Towards LLM Agents for Earth Observation

Anonymous Authors[1]

Abstract

Earth Observation (EO) provides critical planetary data for environmental monitoring, disaster management, climate science, and other scientific domains. Here we ask: *Are AI systems ready for reliable Earth Observation?* We introduce **UnivEARTH**, a benchmark of 140 yes/no questions from NASA Earth Observatory articles across 13 topics and 17 satellite sensors. Using Google Earth Engine API as a tool, LLM agents can only achieve an accuracy of 33% because the code fails to run over 58% of the time. Taken together, our findings identify significant challenges to be solved before AI agents can automate earth observation, and suggest paths forward.

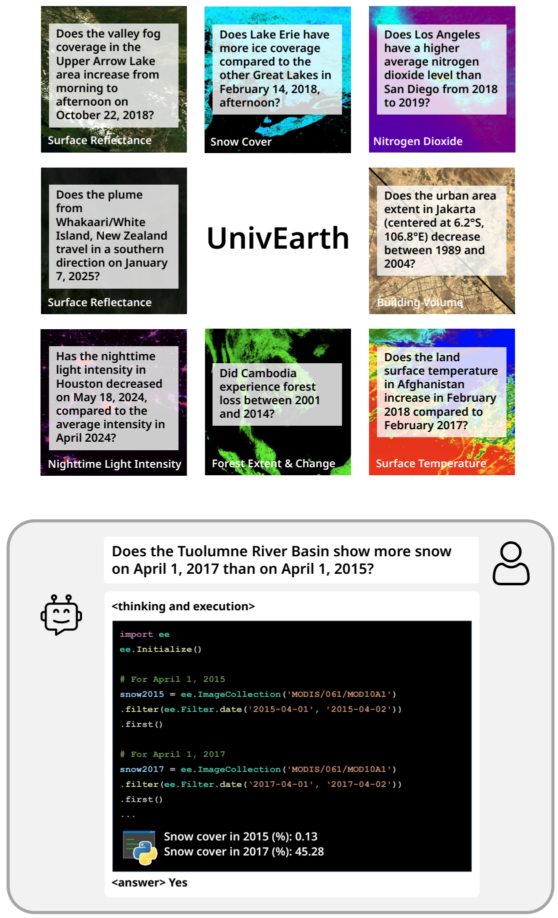

Figure 1. We propose **UnivEARTH** for benchmarking AI agents in Earth Observation.

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.

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. Given recent advances in LLM-based AI agents, we 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 arti-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

cles 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 resulting dataset, **UnivEARTH**, comprises 140 high-quality yes/no questions spanning 13 diverse topics and 17 different sensors and datasets.

We benchmark several off-the-shelf models, including Claude-3.7-Sonnet, DeepSeek-V3, DeepSeek-R1, and o3-mini. We include the option "Data is inconclusive" to allow models to abstain from answering the question if they deem the evidence insufficient. Note that simply answering the question is not enough: we need models to ground their answers in evidence. Thus, we ask models to generate code that uses the Google Earth Engine (Gorelick et al., 2017) Python API to answer the question. Unfortunately, current models fail to produce executable code 58% of the time. As such, in this grounded scenario, the best accuracy is a mere 33.0%. These results indicate that modern AI agents are not ready for reliable earth observation tasks.

In sum, our work makes three key contributions:

- We curate an novel evaluation benchmark of Earth Observation from authoritative sources, with verified answers and guaranteed question answerability.
- We benchmark state-of-the-art LLMs and reveal significant gaps in their ability to answer domain-specific questions and generate reliable analysis as evidence.

2. Benchmark Construction

2.1. 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.

2.2. Data Collection and Validation

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

Data Collection. We downloaded NASA Earth Observatory website articles with the cutoff time of March 10, 2025. Then, we used Claude-3.5-Sonnet to analyze these article texts and generated candidate yes/no question-answer pairs with supporting sentences. We chose the yes/no question format to facilitate evaluation, as assessing the correctness of free-form questions is challenging in scientific domains.

