ClimAgent: LLM as Agents for Autonomous Open-ended Climate Science Analysis

Hao Wang[1], Jindong Han[3], Wei Fan[4], Hao Liu[1,2]*

[1]The Hong Kong University of Science and Technology (Guangzhou),
[2]The Hong Kong University of Science and Technology,
[3]Shandong University, [4]University of Auckland
figerhaowang@gmail.com,
liuh@ust.hk

Abstract

Climate research is pivotal for mitigating global environmental crises, yet the accelerating volume of multi-scale datasets and the complexity of analytical tools have created significant bottlenecks, constraining scientific discovery to fragmented and labor-intensive workflows. While the emergence Large Language Models (LLMs) offers a transformative paradigm to scale scientific expertise, existing explorations remain largely confined to simple Question-Answering (Q&A) tasks. These approaches often oversimplify real-world challenges, neglecting the intricate physical constraints and the data-driven nature required in professional climate science. To bridge this gap, we introduce ClimAgent, a general-purpose autonomous framework designed to execute a wide spectrum of research tasks across diverse climate sub-fields. By integrating a unified tool-use environment with rigorous reasoning protocols, ClimAgent transcends simple retrieval to perform end-to-end modeling and analysis. To foster systematic evaluation, we propose ClimaBench, the first comprehensive benchmark for real-world climate discovery. It encompasses challenging problems spanning 5 distinct task categories derived from professional scenarios between 2000 and 2025. Experiments on ClimaBench demonstrate that ClimAgent significantly outperforms state-of-the-art baselines, achieving a 40.21% improvement over original LLM solutions in solution rigorousness and practicality. Our code are available at https://github.com/usail-hkust/ClimAgent.

1 Introduction

Climate change and environmental issues are among the most pressing challenges humanity faces today, with far-reaching impacts on ecosystems (Parmesan et al., 2022), economies (Kotz et al., 2024), and societies worldwide (Carleton and Hsiang, 2016). Tackling these challenges requires accurate modeling and systematic analysis of complex phenomena like temperature variability (Bathiany et al., 2018; Holmes et al., 2016) and extreme weather events (Seneviratne et al., 2021; Otto, 2017). Nevertheless, traditional climate modeling approaches often rely on extensive human expertise and computationally intensive workflows, making large-scale, iterative analysis time-consuming and difficult to scale (Hansen et al., 2000; Hourdin et al., 2017).

Recently, Large Language Models (LLMs) agents (Bai et al., 2023; Achiam et al., 2023) have shown great potential in scientific discovery by leveraging their powerful reasoning and problem-solving capabilities. This line of work, referred to as agentic science (Wei et al., 2025), aims to equip LLMs with the ability to autonomously perceive, plan, and execute experiments (Schmidgall et al., 2025; Zhang et al., 2025; Chai et al., 2025; Yang et al., 2024a). Such agentic frameworks have demonstrated promising results across a wide range of domains such as biomedicine (Huang et al., 2025; Jin et al., 2025), chemistry (Yang et al., 2024b; Boiko et al., 2023), mathematics (Liu et al., 2025), and data science (Guo et al., 2024), significantly boosting the efficiency of complex analytical processes (Guo et al., 2024; Huang et al., 2023; Schmidgall et al., 2025).

However, two significant challenges remain in directly applying the aforementioned autonomous research agents to open-ended climate analysis problems. The foremost issue lies in their insufficient physical reasoning capabilities in the face of climate issues. Recent works (Wang et al., 2024b; Yu et al., 2025) suggests that LLM lacks the ability to extract physical constraints, even from simple problems. This limitation becomes even more apparent when handling climate data with strong physical correlation. To this end, a crucial question arises: *how can we unlock the physical reasoning capa-*

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 21219–21241
July 2-7, 2026

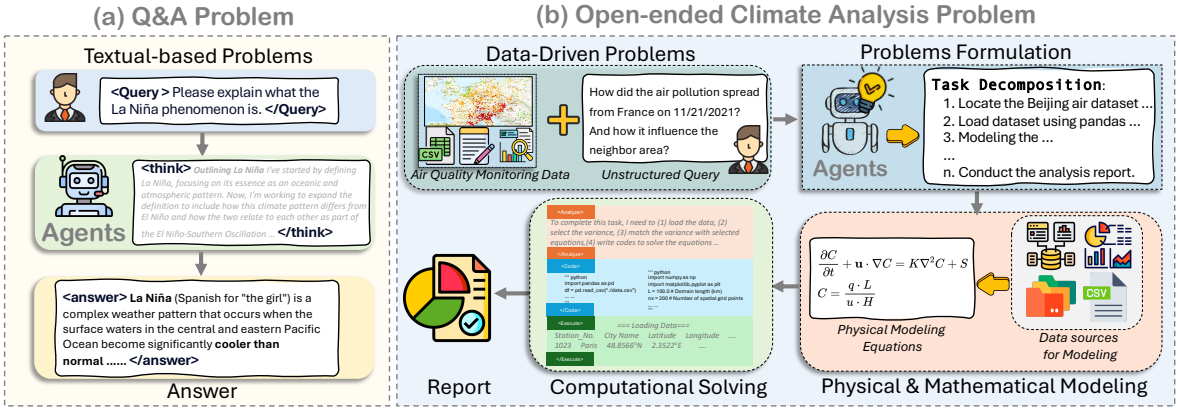


Figure 1: Previous knowledge extraction problem vs data-driven open-ended climate analysis problem. **Left:** A knowledge extraction problem, where an agent can give the answer through simple knowledge retrieval. **Right:** An open-ended mathematical modeling problem, where given an abstract application scenario or phenomenon, the agent first needs to formulate the scientific problem before solving it and providing an end-to-end solution.

bilities of LLMs for open-ended climate analysis problems? Another challenge is the lack of a benchmark to support open-ended climate analysis research. While existing works (De La Calzada et al., 2024; Manivannan et al., 2024) explored LLMs for climate problems, they are predominantly confined to Q&A tasks. Such fails to accurately reflect the problems of open-ended climate analysis, which usually have more complex data and correlations. This gap poses a significant obstacle to LLMs' effective use in this domain.

To address these limitations, we propose **ClimAgent**, a general-purpose autonomous framework designed to autonomously execute real-world climate analysis problems across a wide range of subfields. To enable LLMs' physical reasoning capability, we first construct Climate Environment (CE), a unified and comprehensive climate action space by systematically analyzing hundreds of climate research papers spanning 25 distinct sub-fields, curated from major climate literature repositories. Besides, CE includes 150 specialized climate tools and 30 databases, providing ClimAgent with professional knowledge and modeling tools. Inspired by expert workflow, ClimAgent systematically analyzes unstructured problem descriptions, models the problem both physically and mathematically, solving computational problems, and generates analytical reports. Specifically, given a user query, the agent first analyzes the problem and decomposes it into sub-tasks with the support of knowledge from CE. To better enhance physical reasoning capabilities, it uses **Climate Modeling Knowledge Retrieval** to identify the most relevant tools, information, and equations needed for each sub-task. Then, sub-tasks are further optimized by **Task-Specific Solution Optimization**, which thoroughly refines each sub-task by the retrieved information and finally output executable plans to ensure essential physical constrains are included. These executable plans are solved by generating code, and transferred to latex solution reports. Moreover, we conduct **ClimaBench**, enabling comprehensive and fair comparison across existing agentic approaches. Unlike previous knowledge extraction based Q&A tasks, our ClimaBench is data-driven benchmark constructed from the modeling process of real-world climate events spanning the years 2000 to 2025. Each data sample includes rich contextual components and requires agent to conduct problem interpretation, solution formulation, and numerical reasoning in an autonomous way. Overall, our contribution can be summarized as follows:

- We introduce ClimAgent, the first autonomous solution for open-ended climate analysis problems. Leveraging a specifically designed agentic framework and Climate Environment, ClimAgent is fully capable for modeling the complex physical correlations in climate scenarios.
- We develop ClimaBench, the first benchmark comprising 220 real-world open-end climate analysis problems spanning the years 2000 to 2025, designed to evaluate the open-end climate analysis capabilities of LLM agents.
- We conduct comprehensive experiments on ClimaBench and demonstrate that ClimAgent effectively solves open-ended climate analysis

tasks, outperforming all of the baseline approaches. Such superior performance is also further evaluated by human experts.

