Towards LLM Agents for Earth Observation

Chia-Hsiang Kao[1], Wenting Zhao[1], Cheryl Lam[1], Aarush Umap[1], Shreelekha Revankar[1], Samuel Speas[1], Snehal Bhagat[1], Rajeev Datta[1], Cheng Perng Phoo[1], Utkarsh Mall[3], Carl Vondrick[2], Kavita Bala[1], Bharath Hariharan[1]

[1]Cornell University, [2]Columbia University, [3]MBZUAI

Correspondence: ck696@cornell.edu

Abstract

Earth Observation (EO) provides critical planetary data for environmental monitoring, disaster management, climate science, and other scientific domains. In this work we ask: *Are AI systems ready for reliable Earth Observation?* To answer this, we introduce **UnivEARTH**, a coding benchmark of 408 yes/no questions from NASA Earth Observatory articles across 7 various topics and over 15 satellite instruments and sources. Using Google Earth Engine API as a tool in a zero-shot setup, LLM agents achieve an accuracy of 40.0% where the code fails to run over 44% of the time. To better understand LLM agent behavior, we also analyze the impact of using the JavaScript API versus Python and the effect of providing documentation. Furthermore, we find that using a reflexion (Shinn et al., 2023) framework significantly reduces errors: Claude-4.5-Sonnet, Gemini-2.5-Pro, and GPT-5 accuracies rise to around 60%. However, these results remain only marginally above random chance. Taken together, our findings identify significant challenges to be solved before AI agents can automate earth observation, and suggest paths forward. Our dataset is publicly available.[1]

1 Introduction

In a range of academic disciplines from plant science to anthropology, scientists routinely find the need to analyze planetary data: data about land use, earth surface reflectance, chlorophyll content, and so on. This planetary data is collated and processed from a multitude of "Earth Observation" satellites, and the scientific process involves carefully choosing the right sensor, product, location, and time. Automating this analysis would thus significantly accelerate the scientific process.

Our goal in this paper is to explore AI systems that can automate the task of earth observation in these scientific workflows and thus accelerate the scientific process. While specialized automatic systems for specific earth observation tasks have been deployed for years (Watch, 2002; Giglio et al., 2016; Wu et al., 2018), they lack the flexibility needed for general-purpose, customized queries.

Recent efforts to combine agentic frameworks with remote sensing have largely relied on pre-trained models and predefined toolsets (Liu et al., 2024; Xu et al., 2024; Shabbir et al., 2025; Feng et al., 2025). While these approaches simplify benchmarking, they often fall short of capturing "in-the-wild" complexities. Real-world analysis introduces significant challenges, including the ambiguity of diverse data sources (e.g., Terra vs. Aqua, Landsat-8 vs. 9), sensor-specific spectral bands, and constraints such as revisit rates and cloud cover. By relying on curated environments, prior studies often fail to test applicability in more real-world, dynamic settings. We therefore ask: ***Are AI systems ready for reliable Earth Observation?***

With these desiderata in mind, we begin by introducing **UnivEARTH** : a question-answering (QA) benchmark designed to evaluate LLMs for earth observation. There are two challenges in building such a benchmark: (1) we need to know the kind of questions that one might ask about earth observation data and the corresponding answers, and (2) we need to ensure that the evidence or data needed to support the answer exists and is available. Unlike existing benchmarks, such data of questions, answers, and supporting evidence is not freely available. We address this challenge by leveraging a unique public resource: articles from the *NASA Earth Observatory* (NASA). Each article walks through conclusions derived from observations from satellite imagery. We rigorously curate question-answer pairs from these articles by: (1) leveraging LLMs alongside manually curated QA examples, (2) verifying question answerability through Google Earth Engine (GEE), and (3) careful independent review of each example. The re-

[1]https://iandrover.github.io/2025_univearth/

Findings of the Association for Computational Linguistics: ACL 2026, pages 2597–2611
July 2-7, 2026 ©2026 Association for Computational Linguistics

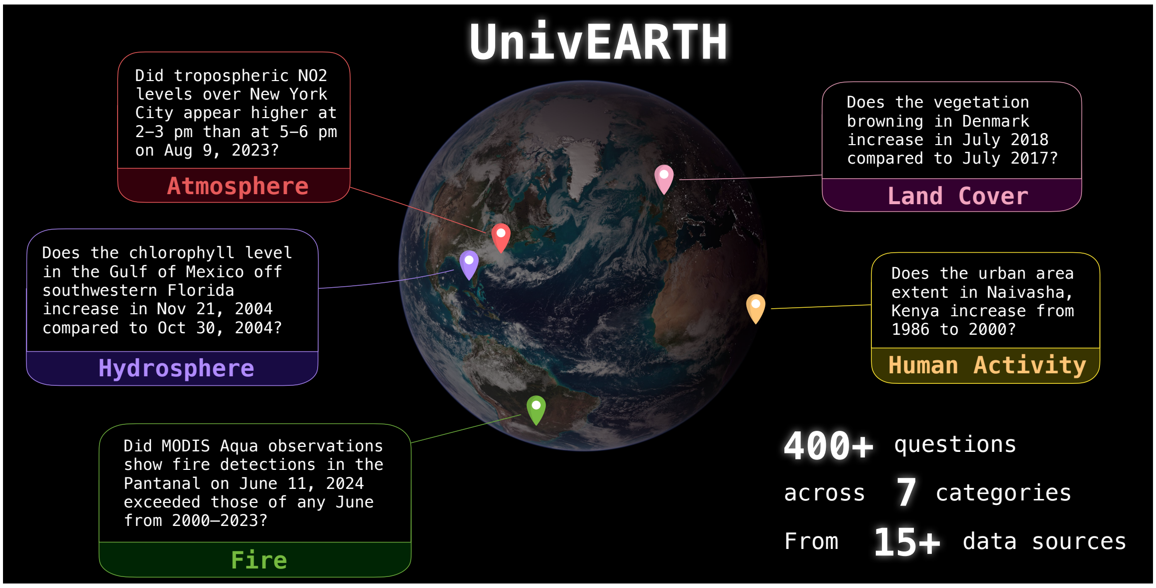

Figure 1: We propose **UnivEARTH** for benchmarking AI agents in Earth Observation.

sulting dataset, **UnivEARTH**, comprises 408 high-quality yes/no questions spanning 7 diverse topics and 15 main sources.