Our prompting strategy had two key components: (1) *Filtering*: We instructed the LLM to reject unsuitable articles, including those regarding sensor specifications, general introductions, or non-satellite imagery, transient observations (e.g., wind speed, tides); (2) *Format Standardization*: We prompted the LLM to focus on yes/no format with spatial and/or temporal comparison. We also conducted an initial editing pass of each question-answer pair to ensure location precision. Additionally, since we asked LLMs to process only text inputs and not images, we manually added new questions based on figures included in the articles. This step was crucial because many Earth Observatory articles convey significant information through the included imagery.

Question 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 Google Earth Engine are presented in Appendix-D. This was necessary because we found that some articles describe phenomena using sensors or products not available in GEE. [1] To filter out such questions, we wrote test implementations using the JavaScript code editor on the Google Earth Engine platform, verified dataset availability, and in some cases identified alternative data sources that could answer similar questions. Thus, all questions in our benchmark can be reasonably answered using the Google Earth Engine platform and available datasets.

Dataset 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-E for the reviewer instruction document.

We recruited four reviewers, with each reviewer evaluating

[1] As an example, a June 2016 article discussed global average carbon dioxide concentrations measured by the Orbiting Carbon Observatory-2 (OCO-2) from September 2014 to September 2015. The OCO-2 dataset is not available in GEE, making any question from this article impossible to answer using the GEE API.

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 exact match. Following this initial assessment, we iteratively revised ambiguous questions with each reviewer until we reached complete agreement on Q1.

2.3. Dataset Statistics

Table 2 in the Appendix A presents statistics of **UnivEARTH**. The dataset composition reflects both article availability and sensor characteristics. Topic statistics are based on the (potentially multiple) tags provided with each article. The dominant topics are *land*, *water*, *human presence* and *atmosphere*, representative of the typical use-cases of Earth Observation data. Example questions and supporting sentences are in Table 3 in the Appendix A.

2.4. Relevance to Science and Real-World Impact

UnivEARTH captures phenomena with significant real-world relevance and active scientific interest. For instance, one 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.

3. Benchmarking SoTA Agents with UnivEARTH

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

Experimental Setup. For each question, we provided the LLM with three options: “Yes”, “No”, or “Inconclusive”. The third option is to allow the LLM to abstain when it is not sure (we evaluate models without this third option in Appendix-B.1). Our primary metric was **Accuracy**: the fraction of questions where the generated answer matched the ground truth. We also measure the **Rejection Rate**, i.e., the proportion of times the model abstained, and the **Selective Accuracy**, which is the Accuracy restricted to cases where the model did not abstain. We benchmark LLM agents, including ChatGPT-4o-mini (Hurst et al., 2024), ChatGPT-o3-mini, Claude-3.5-Haiku, Claude-3.5-Sonnet, Claude-3.7-Sonnet (Anthropic, 2024), DeepSeek-V3 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025), Qwen2.5-72B-Instruct (Yang et al., 2024a), Qwen2.5-Coder-32B-Instruct (Hui et al., 2024) , and Llama-3.3-70B-Instruct (Grattafiori et al., 2024).

3.1. Answering Questions With Google Earth Engine

For question-answering with Google Earth Engine access, we evaluated three frameworks: zero-shot, few-shot, and reflexion-based approaches. In the zero-shot approach, LLMs were instructed to first reason through the problem and then generate appropriate code. For few-shot learning, we provided LLM agents with three question-code examples in a multi-turn conversation format. These examples were drawn from outside the benchmark dataset to prevent contamination. In the reflexion framework (Shinn et al., 2023), we implemented a 3-round reflection process where each round’s question, code, execution results, and errors (if any) were fed back to the LLM agent for reflection, which informed the next round of code generation. After obtaining code scripts, we ran them locally to determine answers, which are parsed by GPT-4o-mini to derive the answers.

Since the models now have access to data, abstention doesn’t make sense. However, there may still be scenarios where the LLM fails to produce an answer: either the code was incorrect, or the data requested by the code was not available (because of sensor availability, revisit frequency, etc.). We therefore replace abstention with failure, which captures both these scenarios.