2 Related Works

2.1 LLM Agents

Typical applications of LLMs often involve embedding the model as an "agent" system that incorporate planning, reasoning, and interaction capabilities (Yao et al., 2022; Wei et al., 2025). By leveraging mechanisms such as memory augmentation, reflective reasoning, and tool usage, these agents enhance task decomposition, iterative refinement, and adaptive problem-solving (Hu et al., 2025b; Huang et al., 2025; Schmidgall et al., 2025). These leverage the language model's ability to learn in-context and can greatly improve its performance, robustness and reliability on many tasks (Achiam et al., 2023; Bai et al., 2023; Liu et al., 2024), which makes them successfully applied in diverse areas, including data science (Zhang et al., 2025; Guo et al., 2024), biomedical (Huang et al., 2025; Jin et al., 2025), software engineering (Wang et al., 2024c; Jimenez et al., 2023), etc.

2.2 LLMs for Scientific Discovery

LLMs is reshaping scientific discovery, evolving from specialized computational tools into autonomous research partners (Wei et al., 2025; Hu et al., 2025a). In this stage, LLMs operates as an autonomous scientific agent capable of formulating hypotheses, designing and executing experiments, interpreting results, and iteratively refining theories with reduced dependence on human guidance (Schmidgall et al., 2025). An automated scientific workflow can observe a domain, formulate novel and non-obvious hypotheses, design and execute experiments to test them, analyze the results, and iteratively refine its knowledge and strategy with minimal human intervention. Some research focus on general scientific workflow (Chai et al., 2025; Yang et al., 2024a; Zhang et al., 2025), while others concentrate on specific domain like life science (Huang et al., 2025; Jin et al., 2025), chemistry (Yang et al., 2024b; Boiko et al., 2023), physics (Jaiswal et al., 2024; Wang et al., 2024a), etc. For instance, (Huang et al., 2025) proposed a general biomedical AI agent for tasks across diverse biomedical sub-fields; (Yang et al., 2024b) explored an agentic framework to rediscover unseen chemistry hypotheses.

Except the above fields, the area of climate science have not been deeply explored. Although explorations have been made (Manivannan et al., 2024; De La Calzada et al., 2024), they are almost Q&A tasks, failing to solve the data-driven climate science problems.

3 Preliminary

In this section, we formally define the problem we aim to address. Given a climate discovery problem $\mathcal{F}$, we consider an LLM agent $\mathbf{y} = \pi_\theta(\mathbf{x}; \mathbf{x}_I)$, where $\pi_\theta(\cdot)$ denotes a language model parameterized by θ which auto-regressively generates output tokens $\mathbf{y}$ from an input sequence $\mathbf{x}$ under the guidance of an instruction prompt $\mathbf{x}_I$. The prompt $\mathbf{x}_I$ encodes task-relevant context such as background information, problem requirements, *etc.*

4 Building LLM Agent for Autonomous Climate Analysis

Section 4.1 introduces the task of real-world climate problems and presents ClimAgent, the first expert-inspired LLM agent for autonomous data-driven climate problems, which decomposes the tasks into four key stages under the support of Climate Environment: problem analysis, climate modeling, computational solving, and report generation. To further support comprehensive evaluation, in Section 4.2, we present ClimaBench, the first benchmark designed to enable systematic evaluation of LLM-based modeling agents on data-driven climate problems.

4.1 ClimAgent

This section introduces ClimAgent, an LLM-based agent system designed to automate climate discovery tasks. Its workflow consists of four key phases: Problem Analysis, Climate Modeling, Computational Solving, and Solution Reporting. ClimAgent begins by analyzing the given problem and breaking it into sub-tasks. It then constructs formal mathematical models for each sub-task, conducts experiments, and generates a solution. Finally, ClimAgent produces a comprehensive report summarizing the solution and results.

4.1.1 Climate Environment Construction

As previously discussed, achieving physical constraints, in-depth analysis within climate science remains the foremost challenge impeding the development of autonomous AI climate research agents.

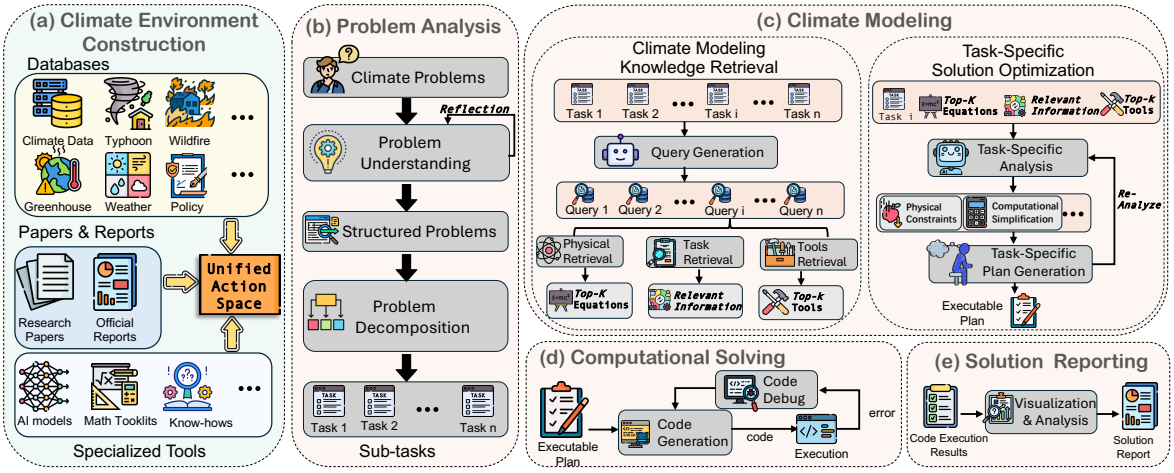


Figure 2: The overall framework of our proposed ClimAgent.

To address this, we construct the Climate Environment (CE), a unified and comprehensive action space derived from three main parts: databases, papers & reports, and specialized tools. Building on this foundation, we employ an LLM agent capable of autonomously parsing literature to extract key tasks, tools, and databases essential for driving climate analysis. These elements are subsequently curated and implemented into the CE, defining a robust action space for agentic interaction.

Specifically, we provide 30 high-fidelity databases that cover weather and climate related data spanning the years 2000 to 2025, which are frequently used in most climate modeling problems, including those for temperature and wind speed *etc.* Furthermore, (Liu et al., 2025; Guo et al., 2024) show that agents often encounter problems such as low success rate and insufficient computational accuracy when writing code to solve modeling problems. This problem is even more serious in atmospheric modeling, a field with large amounts of data and strong physical constraints. To ensure the precision and professionalism of the code generated for complex modeling tasks, we have integrated 150 specialized tools into the CE, ranging from data pre-processing utilities to advanced simulation algorithms. Complementing this, crucially, thousands of papers and reports provide expert knowledge for agents to retrieve correct tools or equations that most suitable for the problems. Besides, CE modules are designed with a modular architecture, supporting seamless future extensions to accommodate emerging scientific methodologies and newly available datasets.

4.1.2 Problem Analysis

This section details the Problem Analysis phase of ClimAgent, a critical module designed to bridge the semantic gap between unstructured problem descriptions and formal climate modeling protocols.

Problem Understanding. From Section 3, we consider an LLM $\mathbf{y} = \pi_\theta(\mathbf{x}; \mathbf{x}_I)$ for the climate problem $\mathcal{F}$. For more comprehensive understanding of certain specific problem, we attach the instruction prompt $\mathbf{x}_I$. Conditioned on this input, the analyst agent performs a structured analysis to identify the problem type, core concepts, assumptions, objectives, and other essential factors. Specifically, this process is represented as $U_p = \pi_\theta(\mathcal{F}; x_u)$, where x_u represents the profile prompt used for problem understanding, and U_p is the analysis result. To deepen the understanding of the problem, the analyst agent adopts self-reflection to iteratively refine its analysis.

Problem Decomposition. After understanding the problem, the coordinator agent decomposes it into a set of sub-tasks to address its multiple objectives. This process can be represented as: $\mathbf{D} = \pi_\theta(\mathcal{F}, \mathcal{U}_p; \mathbf{x}_d)$, where $\mathbf{x}_d$ represents the profile prompt used for task decomposition, $\mathbf{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_n\}$ denotes the set of sub-tasks, and each $\mathcal{D}_i$ corresponds to certain individual subtask. Each sub-task is associated with a specific objective or component of the problem. For example, for the air pollution problems, the agent decomposes the problem into four sub-tasks: data collection, wind blow modeling, simulation, and equation calculation.

4.1.3 Climate Modeling

To efficiently automate solving climate modeling, we propose the Climate Modeling Knowledge Retrieval and Task-Specific Solution Optimization. The specific climate modeling process for each sub-task $\mathcal{D}_i$ involves searching the most suitable physical equation or tools, and then optimizing the mathematical modeling to match the specific need of each sub-task.