We benchmark several state-of-the-art models, including Gemini (Team et al., 2023), Claude, GPT, DeepSeek (Guo et al., 2025), and Kimi (Team et al., 2025). In our setup, simply answering the question is not enough: the models must ground their answers in evidence. We ask models to generate executable code using the Google Earth Engine (Gorelick et al., 2017) API to derive quantitative results that support their answers. Our evaluation reveals significant limitations in current models. We find that agents fail to produce executable code a significant portion of the time, resulting in a best-case baseline accuracy of only 40.0%. To better understand these failures, we analyze performance across various dimensions. We find that: (1) the choice of API language (Python vs. JavaScript) slightly impacts model performance, while logic errors persist across both; (2) providing specific documentation on sensor information (e.g., Collection IDs) drastically reduces syntax errors and hallucinations; and (3) implementing a *reflexion* mechanism (Shinn et al., 2023), allowing agents to debug based on error logs, significantly recovers performance, boosting accuracy to over 60%.

In sum, our work makes three key contributions:

- We curate **UnivEARTH**, a novel evaluation benchmark for Earth Observation derived from authoritative NASA sources with verified answers and executable ground truth.
- We benchmark state-of-the-art LLMs, revealing that while they possess general knowledge, they struggle with the specific syntax and data retrieval logic required for EO analysis.
- We demonstrate that advanced agentic strategies, specifically the injection of precise documentation and the use of error-based reflexion, can bridge the gap between generation failures and reliable scientific analysis.

2 Related Work

LLMs for Scientific Applications. Scientific question answering has garnered significant attention, demonstrated by the development of benchmarks across various domains. General scientific QA benchmarks assess reasoning across multiple scientific disciplines (Saikh et al., 2022; Hendrycks et al., 2020; Wang et al., 2023; Liang et al., 2024; Feng et al., 2024; Wang et al., 2024b), while specialized benchmarks focus on specific areas such as medicine and biology (He et al., 2020; Li et al., 2024b), chemistry and material science (Jablonka et al., 2024; Alampara et al., 2024; Chen et al., 2025b), and remote sensing (Wang et al., 2024a; Danish et al., 2024; Li et al., 2024a).

Many of these prior benchmarks rely on models' internal knowledge, which may not be sufficiently rigorous in a scientific domain. In contrast, **UnivEARTH** demands grounding answers in evidence derived from satellite imagery and products, requiring more interpretable and explicit reasoning. In this vein, our work is similar to prior work on leveraging existing tools or databases (M. Bran et al., 2024; Fossi et al., 2024; Campbell et al., 2025; Laurent et al., 2024), but requires models to navigate a much larger repertoire of data sources (here, sensors and products). These capabilities are a necessary first step if one seeks to automate discovery in the earth sciences, as prior work has sought to do for chemistry (Zheng et al., 2025; Chen et al., 2025a), biology (Swanson et al., 2024), or material science (Strieth-Kalthoff et al., 2024).

Code Generation and Tool-Using AI. Outside of scientific applications, several benchmarks evaluate code generation capabilities, including SWE-bench (Jimenez et al., 2023), SWT-Bench (Mündler et al., 2024), LiveCodeBench (Jain et al., 2024), and SWE-bench Multimodal (Yang et al., 2024). These benchmarks primarily focus on general software engineering tasks rather than domain-specific scientific applications. In the context of data analysis, text-to-SQL benchmarks like Spider (Yu et al., 2018), SEDE (Hazoom et al., 2021), BIRD (Li et al., 2023), and Spider 2.0 (Lei et al., 2024) evaluate models' ability to translate natural language questions into database queries.

Earth Observation With AI Agents. Recently, there has been a surge in research applying agentic frameworks to remote sensing and satellite imagery analysis. For instance, Change-Agent (Liu et al., 2024) and RS-Agent (Xu et al., 2024) integrate LLMs with pre-trained vision models to address tasks such as change detection, object detection, and scene classification. Similarly, ThinkGeo (Shabbir et al., 2025) and EarthAgent (Feng et al., 2025) utilize predefined functions and tools to execute tasks and benchmark their performance. However, these approaches often rely on highly curated environments that fail to capture 'in-the-wild' complexities. Real-world scenarios introduce significant ambiguities, including diverse data sources (e.g., Terra/Aqua, Landsat-8/9, or Sentinel-2), sensor-specific spectral bands, and availability constraints such as revisit rates and cloud cover. Previous work simplifies the complexities of real-world problems and thus limits its applicability in more dynamic, real-world settings. **UnivEARTH** extends this paradigm to a realistic Earth Observation domain, designed to handle the complexities of accessing and analyzing satellite data.

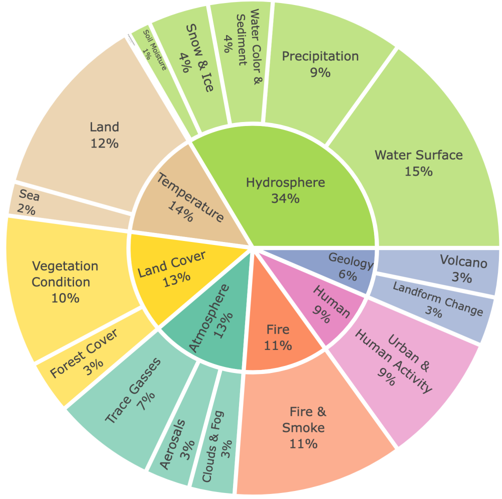

Figure 2: Hierarchical distribution of Earth science topics and sub-categories in **UnivEARTH** . The inner ring represents the parent domain, while the outer ring details specific phenomena.

3 Benchmark Construction

3.1 Data Source

EO science relies heavily on the analysis of remotely sensed data to investigate changes and phenomena. This characteristic makes it particularly suitable for automation through AI agents that can process and analyze large volumes of imagery data. However, benchmarking such AI capabilities requires high-quality question-answer pairs that are both scientifically sound and verifiable through available data sources. To develop our benchmark, we identified NASA's Earth Observatory website (NASA) as an authoritative primary source. Since its inception in May 1998, this platform has published articles covering diverse topics including air quality, climate change, human impact monitoring, and natural events. These articles, authored by NASA Earth Observatory's science writers, provide reliable scientific reporting based on imagery analysis and research findings.

3.2 Data Collection and Validation

The curation pipeline comprises of three stages: annotation, verification, and review.

Annotation. We focus on NASA Earth Observatory website articles with the cutoff time of September 30, 2025. Three annotators generate candidate yes/no question-answer pairs with supporting sentences. We chose the yes/no question format because assessing the correctness of free-form answers is challenging and less objective in scien-

Table 1: Distribution of the main data sources grouped by sensing modality and instrument type.

