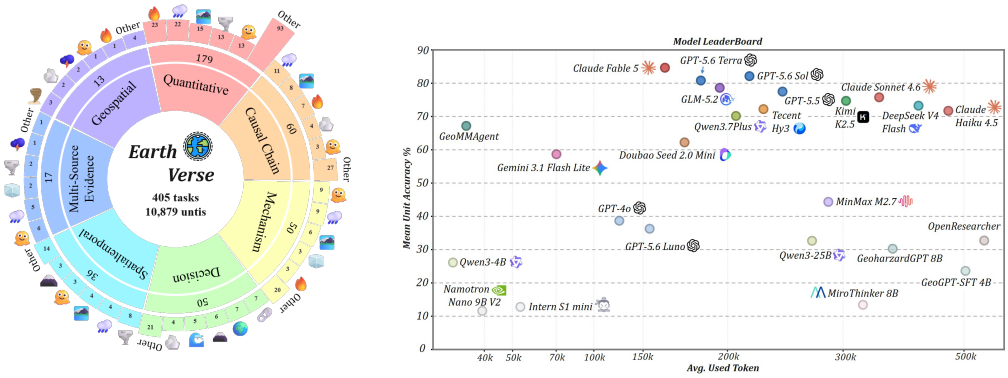
EARTHVERSE: BENCHMARKING SCIENTIFIC AGENTS ACROSS DYNAMIC EARTH SYSTEMS AND NATURAL HAZARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth-system analysis reconstructs changing physical processes from observations that differ in source, scale, timing, and modality. Natural hazards make this work consequential because incomplete evidence can change estimates of severity, exposure, and mechanism. We introduce **EarthVerse**, a benchmark that evaluates scientific agents through package-scoped investigations. Its 405 reproducible tasks are grounded in 199 documented events and 19 hazard families. Agents inspect heterogeneous event packages, choose compatible evidence, execute transparent calculations, reconcile source differences, and preserve provenance in the final answer. We provide executable ground truth that decomposes each task into fine-grained answer units, together with task-specific rubrics that assess the supporting research process while allowing multiple valid paths. We evaluate 25 model and agent systems under a controlled tool-using protocol, then use controlled studies to locate failures in evidence access, tool selection, memory, reasoning, interaction, and scientific execution. Across systems, the best mean answer-unit accuracy is 84.65%, while the highest Strict@95 is only 34.81%. The gap shows that current agents often complete individual steps without maintaining a consistent chain across evidence, scales, units, calculations, and physical interpretation. EarthVerse provides a reproducible basis for measuring end-to-end scientific reliability in dynamic Earth systems. Our code is available at code.



(a) Scientific coverage. **(b)** Leaderboard landscape.

Figure 1: **Coverage and model landscape.** Task and answer-unit coverage across hazards and capabilities (left); mean answer-unit accuracy and token use across 25 systems (right).

1 INTRODUCTION

Earth-system science asks how the atmosphere, land, ocean, ecosystems, and human activity shape one another. These processes govern water, food, health, infrastructure, and climate risk. Natural hazards make the question immediate: weather, climate, and water extremes have caused

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

extensive mortality and economic loss, while compound events can propagate across sectors and regions (World Meteorological Organization, 2023; 2025; Intergovernmental Panel on Climate Change (IPCC), 2022; Zscheischler et al., 2018). Timely analysis matters, but so does scientific traceability. A plausible explanation is not enough when it may inform disaster assessment or response.

Natural-hazard analysis applies Earth-system reasoning to decisions about event severity, exposure, physical mechanism, and response. It requires researchers to reconstruct an evolving process from observations collected for different purposes and at different scales.

The difficulty begins with observation. No single instrument records an evolving Earth system in full. Agency reports, stations, satellites, reanalyses, maps, and impact databases measure different variables over different footprints and time windows. Combining them requires scientific judgment before calculation: the analyst must determine which records describe the same process, which scale is appropriate, and what uncertainty remains. Work on machine learning for Earth systems and model–data integration has made these records easier to use, but has not removed the need to reconcile them (Reichstein et al., 2019; Markonis et al., 2021; Gettelman et al., 2022; Cui et al., 2026).

Foundation models and tool-using agents now make a broader form of analysis possible. Systems such as Google Earth AI combine imagery, population, and environmental models through a reasoning agent, while scientific agents can inspect files, execute code, and revise a result after new observations arrive (Bell et al., 2025; Yao et al., 2023; Chen et al., 2025; Hu et al., 2025; Talemi et al., 2026). This changes the evaluation target. A system has to decide what evidence to seek, bind each value to the correct source and scale, and let tool feedback alter its account when the evidence disagrees.

Most Earth-science benchmarks begin after that decision has been made. They test knowledge, figure interpretation, remote sensing, or geospatial reasoning over a supplied passage, image, or named asset (Zhang et al., 2025; Xu et al., 2026; Zhao et al., 2026). Newer multimodal and agent benchmarks broaden the interface, but generally continue to declare the observation or data layer in advance (Xiao et al., 2026; Li et al., 2025). Frontier systems already approach or exceed published expert references on some of these formats. Such results do not show whether the same system can construct a compatible evidence base for an open question, carry calculations across several sources, and preserve support for every part of its conclusion.

EarthVerse contains 405 reproducible investigations grounded in 199 real disasters and extreme events. Each task opens an event collection assembled from 80 external source and access families. A typical event contains about 34 files, and solving a task requires evidence from more than seven distinct sources on average. Every file is connected to the event, so source choice remains a scientific judgment. The difficulty lies in deciding which records answer the question, whether their time windows and spatial support are compatible, and how much weight each source should carry. Working through these materials takes sustained interaction. Across the leaderboard, each task takes about 17 model rounds on average. Complete evaluations of some higher-cost frontier systems exceed $2,500.

Figure 2 shows the global event collection and representative investigations.

EarthVerse uses transparent mathematics: aggregation, ratios, thresholds, spatial overlap, weighted indices, and bounded counterfactuals. The difficulty lies in combining these operations with compatible evidence. A wrong time window changes the total, which can alter both the comparison and the physical explanation. Tasks therefore combine process reconstruction, quantitative analysis, competing hypotheses, rankings, and open synthesis. Fine-grained answer units expose each obligation, while Strict@95 measures whether the resulting claim–evidence state is nearly complete.

A heat-wave investigation illustrates this distinction. For the 2021 Pacific Northwest event, a solver must distinguish a direct hourly record from a lower-resolution aggregate, identify the matching heat window, and compute heat and humidity statistics over the same temporal support. Retrieving a plausible file is only the first step. The choice of record determines what can be calculated, and that calculation must remain attached to the evidence used to explain the event.

Answer units make these dependencies visible in evaluation. One investigation can require a time window, several quantities, a comparison of sources, and a physical mechanism. A single fluent sum-

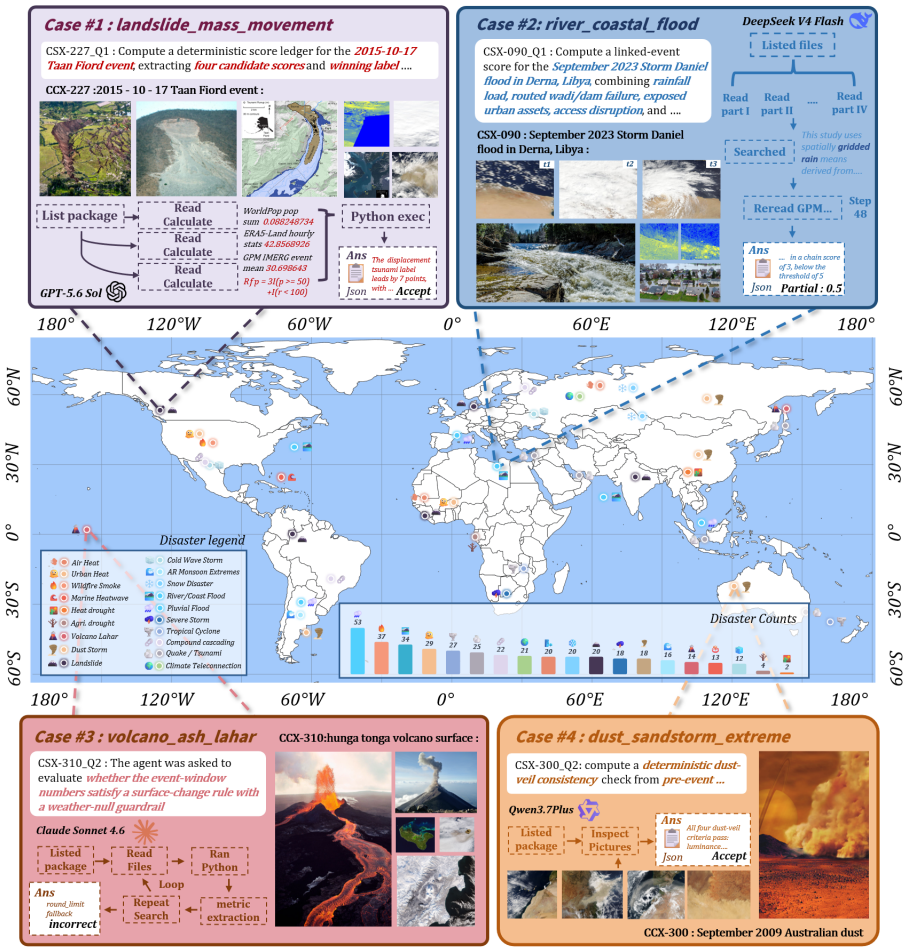


Figure 2: **Benchmark overview.** Global event coverage and four representative investigations. Each case follows a different model trajectory, and performance is scored against the answer units defined for that task. In Case 2, the linked-event score is a deterministic composite derived from package evidence.

mary can conceal an omission in any of these parts. EarthVerse therefore reports partial completion and near-complete task reliability separately, while the process rubric checks whether the recorded work supports the answer.

The results show how demanding this composition is. Across 25 systems, the best mean answer-unit accuracy is 84.65%, whereas the highest Strict@95 is only 34.81%. Even the best system on Strict@95 falls short of near-complete coverage on roughly two thirds of investigations. This is the benchmark's central reliability gap: an agent may recover most local facts and calculations while one unsupported source, mismatched scale, or missing mechanism breaks the account as a whole.

Controlled studies separate evidence-access failures from errors that persist after relevant observations are available. They test reasoning effort, evidence guidance, and final review. External benchmarks provide a complementary test: improvements with supplied observations do not establish that a system can find and reconcile evidence for a new investigation.

Contributions. We contribute **(i)** EarthVerse, a benchmark of reproducible multi-source investigations with executable answer units and task-specific process rubrics; **(ii)** a controlled evaluation of 25 model and agent systems that separates average scientific competence from reliable completion; and **(iii)** controlled experiments across evidence localization, tool selection, memory, reasoning, interaction modes, and professional scientific environments that identify the main failure points along the research chain.

We measure how often agents complete an investigation, then use controlled interventions to separate evidence-selection failures from reasoning errors. Comparisons on supplied-observation benchmarks test how much difficulty arises from constructing the evidence base.

2 Related Work

Earth-science benchmarks often start from a supplied passage, figure, image, or named data layer. EarthSE, MSEarth, and GeoMMBench cover scientific question answering and geospatial reasoning (Xu et al., 2026; Zhao et al., 2026; Xiao et al., 2026); CrisisMMD and xBD provide disaster observations to interpret (Alam et al., 2018; Gupta et al., 2019). DORA uses expert-authored operational tasks and replayable tools with declared multimodal assets (Wang et al., 2026). Prithvi-EO-2.0 and TerraMind learn representations across sensors and tasks, while GEO-Bench-2 tests transfer between sensing regimes (Szwarcman et al., 2026; Jakubik et al., 2025; Simumba et al., 2025). EarthVerse also supplies an evidence package, but leaves agents to identify the relevant records and reconcile them into an event-level account.

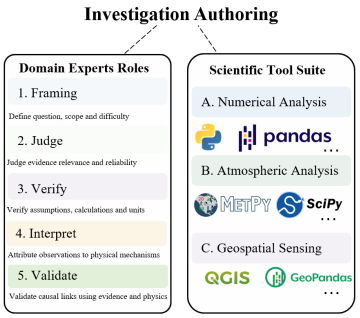


Figure 3: Expert roles and scientific tools for investigation authoring.

Tool-using agents extend evaluation to retrieval, computation, and revision. ReAct couples reasoning with actions, while AgentBench, GAIA, and OSWorld evaluate multi-step interaction in executable environments (Yao et al., 2023; Liu et al., 2024; Mialon et al., 2024; Xie et al., 2024). ScienceAgentBench, SciAgent, and SciAgentGym examine scientific tasks, data, and functions (Chen et al., 2025; Ma et al., 2024; Shen et al., 2026). BLADE and DiscoveryBench account for multiple valid analysis paths and evaluable facets of discovery (Gu et al., 2024; Majumder et al., 2025). These settings motivate assessing the research process alongside the endpoint. EarthVerse applies this distinction to investigations in which agents must select observations from a heterogeneous package before they can support a scientific claim.

Geospatial and deep-research systems address different parts of this workflow. ThinkGeo tests tool execution over remote-sensing imagery; Earth-Agent, OpenEarthAgent, Spatial-Agent, and GeoMMAgent coordinate geospatial operations (Shabbir et al., 2025; Feng et al., 2026; Shabbir et al., 2026; Bao et al., 2026; Xiao et al., 2026). DeepResearcher, OpenResearcher, and MiroThinker emphasize iterative retrieval and long research trajectories (Zheng et al., 2025; Li et al., 2026; MiroMind Team et al., 2025; Yu et al., 2026). Intern-S1, GeoGPT, and GeohazardGPT develop scientific terminology and process knowledge through pretraining (Bai et al., 2025; Zhang et al., 2024; Ge et al., 2025). EarthVerse evaluates how these capabilities work together: answer units check the findings, and process rubrics check their evidential and computational support without prescribing a tool sequence. Appendix H reviews the broader literature and benchmark distinctions.

3 The EarthVerse Benchmark

EarthVerse contains 405 investigations from 199 documented disasters and extreme events, covering 19 hazard families and 6,709 local files. Event packages preserve heterogeneous observations without adding unrelated files merely to obstruct retrieval. Computation uses checkable operations such as aggregation, normalization, thresholds, spatial overlap, and bounded counterfactuals.