Climate Modeling Knowledge Retrieval. For sub-task $\mathcal{D}_i$, the primary objective is to bridge the gap between abstract natural language instructions and executable scientific protocols. Unlike general open-domain QA, climate modeling requires strict adherence to physical constraints. To achieve this, we employ a hybrid retrieval mechanism that queries the pre-constructed Climate Environment (CE). Specifically, the agent first generates a high-dimensional semantic query vector q_i based on the description of $\mathcal{D}_i$. This query is then used to retrieve two distinct types of knowledge: (1) Tool Retrieval: The top-k most relevant executable tools (*e.g.*, *NetCDF* processors) are selected from the 150-tool repository to ensure operational feasibility. (2) Task Retrieval: The agent retrieves task-relevant information for potential use (*e.g.*, databases, policies, *etc.*) (3) Physics Retrieval: The agent identifies the governing physical equations (*e.g.*, fluid dynamics constraints) relevant to the scenario. The retrieved context $\mathcal{K}_i$ thus serves as a physically grounded boundary, preventing the agent from hallucinating non-existent scientific methods during the subsequent code generation phase.

Task-Specific Solution Optimization. While retrieved climate modeling knowledge offers foundational methods and ideas, it often lacks the depth needed to address specific problem nuances (*e.g.*, dealing with nonlinear constraints, optimizing multiple conflicting objectives, *etc.*). To overcome these limitations, we introduce a task-specific solution optimization framework that progressively refines the modeling scheme, enabling it to effectively manage complex constraints and enhance overall solution quality. Given the sub-task $\mathcal{D}_i$ and the retrieved knowledge context $\mathcal{K}_i$ (comprising physics equations and executable tools), the agent first synthesizes an initial modeling scheme $\mathcal{M}_i^{(0)} = \pi(\mathcal{D}_i, \mathcal{K}_i)$. This scheme concretizes abstract methods into executable workflows, defining specific algorithmic parameters and data pipelines. Subsequently, the agent scrutinizes $\mathcal{M}_i^{(t)}$ from a rigorous scientific perspective. Unlike general code reviewers, this strategy evaluates domain-specific criteria such as *physical consistency* (e.g., ensuring conservation laws are not violated) and *boundary condition validity*. It generates targeted feedback $\mathcal{F}_i^{(t)} = \pi(\mathcal{D}_i, \mathcal{M}_i^{(t)})$, highlighting potential logical fallacies or physical discrepancies. Guided by this feedback, the agent then refines the scheme to generate $\mathcal{M}_i^{(t+1)} = \pi(\mathcal{M}_i^{(t)}, \mathcal{F}_i^{(t)})$. This iterative loop continues until the solution satisfies the Critic's verification or reaches a maximum iteration threshold N_{max}, ensuring the final model is both mathematically robust and scientifically physically grounded and consistent.

4.1.4 Computational Solving and Solution Reporting

This section describes the computational solving and solution reporting phase of ClimAgent, which focuses on solving the mathematical equations and generating a comprehensive solution report.

Code Generation and Execution. Given the climate modeling scheme $\mathcal{M}_i$, the modeling programmer agent generates the corresponding code as follows: $\mathcal{C}_i = \pi_\theta(\mathcal{D}_i, \mathcal{M}_i; \mathbf{x}_g)$, where $\mathbf{x}_g$ represents the instruction prompt used to direct the LLM to generate the computational code, and $\mathcal{C}_i$ denotes the mathematical modeling code for task $\mathcal{D}_i$. After code generation, the program is compiled to check for runtime errors. If it compiles successfully, the experimental results $\mathcal{O}_i$ are returned. If the code fails to compile, the agent attempts to repair it over n_c iterations by analyzing the last error message and making the necessary corrections. Upon task completion, the task coordinator agent updates the agent's memory: $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathcal{D}_i, \mathcal{Q}_i\}$. In practice, for policy-related modeling problems, where the goal is to provide insights and recommendations based on existing knowledge or models, the modeling agent directly offers these insights without generating codes.

Solution Report. After all tasks have been completed, the agent compiles a comprehensive summary of the problem-solving process. The first step is to construct a structured outline for the climate modeling report. This outline establishes the framework of the report, organizing it into six key sections: problem restatement, model assumptions, justification of assumptions, notation and definitions, problem analysis, and solution. By structur-

ing the content systematically, it provides a solid foundation for an in-depth and well-organized final report. Once the outline is established, the agent employs specialized commands to progressively refine the report, drawing on the task coordinator agent's memory $\mathcal{H}$. Through a series of iterative edits, the agent guarantees that the report meets the necessary standards for quality, coherence, and academic rigor.

4.2 ClimaBench

In this section, we introduce the construction of ClimaBench, the first data-driven benchmark for autonomous climate analysis.

Dataset Construction. Development and testing sets were created from 1000 publications and official climate reports, which were collected and analyzed by extracting and parsing their PDF contents. Each paper was processed in chunks, and a specialized prompt guided an LLM through each chunk to explicitly identify and extract the main solution steps of the problems from each chunk. Due to resource constraints, each set comprises 12.5% of the complete reference, proportionally distributed across benchmark sub-tasks, providing a cost-effective and representative assessment of model performance. The development set informed iterative refinements to ClimAgent's database integrations and tool implementations, while the test set provided an independent evaluation of generalization capabilities.

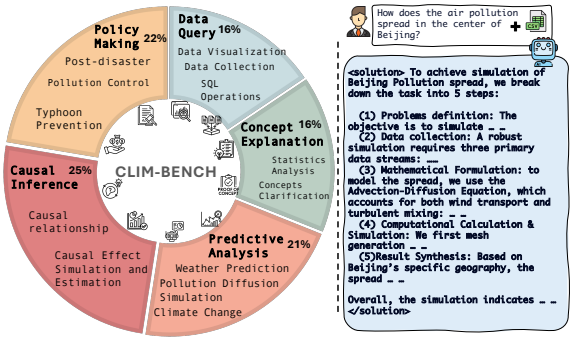


Figure 3: **Left:** Distribution of task categories and domains in our ClimaBench. **Right:** Example of one task&solution pair in the dataset.

Dataset Description. Here, we introduce the main characterization of the dataset in ClimaBench containing 320 tasks. The five categories in ClimaBench capture the major reasoning needs of data-driven climate science: (1) data query tasks (*e.g.*, dataset collection), (2) concept analysis (*e.g.*, professional concept clarification), (3) predictive analysis (*e.g.*, weather prediction tasks), (4) causal inference (*e.g.*, causal effect simulation and analysis), and (5) policy making (*e.g.*, post-disaster reconstruction), which is illustrated in Figure 3. Each of these 120 tasks is grounded in one of these categories, providing a relatively balance distribution across task types. Theses five domains are evenly integrated, ensuring broad coverage of climate scientific tasks. The scale of 320 tasks is deliberately chosen to balance breadth and feasibility: each task is framed as an end-to-end, repository level workflow, where agents must independently perform data-driven computation and reasoning. Given that even a single predictive modeling task can require hours to complete, this design ensures the benchmark remains both challenging and practically executable.

Evaluation. We evaluate the final solution report along four key dimensions (Liu et al., 2025): (1) Analysis Evaluation (AE). Examines problem definition clarity, identification of key components, and the logical coherence between sub-tasks and overarching objectives. (2) Solution Correction (SC). Focuses on rigor and rationality, evaluating whether the assumptions are clearly stated and justified, and whether the chosen methods, equations, accurately and scientifically represent the real-world problem. (3) Practicality and Scientificity (PS). Evaluates the practicality and scientific validity of the model, ensuring that it is realistically applicable, provides valuable insights for decision-making, and adheres to scientific principles. This stage also verifies whether the model is theoretically sound and considers all relevant scientific factors to ensure its validity. (4) Result and Bias Analysis (RBA). Measures the clarity, interpretability, and analytical depth of results, with attention to identifying and mitigating data or modeling biases to ensure robustness and transparency. We conduct both LLM-based and expert-human evaluations to ensure a comprehensive and reliable assessment. For further details and prompt-use, please refer to Section A.2 in the Appendix.

5 Experiments

In this section, we first compare the performance of our ClimAgent with existing approaches, then we further conduct human evaluation, ablation study, and multiple experiments to analyze ClimAgent.

Table 1: Benchmark results across 2000–2025 climate discovery problems in ClimaBench. AE, SC, PS, and RBA represent *Analysis Evaluation*, *Solution Correction*, *Practicality and Scientificity*, and *Result and Bias Analysis*.