All statistics in Table 1 represent averages across 8 trials. The best overall accuracy was only $\sim$33%. The reason for this low accuracy was that for all LLMs, for the majority of trials, the code failed to produce an answer (either failed to run or accessed unavailable data). Even when the code did run, it occasionally gave an incorrect answer ($\sim$ 20% of the time for the best models). This low accuracy suggests that existing LLMs are not capable yet of producing code for answering EO questions. One possible reason is that this domain of coding questions is less well represented in the pre-training data. **UnivEARTH** can thus serve as a practically relevant out-of-domain evaluation for future research into overcoming these limitations. Also, in Table 4 in the Appendix B we showed the results averaged over three trials without Internet Access.

3.2. The Impact of Data Utilization

One of the key challenges with using Google Earth Engine is the need to choose from over 400 imagery collections. We therefore hypothesized that model’s ability to correctly make this choice may be an important factor in their performance.

To test this hypothesis, we looked at the number of unique

Table 1. Comparison of Language Models With Access to Google Earth Engine. The table shows performance across three different metrics: accuracy (%), failure rate (%), and selective accuracy (%) for zero-shot (ZS), 3-shot (TS), and code reflexion (Rfx) scenarios. For Accuracy, values greater than 25% and 30% are shown in **bold** and grayed, respectively.

	Google Earth Engine Access								
	Accuracy (%)			Failure Rate (%)			Selective Accuracy (%)		
Model	ZS	TS	Rfx	ZS	TS	Rfx	ZS	TS	Rfx
4o-mini	8.3	13.1	5.8	89.1	83.2	90.3	68.5	72.2	54.5
o3-mini	**25.7**	**33.0**	**25.1**	70.0	60.7	69.7	81.0	78.9	81.5
Claude-3.5-Haiku	14.9	12.2	13.0	80.6	81.3	81.5	70.2	60.5	66.1
Claude-3.5-Sonnet	**27.0**	23.9	**27.8**	67.5	70.1	66.5	80.8	75.5	79.3
Claude-3.7-Sonnet	**32.4**	**30.6**	**33.0**	61.3	61.2	59.8	81.6	79.3	81.3
DeepSeek-V3	**28.4**	**32.8**	24.3	64.3	58.1	68.9	73.7	79.1	72.4
DeepSeek-R1	19.6	22.8	15.3	75.4	70.8	80.1	77.9	73.8	76.4
Qwen2.5-72B-Instruct	18.7	22.5	15.3	73.9	66.1	77.8	68.3	63.2	63.7
Qwen2.5-Coder-32B-Instruct	10.6	18.4	8.1	83.4	69.8	88.0	65.5	56.4	60.7
Llama-3.3-70B-Instruct	2.8	6.5	2.6	96.7	91.9	96.4	81.7	74.7	65.6

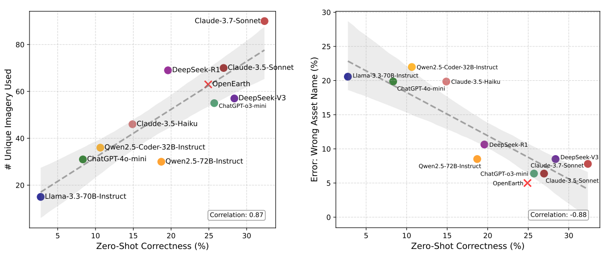

Figure 2. Zero-shot accuracy is correlated with number of unique imagery collections used (left) and negatively with the fraction of times the "Wrong Asset Name" error is encountered (right).

imagery collections accessed by different LLM agents and whether higher-performing models leverage a more diverse range of data sources. As shown in Figure 2 (left), our analysis indeed reveals a strong correlation (r = 0.87) between zero-shot accuracy and number of unique imagery collections queried. This suggests that superior models excel in recalling and applying a wider range of effective imagery collection names. Interestingly, Qwen2.5-72B-Instruct appears as an outlier, achieving nearly 20% accuracy while utilizing relatively few imagery collections, perhaps because it is more effective at using the collections it does access.