3.1 Design principles and investigative practice

Each investigation begins with an event archive. The prompt states a scientific decision but does not disclose the relevant files, paths, or answer-identifying quantities. A task requires 7.26 distinct sources on average from a package with a median of 34 candidates. Agents must compare records with different scales, time windows, variables, and units, then revise their event account when new evidence conflicts with it.

Experts frame and verify the investigations with scientific tools (Figure 3). Answer units specify what must be established; process rubrics check source choice, calculation, revision, and causal reasoning. Alternative report-first or time-series-first paths can receive full credit if their evidence supports the required quantities and mechanisms. Versioned packages make this process reproducible without live feeds, hidden specialist knowledge, or a prescribed tool order.

3.2 Task construction and answer contracts

Each event contributes two tasks, with a third for seven flagship events. The public prompt states a scientific decision but does not identify the relevant files. Every task includes an expert solution, structured ground truth, a deterministic `compute_gt.py` program, and a task-specific process rubric.

The task and evaluation flow is

$$(q_i, \mathcal{D}_i) \xrightarrow{\pi} \tau_i = (a_1, o_1, \ldots, a_T, o_T) \longrightarrow \hat{y}_i,$$
$$E_i = \left[f(\hat{y}_i, y_i^\star), g(\tau_i; s_i, B_i) \right]. \tag{1}$$

Here q_i is the question, $\mathcal{D}_i$ the evidence collection, and π the policy producing trajectory τ_i and answer $\hat{y}_i$. Function f compares the answer with ground truth $y_i^\star$; g evaluates the trajectory against expert analysis s_i and rubric B_i. The 10,879 answer units cover numerical results, orderings, mechanisms, and qualified conclusions, allowing partially correct investigations to be distinguished from nearly complete ones.

A worked heat-wave example appears in Appendix F.

3.3 Capabilities and research environment

Each task has one to three manually reviewed labels from seven capabilities: physical-mechanism reasoning, spatiotemporal reconstruction, quantitative calculation, multi-source synthesis, causal-chain reasoning, ranking or decision, and remote-sensing or geospatial interpretation. Labels can overlap; capability-conditioned scores are diagnostic and do not alter the overall metric. Figure 1(a) shows their coverage, and Figure 4 shows how task metadata connects to observations and the research trajectory.

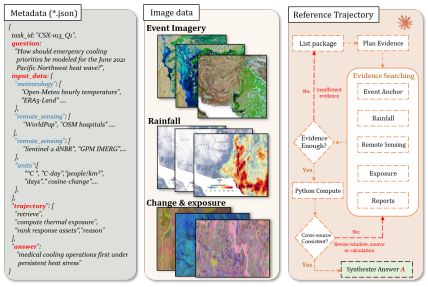


Figure 4: **Task interface.** Metadata, observations, and an auditable research trajectory.

All systems share six operations: discover files, read a source, search local text, inspect structured data, execute scoped Python, and finalize the answer. The registry supports custom tools and retains 170 reusable scientific operations. This common interface leaves the analysis strategy to the solver, while structured returns distinguish execution failures from scientific disagreement. Appendix C.2 gives the execution settings.

3.4 Quality assurance and benchmark comparison

Construction and review. Construction proceeds through event verification, package assembly, task authoring, executable validation, and release review (Figure 5). Automated tools assist, but people check and correct every released task. Review involved six doctoral-level Earth-science experts and twenty undergraduates, who contributed more than 50 hours per person on average. Package-local data must reproduce deterministic ground-truth values, and the public question must match the hidden scoring target. Appendix C.6 details human and agent review; Appendix C.7 lists the seven release gates.

A separate LLM pairwise comparison of complete answers and traces broadly follows the Core ranking (Figure 7), supporting its validity beyond the count of satisfied scoring fields.

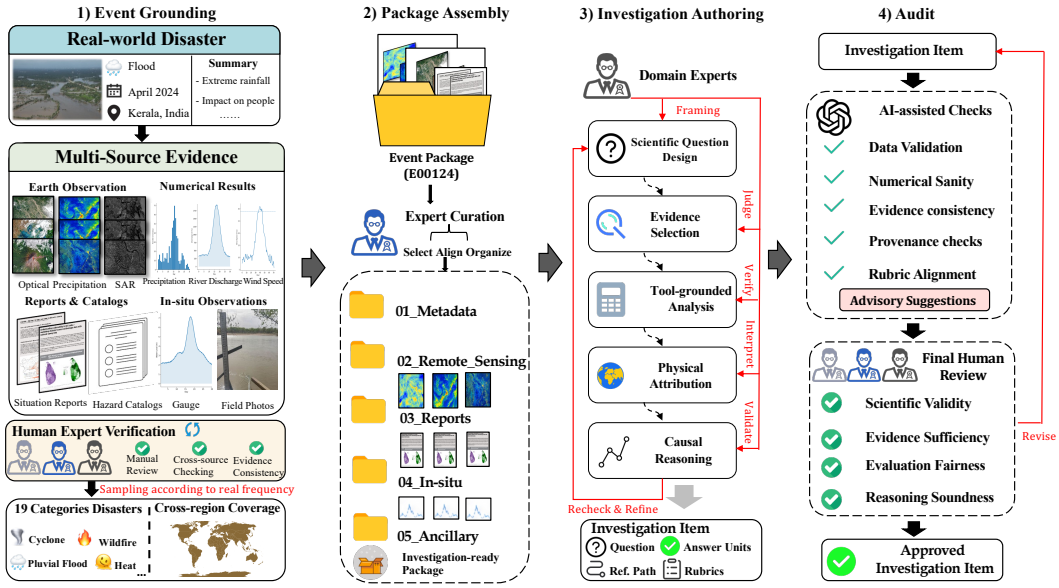


Figure 5: **Benchmark construction.** Event grounding, package assembly, investigation authoring, and expert audit.

Comparison with prior benchmarks. Table 1 compares evidence, analytical work, expert review, and scoring. EarthVerse requires selecting and reconciling records within an event package.

Table 1: **Benchmark comparison: scientific work and evaluation coverage.**

Benchmark	Start evidence	Scientific work	Tools	Find evidence	Multi-source	Compute	Unit score	Trace score	Human review
EarthSE	Paper	Sci. QA	✗	✗	✗	◑	✗	✗	◑
MSEarth	Figure	Figure QA	✗	✗	✗	✗	✗	✗	◑
GeoMMBench	Image	Geo reasoning	✗	✗	✗	✗	✗	✗	◑
ThinkGeo	RS image	Tool use	✓	◑	◑	✓	✗	✓	◑
Earth-Agent	EO assets	EO planning	✓	◑	✓	✓	✗	✓	◑
OpenEarthAgent	Imagery	Multimodal	✓	◑	◑	✓	✗	✓	✗
DORA	Layers	Crisis analysis	✓	◑	✓	◑	◑	✓	◑
EarthVerse	**Package**	**Event synthesis**	✓	✓	✓	✓	✓	✓	✓

✓ full ◑ partial ✗ absent/unreported *EarthVerse scale:* 34 files/package; 7.26 sources and 26.86 units/task.

4 Evaluation Protocol

EarthVerse fixes the scientific question, visible evidence boundary, execution budget, and answer contract across systems. During a run, the system selects evidence, performs calculations, records provenance, and submits a structured scientific answer through the same interface. Runs share limits of 48 turns, 1,680 seconds for research, and 1,800 seconds overall; reference solutions, rubrics, and ground truth remain hidden from solvers. The common controller caps tool observations at 12,000 characters and package-scoped Python calls at 60 seconds. A fixed `gpt-4.1` endpoint scores answers and traces against expert-reviewed targets. Human reviewers verify task validity and replay deterministic ground-truth calculations before release. Appendix C documents configuration and audit details.

4.1 Outcome, Process, and Reliability Metrics

For task i, let U_i be the percentage of required answer units satisfied and H_i the holistic answer judgment, both on a 0–100 scale. Answer correctness is $A_i = \min(H_i, U_i)$: a fluent answer cannot outrank the scientific obligations it actually completes. Let P_i be the task-specific process-rubric

score on the same scale. The primary score is

$$\text{Combined Core} = \frac{1}{2N} \sum_{i=1}^{N} (A_i + P_i) \,. \tag{2}$$

It therefore gives equal weight to endpoint correctness and the documented research process.

Answer units are necessary because one investigation may require a time window, several quantities, a source comparison, and a physical mechanism. A single holistic grade would hide which obligation failed and could let a fluent conclusion mask missing scientific work.

Mean unit accuracy averages U_i across tasks and captures partial completion of required scientific claims and output fields; numerical units use task-specific tolerances. Strict@95 is the percentage of tasks with $U_i \geq 95$ and measures near-complete task reliability rather than average progress. Capability-conditioned scores apply the same Core calculation to expert-tagged task slices. Trajectory statistics report interaction, evidence access, computation, latency, and token demand, but never enter the score. Solver failures receive zero; judge-service failures are rerun until a valid judgment is obtained.

Path pluralism and diagnostic validity. EarthVerse does not prescribe a tool order. Earth-Agent and DORA report final-answer and reference-trajectory metrics; DORA treats answers as primary and trajectory matching as diagnostic, allowing multiple valid paths (Feng et al., 2026; Wang et al., 2026). In multi-source investigations, report-first and time-series-first analyses may reach the same well-supported conclusion. EarthVerse therefore scores what the system establishes and whether the observed evidence and calculations support it. Tools and traces serve reasoning, review, and diagnosis; reproducing a prescribed path earns no credit by itself.

4.2 Evaluated systems

Table 2 compares hosted systems, open-weight models, and research frameworks through one package interface. Some backbones are text-only, while others support native multimodal input; the shared leaderboard does not grant extra evidence to either group. All systems receive the same question, package boundary, answer contract, and execution budget. Rankings compare model–controller systems, including the native harnesses identified in Table 2, rather than isolated backbones. Appendix C.1 gives the complete roster, modality treatment, controller exceptions, and judge configuration.

5 Main Evaluation

Section 5.1 reports overall performance. Section 5.2 examines reasoning, evidence access, and transfer to supplied-observation benchmarks. Appendix B.1 collects the main findings.

5.1 Overall system performance

Setting. Table 2 compares 25 systems under the package-scoped interactive protocol. Harness identifies the controller; systems are grouped by family and ordered by Core. Trajectory columns are per-task means and do not affect the quality scores defined in Section 4.

Mean competence versus conjunctive reliability. Claude Fable 5 leads on Core (84.97), while GPT-5.6 Sol leads on Strict@95 (34.81%). Fable satisfies 84.65% of answer units on average but reaches 95% completion on only 25.93% of tasks. Missing windows, provenance, mechanisms, or unit conversions can invalidate a calculation and its interpretation together. High average accuracy therefore leaves many investigations with consequential gaps.

Interaction and domain knowledge. Most systems share the EarthVerse controller; framework rows retain their own controllers. Neither domain specialization nor interaction volume guarantees quality. OpenResearcher uses 39.73 rounds and 538.7k tokens per task yet trails Fable by 54.18 Core points; GeoMMAgent uses five rounds and reaches only 12.35% Strict@95. Earth-specialized

Table 2: **EarthVerse leaderboard.** Harness names the controller. Scores are percentages; failed, timed-out, or malformed runs remain zero, and batch reads may cover multiple files. Within each family, Core increases downward; **bold blue** and underlined teal mark the top two outcomes.

System configuration		Outcome					Trajectory mean per task						
System	Harness	Answer	Process	Core	Strict@95 (%)	Unit acc.	Rounds	Tool calls	Reads	Files	Python	Latency (s)	Tokens (k)
Agent frameworks (own controllers over the same package interface)													
OpenResearcher	OpenResearcher	30.51	31.06	30.79	4.94	32.74	39.73	33.62	15.00	17.03	5.38	521	538.7
GPT-5.5	GeoMMAgent	65.86	70.84	68.35	12.35	67.18	5.00	8.28	11.93	8.59	0.99	221	37.9
Open-weight and Earth-specialized models													
Intern-S1-mini	EarthVerse	10.78	12.67	11.72	0.25	11.60	4.99	3.43	1.80	2.39	0.07	57	39.4
Nemotron Nano	EarthVerse	12.30	14.01	13.15	0.25	12.80	2.59	2.07	1.26	1.77	0.15	279	53.2
MiroThinker 8B	EarthVerse	12.79	13.62	13.21	2.96	13.42	27.30	25.98	6.07	11.42	2.00	367	318.0
GeoGPT	EarthVerse	22.15	27.32	24.74	0.25	23.61	30.41	29.37	19.33	11.86	2.65	565	502.1
Qwen3-4B	EarthVerse	25.11	24.94	25.02	0.00	26.13	5.06	3.79	2.55	3.66	0.07	132	37.7
GeohazardGPT	EarthVerse	28.88	33.33	31.10	0.49	30.27	30.33	29.34	13.38	15.70	8.44	456	333.7
Qwen3-235B	EarthVerse	31.98	34.64	33.31	0.99	32.66	13.47	12.80	5.33	7.41	3.50	195	274.3
Hosted general-purpose systems													
GPT-5.6 Luna	Codex	35.68	35.63	35.66	11.11	36.28	6.58	5.39	2.37	4.29	1.05	189	156.9
GPT-4o	EarthVerse	37.35	41.11	39.23	0.49	38.69	18.81	8.95	4.04	6.84	0.69	55	127.4
MiniMax M2.7	EarthVerse	43.66	45.77	44.71	7.41	44.35	17.27	16.08	7.11	8.94	2.22	205	285.8
Gemini 3.1 Flash Lite	EarthVerse	57.67	61.66	59.66	2.96	58.67	7.37	6.10	3.59	5.69	1.08	24	70.1
Doubao Seed 2.0 Mini	EarthVerse	61.14	64.84	62.99	5.93	62.22	13.71	12.70	7.21	8.86	1.08	113	175.1
Qwen 3.7 Plus	EarthVerse	69.33	71.37	70.35	18.27	70.16	17.15	15.72	6.51	11.25	3.31	313	204.3
Claude Haiku 4.5	EarthVerse	70.76	73.70	72.23	14.07	71.70	21.31	19.86	10.08	11.72	3.96	324	477.0
Tencent Hy3	EarthVerse	71.27	74.26	72.77	16.30	72.22	19.47	17.77	8.32	12.70	4.41	311	208.7
DeepSeek V4 Flash	EarthVerse	72.32	74.77	73.54	20.25	73.25	27.92	26.81	11.00	16.38	4.07	333	389.7
Kimi K2.5	EarthVerse	73.89	76.76	75.32	14.57	74.70	24.25	22.67	10.01	16.00	4.79	439	302.9
Claude Sonnet 4.6	EarthVerse	74.88	77.44	76.16	20.00	75.76	17.54	16.54	7.72	11.82	4.49	343	331.0
GPT-5.5	EarthVerse	76.85	79.28	78.06	17.04	77.48	15.17	14.20	6.83	10.92	3.74	210	224.9
GLM-5.2	EarthVerse	78.14	80.19	79.17	22.96	78.62	18.32	17.27	7.52	13.36	4.31	259	196.1
GPT-5.6 Terra	Codex	80.12	80.00	80.06	27.16	80.81	12.18	11.13	4.42	9.46	2.71	173	182.6
GPT-5.6 Sol	Codex	81.43	81.94	81.68	**34.81**	82.12	13.88	12.79	4.88	9.65	2.87	205	205.9
Claude Fable 5	Claude Code	**83.99**	**85.95**	**84.97**	25.93	**84.65**	14.45	13.45	5.92	10.02	2.61	247	161.6

pretraining supports terminology and process recognition (Bai et al., 2025; Zhang et al., 2024; Ge et al., 2025), but GeohazardGPT's 30.33 rounds and 8.44 Python calls yield 31.10 Core. These systems still struggle with evidence selection and revision. Capability slices in Appendix D.2 show further gaps in causal reasoning and ranking even where calculation or remote sensing is stronger.