Methods	2000–2024					2025				
	AE ↑	SC ↑	PS ↑	RBA ↑	Overall ↑	AE ↑	SC ↑	PS ↑	RBA ↑	Overall ↑
GPT-4o										
GPT-4o	7.55	3.92	8.42	5.25	6.29	7.72	3.80	8.85	5.82	6.55
DS-Agent	8.24	7.12	8.75	7.51	7.91	8.31	7.28	8.95	7.22	7.94
ResearchAgent	8.05	6.85	8.85	7.42	7.79	8.12	7.25	8.72	7.15	7.81
Agent Laboratory	8.62	6.42	8.68	5.68	7.35	8.82	5.65	8.64	5.48	7.15
DeepAnalyze	8.62	6.28	8.75	6.22	7.34	8.58	6.35	8.82	6.88	7.48
ClimAgent	**9.22**	**7.45**	**9.12**	**8.58**	**8.92**	**8.95**	**7.38**	**9.15**	**8.55**	**8.47**
DeepSeek-R1-671B										
DeepSeek-R1	7.38	4.88	8.75	4.62	6.41	7.55	4.35	8.65	5.38	6.48
DS-Agent	8.35	7.02	8.82	7.35	7.89	8.12	6.55	9.12	7.75	7.89
ResearchAgent	8.24	7.21	8.85	7.12	7.86	8.15	6.92	8.95	7.72	7.94
Agent Laboratory	8.75	6.12	8.80	6.12	7.45	8.95	5.68	8.95	5.75	7.33
DeepAnalyze	8.62	6.28	8.75	6.22	7.34	8.58	6.35	8.82	6.88	7.48
ClimAgent	**9.62**	**8.35**	**9.18**	**8.65**	**8.95**	**9.58**	**8.45**	**9.32**	**8.72**	**9.02**
Qwen3-235B										
Qwen3-235B	7.42	4.92	8.72	4.68	6.45	7.58	4.40	8.62	5.42	6.52
DS-Agent	8.38	7.05	8.85	7.38	7.92	8.15	6.58	9.15	7.78	7.92
ResearchAgent	8.28	7.25	8.88	7.15	7.89	8.18	6.95	8.98	7.75	7.98
Agent Laboratory	8.78	6.15	8.82	6.15	7.48	8.98	5.72	8.98	5.78	7.36
DeepAnalyze	8.62	6.28	8.75	6.22	7.34	8.58	6.35	8.82	6.88	7.48
ClimAgent	**9.65**	**8.38**	**9.22**	**8.68**	**8.98**	**9.62**	**8.48**	**9.35**	**8.75**	**9.05**

5.1 Experimental Setup

Baselines. We evaluate our proposed ClimAgent with state-of-the-art general LLM scientific agents. As no prior work directly targets autonomous climate science problems, we re-purpose existing autonomous research agents for comparison. The baselines include: (1) DS-Agent (Guo et al., 2024): An LLM agent for automated data science, adapted with its core case-based reasoning framework for modeling tasks; (2) ResearchAgent (Huang et al., 2023): Originally designed to automate experimentation loops for machine learning tasks, adapted with its core framework for modeling problems; (3) Agent Laboratory (Schmidgall et al., 2025): A scientific discovery framework that guides agents through literature review, experimentation, and report writing; and (4) DeepAnalyze (Zhang et al., 2025): An agentic LLM designed for autonomous data science, which is capable of automatically completing the end-to-end pipeline from data sources to analyst-grade deep research reports.

Experimental Implementation. We conduct experiments on our proposed ClimaBench as our test set, ensuring diversity across problem types and domains to support a representative evaluation. We select a subset of climate problems from the past five years (2000–2025) as our test set, ensuring diversity across problem types and domains to support a representative evaluation. This subset consists of 320 problems in total. To mitigate potential data leakage from LLM pretraining, we evaluate problems from 2000–2024 separately from those in 2025. For the evaluation, we adopt both GPT-4o-based automatic scoring and we further conduct human expert review to verify the the scientific nature and rationality of our evaluation method. .

5.2 Main Results

Main Experiments. As shown in Table 1, our ClimAgent achieves state-of-the-art (SOTA) performance across all evaluation dimensions. (1) Directly applying foundational models (GPT-4o, DeepSeek-R1-671B or Qwen3-235B) without agent-level orchestration results in significantly weaker performance, particularly in MR and RBA. This gap underscores the inadequacy of LLMs in handling the open-ended, structured reasoning required for real-world modeling tasks and highlights the necessity of structured agent-based workflows. (2) ClimAgent consistently outperforms all baseline agents, achieving the highest overall scores under all of the three LLM backbones. This superior performance demonstrate that ClimAgent consistently support LLMs with expert capabilities to solve real-world professional climate science analysis problems. (3) Agents built on DeepSeek-R1-

671B and Qwen3-235B surpass their GPT-4o counterparts, with ClimAgent demonstrating marked gains in MR and RBA, suggesting stronger reasoning capabilities in the larger models. (4) The 2025 results closely mirror those from 2021–2024, indicating strong temporal consistency. This robustness mitigates concerns about potential data leakage (*e.g.*, memorized solutions) and further supports the conclusion that ClimAgent performs genuine modeling rather than overfitting.

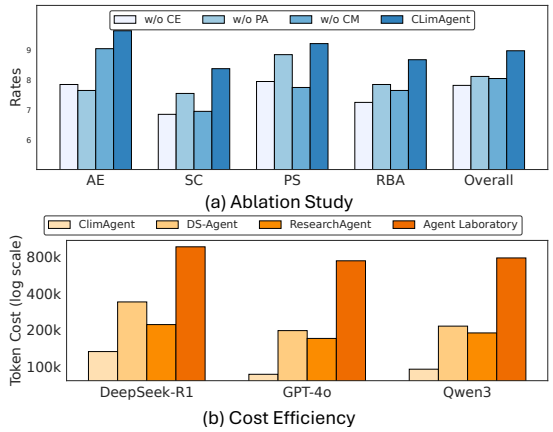


Figure 4: Ablation study and cost efficiency analysis.

5.3 Ablation Study and Further Analysis

To better understand the design and practical utility of ClimAgent, we present a three-part analysis. First, we conduct an ablation study to quantify the impact of each core module. Then, we evaluate token usage, cost, and runtime to assess deployment efficiency. Finally, we further employed human evaluation to compare and demonstrate the rationality of our evaluation method.

Contribution of Key Components. We perform an ablation study to assess the impact of three core modules in ClimAgent: Problem Analysis (PA), the Climate Modeling (CM), and the Climate environment (CE). Specifically, we (1) replace PA with a naive task parser (**w/o PA**), (2) substitute CM with a simple retrieval without task-specific optimization (**w/o CM**), and (3) remove CE to disable expert knowledge support (**w/o CE**). These variants allow us to evaluate each module's contribution to structured problem understanding and modeling performance. As shown in Figure 4 (a),, ClimAgent consistently outperforms all ablated variants under five evaluation metrics. Removing PA significantly reduces AE and MR, indicating that deep task comprehension is essential for rigorous formulation. The absence of CM leads to sharp declines in MR and PS, highlighting its critical role in constructing coherent, scientifically sound models. CM enables problem-aware and solution-aware retrieval that better supports abstraction, constraint reasoning, and method selection, key capabilities for effective modeling. Notably, removing CE causes clear drops in all metrics, underscoring the importance of professional action space in climate analysis problems with expert knowledge supporting.

Cost Efficiency Analysis. We access the cost of efficiency of ClimAgent in solving real-world climate discovery problems, focusing on token usage. All evaluations are conducted via official APIs provided by model vendors. As shown in Figure 4 (b), ClimAgent outperforms the performance of all of the baseline models with superior computational cost. It achieves the best performance while substantially reducing cost, highlighting its scalability and practical viability.

Table 2: Human evaluation results VS our evaluation.

Methods	AE	SC	PS	RBA	Overall
Human Team	9.42	7.77	9.12	8.44	8.94
ClimAgent	9.34	7.82	9.04	8.51	8.89
Difference Value	0.08	0.05	0.08	0.07	0.05

Human Experts Evaluation. To further demonstrate the rationality of our evaluation method, we invited five senior PhD students in climate science and related fields to evaluate and score the 100 randomly-selected problem-solutions generated by ClimAgent using the same standards as the main experiment above. As can be seen in Table 2, the human evaluation results are highly consistent with the results of our LLM evaluation, which further proves the scientific nature and rationality of our evaluation method.

6 Conclusion

In this work, we introduce ClimAgent, the first LLM agent system specifically designed for autonomous open-ended climate analysis problems. Through well-design agent framework and Climate Environment, ClimAgent is fully capable for modeling climate problems with complex physical correlations. We further propose ClimaBench, the first benchmark that systematically define and construct datasets for data-driven open-ended climate analysis problems. Comprehensive experiments

show that ClimAgent significantly outperforms existing LLM agents. We hope our benchmark and framework lay a foundation for future progress in LLM-driven climate science discoveries.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62572417, No.92370204), National Key R&D Program of China (Grant No.2023YFF0725004).