Why do models struggle with using the many Earth Engine collections? We found that the underlying reason was their failure to recall the correct name of the collection. This is shown in Figure 2 (right): we observed a strong negative correlation between model performance and the *wrong assert name* error mode. This observation suggests that our synthetic dataset which filters out incorrect code might help improve the model's ability to recall correct names.

4. Conclusion

Earth observation is critical for earth science, yet automating these complex workflows remains challenging. Our evaluation of state-of-the-art LLM agents through **UnivEARTH** reveals significant limitations in their ability to perform scientific Earth Observation tasks reliably, with even the best models achieving only 49.0% accuracy without internet access. When asked to produce evidence in the form of GEE code, accuracy further drops to 33.0% because of models' inability to correctly navigate the many data sources. This shows that current AI systems fall short of reliably facilitating earth science applications. Nevertheless, our work demonstrates a promising path forward: through fine-tuning a smaller open-source model on specialized synthetic data, we achieved 25.0% accuracy with Llama-3.1-8B – comparable to larger commercial models at a much smaller computational cost. Our analysis reveals that performance correlates strongly with knowledge of diverse Earth observation data sources ($r = 0.87$), suggesting that domain expertise remains crucial for these specialized scientific tasks.

Limitations. We acknowledge some limitations of **UnivEARTH** . First, it comprises few (140) questions, similar to prior work such as HumanEval (Chen et al., 2021). A future version could benefit from an expanded question set. Second, the current benchmark does not include unanswerable questions (Rajpurkar et al., 2018; Kim et al., 2022) where the ground truth answer is "inconclusive".

Impact Statement This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here

References

Alampara, N., Schilling-Wilhelmi, M., Ríos-García, M., Mandal, I., Khetarpal, P., Grover, H. S., Krishnan, N., and Jablonka, K. M. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv:2411.16955*, 2024.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf, 2024.

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

Campbell, Q., Cox, S., Medina, J., Watterson, B., and White, A. D. Mdcrow: Automating molecular dynamics workflows with large language models. *arXiv preprint arXiv:2502.09565*, 2025.

Chen, B., Xiao, X., Li, X., Pan, L., Doughty, R., Ma, J., Dong, J., Qin, Y., Zhao, B., Wu, Z., et al. 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, 2017.

Chen, C., Park, T., Wang, X., Piao, S., Xu, B., Chaturvedi, R. K., Fuchs, R., Brovkin, V., Ciais, P., Fensholt, R., et al. China and india lead in greening of the world through land-use management. *Nature sustainability*, 2(2):122–129, 2019.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chen, T., Anumasa, S., Lin, B., Shah, V., Goyal, A., and Liu, D. Auto-bench: An automated benchmark for scientific discovery in llms. *arXiv preprint arXiv:2502.15224*, 2025a.

Chen, X., Wang, T., Guo, T., Guo, K., Zhou, J., Li, H., Song, Z., Gao, X., and Zhang, X. Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 8(1):4, 2025b.

Danish, M. S., Munir, M. A., Shah, S. R. A., Kuckreja, K., Khan, F. S., Fraccaro, P., Lacoste, A., and Khan, S. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

Feng, K., Ding, K., Wang, W., Zhuang, X., Wang, Z., Qin, M., Zhao, Y., Yao, J., Zhang, Q., and Chen, H. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Fossi, G., Boulaimen, Y., Outemzabet, L., Jeanray, N., Gerart, S., Vachenc, S., Giemza, J., and Raieli, S. Swiftdossier: Tailored automatic dossier for drug discovery with llms and agents. *arXiv preprint arXiv:2409.15817*, 2024.

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

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jaafar, H. and Mourad, R. 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, 2021.

Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Kim, N., Htut, P. M., Bowman, S. R., and Petty, J. (qa)2: Question answering with questionable assumptions. *arXiv preprint arXiv:2212.10003*, 2022.

Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., Ponnapati, M., White, A. D., and Rodriques, S. G. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Lei, F., Chen, J., Ye, Y., Cao, R., Shin, D., Su, H., Suo, Z., Gao, H., Hu, W., Yin, P., et al. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *arXiv preprint arXiv:2411.07763*, 2024.

