviewing a large-language model on climate finance, Finance Res. Lett. 53 (2023) 103617.

[18] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172, https://doi.org/10.1038/s41586-023-06291-2.

[19] E. Agathokleous, C.J. Saitanis, C. Fang, Z. Yu, Use of ChatGPT: what does it mean for biology and environmental science? Sci. Total Environ. 888 (2023) 164154 https://doi.org/10.1016/j.scitotenv.2023.164154.

[20] M.C. Rillig, M. Ågerstrand, M. Bi, K.A. Gould, U. Sauerland, Risks and benefits of large language models for the environment, Environ. Sci. Technol. 57 (2023) 3464.

[21] Y. Wu, M. Xu, S. Liu, Generative artificial intelligence: a new engine for advancing environmental science and engineering, Environ. Sci. Technol. 58 (2024) 17524–17528, https://doi.org/10.1021/acs.est.4c07216.

[22] J.-J. Zhu, M. Yang, J. Jiang, Y. Bai, D. Chen, Z.J. Ren, Enabling GPTs for expert-level environmental engineering question answering, Environ. Sci. Technol. Lett. (2024), https://doi.org/10.1021/acs.estlett.4c00665.

[23] J.-J. Zhu, J. Jiang, M. Yang, Z.J. Ren, ChatGPT and environmental research, Environ. Sci. Technol. 57 (2023) 17667.

[24] Y. Ren, T. Zhang, X. Dong, W. Li, Z. Wang, J. He, et al., WaterGPT: training a large language model to become a hydrology expert, Water 16 (2024) 3075, https://doi.org/10.3390/w16213075.

[25] D. Thulke, Y. Gao, P. Pelser, R. Brune, R. Jalota, F. Fok, et al., Climategpt: towards Ai Synthesizing Interdisciplinary Research on Climate Change, 2024. URL Httpsarxiv Orgabs240109646 n.d.

[26] Z. Bi, N. Zhang, Y. Xue, Y. Ou, D. Ji, G. Zheng, et al., OceanGPT: a large language model for ocean science tasks, ArXiv Prepr ArXiv231002031 (2023).

[27] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, Z. Wang, Adversarial robustness: from self-supervised pre-training to fine-tuning. Proc. IEEECVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 699–708.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (2019) 9.

[29] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, et al., Instruction tuning for large language models: a survey, ArXiv Prepr ArXiv230810792 (2023).

[30] S. Schmidgall, C. Harris, I. Essien, D. Olshvang, T. Rahman, J.W. Kim, et al., Evaluation and mitigation of cognitive biases in medical language models, npj Digit. Med. 7 (2024), https://doi.org/10.1038/s41746-024-01283-6.

[31] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, et al., Lamm: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, Adv. Neural Inf. Process. Syst. 36 (2024).

[32] OpenAI, GPT-4o, OpenAI's New Flagship Model that Can Reason across Audio, Vision, and Text in Real Time, 2024.

[33] M. Kipp, From GPT-3.5 to GPT-4.o: a leap in AI's medical exam performance, Information 15 (2024), https://doi.org/10.3390/info15090543.

[34] Z. Sun, J. Yang, N. Zhang, H. Wu, C. Li, GPT-4o is more like a real person: potentials in surgical oncology, Int. J. Surg. (2024) 10–1097.

[35] L. Wang, Y. Mao, L. Wang, Y. Sun, J. Song, Y. Zhang, Suitability of GPT-4o as an evaluator of cardiopulmonary resuscitation skills examinations, Resuscitation (2024) 110404.

[36] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T.B. Hashimoto, Benchmarking large language models for news summarization, Trans. Assoc. Comput. Linguist 12 (2024) 39–57.

[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, et al., Llama 2: open foundation and fine-tuned chat models, ArXiv Prepr ArXiv230709288 (2023).

[38] Meta, Llama 3.1: A Multilingual Large Language Model Pretrained and Fine-Tuned with Instructions, 2024.