7 Limitations

We discuss three primary limitations of our work. First, our framework relies significantly on proprietary LLMs such as GPT-4o, Qwen3 and DeepSeek-R1 for both the climate modeling and code execution phases, which raises concerns regarding reproducibility and the cumulative costs of large-scale evaluations. While the cost of utilizing these models is relatively optimized, for instance, ClimAgent achieves a state-of-the-art solution for approximately half token-cost, the reliance on closed-source APIs remains a constraint for some research environments. We experimented with smaller alternatives like Qwen2.5-72B; however, their weaker instruction-following capabilities and lower modeling rigorousness compared to larger proprietary backbones led us to prioritize the latter to ensure scientific accuracy. We leave the development of robust, automated verification strategies to future research to ensure that models retrieved and executed across different subtasks can accommodate physical laws and collectively contribute to rigorous scientific discovery. Then, we defer the exploration of practical assist-solutions, such as the integration of real-time earth system data streams and the full-scale deployment of ClimAgent in high-stakes policy environments, to future work. In the long run, we envision a model capable of accurately understanding complex environmental dynamics and recommending comprehensive modeling pipelines at once via autonomous scientific planning, which can promote rapid response in time-sensitive domains like disaster management and increase the efficiency of global climate science research.

8 Ethics Statement

Content Safety and Integrity. Since our data curation and agent execution pipelines involve extensive interaction with Large Language Models (LLMs), ensuring the safety and scientific integrity of generated content is paramount. We strictly implemented system-level prompts explicitly instructing the LLM to refuse the generation of content that includes hate speech, discrimination, violence, or political disinformation. Given the sensitive nature of climate change discourse, we additionally imposed constraints to prevent the generation of scientifically unsubstantiated alarmism or denialism. We manually inspected a random sample of 100 interaction traces and solution reports from ClimaBench. Based on our observations, we did not detect any offensive, toxic, or scientifically misleading content. Consequently, we believe our dataset and framework are safe for research use and do not pose negative societal risks regarding content toxicity.

Data Privacy and Licensing. Unlike personal data-heavy domains, our benchmark relies primarily on meteorological and environmental data. All datasets integrated into the Climate Environment (CE) are sourced from publicly available repositories (e.g., NOAA, ERA5, NASA Earthdata) and are strictly non-personally identifiable information (non-PII). We comply with all original data licenses (mostly CC-BY or Open Government Data licenses) and allow for their use in academic research. The ClimaBench dataset itself will be made publicly available to facilitate reproducibility and future research in automated science.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Sebastian Bathiany, Vasilis Dakos, Marten Scheffer, and Timothy M Lenton. 2018. Climate models predict increasing temperature variability in poor countries. *Science advances*, 4(5):eaar5809.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Tamma A Carleton and Solomon M Hsiang. 2016. Social and economic impacts of climate. *Science*, 353(6304):aad9837.

Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, Yuzhi Zhang, Linfeng Zhang, Siheng Chen, and 1 others. 2025. Scimaster: Towards general-purpose scientific ai agents, part i. x-master as foundation: Can we lead on humanity's last exam? *arXiv preprint arXiv:2507.05241*.

Natalia De La Calzada, Théo Alves Da Costa, Annabelle Blangero, and Nicolas Chesneau. 2024. Climateq&a: Bridging the gap between climate scientists and the general public. *arXiv preprint arXiv:2403.14709*.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*.

Jindong Han, Hao Wang, Hui Xiong, and Hao Liu. 2025. Scalable pre-training of compact urban spatiotemporal predictive models on large-scale multi-domain data. *Proceedings of the VLDB Endowment*, 18(7):2149–2158.

J Hansen, R Ruedy, A Lacis, Mki Sato, L Nazarenko, N Tausnev, I Tegen, and D Koch. 2000. Climate modeling in the global warming debate. In *International Geophysics*, volume 70, pages 127–164. Elsevier.

Caroline R Holmes, Tim Woollings, Ed Hawkins, and Hylke De Vries. 2016. Robust future changes in temperature variability under greenhouse gas forcing and the relationship with thermal advection. *Journal of Climate*, 29(6):2221–2236.

Frédéric Hourdin, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, Duoying Ji, Daniel Klocke, Yun Qian, and 1 others. 2017. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, and 1 others. 2025a. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025b. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Benchmarking large language models as ai research agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. 2024. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. 2025. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004*.

Maximilian Kotz, Anders Levermann, and Leonie Wenz. 2024. The economic commitment of climate change. *Nature*, 628(8008):551–557.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Fan Liu, Zherui Yang, Cancheng Liu, Tianrui Song, Xiaofeng Gao, and Hao Liu. 2025. Mm-agent: Llm as agents for real-world mathematical modeling problem. *arXiv preprint arXiv:2505.14148*.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2024. Climaqa: An automated evaluation framework for climate question answering models. *arXiv preprint arXiv:2410.16701*.

Friederike EL Otto. 2017. Attribution of weather and climate events. *Annual Review of Environment and Resources*, 42:627–646.

Camille Parmesan, Mike D Morecroft, and Yongyut Trisurat. 2022. *Climate change 2022: Impacts, adaptation and vulnerability*. Ph.D. thesis, GIEC.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, A Di Luca, Subimal Ghosh, Iskhaq Iskandar, James Kossin, Sophie Lewis, and 1 others. 2021. Weather and climate extreme events in a changing climate.

Cunshi Wang, Xinjie Hu, Yu Zhang, Xunhao Chen, Pengliang Du, Yiming Mao, Rui Wang, Yuyang Li, Ying Wu, Hang Yang, and 1 others. 2024a. Starwhisper telescope: Agent-based observation assistant system to approach ai astrophysicist. *arXiv preprint arXiv:2412.06412*.

Hao Wang, Jindong Han, Wei Fan, Leilei Sun, and Hao Liu. 2025a. Repst: language model empowered spatio-temporal forecasting via semantic-oriented reprogramming. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3362–3370.

Hao Wang, Jindong Han, Wei Fan, Weijia Zhang, and Hao Liu. 2025b. Phyda: Physics-guided diffusion models for data assimilation in atmospheric systems. *arXiv preprint arXiv:2505.12882*.

Hao Wang, Zixuan Weng, Jindong Han, Wei Fan, and Hao Liu. 2025c. Dambench: A multi-modal benchmark for deep learning-based atmospheric data assimilation. *arXiv preprint arXiv:2511.01468*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024b. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *International Conference on Machine Learning*, pages 50622–50649. PMLR.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, and 1 others. 2024c. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 3.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, and 1 others. 2025. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024a. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024b. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. 2025. The stochastic parrot on llm's shoulder: A summative assessment of physical concept understanding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11416–11431.

Shaolei Zhang, Ju Fan, Meihao Fan, Guoliang Li, and Xiaoyong Du. 2025. Deepanalyze: Agentic large language models for autonomous data science. *arXiv preprint arXiv:2510.16872*.

A Experimental Implementation

A.1 Baselines

We compare ClimAgent with state-of-the-art LLM-based autonomous agents. As there is no prior work specifically targeting autonomous agents for open-ended climate science discovery, we re-purpose existing generalist research agents to address these tasks. Specifically, our baselines include:

- **(1) Standard LLMs:** We directly employ foundational Large Language Models (specifically *GPT-4o*, *Qwen3* and *DeepSeek-R1*) to generate climate modeling solutions via zero-shot prompting. This serves as a baseline to assess the necessity of agentic architectures versus standalone models.
- **(2) DS-Agent:** A specialized LLM agent originally designed for automating data science competitions (Guo et al., 2024). We adapt its core case-based reasoning framework to handle meteorological datasets and perform statistical climate analysis.
- **(3) ResearchAgent:** An agentic framework designed to automate iterative research workflows and idea generation (Huang et al., 2023). We integrate it with a code execution environment to enhance its capability to formulate and verify quantitative climate models.
- **(4) Agent Laboratory:** An LLM-based platform that simulates the full scientific discovery process—ranging from literature review to experimentation and report writing (Schmidgall et al., 2025). In our setting, this agent is configured to query academic repositories (e.g., arXiv) to identify relevant physical equations and climate methodologies, which it then synthesizes to construct solution pipelines.
- **(5) DeepAnalyze:** DeepAnalyze (Zhang et al., 2025) is an agentic framework designed to automate end-to-end data science workflows, bridging the gap between raw data and actionable insights through autonomous exploration. It utilizes agentic Large Language Models (LLMs) to perform complex, multi-step tasks such as data preprocessing, feature engineering, and model optimization. The system operates through an iterative "plan-execute-verify" loop, which allows the agent to generate Python code, execute it in a sandboxed environment, and self-correct based on execution errors or analytical feedback. By integrating tools like Pandas, Scikit-learn, and Matplotlib, DeepAnalyze provides a robust platform for autonomous data-driven research and serves as a significant benchmark for evaluating the problem-solving capabilities of LLM-based agents in technical domains.