Computational cost. Complete evaluations of some higher-cost frontier systems exceed $2,500. GPT-5.6 Terra retains 98% of Sol's Core at less than half the cost. Cost depends on both pricing and controller behavior; Appendix D.4.3 gives the full comparison.

5.2 Reasoning, evidence access, and benchmark transfer

We use controlled GPT-5.5 experiments to test when extra reasoning helps and which evidence constraints limit it (Table 3). The reasoning study combines historically difficult tasks with a fixed-seed sample of the remaining release. Oracle comparisons reuse the same task IDs and Top-20 router outputs; 95% confidence intervals use 5,000 paired task-level bootstrap resamples with seed 2026. These diagnostic subsets do not estimate full-leaderboard effects (Appendix C.5). External benchmarks test whether the difficulty persists with supplied observations.

Reasoning and review. With fixed evidence, high effort lowers Core from 43.24 to 26.17 for compressed input and from 29.44 to 8.90 for all-text input. Interactive access reverses this pattern: Core rises from 51.30 to 73.80 as effort increases from none to high. Final review raises xhigh Core from 68.58 to 86.29, but slightly lowers high-effort Core from 73.80 to 72.27. More work alone is insufficient: medium-effort interaction uses 17.20 rounds and 206.4k tokens for 63.13 Core, whereas high effort uses 14.96 rounds and 174.3k tokens for 73.80. These results favor reasoning that can retrieve evidence, recompute quantities, and check unresolved dependencies.

Evidence access. Answer-free oracles reveal relevant files, evidence roles, useful tool classes, or a plan without disclosing target answers. Relevant-file localization and an evidence map improve Core by 14.72 and 11.34 points; only these paired confidence intervals exclude zero. A tool subset

Table 3: **When reasoning helps.** (a) Core across all five effort settings; compressed and all-text modes use fixed evidence. (b) All four answer-free oracles, relative to the Top-20 router (Core 66.39); paired 95% bootstrap CIs. Complete outcomes, resources, capability scores, and other interventions appear in Appendix D.1.

(a) Reasoning effort: Core

Mode	None	Low	Med.	High	Xhigh
Compressed	**43.24**	27.16	24.29	26.17	16.67
All text	**29.44**	14.73	10.99	8.90	13.50
Interactive	51.30	58.51	63.13	**73.80**	68.58
+ review	55.09	63.72	68.19	72.27	**86.29**

(b) Answer-free oracles

Oracle	ΔCore	95% CI
Relevant files	**+14.72**	[+5.48, +24.69]
Evidence map	**+11.34**	[+2.34, +21.25]
Tool subset	+2.19	[−3.55, +7.28]
Plan	−0.54	[−7.07, +4.35]

adds 2.19 points, while planning slightly lowers the score. Changes to tool presentation, equivalent units, and optional meteorological access do not show the same reliable gain. File localization improves spatiotemporal Core from 64.40 to 91.47, physical-mechanism Core from 65.15 to 80.61, and quantitative-calculation Core from 73.43 to 87.57. Multi-source Core changes little from its high baseline. Missing, truncated, or conflicting evidence lowers the point estimates, but the paired intervals remain wide. Selecting the right observations is the clearest bottleneck in these tests.

Transfer to supplied-observation benchmarks. We evaluate Earth-Bench, EarthSE, MSEarth, and GeoMMBench through three stages: direct GPT-5.5 responses, corrections to erroneous items or judgments, and the EarthVerse procedure. The final stage gives an answer-free guide to variables, calculations, and evidence checks; it contains no target label, item answer, or item-specific evidence. Figure 6 includes published references (Zhao et al., 2026; Xiao et al., 2026). Scores use each benchmark's native metric and should be compared within benchmarks. Complete metrics and source-paper rows accompany the release.

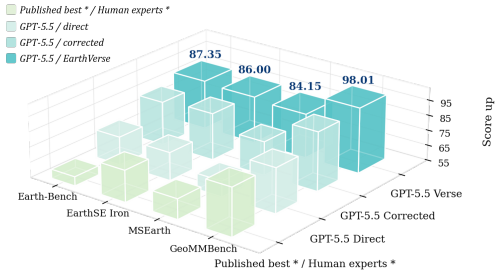


Figure 6: **Cross-benchmark protocol sensitivity.** Benchmark-native scores across three GPT-5.5 stages and published references.

Direct GPT-5.5 already exceeds the published Earth-Bench reference. EarthSE multiple-choice and free-response performance is high, while fill-in accuracy improves from 20.29 to 61.90 after schema and answer-granularity corrections. On MSEarth and GeoMMBench, the corrected and EarthVerse stages close or reverse the gap to the displayed references. GeoMMBench accuracy increases from 84.52 to 98.01, above the published GeoMMAgent score of 88.40 and human reference of 86.50. The guide clarifies the task without supplying answers or new Earth-science knowledge. These gains concern interpretation of selected observations; EarthVerse additionally requires source selection, alignment, verification, and revision, and its Strict@95 remains below 35%.

6 Conclusion

EarthVerse evaluates scientific agents on investigations that require them to select and reconcile heterogeneous observations. The results expose a gap between recovering individual facts and completing a reliable analysis. A mismatched time window or unsupported source can change the calculation and the conclusion that follows from it.

Evidence localization, selective tool routing, and targeted review help agents revise weak analyses. More reasoning over a fixed evidence snapshot does not provide the same benefit. Reliable systems need to keep claims attached to their sources, check whether a method applies, and distinguish current values from those already replaced. Strong results on supplied-observation benchmarks show that interpreting evidence is often easier than assembling it. EarthVerse makes this distinction measurable and provides a basis for improving the full research process.

AI USE STATEMENT

This study evaluates AI systems and uses AI-assisted checks and tests during benchmark construction. LLM agents explored event packages, proposed candidate analyses and scoring checks, and tested tasks through replay, calculation checks, and discrepancy analysis. Human experts reviewed and corrected the tasks, verified that package evidence supported the questions, and checked consistency among the reference solutions, executable ground truth, answer units, and process rubrics before release (Appendix C.6). The evaluation uses a fixed LLM judge under the scoring and audit protocol in Appendix C.4. Generative AI tools also assisted literature retrieval, drafting and editing portions of the manuscript, and LaTeX formatting. The authors take responsibility for the final manuscript, benchmark, and reported results.

ETHICS STATEMENT

EarthVerse draws on publicly available source data and follows the providers' published usage guidelines and applicable terms. Event packages retain source records for traceability. Human experts reviewed and corrected the tasks, checked the supporting evidence, and verified consistency between reference answers and scoring criteria. Some source materials contained file corruption or spatial misalignment. During review, task requirements and scoring targets were checked to avoid reliance on known defective evidence, and affected answer units were removed where necessary. Diverse sources and package-specific calculations help reduce opportunities for memorization and shortcut answering. Because disaster analysis can involve fatalities and affected communities, model outputs require care to avoid unsupported casualty claims, insensitive language, or irresponsible recommendations. The benchmark data are intended for academic research, and outputs should undergo expert review before informing disaster-related decisions.

REPRODUCIBILITY STATEMENT

The main text describes task construction, expert review, and the evaluation protocol. Appendix C provides the experimental configurations and scoring procedures, and Appendix E documents prompts and execution contracts. Our anonymous repository includes the evaluation code, scientific tools, and all 405 tasks with reference answers and structured ground truth. To keep the repository manageable, we selected and uploaded 10 example event packages covering 22 tasks. Upon acceptance, we commit to publicly releasing all event packages, synthetic data and the prompts and instructions used to construct them, reference solutions, ground-truth computation scripts, and human scoring guidelines. We will also release the recorded evaluation trajectories for all evaluated models, including tool calls and outputs, to support independent reproduction, inspection, and further research by the open-source community.

REFERENCES

Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pp. 465–473. AAAI Press, 2018. doi: 10.1609/icwsm.v12i1.14983.

Anthropic. Model system cards. `https://www.anthropic.com/system-cards`, 2026.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-S1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

Riyang Bao, Cheng Yang, Dazhou Yu, Zhexiang Tang, Gengchen Mai, and Liang Zhao. SpatialAgent: Agentic geo-spatial reasoning with scientific core concepts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14896–14911, 2026. doi: 10.18653/v1/2026.acl-long.679. URL `https://aclanthology.org/2026.acl-long.679/`.

Aaron Bell, Amit Aides, Amr Helmy, Arbaaz Muslim, Aviad Barzilai, Aviv Slobodkin, Bolous Jaber, et al. Earth AI: Unlocking geospatial insights with foundation models and cross-modal reasoning. *arXiv preprint arXiv:2510.18318*, 2025.

ByteDance Seed Team. Seed 2.0 official launch. https://seed.bytedance.com/en/blog/seed2-0-%25E6%25AD%25A3%25E5%25BC%258F%25E5%258F%2591%25E5%25B8%2583, 2026.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Zhiqing Cui, Binwu Wang, Qingxiang Liu, Yeqiang Wang, Zhengyang Zhou, Yuxuan Liang, and Yang Wang. Augur: Modeling covariate causal associations in time series via large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 764–787. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.32. URL https://aclanthology.org/2026.acl-long.32/.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7132–7142, 2025. doi: 10.1109/ICCV51701.2025.00670.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hQ0K2Hhq7H.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, and Weijia Li. Earth-Agent: Unlocking the full landscape of Earth Observation with agents. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5b4a459db23e6db9be2a128380953d96-Abstract-Conference.html.

Qi Ge, Pengfa Li, Yinhao Dai, Jin Li, Ni An, Yang Yu, Qing Lv, and Hongyue Sun. GeohazardGPT: Towards large language models for geohazards. *Under review*, 2025.

Andrew Gettelman, Alan J. Geer, Richard M. Forbes, Greg R. Carmichael, Graham Feingold, Derek J. Posselt, Graeme L. Stephens, Susan C. van den Heever, Adam C. Varble, and Paquita Zuidema. The future of Earth System Prediction: Advances in model–data fusion. *Science Advances*, 8(14):eabn3488, 2022. doi: 10.1126/sciadv.abn3488.

Google DeepMind. Gemini 3.1 Flash-Lite. Google DeepMind model card, https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, 2026.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13936–13971, 2024. doi: 10.18653/v1/2024.findings-emnlp.815. URL https://aclanthology.org/2024.findings-emnlp.815/.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric T. Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew E. Gaston. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10–17, 2019. URL https:

//openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/Gupta_Creating_xBD_A_Dataset_for_Assessing_Building_Damage_from_Satellite_CVPRW_2019_paper.html.

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

Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2022. doi: 10.1017/9781009325844. URL https://www.ipcc.ch/report/ar6/wg2/.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. TerraMind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7383–7394, 2025. doi: 10.1109/ICCV51701.2025.00693.

Yigeng Jiang, Tengchao Yang, Taoyong Cui, Jiaxing Wan, Yuan Wang, Weida Wang, Zhiyu Liu, Chuyi Peng, Binzhao Luo, Maoli Gao, et al. Deep research in physical sciences: A multi-agent framework and comprehensive benchmark. *arXiv preprint arXiv:2606.18648*, 2026.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Chenyue Li, Wen Deng, Mengqian Lu, and Binhang Yuan. AtmosSci-Bench: evaluating the recent advance of large language model for atmospheric science. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4850.

Zhuofeng Li, Dongfu Jiang, Xueguang Ma, Haoxiang Zhang, Ping Nie, Yuyu Zhang, Kai Zou, Jianwen Xie, Yu Zhang, and Wenhu Chen. OpenResearcher: A fully open pipeline for long-horizon deep research trajectory synthesis. *arXiv preprint arXiv:2603.20278*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. SciAgent: Tool-augmented language models for scientific reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15701–15736, 2024. doi: 10.18653/v1/2024.emnlp-main.880.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Yannis Markonis, Christoforos Pappas, Martin Hanel, and Simon Michael Papalexiou. A cross-scale framework for integrating multi-source data in Earth System Sciences. *Environmental Modelling & Software*, 139:104997, 2021. doi: 10.1016/j.envsoft.2021.104997.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

MiniMax. MiniMax M2.7: Early echoes of self-evolution. https://www.minimax.io/news/minimax-m27-en, 2026.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, Wenhan Dou, et al. MiroThinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*, 2025.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, 2026a.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, 2026b.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dHng2O0Jjr.

Qwen. Qwen3.7-Plus. Official model directory, https://chat.qwen.ai/legal-agreement/models, 2026.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. doi: 10.1109/ACCESS.2021.3090981.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth System Science. *Nature*, 566:195–204, 2019. doi: 10.1038/s41586-019-0912-1.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997.

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

Akashah Shabbir, Muhammad Umer Sheikh, Muhammad Akhtar Munir, Hiyam Debary, Mustansar Fiaz, Muhammad Zaigham Zaheer, Paolo Fraccaro, Fahad Shahbaz Khan, Muhammad Haris Khan, Xiao Xiang Zhu, et al. OpenEarthAgent: A unified framework for tool-augmented geospatial agents. *arXiv preprint arXiv:2602.17665*, 2026.