[39] C. Deng, T. Zhang, Z. He, Y. Xu, Q. Chen, Y. Shi, et al., K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization, 2023.

[40] OpenAI, GPT-4o Mini: A Small Model by OpenAI with Advanced Text Intelligence and Multimodal Reasoning Capabilities, 2024.

[41] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, et al., Qwen2.5 Technical Report, 2025.

[42] S. Shankar, J. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, I. Arawjo, Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. Proc. 37th Annu. ACM Symp. User Interface Softw. Technol., 2024, pp. 1–14.

[43] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, Proc. 40th Annu. Meet. Assoc. Comput. Linguist. (2002) 311–318.

[44] C.-Y. Lin, Rouge: a package for automatic evaluation of summaries, Text Summ. Branches Out (2004) 74–81.

[45] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Adv. Neural Inf. Process. Syst. 36 (2023) 46595–46623.

[46] X. Wu, P.P. Saraf, G.-G. Lee, E. Latif, N. Liu, X. Zhai, Unveiling scoring processes: dissecting the differences between llms and human graders in automatic scoring, ArXiv Prepr ArXiv240718328 (2024).

[47] Y. Huang, L. Guo, W. Guo, Z. Tao, Y. Lv, Z. Sun, et al., EnviroExam: benchmarking environmental science knowledge of large language models, ArXiv Prepr ArXiv240511265 (2024).

[48] J. Bulian, M.S. Schäfer, A. Amini, H. Lam, M. Ciaramita, B. Gaiarin, et al., Assessing large language models on climate information, ArXiv Prepr ArXiv231002932 (2023).

[49] J. Guo, N. Li, M. Xu, Environmental Large Language Model Evaluation (ELLE) dataset: a benchmark for evaluating generative AI applications in eco-environment domain, ArXiv Prepr ArXiv250106277 (2025).

[50] R. Azamfirei, S.R. Kudchadkar, J. Fackler, Large language models and the perils of their hallucinations, Crit. Care 27 (2023) 120.

[51] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, et al., A comprehensive survey of hallucination mitigation techniques in large language models, ArXiv Prepr ArXiv240101313 6 (2024).

[52] M. Renze, The effect of sampling temperature on problem solving in large language models. Find. Assoc. Comput. Linguist. EMNLP 2024, Association for Computational Linguistics, 2024, pp. 7346–7356, https://doi.org/10.18653/v1/2024.findings-emnlp.432.

[53] H. Huang, B. Xu, X. Liang, K. Chen, M. Yang, T. Zhao, et al., Multi-view fusion for instruction mining of large language model, Inf. Fusion 110 (2024) 102480.

[54] R. Vavekanand, K. Sam, Llama 3.1: an In-Depth Analysis of the Next-Generation Large Language Model, 2024.

[55] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., Lora: low-rank adaptation of large language models, ArXiv Prepr ArXiv210609685 (2021).

[56] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, et al., LlamaFactory: unified efficient fine-tuning of 100+ language models, in: Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. Vol. 3 Syst. Demonstr., Association for Computational Linguistics, Bangkok, Thailand, 2024.

[57] S.M. Jain, Hugging face. Introd. Transform. NLP Hugging Face Libr. Models Solve Probl., Springer, 2022, pp. 51–67.

[58] TechCrunch, OpenAI Unveils GPT-4o Mini: A Small AI Model Powering ChatGPT, with Parameters on Par with Llama 3 8B, Claude Haiku, and Gemini 1.5 Flash, 2024.

[59] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, et al., Parameter-efficient fine-tuning of large-scale pre-trained language models, Nat. Mach. Intell. 5 (2023) 220–235.

[60] S.A. Vaghefi, D. Stammbach, V. Muccione, J. Bingler, J. Ni, M. Kraus, et al., ChatClimate: grounding conversational AI in climate science, Commun. Earth Environ. 4 (2023) 480.

[61] Y. Zhang, C. Wei, Z. He, W. Yu, GeoGPT: an assistant for understanding and processing geospatial tasks, Int J Appl Earth Obs Geoinformation 131 (2024) 103976.