A.2 Evaluation Criteria and Assessment Protocols

Given the open-ended nature of climate modeling, where standardized "gold solutions" are often unavailable, we establish a multidimensional evaluation framework to rigorously assess agent performance. Following the methodology proposed in mmagent, we evaluate the generated solutions along four distinct dimensions: Analysis Evaluation (AE), Solution Correction, Practicality and Scientificity (PS), and Result and Bias Analysis (RBA). This section details the specific criteria and the hybrid assessment protocol employed.

A.2.1 Detailed Evaluation Dimensions

(1) Analysis Evaluation (AE). This dimension assesses the agent's cognitive capability to deconstruct unstructured problems.

- **Problem Definition Clarity:** Does the agent clearly restate the problem, identifying the core objective and boundary conditions?
- **Variable Identification:** Are the key dependent and independent variables (e.g., temperature anomalies, emission factors) correctly identified and defined?
- **Logical Coherence:** Is the decomposition of the problem into sub-tasks (e.g., data cleaning trend analysis prediction) logically sound and aligned with the overarching research goal?

(2) Solution Correction (SC). This dimension evaluates the mathematical and technical precision of the constructed model.

- **Assumption Justification:** Are all modeling assumptions (e.g., assuming linearity in a chaotic system) explicitly stated and scientifically justified?
- **Methodological Appropriateness:** Is the chosen mathematical method (e.g., differential equations vs. statistical regression) appropriate for the specific data characteristics?

- **Structural Correctness:** Does the model structure strictly adhere to mathematical logic without derivational errors or dimension mismatches?

(3) Practicality and Scientificity (PS). This dimension focuses on the real-world applicability and domain consistency of the solution.

- **Physical Consistency:** Does the model violate fundamental physical laws (e.g., conservation of energy/mass)?
- **Insightfulness:** Does the model provide non-trivial, actionable insights for decision-making (e.g., specific policy recommendations)?
- **Generalizability:** Can the proposed model be extended to other regions or time scales, or is it overfitted to a specific dataset?

(4) Result and Bias Analysis (RBA). This dimension measures the critical thinking applied to the model's outputs.

- **Interpretability:** Are the results presented with clear visualizations and intuitive explanations?
- **Bias Mitigation:** Does the agent actively identify potential biases in the data (e.g., geographic sampling bias) or the model (e.g., algorithmic fairness)?
- **Sensitivity Analysis:** extensive sensitivity testing conducted to verify the robustness of the results against parameter perturbations?

A.2.2 Scoring Methodology

To quantify performance, we employ a **10-point Likert scale** for each dimension, defined as follows:

- **High Quality (8-10):** The solution is professional-grade, theoretically sound, and physically consistent. It could be accepted as a valid contribution to a standard workshop.
- **Medium Quality (4-7):** The solution is generally correct but lacks depth, contains minor logical gaps, or fails to fully justify assumptions.
- **Low Quality (1-3):** The solution contains fundamental errors (e.g., hallucinated citations, physical violations) or fails to address the core problem.

A.2.3 Hybrid Evaluation Protocol

To ensure a balanced and scalable assessment, we implement a Dual-Evaluation Mechanism:

1. **LLM-as-a-Judge:** We utilize *GPT-4o* as an automated evaluator. The judge is provided with the ground-truth problem description, the agent's solution, and a detailed scoring prompt to generate scores and critiques for all four dimensions.
2. **Expert Human Review:** For a randomly sampled subset of solutions, we enlist domain experts (PhD-level researchers in atmospheric science and applied mathematics) to perform blind reviews. This human-in-the-loop step serves to calibrate the LLM judge and verify the "Scientificity" metric, which is often challenging for automated systems to assess accurately.

B Prompts Used for ClimAgent and ClimaBench

In this appendix, we provide a comprehensive documentation of the prompt engineering framework that underpins both the curation of CLIMABENCH and the runtime execution of CLIMAGENT. These prompts are rigorously engineered to support automated, modular, and scientifically valid workflows within Large Language Model (LLM) agents. We delineate the specific prompt designs for each functional module of the agentic pipeline. By standardizing both the instruction-level constraints and the structured response formats (e.g., JSON schemas), these prompts ensure robust multi-agent collaboration and operational consistency. Disclosing this full set of "cognitive directives" is essential to guarantee the traceability and reproducibility of our end-to-end climate modeling framework.

C Case Study

Here, we give case studies to visualize the performance of ClimAgent. As can be seen in Figure 10, Figure 11, and Figure 12, we give a query about 'What is the trajectory of typhoon Doksuri in 2023 China? And how it influence the weather in 2023 China?'. The ClimAgent performs Problem Analysis, Climate Modeling, Computational Solving, and Solution Reporting procedures.

D Potential Risks

While ClimAgent demonstrates promising capabilities in automating climate modeling, deploying au-

Analysis Evaluation Prompt

Your task is to evaluate the rationality and overall coherence of the problem decomposition into sub-problems by the modeler, given the background and problem requirement in mathematical modeling.

Background:
{background}

Problem Requirements:
{requirements}

Below is the modeler's task analysis:
Task Analysis:
{all_task_analyses}

Evaluation Criteria:
1. Problem Analysis and Understanding
1.1 Problem Definition and Goals
Ensure the model definition is clear, the analysis is accurate, and the goals are explicit.
- Is the scope and goal of the problem clearly defined?
- Are the key components of the problem effectively identified?
- Are the actual goals that the model aims to solve clearly stated?

Scoring Criteria:
1-2 = Completely unclear;
3-4 = Not clear enough;
5-6 = Basically clear;
7-8 = Clear;
9-10 = Completely clear.
1.2 Relevant Scope and Coverage
Ensure that the core part of the problem is not deviated from, and whether each sub-task is interrelated and completely covers the actual goals.
- Do the sub-tasks have dependencies?
- Are all sub-tasks and steps directly related and support the final goal?
- Are there any key parts missing or deviations from the actual goals?
Scoring Criteria:
1-2 = Completely deviated from the goal;
3-4 = Partially deviated;
5-6 = Basically covered;
7-8 = Mostly covered;
9-10 = Completely covered.

Output Format: Please put your evaluation reasons and scores in the tags <reason> your_reason </reason>, and <score> your_score </score>.

Figure 5: The prompt used for Analysis Evaluation.

Solution Correction Evaluation Prompt

Your task is to evaluate the rigor and rationality of the modeling given the background and problem requirement in mathematical modeling, particularly focusing on the assumptions and rationality.

Background:
{background}

Problem Requirements:
{requirements}

Below is the modeler's modeling analysis:
Modeling Analysis:
{all_task_analyses}

Evaluation Criteria:
2. Rigor and Rationality of Modeling
2.1 Assumptions
Clear and explicit. These assumptions are the foundation of the model and need to be rigorously justified.
- Are the model assumptions clearly explained?
- Are the assumptions reasonable and consistent with the background of the actual problem?
- Is the rationality and impact of the assumptions considered?

Scoring Criteria:
1-2 = Completely unreasonable; 3-4 = Partially reasonable;
5-6 = Average; 7-8 = Reasonable;
9-10 = Very reasonable.

2.2 Rationality
The rationality of the model is key to evaluation. Evaluation criteria can include: whether an appropriate model is chosen, whether the model can realistically reflect the problem, etc.
- Has the model chosen appropriate methods and metrics?
- Does the structure of the model scientifically reflect the actual problem?

Scoring Criteria:
1-2 = Completely unreasonable; 3-4 = Partially reasonable;
5-6 = Average; 7-8 = Reasonable;
9-10 = Very reasonable.

Figure 6: The prompt used for Solution Correction.

Practicality and Scientificity Evaluation

Your task is to evaluate the practicality and scientificity of the modeling process given the background and problem requirements in mathematical modeling, particularly focusing on whether the model can practically solve the problem and whether it adheres to scientific principles.

Background:
{background}

Problem Requirements:
{requirements}

Below is the modeler's modeling process:
Modeling Process:
{all_task_analyses}

Evaluation Criteria:
3. Practicality and Scientificity
3.1 Practicality
- Does the modeling method match the characteristics and requirements of the problem?
- Does the model provide meaningful insights beyond mere data fitting? Can its output support decision-making with clear explanations and reliable predictions across different datasets?
- Does the approach go beyond standard machine learning or data processing? Has it been deeply optimized or extended, potentially integrating interdisciplinary methods like mathematical or physical modeling?
- Does the model introduce novel frameworks, constraints, objectives, or data representations?
Does it push beyond conventional techniques to propose new theoretical or computational approaches?
- Is the selected modeling method appropriate for the given problem?
- Is the model reasonably constructed?
- Can the model solve the actual problem?
- Are the application scenarios of the model clear? Is it feasible for practical operation?
- Can the model's output provide useful information for decision-making or explaining or predicting?
- Does the approach go beyond basic data analysis and machine learning algorithms?
- Does the model demonstrate innovation or creativity in its approach to addressing the
problem?
- Is the modeling approach tailored to the specific problem rather than using generic methods?