Lei, Y., Yao, T., Sheng, Y., Yang, K., Yang, W., Li, S., Zhou, J., Jiang, Y., and Yu, Y. Unprecedented lake expansion in 2017–2018 on the tibetan plateau: Processes and environmental impacts. *Journal of Hydrology*, 619:129333, 2023.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., et al. 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, 2023.

Li, K., Dong, F., Wang, D., Li, S., Wang, Q., Gao, X., and Chua, T.-S. Show me what and where has changed? question answering and grounding for remote sensing change detection. *arXiv preprint arXiv:2410.23828*, 2024a.

Li, M., Weng, B., Yan, D., Bi, W., and Wang, H. 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, 2022.

Li, Z., Yang, X., Choi, K., Zhu, W., Hsieh, R., Kim, H., Lim, J. H., Ji, S., Lee, B., Yan, X., et al. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*, 2024b.

Liang, Z., Guo, K., Liu, G., Guo, T., Zhou, Y., Yang, T., Jiao, J., Pi, R., Zhang, J., and Zhang, X. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6 (5):525–535, 2024.

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

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

Nasiri, V., Deljouei, A., Moradi, F., Sadeghi, S. M. M., and Borz, S. A. 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, 2022.

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nature Food*, 3(1):19–28, 2022.

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

Richey, A. S., Thomas, B. F., Lo, M.-H., Reager, J. T., Famiglietti, J. S., Voss, K., Swenson, S., and Rodell, M. Quantifying renewable groundwater stress with grace. *Water resources research*, 51(7):5217–5238, 2015.

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

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

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

Senay, G. B., Friedrichs, M., Morton, C., Parrish, G. E., Schauer, M., Khand, K., Kagone, S., Boiko, O., and Huntington, J. 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, 2022.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

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

Strieth-Kalthoff, F., Hao, H., Rathore, V., Derasp, J., Gaudin, T., Angello, N. H., Seifrid, M., Trushina, E., Guy, M., Liu, J., et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384(6697): eadk9227, 2024.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pp. 2024–11, 2024.

Wang, J., Zheng, Z., Chen, Z., Ma, A., and Zhong, Y. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024a.

Wang, L., Hu, Y., He, J., Xu, X., Liu, N., Liu, H., and Shen, H. T. 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*, pp. 19162–19170, 2024b.

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

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A. R., Zhang, S., Sun, Y., and Wang, W. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

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

Wu, H., Gu, G., Yan, Y., Gao, Z., and Adler, R. F. Global flood monitoring using satellite precipitation and hydrological modeling. *Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting*, pp. 87–113, 2018.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Yang, J., Jimenez, C. E., Zhang, A. L., Lieret, K., Yang, J., Wu, X., Press, O., Muennighoff, N., Synnaeve, G., Narasimhan, K. R., et al. Swe-bench multimodal: Do ai systems generalize to visual software domains? *arXiv preprint arXiv:2410.03859*, 2024b.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

Yue, L., Li, B., Zhu, S., Yuan, Q., and Shen, H. 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, 2023.

Zheng, Z., Florit, F., Jin, B., Wu, H., Li, S.-C., Nandiwale, K. Y., Salazar, C. A., Mustakis, J. G., Green, W. H., and Jensen, K. F. Integrating machine learning and large language models to advance exploration of electrochemical reactions. *Angewandte Chemie*, 137(6):e202418074, 2025.

Zhou, G., Liu, W., Xie, C., Song, X., Zhang, Q., Li, Q., Liu, G., Li, Q., and Luo, B. Accelerating thermokarst lake changes on the qinghai–tibetan plateau. *Scientific Reports*, 14(1):2985, 2024a.

Zhou, Y., Liu, B., Cui, Y., Wang, X., Cao, M., Zhang, S., Xiao, X., and Dong, J. 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, 2024b.

A. More Dataset Description

Table 2. Topics and Product Distribution in **UnivEARTH** . The left table shows the distribution of topics and their associated key words, while the right table presents the distribution of satellite products following Google Earth Engine nomenclature.