Yujiong Shen, Yajie Yang, Zhiheng Xi, Binze Hu, Huayu Sha, Qiyuan Peng, Jiazheng Zhang, Junlin Shang, Jixuan Huang, Yutao Fan, Jingqi Tong, Ming Zhang, Shihan Dou, Zhenfei Yin, Xingjun Ma, Lei Bai, Tao Gui, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. SciAgentGym: Benchmarking multi-step scientific tool-use in LLM agents. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66785.

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longepe, Xiao Xiang Zhu, Hannah Kerner, et al. GEO-Bench-2: From performance to capability, rethinking evaluation in geospatial AI. *arXiv preprint arXiv:2511.15658*, 2025.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth Observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2025.3642610. URL https://ieeexplore.ieee.org/document/11296896/.

Niloufar Alipour Talemi, Julia Boone, and Fatemeh Afghah. Agentic AI in remote sensing: Foundations, taxonomy, and emerging systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 786–799, 2026. doi: 10.1109/WACVW68408.2026.00088. URL https://openaccess.thecvf.com/content/WACV2026W/GeoCV/html/Talemi_Agentic_AI_in_Remote_Sensing_Foundations_Taxonomy_and_Emerging_Systems_WACVW_2026_paper.html.

Tencent. Tencent unveils Hy3 preview; model enhances agent capabilities and real-world usability. https://www.tencent.com/en-us/articles/2202320.html, 2026.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. XLRS-Bench: Could your multimodal LLMs understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025. doi: 10.1109/CVPR52734.2025.01336.

Junjue Wang, Weihao Xuan, Heli Qi, Pengyu Dai, Kunyi Liu, Hongruixuan Chen, Zhuo Zheng, Junshi Xia, Stefano Ermon, and Naoto Yokoya. Can LLM agents respond to disasters? benchmarking heterogeneous geospatial reasoning in emergency operations. *arXiv preprint arXiv:2605.11633*, 2026.

World Meteorological Organization. Atlas of mortality and economic losses from weather, climate and water-related hazards, 1970–2021. Technical report, World Meteorological Organization, 2023. URL https://public.wmo.int/publication-series/atlas-of-mortality-and-economic-losses-from-weather-climate-and-water-related-hazards-1970-2021.

World Meteorological Organization. State of the global climate 2024. Technical Report WMO-No. 1368, World Meteorological Organization, 2025. URL https://wmo.int/publication-series/state-of-global-climate/state-of-global-climate-2024.

Aoran Xiao, Shihao Cheng, Yonghao Xu, Yexian Ren, Hongruixuan Chen, and Naoto Yokoya. GeoMMBench and GeoMMAgent: Toward expert-level multimodal intelligence in geoscience and remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34843–34853, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094, 2024. doi: 10.52202/079017-1650.

Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, Xiaoyu Yue, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. EarthSE: A benchmark evaluating earth scientific exploration capability for large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jyYE06FL8G.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Tao Yu, Yiming Ding, Shenghua Chai, Minghui Zhang, Zhongtian Luo, Xinming Wang, Xinlong Chen, Zhaolu Kang, Junhao Gong, Yuxuan Zhou, Haopeng Jin, Zhiqing Cui, Jiabing Yang, YiFan Zhang, Hongzhu Yi, Zheqi He, Xi Yang, Yan Huang, and Liang Wang. Omni-DeepSearch: A benchmark for audio-driven omni-modal deep search. *arXiv preprint arXiv:2605.08762*, 2026. URL https://arxiv.org/abs/2605.08762.

Z.ai. GLM-5.2: Built for long-horizon tasks. https://z.ai/blog/glm-5.2, 2026.

Yifan Zhang, Cheng Wei, Zhengting He, and Wenhao Yu. GeoGPT: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131:103976, 2024. doi: 10.1016/j.jag.2024.103976.

Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, et al. Large multimodal models evaluation: a survey. *Science China Information Sciences*, 68(12):221301, 2025. doi: 10.1007/s11432-025-4676-4.

Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. MSEarth: A multimodal benchmark for earth science phenomenon discovery with MLLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5270–5301, 2026. doi: 10.18653/v1/2026.acl-long.239. URL https://aclanthology.org/2026.acl-long.239/.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

Jakob Zscheischler, Seth Westra, Bart J. J. M. van den Hurk, Sonia I. Seneviratne, Philip J. Ward, Andy Pitman, Amir AghaKouchak, David N. Bresch, Michael Leonard, Thomas Wahl, and Xuebin Zhang. Future climate risk from compound events. *Nature Climate Change*, 8(6):469–477, 2018. doi: 10.1038/s41558-018-0156-3.

APPENDIX GUIDE

A **Appendix A Controlled Studies** 17

How reasoning, evidence access, memory, tools, and execution conditions change research quality.

Answer mode • reasoning effort • routing • context • robustness • scientific environments

B **Appendix B Discussion, Broader Impact, and Limitations** 20

What the results imply for scientific agents, disaster analysis, oversight, and responsible use.

Research position • broader impact • system design • limitations

C **Appendix C Experimental Setup and Reproducibility** 22

How models, modalities, inference, scoring, study manifests, and release decisions are controlled.

System roster • modality interface • main configuration • sampling • audit

D **Appendix D Extended Diagnostics** 28

Where capability, trajectory, error-path, resource, intervention, and tool-compliance analyses are reported.

Capability scores • error paths • costs • ablations • uncertainty

E **Appendix E Prompts and Execution Contracts** 37

What the solver, verifier, judge, and benchmark-construction reviewers are instructed to do.

Runtime visibility • solver contract • final review • judging • construction review

F **Appendix F Representative Case Set** 40

Which systems and tasks were selected to represent the benchmark's main scientific workflows and failure modes.

Five successful cases • one partial failure • system family • task scope • score

G **Appendix G Detailed Research Trajectories** 41

How selected systems discover evidence, calculate results, recover from errors, and finalize an answer.

Full task • ordered trace • computed ledger • evidence reconciliation • failure analysis

H **Appendix H Detailed Related Work** 48

Scientific benchmarks, disaster intelligence, tool-using agents, and geospatial research systems.

Supplied observations • scientific agents • evidence construction

Reading routes

Experiments A + D **Reproduction** C + E **Interpretation** B + F + G

A CONTROLLED STUDIES

We vary answer mode, reasoning effort, evidence localization, tool routing, context, stopping, and professional scientific execution. Unless noted, every condition uses GPT-5.5 under the protocol in Section 4.2, with identical task manifests within each experiment.

A.1 REASONING EFFORT AND ANSWER MODE

We cross four answer modes with five reasoning settings on a fixed, difficulty-stratified manifest. Compressed direct uses a bounded evidence summary, while all-text direct serializes every readable file. Both interactive conditions use the same EarthVerse harness; the review variant adds one evidence-and-calculation audit before finalization. The compact view below retains the outcome columns so that each reasoning curve can be compared without the denser interaction and resource diagnostics in Table 7(a).

A.2 TOOLS, CONTEXT, AND INTERACTION BUDGET

On one fixed paired manifest, we vary catalog size, semantic routing, tool composition, context retention, and interaction budget. Every condition uses the default medium reasoning setting and the same answer and process scores. Appendix D.5 reports trajectory diagnostics.

A.2.1 TOOL-CATALOG SCALE

Nested fixed catalogs expose increasingly broad subsets of the same 170-tool registry (panel (a) of Table 5). The six-tool condition contains package discovery, reading, search, tabular inspection, Python execution, and finalization; each larger catalog adds domain tools without changing the task prompt.

The full catalog loses 9.80 Core points relative to the six-tool foundation while using 19% more tokens. The decline is not monotonic, which rules out a simple capacity penalty. Extra tools add nearby analytical paths with different assumptions about variables, units, aggregation, and applicability. When the controller cannot retain why it chose one path, later outputs from another tool can enter the same ledger as if they were comparable. Catalog breadth then creates scientific ambiguity. Reporting a tool count without the navigation policy misses this failure mode.

A.2.2 SEMANTIC ROUTING AND SCIENTIFIC COMPOSITION

The routing study selects tools from task text and package metadata before interaction (panel (b) of Table 5). Top-k routing exposes the highest-scoring individual tools, whereas workflow routing first selects a domain workflow. A separate composition study compares an expert-only suite, a mixed general/expert suite, and atmospheric Python and Fortran suites.

Top-20 routing recovers 11.21 Core points over the full catalog with 34% fewer tokens. Its advantage comes from narrowing the next choice without deciding the investigation in advance. Workflow routing is more brittle because it can commit to a domain before the evidence warrants that commitment; the expert-only suite has the opposite problem, lacking ordinary file and calculation operations that connect specialist output to a scored claim. Routing should act as a revisable prior: preserve a small general core, introduce specialist methods when their required variables appear, and allow the evidence to overturn the initial route.

A.2.3 CONTEXT RETENTION

We vary the amount of tool history preserved verbatim before older content is summarized (panel (c) of Table 5). Aggressive compression keeps a narrow recent window; short and long policies progressively retain more observations; full history disables summarization within the tested budget.

Long bounded retention gains 13.66 Core points over aggressive compression and 5.27 over full replay. Too little history forces the agent to rediscover files and often severs intermediate values from their provenance. Full replay retains rejected hypotheses and superseded numbers alongside

Table 4: **Outcome view of the reasoning sweep.** GPT-5.5; **bold blue** and shading mark the best setting in each access mode.

Compressed direct (selected evidence)

Effort	Answer	Process	Core	Unit acc.
none	**41.37**	**45.10**	**43.24**	**42.21**
low	24.99	29.32	27.16	25.43
medium	22.88	25.69	24.29	23.26
high	24.17	28.17	26.17	24.28
xhigh	16.67	16.67	16.67	16.67

All-text direct (budgeted evidence)

Effort	Answer	Process	Core	Unit acc.
none	**27.66**	**31.22**	**29.44**	**28.27**
low	13.26	16.20	14.73	13.36
medium	9.79	12.18	10.99	10.05
high	8.80	9.00	8.90	8.82
xhigh	12.00	15.00	13.50	12.00

EarthVerse interactive

Effort	Answer	Process	Core	Unit acc.
none	48.89	53.70	51.30	49.46
low	56.97	60.05	58.51	57.53
medium	61.45	64.80	63.13	62.35
high	**72.41**	**75.19**	**73.80**	**72.59**
xhigh	69.66	67.50	68.58	69.83

EarthVerse interactive + review

Effort	Answer	Process	Core	Unit acc.
none	53.70	56.48	55.09	54.06
low	61.93	65.50	63.72	62.78
medium	67.16	69.22	68.19	67.91
high	69.82	74.71	72.27	70.73
xhigh	**84.36**	**88.21**	**86.29**	**84.40**

Table 5: **System ablations.** GPT-5.5 at medium reasoning; **bold blue** marks panel bests.

	Outcome			Coverage	Core by capability							Resource
Condition	Answer	Process	Core	Unit acc.	Physical mech.	Spatio-temporal	Quant. calc.	Multi-source	Causal chain	Ranking	RS / geo.	Tokens
(a) Tool catalog: nested subsets of 170 tools												
Fixed tools (6)	**64.34**	**65.62**	**64.98**	**64.69**	63.27	66.22	70.34	97.50	63.35	51.78	58.12	229,557
Fixed tools (10)	60.22	57.10	58.66	60.35	56.74	58.36	61.04	95.22	55.77	58.32	61.25	194,047
Fixed tools (20)	60.58	59.38	59.98	60.93	58.13	70.50	61.67	95.00	54.90	56.50	55.62	232,912
Fixed tools (40)	58.55	61.00	59.77	59.43	58.16	47.81	62.03	90.45	63.63	59.57	54.12	232,993
Fixed tools (70)	58.76	61.25	60.01	59.19	58.54	55.68	63.89	87.95	60.14	54.21	54.62	233,759
Fixed tools (100)	63.96	63.12	63.54	64.11	61.81	58.93	65.84	96.37	65.79	60.82	51.00	278,936
Full catalog (170)	54.48	55.88	55.18	55.46	55.58	55.72	54.96	47.50	52.75	63.15	55.00	273,216
(b) Routing and tool composition												
Top-10 semantic router	64.12	**66.25**	65.19	64.55	64.37	60.64	67.93	80.74	67.50	57.20	59.17	216,227
Top-20 semantic router	**66.90**	65.88	**66.39**	**67.14**	65.15	64.40	73.43	90.00	65.89	50.84	53.94	180,010
Workflow-profile router	56.05	56.75	56.40	56.47	54.74	52.63	58.60	87.95	55.39	55.00	60.62	220,232
Expert-only suite	59.28	60.88	60.08	59.50	58.48	61.08	61.71	90.45	57.79	61.08	57.62	295,490
General + expert suite	59.07	63.00	61.04	59.95	59.25	66.01	63.21	95.00	61.34	49.27	51.38	195,092
Atmospheric Python suite	59.48	60.12	59.80	59.67	58.26	59.64	60.69	89.20	60.94	58.31	49.25	198,802
Atmospheric Fortran suite	57.46	57.88	57.67	57.51	55.81	51.05	61.19	92.95	57.35	53.88	60.75	205,262
(c) Context retained before summarization												
Aggressive compression	48.75	51.62	50.19	49.00	48.07	57.55	53.73	90.45	44.65	45.44	47.00	197,038
Short retained context	56.31	56.38	56.34	56.52	54.16	57.30	55.54	97.72	57.85	60.08	42.25	209,827
Long retained context	**64.32**	**63.38**	**63.85**	**64.75**	62.21	59.43	68.29	95.00	64.72	58.79	48.25	227,536
Full retained history	58.90	58.25	58.58	59.42	56.98	55.39	62.15	89.00	57.25	52.59	61.97	187,700
(d) EarthVerse-harness turn budget under one wall clock												
12-turn budget	44.84	49.50	47.17	44.95	44.65	45.36	46.27	95.00	47.73	57.50	35.13	106,263
24-turn budget	54.10	52.88	53.49	54.32	51.31	57.88	55.46	95.00	50.93	49.13	50.92	161,404
96-turn budget	59.72	60.12	59.92	60.16	57.94	62.76	60.53	97.50	62.10	54.08	44.38	182,798
Wall-clock limited	**62.65**	**64.50**	**63.57**	**62.91**	62.16	62.72	66.66	90.45	67.11	56.91	33.38	221,180

active values, so discarded information can reappear in later calculations. Scientific memory therefore needs a compact claim–evidence state that binds claims to sources, units, temporal and spatial support, and unresolved conflicts while archiving replaced values.