Category – Instrument / Source	Count
Multispectral Optical Imaging	
Landsat Series (4–9)	162
MODIS System (Terra/Aqua)	140
VIIRS	60
SeaWiFS	3
Sentinel-2	2
Atmospheric Composition	
Sentinel-5P (TROPOMI)	16
TOMS	4
OMI-Aura	4
TEMPO	2
Hydrology & Cryosphere	
GRACE	11
GPM	8
CHIRPS	7
SMAP	4
Weather Satellites & Models	
GEOS Model	17
ERA5 Reanalysis	3
Total	**443**

tific domains. During this process, we disregard articles about sensor specifications, general introductions, non-satellite imagery, transient observations (e.g., wind speed, tides), or statements without clear sources. Table 2 shows some example questions and supporting sentences.

Verification. We examined whether questions derived from NASA articles can be answered using the data available in Google Earth Engine (GEE) (Gorelick et al., 2017). Background details on GEE are presented in Appendix-B. This verification step was critical for ensuring the questions in the benchmark are approachable with existing data catalog hosted in GEE. This was necessary because some articles describe phenomena using sensors or products not available in GEE, making it impossible for agents to answer them. For example, articles (e.g., on December 2015, June 2016, or May 2017) about sulfur dioxide often reference the Ozone Monitoring Instrument (OMI) on the Aura satellite. While OMI has monitored sulfur dioxide and other air pollutants since 2004, its sensor data is not available on GEE. The alternative Sentinel-5P NRTI sulfur dioxide datasets only begin in 2018, creating a temporal constraint.

Review. Following verification, we recruited reviewers to evaluate the quality and clarity of the questions. These reviewers were asked to: (Q1) Provide a yes/no answer to each question based on the text and image of the article; (Q2) Assess whether the answer was supported by the text in the corresponding NASA article; (Q3) Evaluate whether the answer was supported by imagery in the article; and (Q4) Assess if location information needs verification through external sources. The fourth assessment point was included because some questions, particularly those manually edited or designed, required geographical review. In these cases, reviewers were permitted to use Google Maps to verify geo-locations. Please refer to Appendix. D for the reviewer instruction document.

In the first version of the dataset, composed of 140 questions, four reviewers were recruited, with each reviewer evaluating half of the dataset. The initial review showed inter-reviewer agreement rates of 90.1%, 73.2%, 78.9%, and 81.7% for Q1, Q2, Q3, and Q4, respectively. The agreement rate is computed as an exact match. Following this initial assessment, we iteratively revised ambiguous questions with each reviewer until we reached complete agreement on Q1. In the extended version of the dataset, composed of 268 questions, each question is reviewed by two reviewers, and the question-answer pairs are iteratively revised until reviewers and annotators reach consensus. Because the review process focused on verifying the correctness of answers with respect to the source NASA articles rather than domain-specific expertise, we recruited seven reviewers who were undergraduate or graduate students in computer science.

3.3 Dataset Statistics

Figure 2 illustrates the distribution of topics within the dataset. The largest category is *hydrosphere* (34%), encompassing questions regarding surface water extent, precipitation, water color, snow, and moisture. Other major categories include temperature (14%, primarily land surface temperature), land cover (13%, such as vegetation and forest cover), and atmosphere (13%, including trace gases, aerosols, and clouds). Additional domains cover fire, human activities (e.g., urban expansion), and geology. This broad range of tasks presents a holistic view of Earth Observation, marking a prominent distinction from previous works that primarily focus on object detection or scene classification. Complementing this, Table 1 details the major data sources used in our dataset. The majority of observations are derived from Landsat and the MODIS system (onboard the Terra and Aqua satellites), and VIIRS.

Table 2: Examples of **UnivEARTH** .

Topic	Example	Supporting Sentences
Atmosphere	Did nitrogen oxide concentrations in the Northern Hemisphere increase from 2019 to 2020?	The annual growth rate for 2020 was the highest scientists had recorded since systematic annual methane measurements began in 1983—an increase of 15 parts per billion, which was exceeded again in 2021.
Land Cover	Does forest cover decrease in Argentina's Salta Province from December 2000 to December 2019?	The images above show deforestation over a span of two decades around the Salta Province of northern Argentina. The image from December 18, 2000, shows a mix of cleared land and greener areas. The image from December 24, 2019, shows much of the forest replaced by large fields.
Hydrosphere	Does Lake Erie have more ice coverage compared to the other Great Lakes in February 14, 2018 afternoon?	On the same date last year, total ice cover was 9.7 percent. Lake Erie was the iciest of the five lakes, with 93.3 percent iced over.

3.4 Relevance to Science and Real-World Impact

UnivEARTH captures phenomena with significant real-world relevance and active scientific interest. For instance, our benchmark includes questions about red tides in the Gulf of Mexico derived from a December 2004 article, addressing phenomena linked to potential fish mortality and human respiratory irritation. Another question focuses on the number of lakes on the Tibetan Plateau based on a March 2025 article, directly connecting to recent research on accelerated lake formation in this critical region (Li et al., 2022; Lei et al., 2023; Zhou et al., 2024a,b). The benchmark also covers other scientifically relevant topics including chlorophyll concentration and climate patterns in the Pacific Ocean (Wang et al., 2005), the trend of disappearing lakes in Siberia (Smith et al., 2005), lake surface albedo dynamics (Argaman et al., 2012), groundwater depletion in the Indus Basin (Richey et al., 2015), increasing global leaf area (Chen et al., 2019), and global cropland expansion (Potapov et al., 2022). Thus our benchmark provides a sampling of questions that scientists may want answered in the course of their research.

4 Benchmarking SoTA Agents with UnivEARTH

In this section, we evaluate state-of-the-art LLM agents on our benchmark dataset.

Experimental Setup. We adopt a zero-shot setup for evaluation. Given a query, the model is instructed to first reason about the plan for the code and then generate a Google Earth Engine (GEE) Python script. This script is executed, and the resulting output, along with the original question and code, is processed by Gemini-2.5-Flash to derive the final output.

To address the high frequency of code generation failures (discussed subsequently), we employ a specific error taxonomy during the evaluation. We prompt Gemini-2.5-Flash to classify failures into one of the following categories:

- Empty Collection (C1): No images found for the specified date or location in one or both comparison points.
- No Valid Pixels (C2): Images exist, but no valid pixels remain after preprocessing, such as cloud masking.
- Calculation Failure (C3): The result is None, NaN, or otherwise invalid[2].
- Code/Syntax Error (D): Failures caused by incorrect image/band names, syntax errors, invalid methods, or memory issues.

This categorization allows for a detailed diagnosis of model failure modes in real-world scenarios.

Our primary metric is correctness, defined as the fraction of questions where the generated answer matches the ground truth. We also measure the wrong answer rate alongside the distribution of error types. We benchmark a suite of LLM agents, including Gemini-2.5-Pro, Gemini-2.5-Flash (Team et al., 2023), Claude-4.5-Sonnet, Claude-4.5-Haiku, GPT-5, DeepSeek-Reasoner (DeepSeek-AI, 2025), DeepSeek-Chat, and Kimi-k2 (Team et al., 2025).