Topic	Count	Key Words
Land	81	pools, disappearing lakes, seasonal greening, hottest spots, vegetation
Water	47	lake, groundwater, evapotranspiration, chlorophyll, sediment
Human presence	44	cropland, nighttime light, nitrogen dioxide, urban expansion, farms
Atmosphere	29	cloud, aerosols, fog, carbon monoxide, ozone
Heat	20	urban heat island, heat Wave, sea surface temperatures
Life	15	flower, deforestation, urban growth
Floods	12	rainfall, flood
Severe storms	10	floodwaters, rainfall
Snow/ice	10	frozen lake, ice cover, winter snow
Fires	7	fire, burn scar, fire season
Drought	6	rainfall anomaly, soil moisture anomaly, worst drought
Volcanoes	6	lava flows, plume
Water color	5	color changes, phytoplankton, bloom

Product	Count
MODIS	53
Landsat	43
VIIRS	13
Sentinel-5	5
GRACE	4
TRMM	3
SMAP	3
Aura	2
CHIRPS	2
Combined	2
GPM	2
SeaWiFS	2
EO-1	1
GHSL	1
Hansen	1
Sentinel-1	1
Sentinel-6	1
TEMPO	1

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. Benchmarking SoTA Agents Against UnivEARTH

B.1. Baseline Without Internet Access

In Table 4 we showed the results averaged over three trials. We found that even the best model (i.e., in this case DeepSeek-R1) cannot answer more than half of the questions correctly and that the majority of them are worse than random (33.3%). Intriguingly, all models use the abstention option well: the accuracy on questions they do not abstain on is frequently more than 70%. However, unfortunately, all LLMs abstained on the majority of questions (e.g., Claude-3.7-Sonnet abstained on 82% of questions). The one exception to this was DeepSeek-R1: It abstained the least and achieved the highest accuracy, but still did not answer about half the questions.

B.2. Baseline With Web Search Access

To further assess benchmark difficulty, we examined performance when the LLM agent had access to internet-based information sources but not Google Earth Engine. This condition helps establish whether task performance remains challenging even with access to general online resources. During web search, the LLMs are prompted to generate web

Table 3. 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. [2]
Life	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. [3]
Human Presence	Does Houston show higher nighttime light intensity in December 2012 and 2013 compared to the average light output during the non-December months from 2012 to 2014?	The map compares the nighttime light signals from December 2012 and 2013 to the average light output for the rest of 2012 to 2014. Green shading marks areas where light usage increased in December. [4]
Snow / ice	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. [5]
Drought	Does Somalia show higher soil moisture in April 2019 compared to the average April conditions?	This map shows soil moisture anomalies in April 2019—an expression of drought and how it affects conditions for growing crops. Areas in green had more moisture in the upper layers of soil than the norm for April, while areas in red had less. In Somalia, rainfall was spotty, with just a few measuring stations in the north recording significant accumulations in April. [6]

search queries, and these queries are used to search online, with the results returned to the LLM for question answering. We implemented this using the Serper API[7] to retrieve snippets from the top ten Google Search results for each query. To avoid information leakage, we exclude webpages with domains including "nasa.gov" and "earth.org" where the same articles might be found. With search results available, six out of nine models performed better in both the two-option scenario and three-option scenario in terms of correctness; selective correctness also increased across most models, except for DeepSeek-R1. However, note that in a scientific application, grounding answers in uncurated webpages may not be sufficiently rigorous.

C. More results with OpenEARTH

C.1. Performance

Table 6 presents the performance of our fine-tuned model across different training checkpoints. The results demonstrate significant improvement, with overall accuracy reaching 25.0% after completing the training process. Even after just one epoch of training on our synthetic dataset, the model achieves 17.6% accuracy—a great improvement over the 2.8% accuracy of the much larger Llama-3.3-70B-Instruct model in zero-shot settings (Table 1). This substantial gain demonstrates the effectiveness of domain-specific fine-tuning with synthetic data, despite the absence of human verification for every example.

With additional training epochs, the model's performance continues to improve steadily. Interestingly, this improvement primarily stems from a decreased failure rate rather than increased selective accuracy. This indicates that the model becomes more proficient at executing code successfully and handling time availability scenarios – key capabilities for practical Earth observation applications.