> **Finding. Scientific memory is selective.** Long bounded retention beats both aggressive compression and full replay. The useful state is a compact ledger of active claims, sources, units, and unresolved conflicts, not a verbatim transcript.

A.2.4 Interaction budget

The final study changes only the EarthVerse-harness turn limit. The unbounded condition still uses the common wall-clock budget, so it tests productive stopping rather than unlimited execution.

Twelve turns usually end before discovery, calculation, and verification can all occur. Additional turns improve mean quality, but the gain depends on what those turns accomplish. Extra interaction helps when it closes a named evidence gap and hurts when the agent reopens a settled calculation or follows an unproductive branch. Stopping is therefore a coverage decision: the system should stop when required claims are supported, units and provenance are checked, and material contradictions have been resolved or disclosed.

A.3 Oracle access, discoverability, and evidence robustness

These conditions reuse one paired task set and the Top-20 router baseline. Table 7(b) reports the outcomes; Appendix D.6 retains trajectory and resource diagnostics.

A.3.1 Oracle bottleneck localization

Each oracle removes one information bottleneck while leaving the model, budget, tools, and scoring unchanged. It is a diagnostic upper bound rather than a proposed deployment setting, and none of the conditions reveals the target answer. The relevant-file oracle identifies only the package-relative files used by the reference analysis. The tool-subset oracle names useful tool classes; the planning oracle supplies a short answer-free plan; and the evidence-map oracle adds source roles and intermediate quantities without stating the conclusion.

A.3.2 Tool naming and discoverability

We hold tool implementations fixed and alter only how the catalog is presented (panel (b) of Table 7(b)). Descriptive names add intended-use and exclusion cues; hierarchical discovery chooses a domain before exposing tools; dynamic descriptions retrieve candidate tools at each step; opaque names remove semantic hints.

A.3.3 Evidence perturbation and robustness

Perturbations preserve the question and answer schema while modifying the package evidence (panel (c) of Table 7(b)): irrelevant files are added, a secondary source is made inconsistent, one supporting file is removed, long text is truncated, or units are changed while preserving physical equivalence.

A.3.4 Compact meteorological access

This probe tests access rather than compulsory execution. One condition exposes the same compact shortlist of meteorological operations to every task; the other uses task and package cues to expose at most three. The agent may ignore them in either condition. This separates tool availability from the professional-environment study below, where repeated specialist use is required.

A.4 Professional meteorological environments

This study asks a stricter question: what happens when a task must be completed through a named scientific environment? Each condition requires repeated use of one of 13 libraries or runtimes under the same protocol. Adapters expose only implemented operations and record whether calls succeed, allowing the results to separate useful scientific fit from technically successful but unsuitable method use.

Julia, R climate, MetPy, xclim, and CDO improve Core because their implemented operations match common package structures such as time series, gridded fields, and atmospheric profiles. Julia gives the largest gain (+4.62), but the ranking is less informative than the conditions under which the tools help. They compress many low-level numerical steps into an inspectable transformation without hiding the variables or aggregation. That advantage disappears when the package lacks the temporal support or paired variables assumed by the method.

Required use exposes a failure that optional-tool studies can miss. Pyextremes loses 6.88 Core points because many event packages do not contain a long, homogeneous series suitable for return-level estimation; forecast-verification methods face the same mismatch when paired forecasts and

Table 6: **Required scientific environments.** GPT-5.5; ΔCore is relative to the Top-20 router.

Environment	Outcome			Change	Coverage
	Answer	Process	Core	Δ Core	Unit acc.
Top-20 router (ref.)	66.90	65.88	66.39	0.00	67.14
(a) Python scientific libraries					
MetPy	68.69	70.38	69.53	+3.14	68.80
xclim	67.77	69.62	68.70	+2.31	67.95
thermofeel	66.78	68.75	67.76	+1.37	67.47
PyET	65.34	67.38	66.36	−0.03	65.89
climate-indices	64.50	66.75	65.62	−0.77	65.58
xskillscore	63.37	64.12	63.75	−2.64	63.90
pyextremes	59.02	60.00	59.51	−6.88	59.27
(b) Compiled and domain-language environments					
Julia	**71.77**	70.25	**71.01**	**+4.62**	**71.99**
R climate	68.55	**71.38**	69.96	+3.57	69.44
CDO	66.91	69.50	68.20	+1.81	67.35
NCO	67.39	65.00	66.19	−0.20	67.49
NCL	65.03	65.00	65.01	−1.38	65.27
Fortran	64.46	64.25	64.35	−2.04	64.52

observations are absent. Under a hard requirement, the agent sometimes bends the available data toward the method instead of rejecting the method. A scientific interface needs an explicit refusal path, backed by minimum record length, sampling assumptions, missing-data rules, and a statement of what the output can support.

Finding. Execution success is not scientific validity. A specialist method helps only when its required variables and sampling support exist. Otherwise, a successful call can produce a precise result that the package cannot justify.

B Discussion, Broader Impact, and Limitations

B.1 Findings from the main evaluation

These findings summarize the system comparisons and controlled studies in Section 5.

Finding. Reliability requires the whole evidence chain. High average accuracy can hide one missing source, window, unit, calculation, or mechanism. In disaster analysis, that single break can invalidate an otherwise plausible account.

Finding. Useful interaction changes the scientific state. A tool call matters when its observation changes the chosen source, calculation, or physical account. Longer traces that preserve the same mistake add cost, not quality.

Finding. Reasoning needs an editable claim–evidence state. Additional reasoning helps when it can change the investigation. Interactive retrieval and final review let the system replace a weak source, recompute a quantity, and carry the correction into the final answer. With a fixed evidence snapshot, the same effort can reinforce an early error.

Finding. Long-horizon evidence construction remains unresolved. Answer-free structural guidance brings a frontier model to or above published reference levels on supplied-observation benchmarks, including the displayed human lines. EarthVerse Strict@95 remains below 35% because the agent must construct, check, and revise the evidence base itself.

B.2 RESEARCH POSITION

> **Position.** Earth-science analysis requires systems to maintain the correspondence between heterogeneous observations and scientific claims. Satellite products, station records, reanalyses, and event reports can support a shared physical explanation only when their scales, variable meanings, time windows, units, and transformations remain aligned. EarthVerse evaluates this evolving claim–evidence state: a single broken binding can make a locally correct calculation describe the wrong Earth process (Du et al., 2026; Jiang et al., 2026).

EarthVerse evaluates a complete research loop over a bounded event archive. The agent must decide what counts as evidence, transform it, and revise an event-scale explanation while retaining enough provenance for another analyst to replay the work. This is more constrained than open-ended research, but it begins earlier than supplied-observation QA and asks more of the controller than a tool-execution benchmark.

Strong systems often interpret a selected observation correctly. Reliability falls when they must choose the record, its scale, the compatible variables, and the point at which the analysis is complete. Many failures originate in *evidential binding*: a plausible value is attached to the wrong aggregation, time window, unit, or source role, and later reasoning treats the mismatch as established fact. This explains why stronger reasoning, more tools, and longer traces can all fail to help. They operate on the state the controller has preserved. If that state no longer records what a value means, additional reasoning only makes the wrong account more elaborate.

Complex Earth-system investigations rarely have a unique valid research path. EarthVerse therefore prescribes neither a tool order nor a reference trajectory. Different paths are valid when they recover the required windows, quantities, source roles, and mechanisms, and preserve support for the resulting claims (Gu et al., 2024; Majumder et al., 2025).

B.3 BROADER IMPACT

Scientific practice. Assistants can already reduce the mechanical burden of first-pass hazard analysis: inventorying an archive, locating candidate records, harmonizing units, repeating calculations, and surfacing source disagreement. The experiments also show where this assistance becomes risky. A fluent system may complete each local operation while carrying a scale mismatch through the full analysis. Experts still need to judge comparability, uncertainty, and consequences; package-local provenance gives them something concrete to inspect.

Reliability in high-stakes settings. High average unit accuracy can hide one missing source, time window, calculation, or mechanism. In disaster analysis, that single omission may change estimated severity, exposed population, or the causal account used to justify action. The benchmark does not certify operational readiness. It makes these omissions visible before a result is treated as operational evidence.

Access and evaluation practice. A local package reduces dependence on changing APIs and lets researchers inspect the same evidence. Evaluating frontier systems across hundreds of long trajectories is still expensive and may exclude smaller groups. Cost should be reported alongside quality, and releases should include compact diagnostic subsets, task-level scores, and traces. Subset results should not be presented as full-benchmark estimates.

B.4 SYSTEM DESIGN IMPLICATIONS

For Earth-system analysis, a trajectory is useful only when it preserves the support for each claim. A record of calls without source bindings cannot show whether the final explanation follows from the evidence.

Claim–evidence state. A transcript records chronology but does not identify which claims and values remain active. The controller needs a claim–source graph in which every derived value retains its source, unit, time window, transformation, and relation to competing explanations. Corrections should replace active values without erasing the history that explains the change.

Scientific operations. Scientific tools need to expose applicability as part of their output. Record length, sampling support, missing-data rules, and estimator assumptions determine whether a computed value can support the claim. Training that rewards successful calls alone encourages the model to force data into an available method; justified refusal should receive credit when the inputs are inadequate.

Verification and oversight. Final review should operate on claims rather than prose. For each required conclusion, the system should expose coverage, dimensional checks, source conflicts, and the mechanism tests it rejected. Experts can then inspect the points where judgment entered the analysis and decide whether the result is fit for use. A polished report without this record is not auditable.

B.5 Limitations

Coverage and ecological validity. EarthVerse is a controlled approximation of research. Its 199 packages cover many natural hazards, but not the full range of Earth-system science, long-term attribution, or live forecasting. Package isolation improves reproducibility and blocks web leakage, at the cost of open-ended discovery and institutional data-quality judgments. Masked products also prevent tests of absolute geolocation.

Ground truth and judge validity. Structured ground truth necessarily chooses one decomposition of a scientific answer. Process scores also rely on an LLM judge that compares each trace with a reference solution. Separate answer and process blocks reduce score leakage.

Data provenance and maintenance. The assembled event packages, task formulations, answer units, and derived targets were created for EarthVerse and were not available online before release, which reduces the chance of exact benchmark memorization. Some underlying historical events and source documents may still have appeared in pretraining. Package isolation, hidden scoring artifacts, and package-specific calculations make shortcut answering harder.

Statistical scope of controlled studies. The controlled studies use paired subsets, and several intervals in Appendix D.6 include zero. Large contrasts, such as the relevant-file oracle, support architectural hypotheses. Smaller changes need replication across more packages and backbones. Cross-benchmark revision measures recoverability under feedback, not zero-shot superiority.

C Experimental Setup and Reproducibility

This section collects implementation details that are necessary for reproduction but not for interpreting the main findings. It records the evaluated systems, the common evidence interface, the leaderboard configuration, controlled-study manifests, scoring isolation, and retained audit artifacts.

C.1 Systems, controllers, and model modality

The leaderboard contains 25 complete systems. Hosted endpoints include the OpenAI GPT-4o, GPT-5.5, and GPT-5.6 tiers (OpenAI, 2024; 2026a;b); Anthropic Claude tiers (Anthropic, 2026); Gemini (Google DeepMind, 2026); MiniMax M2.7 (MiniMax, 2026); Seed 2.0 (ByteDance Seed Team, 2026); Qwen3.7 (Qwen, 2026); Tencent Hy3 (Tencent, 2026); DeepSeek-V4 (DeepSeek-AI, 2026); Kimi K2.5 (Kimi Team, 2026); and GLM-5.2 (Z.ai, 2026). The open-weight group includes Qwen3 at 4B and 235B scales (Yang et al., 2025), Nemotron Nano (NVIDIA, 2025), MiroThinker 8B (MiroMind Team et al., 2025), GeohazardGPT, GeoGPT, and Intern-S1-mini (Ge et al., 2025; Zhang et al., 2024; Bai et al., 2025).

Most systems run in the EarthVerse harness, which manages package discovery, tool use, context, recovery, and finalization. GeoMMAgent uses its adapted five-stage coordinator with GPT-5.5 as the worker model in every stage; OpenResearcher retains its research-loop controller (Xiao et al., 2026; Li et al., 2026). These rows therefore describe a model–controller system, not a bare backbone.

The roster mixes text-only models with natively multimodal endpoints. Direct image interpretation is required by only a small fraction of EarthVerse tasks; most remote-sensing evidence is represented

by product metadata, raster statistics, change indices, or other auditable tool outputs. Raw pixels are not selectively sent to models with vision support. This keeps the evidence interface comparable, although the resulting scores should not be read as a standalone test of visual perception.

C.2 Shared experimental configuration

Every run exposes a scientific question and one evidence root while keeping the solution, rubric, ground truth, scoring code, prior outputs, and web evidence hidden. Reads, writes, and Python execution stay inside the active evidence collection. Reports and tables are returned as local text or structured summaries; image and raster operations return recorded metadata and measurements.

The EarthVerse harness is the common controller for the standard model rows. To study the attainable system-level ceiling, we also evaluate the GPT-5.6 family in Codex and Claude Fable 5 in Claude Code, the strongest model-native harnesses available for those systems. GeoMMAgent pairs GPT-5.5 with its five-stage coordinator, while OpenResearcher keeps its own controller. The Harness column in Table 2 makes these differences explicit.

The common execution limit is 48 turns, 1,680 seconds for research, and 1,800 seconds overall. The controller retains 16 recent messages and a 24,000-character research state; individual tool observations are capped at 12,000 characters. Package-scoped Python has a 60-second call limit. Up to two schema repairs are allowed, with 120 seconds reserved for finalization and a separate 120-second final-call timeout. We retain provider-native reasoning and generation controls when available, use endpoint defaults otherwise, and do not impose one temperature across providers. The judge is a fixed `gpt-4.1` endpoint, and randomized procedures use seed 2026. Endpoint access dates and local model revisions are retained in the release registry.

C.3 Package evidence sources and access interfaces

The 199 package manifests contain 6,397 provenance records, 835 raw `source_name` labels, and 6,709 local files from 107 external host domains. Many raw labels are event-specific filenames, response formats, endpoint variants, or repeated queries to the same service rather than distinct sources. We therefore group aliases at the product/API-family level while keeping distinct providers, sensors, product versions, and access semantics separate. This yields 80 external source and access families. We also list nine package-local derivative families used directly in computation; they do not count as additional upstream sources.