[2]For example, Claude-4.5-Sonnet produced a physically implausible value of 1723.29°C by neglecting to apply the 0.01 scale factor inherent to the `NOAA/CDR/OISST/V2_1` dataset.

Table 3: Performance Comparison of Code LLMs on Earth Engine Tasks. The table reports the distribution of outcomes across distinct categories. **Execution Success** indicates the code ran without errors (Correct vs. Wrong Answer). **Failed Execution/Logic** indicates specific failure modes: Empty Collection or Images (C1), No Valid Pixels (C2), Calculation Failure (C3), or Syntax Error (D).

	Execution Success (%)		Failed Execution / Logic (%)			
Model	**Correct**	Wrong Ans.	Empty (C1)	No Pixels (C2)	Calc Fail (C3)	Syntax (D)
Gemini-2.5-Pro	36.9	9.6	15.8	2.7	1.0	34.0
Gemini-2.5-Flash	10.6	5.9	15.0	4.4	3.2	60.8
Claude-4.5-Sonnet	**40.0**	15.7	15.0	4.9	2.9	21.4
Claude-4.5-Haiku	16.7	12.8	14.0	4.9	4.9	46.7
GPT-5	33.1	12.0	4.4	4.9	2.7	42.9
DeepSeek-Reasoner	28.5	13.0	15.0	5.2	4.2	34.2
DeepSeek-Chat	27.0	13.0	17.0	7.4	4.9	30.7
Kimi-k2	17.0	8.4	23.1	3.7	1.2	46.7

Table 4: Comparison of hallucinated vs. correct GEE Collection IDs. Models often struggle with specific naming conventions (e.g., version numbers, tiers, or provider prefixes).

Domain	**Collection ID (✗ Hallucinated / ✓ Correct)**
Landsat	✗ `LANDSAT/LC08/C02/SR`
	✓ `LANDSAT/LC08/C02/T1_L2`
VIIRS	✗ `NOAA/VIIRS/001/VNP14IMGTDL_NRT`
	✓ `NASA/LANCE/SNPP_VIIRS/C2`
TRMM	✗ `NASA/TRMM/3B43V7`
	✓ `TRMM/3B43V7`
MODIS	✗ `MODIS/051/MCD19A2_GRANULAR_AOD`
	✓ `MODIS/061/MCD19A2_GRANULES`
JRC	✗ `JRC/GHSL/P2023A/GHS_POP`
	✓ `JRC/GHSL/P2023A/GHS_SMOD_POP`

4.1 Answering Questions With Google Earth Engine

Table 3 presents the average performance across the dataset. The highest overall accuracy achieved was only 40.0% by Claude-4.5-Sonnet, closely followed by Gemini-2.5-Pro (36.9%) and GPT-5 (33.1%). DeepSeek-Reasoner and DeepSeek-Chat exhibit comparable levels of correctness, whereas Kimi-k2, Claude-4.5-Haiku, and Gemini-2.5-Flash perform the poorest, with accuracy below 20%.

A deeper investigation into failure modes reveals a persistently high rate of syntax and method errors (Category D). Even the strongest model, Claude-4.5-Sonnet, suffered a 21.4% error rate in this category, while all others exceeded 30%. This indicates that current models lack proficiency in utilizing the GEE API. We show some examples of hallucinated collection ID or names in Table 4.

Regarding specific execution failures, the most common issue is that there is no available image for the queried dataset (C1). In contrast, we observed a relatively lower frequency of errors related to empty pixel sets (C2) or calculation failures (C3). Overall, the low accuracy suggests that existing LLMs are not yet capable of reliably producing code for Earth Observation tasks, likely due to the under-representation of this specific domain in pre-training corpora. Consequently, **UnivEARTH** serves as a practically relevant, out-of-domain benchmark for future research aimed at overcoming these limitations.

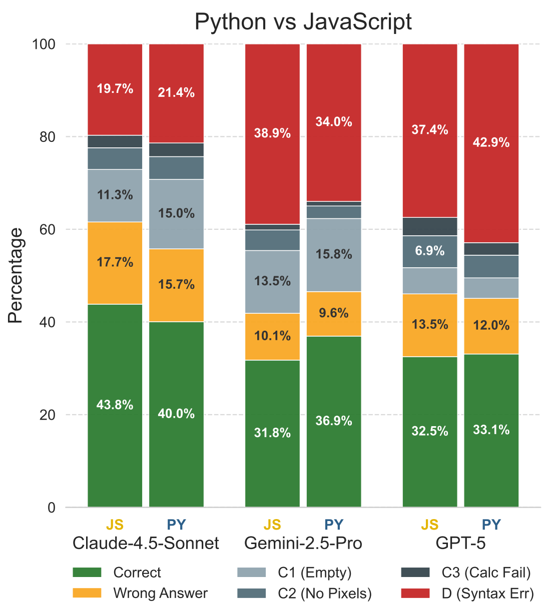

Figure 3: Performance in Python against JavaScript.

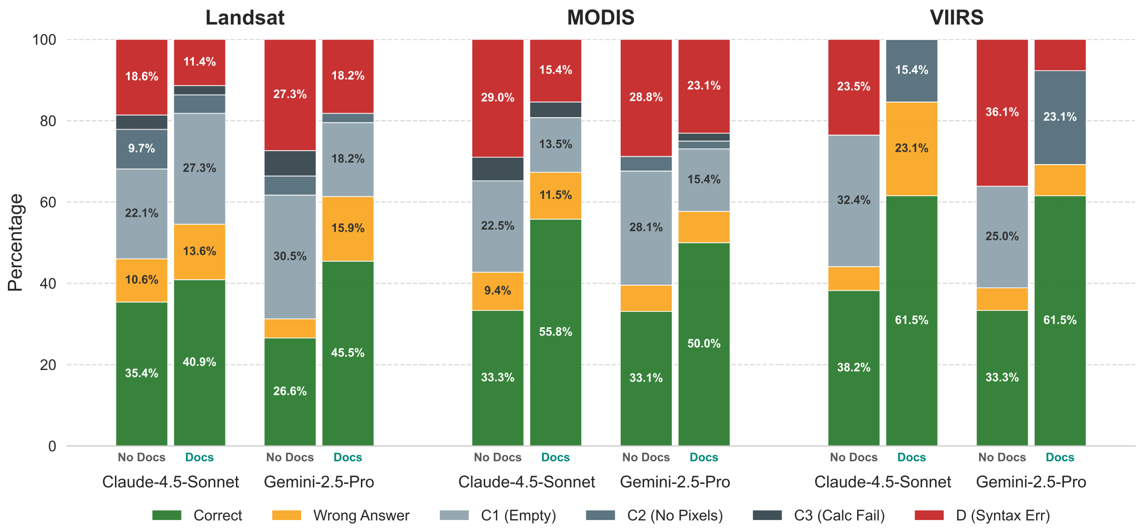

Figure 4: Impact of providing ground-truth Collection IDs (documentation) on performance, stratified by instrument.