[7] https://serper.dev

Table 4. Comparison of Language Models without Internet Access. **2 OPT**: binary choice task (accuracy %); **3 OPT**: three-option task with rejection option (accuracy %, rejection %, and selective accuracy %).

	No Internet			
	2 OPT	3 OPT		
Model	Accuracy	Accuracy	Rejection	Selective Accuracy
ChatGPT				
GPT-4o-mini	50.48	1.67	97.62	70.00
O3-mini	72.14	18.10	78.33	83.52
Claude				
Claude-3.5-Haiku	71.43	23.10	69.76	76.38
Claude-3.5-Sonnet	83.81	41.90	54.29	91.67
Claude-3.7-Sonnet	81.43	17.14	82.38	97.30
DeepSeek				
DeepSeek-V3	69.76	17.38	80.95	91.25
DeepSeek-R1	75.71	49.05	44.52	88.41
Qwen				
Qwen2.5-72B-Instruct	60.24	2.84	97.16	100.00
Qwen2.5-Coder-32B-Instruct	51.43	0.71	99.29	100.00
Llama				
Llama-3.3-70B-Instruct	67.86	27.86	66.19	82.39

D. 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 & 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.

Table 5. Comparison of Language Models with Internet Access. **2 OPT**: binary choice task (accuracy %); **3 OPT**: ternary choice task with rejection option (accuracy %, rejection %, and selective accuracy %).

	Web Search			
	2 OPT	3 OPT		
Model	accuracy	accuracy	Rejection	Selective Accuracy
ChatGPT				
GPT-4o-mini	60.71	18.57	80.00	92.86
O3-mini	75.71	22.14	77.14	96.88
Claude				
Claude-3.5-Haiku	65.71	20.00	75.71	82.35
Claude-3.5-Sonnet	85.00	32.86	65.71	95.83
Claude-3.7-Sonnet	78.57	27.14	72.86	100.00
DeepSeek				
DeepSeek-V3	74.29	25.00	72.86	92.11
DeepSeek-R1	73.57	25.00	71.43	87.50
Qwen				
Qwen2.5-72B-Instruct	75.71	33.57	64.29	94.00
Qwen2.5-Coder-32B-Instruct	60.71	7.86	92.14	100.00
Llama				
Llama-3.3-70B-Instruct	75.71	30.71	65.71	89.58

Table 6. Performance of Our Trained Model Across Different Checkpoints. The table shows zero-shot performance with accuracy (%), failure rate (%), and selective accuracy (%).

Checkpoint	Accuracy	Failure	Selective Accuracy
1	17.63	72.63	57.74
2	19.24	68.04	54.30
3	23.26	64.78	58.87
4	25.04	61.07	57.61

E. Reviewer instructions for UnivEARTH

Below, we detail the instructions given to the reviewers.

E.1. Goal

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

E.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.
5. **Other comments**

E.3. Examples 1

Example 1

Question: Does the Tuolumne River Basin have more snow on April 1, 2017 than on April 1, 2015?
URL: https://earthobservatory.nasa.gov/images/90073/sierra-snowpack-bigger-than-last-four-years-combined

1. **What is the answer to this question?**
   You should answer **(A) Yes**
2. **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.
3. **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"*
4. **Do you need to use Google Maps to check location information?**
   You should answer **No**.
5. **Other comments**
   You don't have to write anything.

E.4. Example 2

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: `https://earthobservatory.nasa.gov/images/87627/on-the-lookout`

1. **What is the answer to this question?**
   You should answer **(B) No**
2. **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"*.
3. **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"*
4. **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."*
5. **Other comments**
   You don't have to write anything.

E.5. Example 3

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: https://earthobservatory.nasa.gov/images/87627/on-the-lookout

1. **What is the answer to this question?**
   You should answer **(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?**
   You should answer **Not Supported**
   You should write *"The text talks about events in 2016, not 2017"*.
3. **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."*
4. **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."*
5. **Other comments**
   You can write *"I think the question is wrong. Please take a look."*

F. 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 empirical 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., 2024b). 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. **UnivEARTH** extends this paradigm to the Earth observation domain for accessing and analyzing satellite data.