PACKAGE SOURCES **Distinct source and access families across 199 event packages**

Event discovery, reports, and authoritative anchors **29 families**

Provider or product	Access form	Role in a package
GDACS	event-list REST/GeoJSON API; RSS archive	Cross-hazard event discovery, dates, alert levels, and identifiers.
NASA EONET v3	REST search by date, category, or keyword	Wildfire, flood, storm, and general natural-event catalog records.
ReliefWeb	public update and report search	Humanitarian situation reports, response updates, and impact narratives.
HDX	CKAN `package_search` API	Candidate humanitarian datasets and country- or event-specific impact layers.
GDELT	event-window article search	Secondary news discovery used only when the package retains the result.
Wikipedia	REST page-summary and event-search interfaces	Event-name cross-checks and compact encyclopedic context.
Wikidata	entity-search API	Structured event and place identifiers used for identity checks.
NASA Earth Observatory	event articles and image pages	Locked event anchors and satellite-based physical narratives.
NASA Scientific Visualization Studio	visualization pages and media API	Official imagery, animations, and explanatory visual products.

PACKAGE SOURCES Distinct source and access families across 199 event packages *continued*

NASA Disasters and Applied Sciences	activation and event pages	Satellite-based disaster-response context and mapped observations.
World Meteorological Organization	reports, topic pages, and bulletins	Authoritative global weather, climate, and extreme-event summaries.
World Meteorological Centre Beijing	event reports and operational bulletins	Regional meteorological summaries and event chronology.
NOAA Climate.gov	event trackers and climate reports	Reviewed US climate-event narratives and physical context.
US National Weather Service and WPC	forecast discussions, storm summaries, and precipitation pages	Operational weather chronology, rainfall totals, and warning context.
NOAA NESDIS and SSD	satellite product and event pages	Operational satellite imagery and environmental monitoring context.
Copernicus EMS	activation and rapid-mapping archives	Event confirmation and links to emergency-mapping products.
UN OCHA	situation reports and humanitarian updates	Reported impacts, needs, access constraints, and response activity.
UNOSAT	satellite-assessment reports and maps	Mapped damage, flood, and exposure evidence.
UN country teams and UN Geneva	country and event updates	Local consequence and response narratives retained as event evidence.
IFRC	emergency appeals and operational updates	Humanitarian consequences, needs, and Red Cross response records.
AHA Centre	regional disaster updates and reports	ASEAN event impacts and coordinated response context.
World Health Organization	health-emergency reports	Mortality, morbidity, service disruption, and public-health consequences.
UNICEF	humanitarian situation reports	Impacts on children, water, health, education, and relief delivery.
FAO	food, agriculture, and livelihood reports	Crop, livestock, food-security, and rural-impact evidence.
World Bank	assessment and recovery documents	Infrastructure, economic loss, and reconstruction context.
World Weather Attribution	event-study pages and reports	Attribution context and documented physical or societal drivers.
South Asian operational agencies	IMD/RSMC New Delhi, DHM Nepal, and DMC Sri Lanka	Official cyclone, rainfall, flood, and warning records.
National meteorological services	BoM, HKO, UK Met Office, MetService, and SAWS	National weather reports, warnings, and climate summaries.
Scientific institutes	ESSL, ICIMOD, NGI/NGU, ICL, OSTI	Hail, glacier, landslide, heat, and other specialist event accounts.

Meteorology, climate, hydrology, ocean, and air quality — 13 families

Provider or product	Access form	Role in a package
ERA5-Land	hourly reanalysis; package-local aggregates/stacks	Temperature, precipitation, wind, and land-surface background.
CHIRPS v2	daily precipitation archive	Event accumulation, wet spells, climatology, and rainfall anomalies.
NASA GPM IMERG V07	Earthdata CMR discovery and precipitation granules	Satellite precipitation totals, timing, and spatial structure.
Open-Meteo Archive	historical hourly/daily REST API	Point temperature, humidity, apparent temperature, rain, snow, and wind.
Open-Meteo Air Quality	historical air-quality REST API	$PM_{2.5}$, PM_{10}, aerosol, and related atmospheric fields.
NASA POWER	daily point REST API	Temperature, corrected precipitation, and 10-m wind cross-checks.
NOAA PSL climate indices	download tables for ONI, SOI, and related modes	Large-scale climate-state and teleconnection context.

PACKAGE SOURCES Distinct source and access families across 199 event packages *continued*

NOAA CPC ONI	ASCII and tabular index archive	Operational Oceanic Niño Index values and episode timing.
IRI and Australian BoM	ENSO monitoring archives	Independent ENSO phase and coupled-ocean background.
GPCC	precipitation-anomaly products	Gauge-based anomaly context alongside CHIRPS.
NOAA OISST v2 high resolution	PSL THREDDS/ERDDAP catalog	Sea-surface temperature and marine-heatwave context.
Indian Ocean Dipole index	DMI archive	Indian Ocean climate-mode state for regional events.
NOAA ocean services	tide gauges and Ocean Prediction Center analyses	Coastal water levels, surge context, and marine storm structure.

Remote sensing and geospatial context — 12 families

Provider or product	Access form	Role in a package
NASA Worldview/GIBS	snapshot API and MODIS/VIIRS browse imagery	Pre-event and in-event true-color or thematic visual context.
MODIS/VIIRS vegetation products	Worldview or archive previews	Vegetation condition and drought-stress context.
MODIS/VIIRS MAIAC AOD	aerosol and smoke previews	Smoke transport and aerosol loading context.
Sentinel-1 GRD	VV pre/post imagery and change products	Flood, deformation, and surface-change comparison.
Sentinel-2 SR Harmonized	surface reflectance and dNBR derivatives	Burn severity and optical before/after consistency.
Google Satellite Embedding / AlphaEarth	annual 64-band embeddings	Annual state vectors, cosine-change statistics, and change rasters.
Copernicus EMS Rapid Mapping	EMSR/GFM maps and archive products	Mapped flood, fire, earthquake, and exposure footprints.
Microsoft Planetary Computer	STAC metadata search	AOI/date discovery for registered Earth-observation assets.
Natural Earth	Admin-0 and physical basemap files	Low-resolution geopolitical and physical reference layers.
geoBoundaries	administrative-boundary API	Open administrative geometries for package AOIs.
GHSL	built-up and settlement layers	Urban extent and settlement structure.
Copernicus DEM and SRTM	elevation and derived slope	Terrain constraints for flood, landslide, and glacier analyses.

Exposure, infrastructure, and human impact — 9 families

Provider or product	Access form	Role in a package
WorldPop GP 100 m	population grids and AOI summaries	Population exposure, density, and resampled 1-km context.
OpenStreetMap Nominatim	geocoding REST API	Event location, fallback coordinates, and AOI checks.
OpenStreetMap Overpass	bounded feature-query API	Roads, buildings, services, shelters, and critical amenities.
OpenStreetMap Geofabrik	regional extracts	Broader road and infrastructure context where bounded queries are insufficient.
FEWS NET and IPC	food-security products and reports	Drought exposure, food-security phases, and response pressure.
NYC Open Data	Socrata 311 service-request API	Flood complaints and city-service stress during urban rainfall.
Harris County and Houston open data	flood-warning and 311 interfaces	Gauge conditions, complaints, and local flood impacts.
Hong Kong public feeds	warnings, incident, and transport records	Local disruption and response evidence for typhoon events.
Brazil CEMADEN and civil defence	alerts and municipal reports	Rainfall-triggered hazard warnings and local consequences.

PACKAGE SOURCES Distinct source and access families across 199 event packages *continued*

Hazard-specific scientific and operational sources — 17 families

Provider or product	Access form	Role in a package
USGS ComCat	earthquake-event REST API	Origin time, magnitude, depth, location, and event products.
USGS ShakeMap and PAGER	event-product feeds	Instrumental intensity, shaking footprint, and loss context.
USGS Did You Feel It?	community-intensity product	Reported shaking distribution and populated-area impacts.
NOAA NCEI/WDS	Global Historical Tsunami Database	Tsunami source, run-up, and historical event records.
NOAA NHC	Tropical Cyclone Reports archive	Track, intensity, rainfall, surge, and impact summaries.
NOAA HRD H*Wind	surface-wind analysis archive	Storm wind structure and peak-wind calibration.
NOAA Coral Reef Watch	5-km thermal-stress products	HotSpot, Degree Heating Week, and bleaching-risk context.
Copernicus GloFAS/GFM/EFAS	flood-monitoring and event products	River-flood extent, timing, and European flood context.
Dartmouth Flood Observatory	Global Flood Monitor/archive	Independent flood occurrence and extent context.
Smithsonian GVP	volcano reports and event pages	Eruption chronology, volcano status, and physical interpretation.
Volcanic Ash Advisory Centres	operational ash advisories	Ash extent, altitude, movement, and aviation context.
GLIMS and Randolph Glacier Inventory	glacier outlines and catalog interfaces	Glacier geometry and inventory context.
NSIDC and ICIMOD	glacier and cryosphere reports	Regional ice change and glacier-hazard evidence.
NASA Global Landslide Catalog / COOLR	landslide catalog search	Occurrence records and landslide-event cross-checks.
USGS landslide releases	event pages, publications, and data releases	Site evidence for landslides and landslide-generated tsunamis.
ESSL and IEM/NWS archives	severe-weather summaries and Local Storm Reports	Hail, wind, tornado, and convective-event observations.
Australian fire and smoke products	official extent and smoke records	Fire perimeter, smoke, and air-quality context.

Package-local derivatives retained for computation — 9 derived families

Provider or product	Access form	Role in a package
Compact event AOI	event-bounding-box derivative	Common spatial support for package layers.
ERA5-Land aggregate and stack	hourly statistics and raster bundle	Compact numerical access without replacing the upstream provenance.
GPM and CHIRPS event accumulations	derived totals and AOI rasters	Comparable precipitation ledgers across satellite and gauge products.
WorldPop summaries	AOI sum and 1-km resample	Deterministic population exposure inputs.
Sentinel-1 change	pre/post statistics and raster	Auditable radar change metrics.
Sentinel-2 dNBR	summary and raster	Auditable burn-severity calculation.
Satellite-embedding change	annual statistics and raster	Normalized annual surface-change evidence.
Bounded OSM slice	package-local JSON	Stable infrastructure and service counts.
Locked event anchor	package-local report copy	Fixed event identity, time window, and provenance for replay.

The inventory is exhaustive at this level of grouping, although each event uses only a hazard- and availability-specific subset. The 80 external rows cover distinguishable upstream sources and access methods; the nine derived rows record transformations reused across packages. Per-package manifests retain the raw label, exact URL, acquisition status, package-relative output, byte count, and checksum for record-level provenance.

C.4 Scoring, isolation, and audit protocol

The judge receives the task, structured ground truth, final answer, reference process, rubric, and trajectory. It scores required answer units separately from evidence selection, calculation, source reconciliation, and mechanism analysis. Combined Core is the mean of these blocks, and Strict@95 records near-complete tasks. Capability columns use overlapping expert tags as diagnostic slices, not additive categories. Controlled comparisons use paired manifests and task-level bootstrap resampling.

Solvers may list, read, search, and compute only within the active event package. The reference solution, rubric, structured ground truth, and ground-truth program remain hidden. Tools record each path and operation before shortening observations, which preserves provenance through context compression. Finalization receives the accumulated research state and required schema but no answer-bearing hint. Released traces cannot reconstruct provider-hidden reasoning, proprietary routing, or undocumented cache accounting.

C.5 Controlled-study sampling design

The reasoning-effort study uses a fixed manifest containing historically difficult tasks and a fixed-seed random draw from the rest of the release, preferring distinct events. Tool, routing, context, and interaction studies use a smaller paired manifest drawn from it. Oracle and robustness conditions reuse the same task IDs and Top-20 router outputs. Paired 95% intervals use 5,000 bootstrap resamples of task-level score differences with seed 2026.

Professional-environment conditions use GPT-5.5 with the default medium reasoning setting and 48 turns. Core libraries retain the paired manifest; input-dependent environments draw a difficulty-balanced compatible manifest with seed 2026. Their compliance contract requires at least two successful, distinct specialist calls, targets three, and counts at most four. These paired ablations are diagnostic and should not be read as exact full-leaderboard effects.

Artifacts retained for audit.

- **Task identity.** Version, manifest, package checksum, and question reproduce the scientific input.
- **System identity.** Endpoint revision, controller, tool profile, and context policy reproduce the executable system.
- **Research trace.** Actions, arguments, observations, errors, and timing reconstruct the investigation, including failed calls.
- **Submission.** Raw text, parsed object, repair attempts, and final status keep malformed or missing answers in the denominator.
- **Reference and judgment.** Ground truth, solution, computation, rubric, judge prompt, and criterion decisions support answer-unit and aggregate recomputation.
- **Paired comparison.** Task IDs, intervention, seed, and bootstrap resamples identify the controlled change.

C.6 Human evaluation and agent-assisted review

Human review begins when a task is conceived and continues through release. The team included six doctoral-level experts in Earth science and twenty undergraduates in related fields. Doctoral reviewers set the scientific direction, checked physical validity, and approved the final scoring specification. Undergraduate reviewers followed a common protocol for source and metadata checks, package replay, calculation verification, and secondary review of answer units. Reviewers contributed more than 50 hours per person on average.

Task direction and scoring checkpoints. For every task, a doctoral-level reviewer first analyzed the event and specified the scientific question, hazard process, spatial and temporal support, admissible evidence, required calculations, and answer fields. LLM agents then explored the package for plausible source combinations, alternative calculations, ambiguities, and likely failure modes. Reviewers used these passes to refine the reference analysis and define the final answer units and

process checks. Agents proposed candidates; human experts decided which checks were scientifically necessary, supported by the package, and suitable for scoring.

Joint verification and correction. Every candidate task underwent agent replay followed by human inspection. Reviewers confirmed that package evidence could answer the public question, that `compute_gt.py` regenerated the structured targets, and that the solution, answer units, and rubric referred to the same quantities and physical claims. Any mismatch returned the task to construction. A small number of sources had minor spatial offsets between the data footprint and the documented disaster location. Human experts removed the affected answer units rather than score a quantity the package could not support.

Iteration scale. For difficult cases, constructing and reviewing a single question consumed more than 20 million tokens across repeated agent passes. These passes covered evidence exploration, draft construction, calculation checks, replay, discrepancy analysis, and revision. Human reviewers used them to align the scientific question, evidence, expert solution, deterministic program, answer units, and process rubric before release.