4.2 Python versus JavaScript

Since Google Earth Engine (GEE) also offers a JavaScript API for rapid prototyping and interactive analysis, we benchmark whether model performance is language-dependent. As illustrated in Figure 3, models exhibit distinct sensitivities to the choice of programming language. Claude-4.5-Sonnet achieves a slight performance gain, rising from 40.0% in Python to 43.8% in JavaScript. Conversely, performance for Gemini-2.5-Pro and GPT-5 degrades by 5.1% and 0.6%, respectively, when using JavaScript. Overall, LLM agents do show a comparable performance regarding API languages.

Investigating the cause, we find that the logic-based wrong answer rate remains consistent across languages. Instead, the divergence stems from execution stability. For Claude, the improvement is driven by a better ability to identify valid image collections in JavaScript (i.e., reduce in type D error). In contrast, the performance drop in Gemini-2.5-Pro models is primarily caused by an increase in syntax errors, indicating a relative unfamiliarity with the GEE JavaScript API compared to its Python counterpart.

4.3 Does Providing Collection ID Reduce Hallucinations?

Given the high prevalence of hallucinated dataset names observed in Table 4, we investigate whether providing specific GEE Collection IDs and metadata in the prompt mitigates these errors. We focus our analysis on questions resolvable via Landsat, MODIS, and VIIRS modalities. To this end, we adopt a two-round dialogue approach. In the first round, the agent is asked to select the appropriate modality ("Landsat", "MODIS", "VIIRS", or "others") for the query. Based on this selection, the associated documentation, including the precise Collection ID and band names, is injected into the prompt for the second round, where the model generates the executable code. If "others" is selected, no additional information is provided.

The stratified results are presented in Figure 4. The provision of documentation yields a clear reduction in Code/Syntax Errors (Category D) across nearly all configurations. For instance, in the MODIS modality, Claude-4.5-Sonnet reduces its syntax error rate from 29.0% to 15.4%. This effect is most pronounced in the VIIRS modality, where Claude-4.5-Sonnet effectively eliminates syntax errors (23.5% to 0%). Consequently, VIIRS accuracy improves significantly from 38.2% to 61.5%.

However, we observe that a reduction in syntax errors does not always translate directly to correctness. As execution success rates rise, other error categories become more visible. In the VIIRS case, while the code runs successfully, the Wrong Answer rate increases from 8% to 23.1% for Claude-4.5-Sonnet. In the case of Gemini-2.5-Pro on VIIRS, the errors change from category C2 (gray) and D (red) to C3 (darker gray). These findings suggest that while documentation enables the model to successfully access the data, it does not necessarily ensure the full correctness of the coding.

In summary, while documentation successfully lowers the barrier to data access (resolving hal-

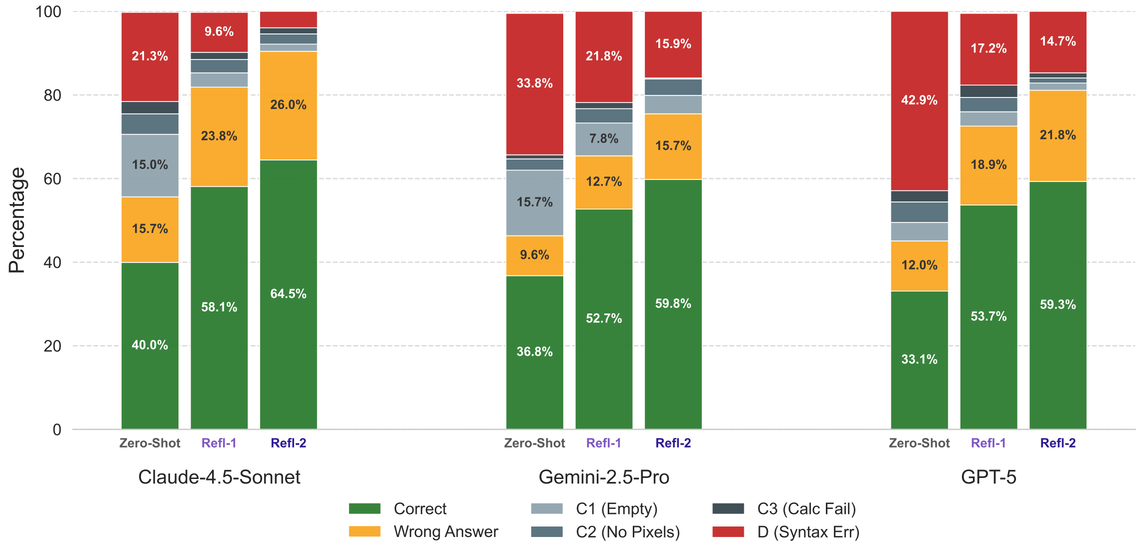

Figure 5: Reflexion improves performance. We observe a consistent increase in correctness across all agents. This gain is primarily driven by the effective mitigation of code and syntax errors (Category D).

lucinations), it exposes secondary challenges in logical data processing. Once the syntax hurdles are cleared, models often produce valid code that fails to retrieve data due to incorrect date filtering or cloud masking logic, shifting the failure mode from an immediate crash (Category D) to an empty result (Category C1).

4.4 Does Reflection Help?

To assess the potential for improvement beyond the zero-shot baseline, we leverage the *reflexion* framework (Shinn et al., 2023). In our experimental setup, the reflection mechanism is triggered specifically when the model fails to produce a valid execution result (i.e., encountering syntax errors, timeouts, or empty returns), rather than on logical correctness alone. During the reflection phase, the agent is presented with the original query, its previous code attempt, and the resulting execution traceback or error log. It is then instructed to debug and regenerate the script. We allow for a maximum of two rounds of such reflection.

As illustrated in Figure 5, the reflexion scheme show substantial performance gains across all evaluated models. Claude-4.5-Sonnet demonstrates the highest overall efficacy, improving its correctness from 40.0% in the zero-shot setting to 64.5% after two rounds of reflection. GPT-5 also exhibits a dramatic recovery, jumping from a baseline of 33.1% to 59.3%, a 79 percentage point increase.

The breakdown of error categories reveals the mechanics of this improvement. The feedback loop is particularly effective at mitigating Code/Syntax Errors (Category D). We also observe a notable reduction in Empty Images (C1) errors, suggesting that the error logs effectively guide the models to correct invalid date filters or choose the correct modalities that have the data for the specific time.