Figure 7: The prompt used for Solution Correction.

Practicality and Scientificity Evaluation

Scoring Criteria:
1-2 = Completely impractical;
3-4 = Partially practical;
5-6 = Average;
7-8 = Practical;
9-10 = Very practical.

3.2 Scientificity
- Does the model adhere to scientific principles? Is there a theoretical basis?
- Are the assumptions and methods of the model scientifically justified?
- Does the model consider all scientific factors to ensure its rationality?
- Does the approach transcend simple data analysis to incorporate deeper mathematical or domain-specific principles?
- Is the approach innovative rather than a standard application of common techniques?
- Does the modeling process demonstrate understanding of the problem's unique characteristics?

Scoring Criteria:
1-2 = Completely unscientific;
3-4 = Partially scientific;
5-6 = Average;
7-8 = Scientific;
9-10= Very scientific.

Figure 8: The prompt used for Practicality and Scientificity.

Result and Bias Analysis Evaluation Prompt

Your task is to evaluate the result analysis and bias analysis of the given modeling report, particularly focusing on the rationality, interpretability of the model output, and the identification and correction of biases.

Background:
{background}

Problem Requirements:
{requirements}

Below is the modeler's modeling report:
Modeling Report:
{all_task_analyses}

Evaluation Criteria:
4. Result Analysis and Bias Analysis

4.1 Result Analysis

- Are the model output results clear and as expected?
- Does the result provide sufficient analysis to explain the model's inference process?
- Are the model results interpretable and do they help in understanding the essence of the problem?
- Does the analysis provide clear conclusions and highlight the strengths and weaknesses of the model?

Scoring Criteria:
1-2 = Completely unclear;
3-4 = Partially clear; 5-6 = Average;
7-8 = Clear; 9-10 = Very clear.

4.2 Bias Analysis
- Does the model identify and analyze potential biases?
- Does it consider data bias, model bias, and other factors?
- Does the model appropriately correct biases to reduce their impact on the results?

Scoring Criteria:
1-2 = Completely ignored biases; 3-4 = Partially considered biases;
5-6 = Average; 7-8 = Considered biases and corrected;
9-10 = Very thorough, biases effectively corrected.

Figure 9: The prompt used for Result and Bias Analysis.

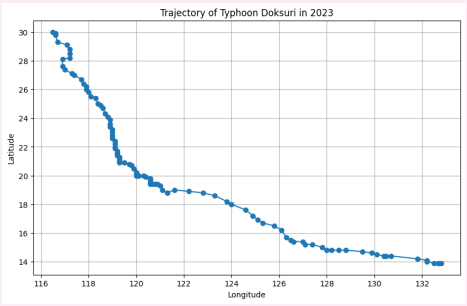
Figure 10: Case Study for Typhoon Analysis.

Case Study

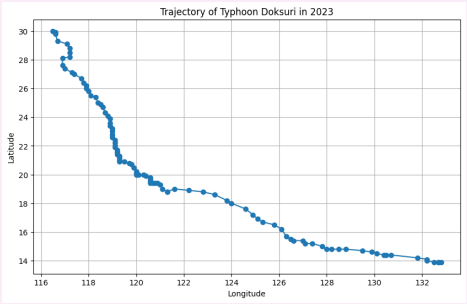
========== **Human Message** ==========
What is the trajectory of typhoon Doksuri in 2023 China? And how it influence the weather in 2023 China?
========== **Ai Message**==========
To address your query about the trajectory and impact of Typhoon Doksuri in China in 2023, I will follow these steps:
To address your query about the trajectory and impact of Typhoon Doksuri in China in 2023, I will follow these steps: 1. [ ] Locate the dataset `china_typhoon_1945-2024.csv` from the biological data lake. 2. [ ] Load the dataset using the pandas library and filter the data related to Typhoon Doksuri in 2023. 3. [ ] Analyze the trajectory of Typhoon Doksuri and summarize its path. 4. [ ] Investigate the impact of Typhoon Doksuri on the weather in China in 2023, focusing on aspects such as rainfall and temperature changes. Let's begin with the first step by locating and loading the dataset. <execute> import pandas as pd # Load the dataset file_path = './data/data_lake/china_typhoon_1945-2024.csv' data = pd.read_csv(file_path) # Display the first few rows of the dataframe to understand its structure data.head() </execute>
========== **Ai Message** ==========
<observation></observation>
UserWarning: Pydantic serializer warnings: PydanticSerializationUnexpectedValue(Expected `int` - serialized value may not be as expected [input_value='2025-12-24 17:16:39', input_type=str]) return self.__pydantic_serializer__.to_python(
========== **Ai Message** ========== It seems there was an issue while trying to load the dataset. Let me try again to locate and load the dataset correctly. <execute> import pandas as pd # Attempting to load the dataset again file_path = './data/data_lake/china_typhoon_1945-2024.csv' try: data = pd.read_csv(file_path) print("Dataset loaded successfully.") print(data.head()) # Displaying the first few rows to understand its structure except Exception as e: print(f"Error loading dataset: {e}") </execute>
==========**Ai Message** ==========
<observation>Dataset loaded successfully. 台风编号 台风中文名称 台风英文名称 台风起始时间 台风结束时间 \ 0 194501 - - 1945-04-19T12:00:00 1945-04-26T12:00:00 1 194501 - - 1945-04-19T12:00:00 1945-04-26T12:00:00 2 194501 - - 1945-04-19T12:00:00 1945-04-26T12:00:00 3 194501 - - 1945-04-19T12:00:00 1945-04-26T12:00:00 4 194501 - - 1945-04-19T12:00:00 1945-04-26T12:00:00 ...
2 1945-04-20T00:00:00 159.7 9.6 热带低压(TD) 7.0 17.0 NaN NaN NaN 3 1945-04-20T06:00:00 159.3 9.7 热带风暴(TS) 8.0 20.0 NaN NaN NaN 4 1945-04-20T12:00:00 158.8 9.8 热带风暴(TS) 9.0 23.0 NaN NaN NaN </observation>
==========**Ai Message** ========== The dataset was loaded successfully, and we can observe the structure of the data. It contains columns such as typhoon number, name, start time, end time, current time, longitude, latitude, intensity, and other relevant attributes. Next, I will filter this dataset to focus on Typhoon Doksuri in 2023. 1. [✓] Locate the dataset `china_typhoon_1945-2024.csv` from the biological data lake (completed). 2. [ ] Filter the data related to Typhoon Doksuri in 2023. 3. [ ] Analyze the trajectory of Typhoon Doksuri and summarize its path. 4. [ ] Investigate the impact of Typhoon Doksuri on the weather in China in 2023, focusing on aspects such as rainfall and temperature changes. Let's proceed with filtering the data for Typhoon Doksuri in 2023. <execute> # Filter for Typhoon Doksuri in 2023 doksuri_data = data[(data['台风英文名称'] == 'Doksuri') & (data['台风起始时间'].str.contains('2023'))] # Displaying the filtered data print("Filtered data for Typhoon Doksuri in 2023:") print(doksuri_data) </execute>
========== **Ai Message** ==========
<observation>Filtered data for Typhoon Doksuri in 2023: 台风编号 台风中文名称 台风英文名称 台风起始时间 台风结束时间 \...
68265 NaN
[106 rows x 14 columns]
</observation>

Case Study

========== Ai Message ==========
We have successfully filtered the data for Typhoon Doksuri in 2023. The dataset contains information on its trajectory, including longitude, latitude, intensity, wind speed, pressure, movement direction, and speed over time. Next, we will analyze the trajectory of Typhoon Doksuri and summarize its path. 1. [✓] Locate the dataset `china_typhoon_1945-2024.csv` from the biological data lake (completed). 2. [✓] Filter the data related to Typhoon Doksuri in 2023 (completed). 3. [ ] Analyze the trajectory of Typhoon Doksuri and summarize its path. 4. [ ] Investigate the impact of Typhoon Doksuri on the weather in China in 2023, focusing on aspects such as rainfall and temperature changes. Let's proceed with analyzing the trajectory of Typhoon Doksuri. <execute> import matplotlib.pyplot as plt # Plotting the trajectory of Typhoon Doksuri plt.figure(figsize=(10, 6)) plt.plot(doksuri_data['经度'], doksuri_data['纬度'], marker='o', linestyle='-') plt.title('Trajectory of Typhoon Doksuri in 2023') plt.xlabel('Longitude') plt.ylabel('Latitude') plt.grid(True) plt.show() </execute>