C.7 Benchmark construction and release gates

Each construction stage produces an auditable artifact and an explicit release decision. A failed gate returns the task to the responsible stage.

Seven release gates.

1. **Event anchoring:** verify identity, hazard type, time window, spatial scale, and event–source agreement.
2. **Package assembly:** build a typed relative-path manifest and confirm that the required evidence roles are readable.
3. **Task authoring:** require package-local resolution without revealing a file path or answer-bearing value.
4. **Reference analysis:** align claims, units, mechanisms, and defensible alternatives; trace every load-bearing claim.
5. **Executable grounding:** regenerate structured ground truth deterministically and pass schema and tolerance checks.
6. **Rubric design:** map the task obligations to five to eight criteria totaling exactly 20 points.
7. **Final review:** align the question, solution, ground truth, program, and rubric before release approval.

D Extended Diagnostics

Table 8 breaks Combined Core down by seven overlapping capability tags. Because a task may carry several tags, the columns are diagnostic views rather than additive partitions.

The strongest systems lead across nearly every tag, but their residual errors concentrate in the connections between results. Among the top five, causal-chain scores trail quantitative calculation by 4.9–7.9 points. A single defensible data path may be enough to compute a total or ratio. An evolving hazard account is harder to stabilize: rainfall, exposure, damage, and response evidence must refer to compatible windows and scales, and the proposed mechanism must survive their disagreement. GeoMMAgent breaks the otherwise stable ordering. Its short, structured search works well on tabular evidence but often bypasses raster intermediates, leaving remote sensing at 58.82. Capability scores thus capture the organization of an investigation as much as the knowledge of its backbone.

D.1 Complete reasoning and evidence-intervention results

Table 7 expands the main-text summary in Table 3. It retains every outcome, interaction, evidence, resource, and capability column, including tool presentation, evidence perturbations, and optional

meteorological access. The following sections provide the corresponding trajectory and error diagnostics.

D.2 Capability-conditioned performance

Table 8: **Core by capability.** Tags overlap; system groups follow Table 2. **Bold blue** marks column bests.

System	Earth-process analysis			Evidence synthesis			
	Physical mech.	Spatio-temporal	Quant. calc.	Multi-source	Causal chain	Ranking	RS / geo.
Agent frameworks							
GeoMMAgent	67.72	72.94	68.85	70.56	66.15	67.00	58.82
OpenResearcher	30.46	35.24	31.35	34.63	25.35	30.41	33.18
Open-weight and Earth-specialized models							
Qwen3-235B	31.90	40.35	34.03	35.51	29.00	29.29	32.32
GeohazardGPT	31.77	31.31	28.04	30.42	32.12	39.04	29.16
Qwen3-4B	24.63	22.71	24.70	24.32	25.89	24.06	32.61
GeoGPT	25.34	25.70	24.66	23.14	23.63	26.28	22.77
MiroThinker 8B	13.51	14.39	14.68	14.52	8.61	6.76	21.96
Nemotron Nano	11.79	14.41	12.30	17.89	12.42	13.26	15.48
Intern-S1-mini	12.04	10.70	11.99	10.45	10.57	13.21	16.50
Hosted general-purpose systems							
Claude Fable 5	**85.15**	**85.51**	**87.14**	**83.45**	**82.20**	**80.98**	**89.99**
GPT-5.6 Sol	81.81	83.31	85.42	82.13	77.54	71.63	88.50
GPT-5.6 Terra	80.83	79.83	82.96	79.06	76.26	76.12	86.27
GLM-5.2	80.25	77.30	80.24	76.85	80.08	78.34	79.27
GPT-5.5	77.21	78.38	81.91	76.79	74.33	75.76	83.77
Claude Sonnet 4.6	75.49	73.60	78.36	77.92	73.43	79.73	75.93
Kimi K2.5	75.07	79.10	77.81	75.34	70.48	70.82	77.94
DeepSeek V4 Flash	73.63	76.33	75.16	72.59	68.91	72.33	77.75
Tencent Hy3	72.67	75.92	73.71	70.32	69.72	70.23	79.41
Claude Haiku 4.5	69.60	72.44	76.09	73.50	67.73	66.76	79.66
Qwen 3.7 Plus	70.41	72.40	71.13	69.27	67.19	68.08	78.26
Doubao Seed 2.0 Mini	63.24	66.43	62.67	59.30	60.42	63.51	70.67
Gemini 3.1 Flash Lite	60.18	61.40	60.38	56.99	56.37	58.35	68.49
MiniMax M2.7	42.76	52.91	48.80	49.53	39.99	24.21	42.28
GPT-4o	39.97	38.86	38.77	34.93	37.56	43.70	48.30
GPT-5.6 Luna	34.74	42.91	38.58	36.48	31.77	21.17	38.21

D.3 Preference structure and diagnostic error paths

Pairwise judgments broadly recover the leaderboard order while exposing uncertainty hidden by Core alone. Figure 7 separates wins, losses, and ties in opponent-balanced comparisons. Leaders win consistently, adjacent frontier systems remain close, and ties are uncommon. The drop beyond the leading group agrees with low Strict@95: partial success is common, while consistently complete research remains rare.

> **Position.** Reliability in dynamic Earth systems depends on revising the claim–evidence state as observations change. A system must detect conflicts, isolate local errors, and carry corrections into derived quantities, causal interpretations, and the final answer. Systems differ not only in how often they err, but also in whether those errors are found, contained, and repaired. Longer traces, more tool calls, and higher execution success do not measure that capacity (Chen et al., 2025; Du et al., 2026).

Figure 8 shows the same separation at task level. Frontier models have a dense high-score mode but retain a long tail toward zero. Agent frameworks cluster at moderately high scores, while open-weight and Earth-specialized systems shift downward and rarely approach complete solutions. Error counts overlap far more than Core scores. Strong systems still make mistakes; their advantage lies in catching or containing them before the final answer. Error frequency therefore describes the research process, while Core records the damage that remains.

Table 7: **Reasoning and evidence interventions with GPT-5.5.** (a) Four access modes and five effort settings; resources are per-task means, and interaction/evidence diagnostics apply to interactive runs. (b) Paired interventions relative to the Top-20 router; intervals are paired 95% bootstrap CIs.

(a) Reasoning effort and evidence access

	Outcome				Interaction		Evidence		Resources	
Effort	Answer	Process	Core	Unit acc.	Rounds	Calls	Files	Refs	Tokens (k)	Latency (s)
Compressed direct										
none	**41.37**	**45.10**	**43.24**	**42.21**	–	–	–	–	13.9	55.6
low	24.99	29.32	27.16	25.43	–	–	–	–	14.9	120.5
medium	22.88	25.69	24.29	23.26	–	–	–	–	15.8	152.4
high	24.17	28.17	26.17	24.28	–	–	–	–	16.6	71.5
xhigh	16.67	16.67	16.67	16.67	–	–	–	–	17.6	88.1
All-text direct										
none	**27.66**	**31.22**	**29.44**	**28.27**	–	–	–	–	81.6	32.1
low	13.26	16.20	14.73	13.36	–	–	–	–	82.7	44.1
medium	9.79	12.18	10.99	10.05	–	–	–	–	83.5	61.0
high	8.80	9.00	8.90	8.82	–	–	–	–	85.4	78.5
xhigh	12.00	15.00	13.50	12.00	–	–	–	–	87.7	94.4
EarthVerse interactive										
none	48.89	53.70	51.30	49.46	9.68	8.68	7.00	9.94	96.9	140
low	56.97	60.05	58.51	57.53	14.70	13.70	9.44	14.70	162.4	259
medium	61.45	64.80	63.13	62.35	17.20	16.20	10.71	18.12	206.4	424
high	**72.41**	**75.19**	**73.80**	**72.59**	14.96	13.96	9.92	17.19	174.3	571
xhigh	69.66	67.50	68.58	69.83	11.25	10.25	10.25	16.50	100.0	674
Interactive with final review										
none	53.70	56.48	55.09	54.06	11.62	9.62	7.42	10.46	130.7	108
low	61.93	65.50	63.72	62.78	18.18	16.18	10.60	17.02	227.0	253
medium	67.16	69.22	68.19	67.91	20.81	18.81	11.88	21.02	266.3	552
high	69.82	74.71	72.27	70.73	20.63	18.63	12.13	21.10	255.1	1,105
xhigh	**84.36**	**88.21**	**86.29**	**84.40**	17.57	15.57	10.14	18.29	217.7	1,493

(b) Evidence-path interventions

	Outcome			Paired effect		Core by capability						
Intervention	Answer	Process	Core	Δ Core	95% CI	Physical mech.	Spatio-temporal	Quant. calc.	Multi-source	Causal chain	Ranking	RS / geo.
Top-20 semantic-router baseline	66.90	65.88	66.39	0.00	–	65.15	64.40	73.43	90.00	65.89	50.84	53.94
Answer-free information oracles												
Relevant-file oracle	**79.46**	**82.75**	**81.11**	**+14.72**	[+5.48, +24.69]	80.61	91.47	87.57	90.45	76.53	60.70	72.09
Evidence-map oracle	76.96	78.50	77.73	+11.34	[+2.34, +21.25]	76.69	90.50	83.83	97.50	73.35	58.16	61.25
Tool-subset oracle	68.53	68.62	68.58	+2.19	[−3.55, +7.28]	67.32	65.67	72.60	92.50	69.54	59.14	59.75
Planning oracle	65.19	66.50	65.84	−0.54	[−7.07, +4.35]	64.31	57.10	71.24	95.00	67.93	54.76	60.25
Tool-interface presentation												
Hierarchical catalog	66.76	67.62	67.19	+0.80	[−9.18, +11.86]	65.81	66.72	67.83	93.50	70.78	56.60	61.67
Descriptive tool names	66.88	66.75	66.82	+0.43	[−3.55, +4.05]	65.07	61.21	72.92	100.00	66.98	57.59	55.23
Dynamic descriptions	61.39	61.75	61.57	−4.82	[−15.02, +4.33]	59.81	61.56	63.81	95.00	61.27	54.58	60.75
Opaque tool names	56.72	59.50	58.11	−8.28	[−23.57, +6.33]	56.16	48.88	58.54	95.22	59.76	54.01	80.00
Evidence perturbations												
Equivalent unit change	67.61	71.12	69.37	+2.98	[−9.75, +15.53]	68.02	66.93	71.44	95.00	71.29	55.47	76.38
Added distractor files	65.91	67.75	66.83	+0.44	[−6.80, +7.27]	65.46	60.47	72.15	92.95	65.16	59.98	70.62
Missing evidence	63.97	64.75	64.36	−2.03	[−12.05, +7.08]	62.49	60.31	67.82	100.00	66.63	54.74	55.12
Truncated evidence	63.17	63.38	63.27	−3.12	[−14.13, +7.47]	61.78	55.67	70.53	91.70	64.24	41.52	68.92
Conflicting evidence	63.03	63.12	63.08	−3.31	[−12.05, +2.97]	61.50	67.97	66.63	92.95	60.46	55.19	54.50
Optional meteorological access												
Compact meteorology suite	62.69	65.38	64.03	−2.36	[−11.92, +5.89]	62.14	61.97	66.67	100.00	65.22	56.15	58.47
Cue-routed meteorology	57.22	59.88	58.55	−7.84	[−19.40, +3.93]	56.74	58.87	59.25	92.95	60.28	50.45	57.25

The family-level flow in Figure 9 sharpens this result. Computation and answer errors occur in every family, yet their consequences differ. Frontier systems more often keep affected tasks in the 40–70 or above-70 bands; open-weight and Earth-specialized systems more often fall below 40. Recovery, cross-source checking, and final synthesis determine whether a local mistake remains local or spreads through the answer.

Figure 10 examines the same pattern within systems. The largest Core losses accompany premature stopping, failed recovery, tool failure, and missing required outputs—failures that interrupt the

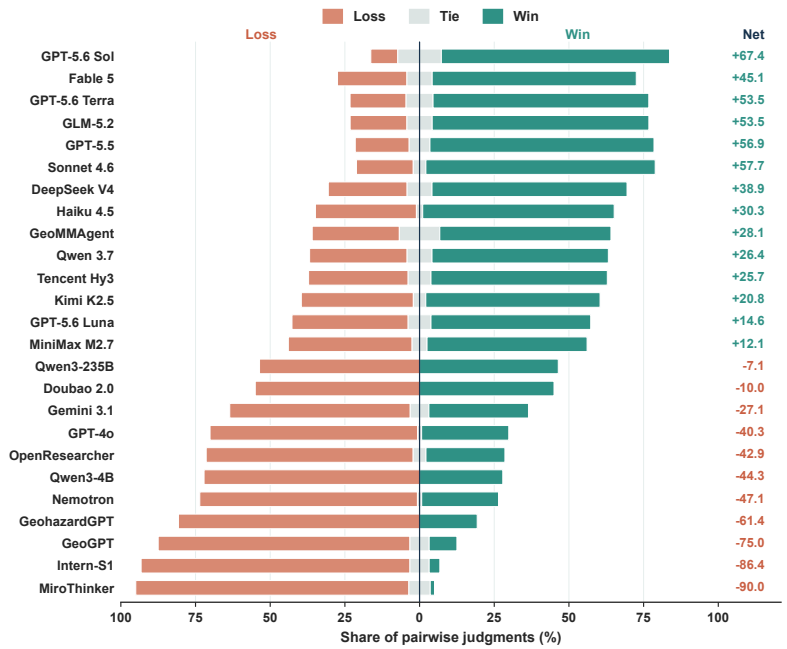


Figure 7: Pairwise win, tie, and loss shares. Net is win share minus loss share.