However, it is worth noting that as syntax and execution errors decrease, the rate of wrong answers increases simultaneously and proportionally (e.g., rising from 15.7% to 26.0% for Claude-4.5-Sonnet). This indicates that while self-reflection allows the code to run to completion, it does not guarantee that the code logic is correct. Nevertheless, the ability to convert fatal errors into executable code significantly boosts the overall ceiling for autonomous Earth Observation analysis.

5 Conclusion

We introduce **UnivEARTH** to investigate how reliable are AI agents in solving Earth observation questions. When asked to produce evidence in the form of Google Earth Engine (GEE) code, the best zero-shot accuracy is 40.0% primarily because of models' inability to correctly navigate the many data sources. Our detailed analysis identifies clear pathways for improvement. We show that the primary bottleneck is not a lack of coding capability, but a lack of domain-specific grounding. By injecting precise documentation regarding Collection IDs, we significantly reduce hallucinations, though this unmasks deeper challenges in downstream logical reasoning. Furthermore, we find that employing iterative self-correction (reflexion) can recover a substantial portion of these failures, boosting performance to 64.5%. By curating **UnivEARTH** , we hope to expose the gap between general-purpose reasoning and domain-specific execution, paving the way for future research into agents that can reason scientifically about the Earth.

Limitations

We acknowledge several limitations of **UnivEARTH** . First, the current benchmark does not include unanswerable questions (Rajpurkar et al., 2018; Kim et al., 2022), where the ground-truth answer is "inconclusive." Second, **OpenEARTH** is limited to a yes–no question format; future versions could be extended to include multiple-choice scenarios. Finally, most questions are mined from NASA websites; future iterations could broaden coverage by sourcing questions from scientific papers.

References

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Krishnan, and Kevin Maik Jablonka. 2024. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv:2411.16955*.

E Argaman, SD Keesstra, and A Zeiliguer. 2012. Monitoring the impact of surface albedo on a saline lake in sw russia. *Land Degradation & Development*, 23(4):398–408.

Quintina Campbell, Sam Cox, Jorge Medina, Brittany Watterson, and Andrew D White. 2025. Mdcrow: Automating molecular dynamics workflows with large language models. *arXiv preprint arXiv:2502.09565*.

Bangqian Chen, Xiangming Xiao, Xiangping Li, Lianghao Pan, Russell Doughty, Jun Ma, Jinwei Dong, Yuanwei Qin, Bin Zhao, Zhixiang Wu, and 1 others. 2017. A mangrove forest map of china in 2015: Analysis of time series landsat 7/8 and sentinel-1a imagery in google earth engine cloud computing platform. *ISPRS Journal of Photogrammetry and Remote Sensing*, 131:104–120.

Chi Chen, Taejin Park, Xuhui Wang, Shilong Piao, Baodong Xu, Rajiv K Chaturvedi, Richard Fuchs, Victor Brovkin, Philippe Ciais, Rasmus Fensholt, and 1 others. 2019. China and india lead in greening of the world through land-use management. *Nature sustainability*, 2(2):122–129.

Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. 2025a. Autobench: An automated benchmark for scientific discovery in llms. *arXiv preprint arXiv:2502.15224*.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Zirui Song, Xin Gao, and Xiangliang Zhang. 2025b. Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 8(1):4.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2024. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, and 1 others. 2025. Earthagent: Unlocking the full landscape of earth observation with agents. *arXiv preprint arXiv:2509.23141*.

Gabriele Fossi, Youssef Boulaimen, Leila Outemzabet, Nathalie Jeanray, Stephane Gerart, Sebastien Vachenc, Joanna Giemza, and Salvatore Raieli. 2024. Swiftdossier: Tailored automatic dossier for drug discovery with llms and agents. *arXiv preprint arXiv:2409.15817*.

Louis Giglio, Wilfrid Schroeder, and Christopher O Justice. 2016. The collection 6 modis active fire detection algorithm and fire products. *Remote sensing of environment*, 178:31–41.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Moshe Hazoom, Vibhor Malik, and Ben Bogin. 2021. Text-to-sql in the wild: A naturally-occurring dataset based on stack exchange data. *arXiv preprint arXiv:2106.05006*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hadi Jaafar and Roya Mourad. 2021. Gymee: a global field-scale crop yield and et mapper in google earth engine based on landsat, weather, and soil data. *Remote Sensing*, 13(4):773.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

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*.

Najoung Kim, Phu Mon Htut, Samuel R Bowman, and Jackson Petty. 2022. (qa)2: Question answering with questionable assumptions. *arXiv preprint arXiv:2212.10003*.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, and 1 others. 2024. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *arXiv preprint arXiv:2411.07763*.

Yanbin Lei, Tandong Yao, Yongwei Sheng, Kun Yang, Wei Yang, Shenghai Li, Jing Zhou, Yaozhi Jiang, and Yifan Yu. 2023. Unprecedented lake expansion in 2017–2018 on the tibetan plateau: Processes and environmental impacts. *Journal of Hydrology*, 619:129333.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, and 1 others. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357.

Ke Li, Fuyu Dong, Di Wang, Shaofeng Li, Quan Wang, Xinbo Gao, and Tat-Seng Chua. 2024a. Show me what and where has changed? question answering and grounding for remote sensing change detection. *arXiv preprint arXiv:2410.23828*.

Meng Li, Baisha Weng, Denghua Yan, Wuxia Bi, and Hao Wang. 2022. Variation trends and attribution analysis of lakes in the qiangtang plateau, the endorheic basin of the tibetan plateau. *Science of The Total Environment*, 837:155595.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024b. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2024. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Niels Mündler, Mark Müller, Jingxuan He, and Martin Vechev. 2024. Swt-bench: Testing and validating real-world bug-fixes with code agents. *Advances in Neural Information Processing Systems*, 37:81857–81887.

NASA. NASA earth observatory articles. https://earthobservatory.nasa.gov.

Vahid Nasiri, Azade Deljouei, Fardin Moradi, Seyed Mohammad Moein Sadeghi, and Stelian Alexandru Borz. 2022. Land use and land cover mapping using sentinel-2, landsat-8 satellite images, and google earth engine: A comparison of two composition methods. *Remote sensing*, 14(9):1977.