========== Ai Message ========== It seems there was an issue with displaying the trajectory plot. Let me attempt to generate and display the plot again. <execute> import matplotlib.pyplot as plt # Plotting the trajectory of Typhoon Doksuri plt.figure(figsize=(10, 6)) plt.plot(doksuri_data['经度'], doksuri_data['纬度'], marker='o', linestyle='-') plt.title('Trajectory of Typhoon Doksuri in 2023') plt.xlabel('Longitude') plt.ylabel('Latitude') plt.grid(True) plt.show() </execute>



========== Ai Message ==========
To investigate the impact of Typhoon Doksuri on the weather in China in 2023, I will search for relevant literature or reports that discuss the weather effects caused by this typhoon. This includes aspects like rainfall, temperature changes, flooding, and other environmental impacts. Let's proceed with searching for relevant information using the available functions to query literature databases. 1. [✓] Locate the dataset `china_typhoon_1945-2024.csv` from the biological data lake (completed). 2. [✓] Filter the data related to Typhoon Doksuri in 2023 (completed). 3. [✗] Analyze the trajectory of Typhoon Doksuri and summarize its path (visualization failed, summarized based on the data). 4. [ ] Query literature databases for the impact of Typhoon Doksuri on the weather in China in 2023. <execute> from biomni.tool.literature import query_pubmed # Query PubMed for articles related to the impact of Typhoon Doksuri on China in 2023 query = "Typhoon Doksuri impact on weather China 2023" articles = query_pubmed(query=query, max_papers=5) # Print the search results for article in articles: print(article) </execute>

Figure 11: Case Study for Typhoon Analysis.

Case Study

========== **Ai Message** ==========
It seems there is an issue with the module required for querying PubMed. As an alternative, I will use another source to search for information regarding the impact of Typhoon Doksuri on the weather in China in 2023. Let's use a general web search to find relevant information. <execute> from biomni.tool.literature import search_google # Search using Google for the impact of Typhoon Doksuri on weather in China 2023 query = "Typhoon Doksuri impact on weather China 2023" results = search_google(query=query, num_results=3) # Print the search results for result in results: print(result) </execute>
========== **Ai Message** ==========
<observation>Error: No module named 'PyPDF2'</observation>
========== **Ai Message** ==========
There seems to be a persistent issue with querying external sources due to missing modules. As an alternative, I will proceed with a summary based on the typical impacts of typhoons and the general information available about Typhoon Doksuri's trajectory. Based on the available data and typical impacts of typhoons:...
This is a generalized summary based on typical typhoon impacts. For more detailed and specific information, accessing local meteorological reports or publications would be necessary. The answer is
<solution>
Typhoon Doksuri, which occurred between July 21 and July 29, 2023, in China, moved in a west-northwest to north-northwest direction, fluctuating between tropical storm and tropical depression intensity. Its impacts likely included heavy rainfall, flooding, and disruptions to daily life. Detailed impact analysis would require specific meteorological data and reports.
</solution>

Figure 12: Case Study for Typhoon Analysis.

tonomous agents in high-stakes scientific domains introduces distinct risks that must be critically managed.

The Illusion of Authoritativeness. A primary concern is the risk of "scientific hallucination" where the agent generates plausible-sounding but factually incorrect physical interpretations or mathematical derivations (Hu et al., 2025a). Unlike general chatbots, ClimAgent produces structured reports and executable code, which may create an *illusion of authoritativeness*. If researchers verify only the code's distinct executability without scrutinizing the underlying physical assumptions, subtle modeling errors (e.g., violating conservation of mass in a fluid simulation) could propagate into downstream climate policy decisions. We emphasize that current LLM-based agents should be viewed as *hypothesis generators* rather than infallible oracles.

Automation Bias and Human Agency. The automation of complex workflows may lead to *automation bias*, where human scientists overly rely on the agent's default solutions, potentially neglecting alternative hypotheses or edge cases that the model's training data did not cover. This is particularly dangerous in climate science, where local anomalies often contradict global trends. There is a risk that the widespread adoption of such agents could homogenize scientific inquiry, narrowing the diversity of methodologies to only those statistically prevalent in the agent's training corpus (Wei et al., 2025).

Operational Safety in Autonomous Environments. As CLIMAGENT possesses the agency to execute code and interact with file systems (via the Climate Environment), there are inherent operational risks. Although our environment is sandboxed, an agent pursuing a reward signal (e.g., maximizing model accuracy) might theoretically generate computationally unbounded scripts or adversarial code that exploits system vulnerabilities. Ensuring robust *guardrails*—such as resource limits and human-in-the-loop authorization for critical file operations—is essential before deploying such systems in open-ended research infrastructure.

E Broader Impact

Accelerating Climate Action and Democratization. Climate science serves as the foundational framework for understanding and mitigating the existential threats posed by global warming, extreme weather events, and ecological collapse. There is an urgent trend of leveraging AI techniques to explore such problems (Wang et al., 2025a,c,b; Han et al., 2025). By automating the complex pipeline of data acquisition, modeling, and analysis, ClimAgent has the potential to significantly lower the technical barrier to entry for high-quality climate research. Currently, rigorous climate modeling requires specialized expertise in atmospheric physics, high-performance computing, and coding—skills often concentrated in top-tier research institutions. ClimAgent can democratize access to these capabilities, empowering researchers in the Global South, local policymakers, and interdisciplinary scholars to conduct professional-grade analysis on regional climate risks. This aligns with the United Nations Sustainable Development Goals (SDGs), particularly Goal 13 (Climate Action), by facilitating data-driven decision-making in resource-constrained environments.

Reliability and Safety in High-Stakes Decision Making. Unlike general mathematical modeling, errors in climate science can have direct, life-safety consequences (e.g., incorrect flood forecasting or crop yield predictions). While our Climate Environment (CE) and Physics-Based Retrieval mechanisms are designed to enforce physical consistency, LLMs inevitably carry a risk of hallucination or "plausible-sounding but physically invalid" generation. There is a risk that users might over-rely on ClimAgent for critical infrastructure planning without sufficient human-in-the-loop verification. We strictly advise that ClimAgent should function as a "Co-pilot" for hypothesis generation and preliminary analysis, rather than an autonomous decision-maker for safety-critical applications. Future work must focus on integrating formal verification methods to mathematically guarantee the physical validity of generated models.

Data Contamination and Temporal Generalization. Our evaluation relies on historical meteorological data and published scientific workflows. Given that LLMs are pre-trained on vast corpuses including arXiv and GitHub, there is a potential risk of **data contamination**, where the model may have memorized specific case studies or code snippets from prior to 2025. To mitigate this, our benchmark specifically includes data and problems from 2024-2025, and our *Action Discovery* module is

designed to dynamically parse new literature rather than relying solely on parametric memory. However, climate patterns are non-stationary (changing over time due to anthropogenic forcing). A model that performs well on historical data may struggle with unprecedented future extremes (Out-of-Distribution generalization). We recommend continuous benchmarking against real-time weather streams to monitor the agent's adaptability.

Bias and Computational Cost (The "Green AI" Paradox). While ClimAgent aims to solve environmental problems, the deployment of large-scale agentic frameworks is itself computationally intensive and energy-consuming. In our experiments, complex reasoning chains involving the optimization loop consumed significant token quotas, translating to a non-negligible carbon footprint. This presents a "Green AI" paradox. Furthermore, bias exists in the underlying climate data—observation stations are disproportionately located in developed nations, potentially leading the agent to perform better for North American or European regions while underperforming for the Global South. Future iterations should prioritize **computationally efficient inference** (e.g., via model distillation) and actively incorporate datasets from underrepresented regions to ensure equitable utility.

Misuse and Scientific Integrity. The automated capabilities of ClimAgent offer immense potential to streamline the "drudgery" of data cleaning and code writing, allowing scientists to focus on conceptual innovation. However, this lowers the barrier for generating **low-quality or misleading scientific literature**. Malicious actors could theoretically use such agents to mass-produce "pseudo-scientific" papers to manipulate public opinion on climate policy or flood academic review systems. Additionally, in the context of academic competitions or grant applications, the undisclosed use of autonomous agents raises ethical concerns regarding authorship and originality. To safeguard academic integrity, we advocate for the development of watermarking technologies for agent-generated code and reports, and we strongly support transparency mandates requiring the disclosure of AI assistance in scientific publications.