Peter Potapov, Svetlana Turubanova, Matthew C Hansen, Alexandra Tyukavina, Viviana Zalles, Ahmad Khan, Xiao-Peng Song, Amy Pickens, Quan Shen, and Jocelyn Cortez. 2022. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nature Food*, 3(1):19–28.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Alexandra S Richey, Brian F Thomas, Min-Hui Lo, John T Reager, James S Famiglietti, Katalyn Voss, Sean Swenson, and Matthew Rodell. 2015. Quantifying renewable groundwater stress with grace. *Water resources research*, 51(7):5217–5238.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Manali Santra, Chandra Shekhar Dwivedi, and Arvind Chandra Pandey. 2024. Quantifying shoreline dynamics in the indian sundarban delta with google earth engine (gee)-based automatic extraction approach. *Tropical Ecology*, 65(3):426–442.

Nazmus Sazib, Iliana Mladenova, and John Bolten. 2018. Leveraging the google earth engine for drought assessment using global soil moisture data. *Remote sensing*, 10(8):1265.

Gabriel B Senay, MacKenzie Friedrichs, Charles Morton, Gabriel EL Parrish, Matthew Schauer, Kul Khand, Stefanie Kagone, Olena Boiko, and Justin Huntington. 2022. Mapping actual evapotranspiration using landsat for the conterminous united states: Google earth engine implementation and assessment of the ssebop model. *Remote Sensing of Environment*, 275:113011.

Akashah Shabbir, Muhammad Akhtar Munir, Akshay Dudhane, Muhammad Umer Sheikh, Muhammad Haris Khan, Paolo Fraccaro, Juan Bernabe Moreno, Fahad Shahbaz Khan, and Salman Khan. 2025. Thinkgeo: Evaluating tool-augmented agents for remote sensing tasks. *arXiv preprint arXiv:2505.23752*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

L C Smith, Yongwei Sheng, GM MacDonald, and LD Hinzman. 2005. Disappearing arctic lakes. *Science*, 308(5727):1429–1429.

Felix Strieth-Kalthoff, Han Hao, Vandana Rathore, Joshua Derasp, Théophile Gaudin, Nicholas H Angello, Martin Seifrid, Ekaterina Trushina, Mason Guy, Junliang Liu, and 1 others. 2024. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384(6697):eadk9227.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. 2024a. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5481–5489.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19162–19170.

Xiaoqiang Wang and Bang Liu. 2024. Oscar: Operating system control via state-aware reasoning and re-planning. *arXiv preprint arXiv:2410.18963*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Xiujun Wang, James R Christian, Ragu Murtugudde, and Antonio J Busalacchi. 2005. Ecosystem dynamics and export production in the central and eastern equatorial pacific: A modeling study of impact of enso. *Geophysical Research Letters*, 32(2).

Global Forest Watch. 2002. Global forest watch. *World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002)*.

Huan Wu, Guojun Gu, Yan Yan, Zhen Gao, and Robert F Adler. 2018. Global flood monitoring using satellite precipitation and hydrological modeling. *Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting*, pages 87–113.

Wenjia Xu, Zijian Yu, Boyang Mu, Zhiwei Wei, Yuanben Zhang, Guangzuo Li, and Mugen Peng. 2024. Rs-agent: Automating remote sensing tasks through intelligent agent. *arXiv preprint arXiv:2406.07089*.

John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, and 1 others. 2024. Swe-bench multimodal: Do ai systems generalize to visual software domains? *arXiv preprint arXiv:2410.03859*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Linwei Yue, Baoguang Li, Shuang Zhu, Qiangqiang Yuan, and Huanfeng Shen. 2023. A fully automatic and high-accuracy surface water mapping framework on google earth engine using landsat time-series. *International Journal of Digital Earth*, 16(1):210–233.

Zhiling Zheng, Federico Florit, Brooke Jin, Haoyang Wu, Shih-Cheng Li, Kakasaheb Y Nandiwale, Chase A Salazar, Jason G Mustakis, William H Green, and Klavs F Jensen. 2025. Integrating machine learning and large language models to advance exploration of electrochemical reactions. *Angewandte Chemie*, 137(6):e202418074.

Guanghao Zhou, Wenhui Liu, Changwei Xie, Xianteng Song, Qi Zhang, Qingpeng Li, Guangyue Liu, Qing Li, and Bingnan Luo. 2024a. Accelerating thermokarst lake changes on the qinghai–tibetan plateau. *Scientific Reports*, 14(1):2985.

Yan Zhou, Bailu Liu, Yaoping Cui, Xinxin Wang, Mengmeng Cao, Sen Zhang, Xiangming Xiao, and Jinwei Dong. 2024b. Annual improved maps to understand the complete evolution of 9 thousand lakes on the tibetan plateau in 1991–2023. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217:134–148.

A More Dataset Description

Our dataset also comprises a large variety (~ 17) of satellite sensors. MODIS (Moderate Resolution Imaging Spectroradiometer) observations are most numerous due to its daily temporal resolution and complementary morning and afternoon observations from MODIS Terra and MODIS Aqua satellites. The second highest is Landsat (Land Remote-Sensing Satellite), which has provided historical coverage dating to 1972, making it valuable for decade-long comparisons and analyses, though its 16-day revisit time limits temporal resolution. VIIRS (Visible Infrared Imaging Radiometer Suite), launched in 2012, offers daily observations with specialized capabilities for nighttime light intensity measurements. Note that in Earth observation, we refer to the instrument as a sensor and its data products as products; for Google Earth Engine (GEE), these are organized as imagery collections.

Other important sensors, though less frequently mentioned in the posts, include: TRMM (Tropical Rainfall Measuring Mission) for precipitation monitoring; GRACE (Gravity Recovery and Climate Experiment) for gravity field measurements; TOMS (Total Ozone Mapping Spectrometer) for atmospheric ozone monitoring; SMAP (Soil Moisture Active Passive) for global soil moisture mapping; GLDAS (Global Land Data Assimilation System) for land surface modeling. These sensors, while appearing less frequently in our dataset, play crucial roles in long-term Earth observation and environmental monitoring.

B Introduction to Google Earth Engine

Google Earth Engine (GEE) (Gorelick et al., 2017) is a cloud-based platform that enables users to perform geospatial analysis at a planetary scale using Google's computational infrastructure. It houses over 90 petabytes of analysis-ready satellite imagery and more than 1,000 curated geospatial datasets spanning 50+ years of historical data, including imagery from satellites such as Landsat, MODIS, and Sentinel, as well as climate and weather datasets, geophysical data, terrain information, and land cover data. Researchers harness this technology for various Earth Observation and applications such as forest mapping (Chen et al., 2017), drought monitoring (Sazib et al., 2018), crop yield estimation (Jaafar and Mourad, 2021), land use and land cover (Nasiri et al., 2022), evapotranspiration (Senay et al., 2022), shoreline analysis (Santra et al., 2024), and water detection (Yue et al., 2023), etc.

Terminology. A **sensor** refers to a device that detects and measures physical properties (like reflectance, temperature, etc.), such as optical cameras, radar, and spectrometers mounted on satellites or aircraft. A **product** is a processed dataset derived from sensor data, typically preprocessed for calibration, quality control, and transformation into specific variables. In GEE specifically, an **imagery collection** is a set of related images grouped together for analysis.

API Usage. GEE provides a JavaScript and Python API that enables users to access and filter the extensive data catalog, apply algorithms for image processing and analysis, perform geospatial computations across multiple processors in parallel. In this paper, the AI agents generate the Python code and execute it. The GEE API script calls the GEE server for the computation. The results, mostly the final statistics, are sent back to the local agents for further deduction and answering.

C Error Analysis

Error Taxonomy (Gemini-2.5-Pro, Zero-Shot Baseline). Table 5 shows the breakdown of failure modes under the zero-shot baseline. This reveals that the dominant failure modes are knowledge-related, not code-related: wrong dataset IDs (14.2%), temporal/spatial mismatch (15.7%), and wrong band names (9.6%) together account for $\sim$40% of all failures, while only 10% are pure syntax errors.

After applying iterative re-planning (OSCAR (Wang and Liu, 2024), 3 rounds), the error profile (Table 6 shifts significantly. Re-planning substantially reduces knowledge-grounding errors (wrong dataset, wrong band, temporal/spatial mismatch) and execution errors. However, the proportion of wrong answers more than doubles from 9.6% to 21.3%: as more questions successfully execute, the remaining errors shift from pipeline failures to reasoning errors where the code runs correctly but produces an incorrect conclusion (*e.g.*, wrong threshold, incorrect comparison logic). This suggests that after resolving surface-level failures, the next frontier is improving the agent's domain reasoning and scientific judgment.

This process-level analysis provides actionable diagnostic insight: (1) *retrieval-augmented tool grounding*, where agents query a structured catalog at inference time rather than relying on memorized schemas, and (2) *improved domain reasoning* through better prompting strategies or domain-specific fine-tuning to address the growing proportion of reasoning errors.

Error Type	Count	Percentage(%)
Correct	150	36.8%
Wrong Dataset (hallucinated asset ID)	58	14.2%
Temporal/Spatial Mismatch (empty collection)	64	15.7%
Wrong Band (non-existent band name)	39	9.6%
Syntax/Runtime Error	41	10.0%
Wrong Answer (correct execution, wrong reasoning)	39	9.6%
Masking/Preprocessing (no valid pixels)	11	2.7%
Numerical Failure (NaN/None result)	4	1.0%

Table 5: Error taxonomy for Gemini-2.5-Pro under the zero-shot baseline.

Error Type	Baseline	After Re-planning	Δ
Correct	36.8%	55.9%	+19.1%
Wrong Dataset	14.2%	10.3%	−3.9%
Wrong Band	9.6%	1.5%	−8.1%
Temporal/Spatial Mismatch	15.7%	2.9%	−12.8%
Syntax/Runtime	10.0%	2.5%	−7.5%
Wrong Answer	9.6%	21.3%	+11.7%

Table 6: Error profile shift after iterative re-planning (OSCAR, 3 rounds).

D Reviewer instructions for UnivEARTH

Below, we detail the instructions given to the reviewers.

D.1 Goal

Given a question and an article about earth science, your task is to provide an answer.

D.2 Evaluation Questions

1. **What is the answer to this question?**
   Please answer (A) Yes, (B) No, or (C) I don't know, or data is not conclusive.

2. **Is the answer to the question being supported by the text from the article?**
   Please copy and paste the relevant texts that you use to derive your answer from the article.
   - *Strongly Supported*: The article explicitly states the answer or provides clear evidence.
   - *Moderately Supported*: The article implies the answer, but requires some inference.
   - *Not Supported*: The article contradicts the answer or provides no relevant information.

3. **Is the answer to the question being supported by the image from the article?**
   If yes, please explain how the image supports the answer to the question.
   - *Strongly Supported*: The article explicitly states the answer or provides clear evidence.
   - *Moderately Supported*: The article implies the answer, but requires some inference.
   - *Not Supported*: The article contradicts the answer or provides no relevant information.

4. **Do you need to use Google Maps to check location information?**
   If yes, please explain why using Google Maps is required. Please answer (A) Yes, (B) No.

Below are the three examples provided to the reviewers.

Example 1

Question: Does the Tuolumne River Basin have more snow on April 1, 2017 than on April 1, 2015?
URL: link
Q1: What is the answer to this question?
You should answer **(A) Yes**
Q2: Is the answer to the question being supported by the text from the article?
You should answer **Strongly Supported**
You should copy the text *"New NASA data show that snowpack in Tuolumne River Basin—a major source of water for San Francisco and California's Central Valley—is currently greater than that of the four previous years combined."* and paste it to the spreadsheet.
Q3: Is the answer to the question being supported by the image from the article?
You should answer **Strongly Supported**
You should explain the reason: *"The image shows greater snow water equivalent in April 1, 2017, compared to April 1, 2015"*
Q4: Do you need to use Google Maps to check location information?
You should answer **No**.
Q5: Other comments
You don't have to write anything.

Example 2

Question: Does Cape Lookout National Seashore show lower turbidity in the region centered at (34.659539, -76.464976) than the region centered at (34.607982, -76.338262) on February 18, 2016?
URL: link
Q1: What is the answer to this question?
You should answer **(B) No**
Q2: Is the answer to the question being supported by the text from the article?
You should answer **Not Supported**
You should write *"The text does not mention that"*.
Q3: Is the answer to the question being supported by the image from the article?
You should answer **Strongly Supported**
You should explain the reason: *"The image shows that (34.659539, -76.464976) had less turbidity than another region"*
Q4: Do you need to use Google Maps to check location information?
You should answer **Yes**
You should explain the reason: *"Neither the image nor the text shows the two geolocations."*
Q5: Other comments
You don't have to write anything.

Example 3

Question: Does Cape Lookout National Seashore show lower turbidity in the region centered at (34.659539, -76.464976) than the region centered at (34.607982, -76.338262) on February 18, 2017?
URL: link
Q1: What is the answer to this question?
You should answer **(C) I don't know, or data is not conclusive**
Q2: Is the answer to the question being supported by the text from the article?
You should answer **Not Supported**
You should write *"The text talks about events in 2016, not 2017"*.
Q3: Is the answer to the question being supported by the image from the article?
You should answer **Not Supported**
You should explain the reason: *"The time period is incorrect."*
Q4: Do you need to use Google Maps to check location information?
You should answer **Yes**
You should explain the reason: *"Neither the image nor the text shows the two geolocations."*
Q5: Other comments
You can write *"I think the question is wrong. Please take a look."*