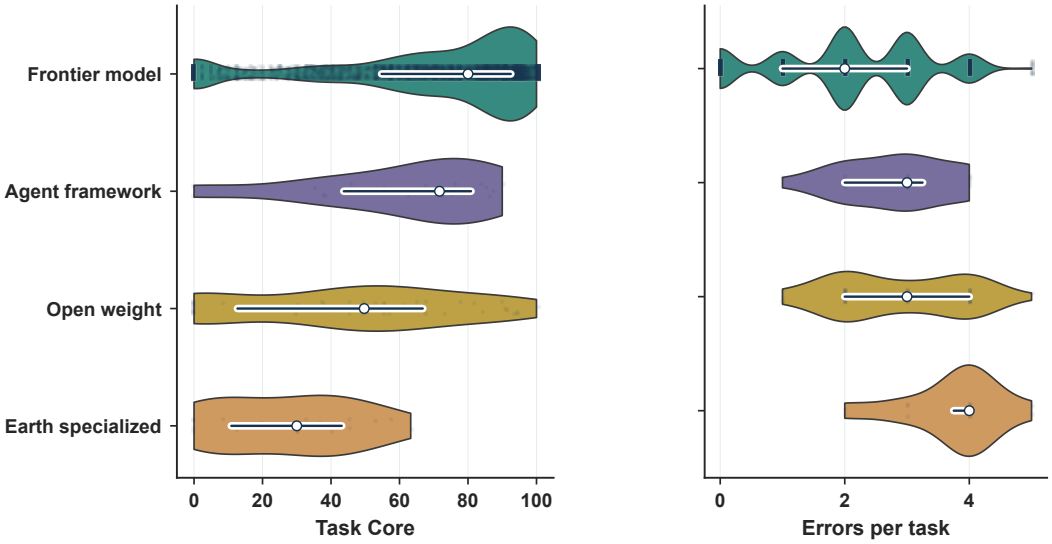


Figure 8: Task Core and diagnosed errors by system family.

research loop itself. Numerical and schema errors are common, but their average effect is smaller because some runs catch or absorb them before finalization. The inset connects missing evidence, output omissions, schema mistakes, and calculation errors into one cluster. These are often successive stages of a failed investigation, not independent defects.

D.4 Reasoning modes and resource accounting

Tables 9–11 report trajectory, latency, token use, and evaluation cost. They separate interactive retrieval from direct serialization and model activity from successful evidence acquisition.

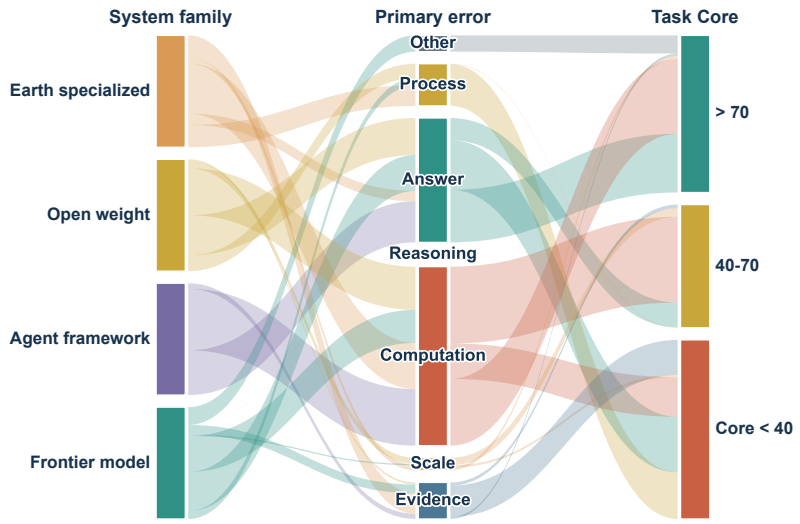


Figure 9: System family, primary error family, and task-level Core band.

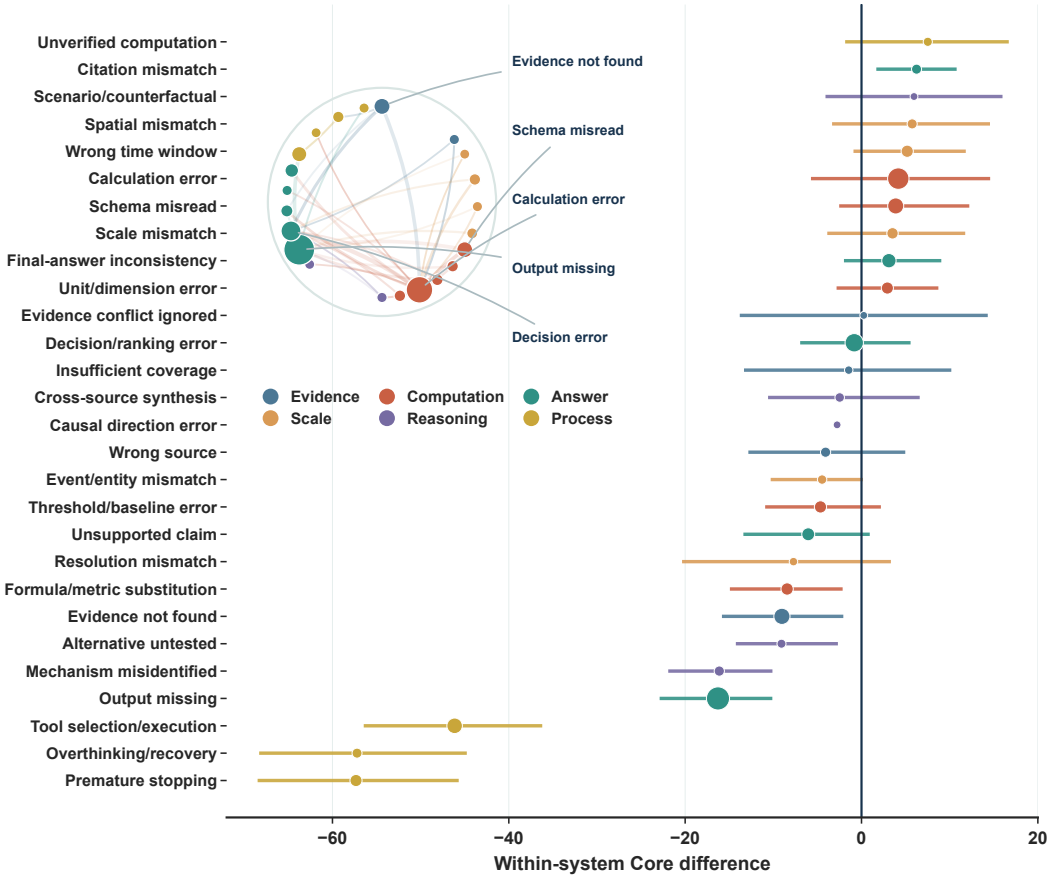


Figure 10: Within-system Core differences and the strongest error co-occurrences.

D.4.1 Interactive trajectory diagnostics

Higher reasoning effort changes the treatment of evidence more than the amount collected. At xhigh, the EarthVerse harness opens about as many files as at medium effort and uses fewer rounds, yet takes almost twice as long. The additional work occurs between observations: the model compares

interpretations, revisits calculations, and decides whether its account is complete. Explicit review intensifies this pattern. From harness-high to review-xhigh, output nearly doubles while file reads barely move. Some of that deliberation repairs omissions; some only restates a settled path. Rounds, tokens, and latency therefore measure different kinds of work, and none is a reliable stand-in for scientific progress.

D.4.2 Direct package-serialization diagnostics

Direct conditions receive a fixed package serialization before generation. Compressed direct includes more files in far fewer tokens than all-text, yet neither mode lets the model return to a source after discovering that a window, unit, or variable was wrong. That restriction changes the task. The model can reason about the supplied snapshot, but it cannot conduct the next measurement implied by its own reasoning. Larger inputs and longer outputs fail to rescue these baselines because the evidence schedule is fixed before the analysis begins. The missing operation is revision of the observation set, not another pass over the same text.

D.4.3 Evaluation cost

Table 11 reports the cost of the evaluated runs for models with a comparable USD list price.

Quality and cost do not form a simple ladder. The strongest frontier evaluations exceed $2,500, while GPT-5.6 Terra retains about 98% of GPT-5.6 Sol's Core at 44% of its cost. The two Flash systems remain below $45 but differ by almost 14 Core points. A long trace from a low-price model may still cost less than a short frontier run, so quality and cost should be reported together when planning replication.

D.5 Tool, memory, and interaction ablations

Table 5 reports outcome and capability scores for each ablation; Table 12 shows how the same runs searched, computed, and consumed context. Token use is descriptive and does not affect scoring.

D.5.1 Research behavior and context consumption

The trajectory diagnostics expose two different kinds of false productivity. Under aggressive compression, the agent lists files more than four times as often as with long retention and records more evidence references, yet receives the lowest process score. It is rebuilding context that the controller discarded, not broadening the analysis. The expert-only suite looks efficient for the opposite reason: it has the highest call success while inspecting the fewest distinct files. Its tools run cleanly, but the investigation is narrow. Activity counts become meaningful only when tied to state changes: a new source should resolve a claim, a computation should test a mechanism, and a repeated read should have a reason.

D.5.2 Capability-conditioned effects

The largest capability losses occur when evidence must remain connected across several steps. With 170 visible tools, multi-source performance falls to 47.50 even though most other catalog configurations stay near or above 88. Aggressive compression loses 20.07 points in causal-chain reasoning relative to long retention. The two interventions damage continuity in different places. A large catalog makes the next action unstable; compression makes the meaning of earlier actions unstable. A single calculation can survive either disturbance because its inputs are local. A physical account cannot: its validity depends on remembering why several observations belong together.

D.6 Oracle, interface, and evidence robustness

Table 7(b) consolidates outcome scores, paired uncertainty, and capability slices for each intervention. Table 13 separately reports trajectory and resource effects. Confidence intervals resample paired task-level Core differences. Oracles provide guidance without answers, while robustness conditions alter only package evidence.

Table 9: **Interactive trajectory diagnostics.** Per-task means; call success is the share of tool calls that return normally.

		Execution			Evidence			Resources				
Mode	Effort	Rounds	Tool calls	Success (%)	Reads	Files	Evidence refs.	Input (k)	Output (k)	Total (k)	Cached (k)	Latency (s)
EarthVerse	none	9.68	8.68	92.08	3.84	7.00	9.94	94.3	2.6	96.9	10.6	140
	low	14.70	13.70	93.67	5.64	9.44	14.70	157.4	5.1	162.4	15.4	259
	medium	17.20	16.20	95.39	6.84	10.71	18.12	198.6	7.9	206.4	17.9	424
	high	14.96	13.96	97.19	5.15	9.92	17.19	164.8	9.5	174.3	13.2	571
	xhigh	11.25	10.25	100.00	4.75	10.25	16.50	90.6	9.4	100.0	15.9	674
Review	none	11.62	9.62	94.26	4.96	7.42	10.46	126.9	3.8	130.7	18.8	108
	low	18.18	16.18	94.94	6.84	10.60	17.02	219.0	8.0	227.0	25.1	253
	medium	20.81	18.81	94.78	6.56	11.88	21.02	253.8	12.5	266.3	25.2	552
	high	20.63	18.63	95.37	6.50	12.13	21.10	239.2	15.9	255.1	22.2	1,105
	xhigh	17.57	15.57	99.29	5.86	10.14	18.29	198.9	18.8	217.7	25.6	1,493

Table 10: **Direct-serialization resources.** Direct modes inject a fixed file set before generation; values are per-task means.

Mode	Effort	Injected files	Evidence refs.	Input	Output	Total	Latency (s)
Compressed	none	17.92	24.56	13,016	878	13,894	55.61
	low	17.92	25.31	13,033	1,857	14,890	120.45
	medium	17.92	25.72	12,838	3,004	15,842	152.36
	high	17.93	26.73	12,958	3,630	16,588	71.47
	xhigh	17.67	25.00	12,878	4,726	17,604	88.14
All text	none	7.57	13.89	80,842	751	81,593	32.06
	low	7.68	14.00	80,886	1,854	82,739	44.10
	medium	7.46	13.90	80,941	2,557	83,498	61.02
	high	8.73	15.13	81,623	3,735	85,359	78.49
	xhigh	8.00	10.00	83,655	4,013	87,668	94.38

Table 11: **Evaluation cost.** Rows descend by total cost.

System	Core	Tokens/task (k)	Price (USD/M)	Cost (USD)
Claude Fable 5	84.97	161.6	50.00	3,271.57
GPT-5.5	78.06	224.9	30.00	2,732.09
GPT-5.6 Sol	81.68	205.9	30.00	2,501.58
Claude Sonnet 4.6	76.16	331.0	15.00	2,011.07
GPT-5.6 Terra	80.06	182.6	15.00	1,109.03
Claude Haiku 4.5	72.23	477.0	5.00	966.02
GPT-4o	39.23	127.4	10.00	515.92
GPT-5.6 Luna	35.66	156.9	6.00	381.31
DeepSeek V4 Flash	73.54	389.7	0.28	44.19
Gemini 3.1 Flash Lite	59.66	70.1	1.50	42.61

D.6.1 PAIRED INTERVENTION UNCERTAINTY

Only the two evidence-localization oracles have intervals that exclude zero: relevant files [+5.48, +24.69] and the evidence map [+2.34, +21.25]. The stable gain comes from knowing where evidence is, not from receiving a generic plan. A plan names familiar analytical moves; localization resolves an event-specific uncertainty that scientific knowledge alone cannot settle. The other interventions

Table 12: **System-ablation trajectories.** Per-task means for the panels in Table 5.

Condition	Tools			Evidence		Actions			Process	Tokens			
	Calls OK	Calls failed	Success (%)	Files	Evidence refs.	Lists	Searches	Python	Process (%)	Input	Output	Total	Cached
(a) Tool-catalog scale													
Fixed tools (6)	15.80	2.15	92.52	12.75	19.80	1.95	4.45	5.45	65.62	219,639	9,918	229,557	25,882
Fixed tools (10)	15.25	1.75	92.22	11.05	17.40	1.30	3.30	3.35	57.10	185,888	8,159	194,047	25,408
Fixed tools (20)	15.55	2.05	91.74	12.05	17.15	1.45	3.10	4.60	59.38	223,666	9,246	232,912	25,267
Fixed tools (40)	15.75	1.95	92.81	11.95	19.70	1.55	4.15	4.10	61.00	224,175	8,818	232,993	24,346
Fixed tools (70)	15.00	1.90	92.39	11.40	18.35	1.25	3.35	2.85	61.25	224,853	8,906	233,759	26,445
Fixed tools (100)	16.35	2.00	92.54	12.00	20.25	1.10	3.90	3.60	63.12	268,527	10,409	278,936	32,294
Full tool catalog (170)	13.90	1.50	95.04	9.55	13.15	1.05	1.85	2.55	55.88	264,983	8,233	273,216	33,331
(b) Semantic routing and tool composition													
Top-10 dynamic router	14.65	2.15	93.16	12.40	18.65	1.40	3.50	4.70	66.25	205,908	10,319	216,227	29,286
Top-20 dynamic router	12.75	1.65	92.98	11.30	16.15	1.35	2.60	4.05	65.88	170,972	9,038	180,010	21,696
Workflow-aware router	15.20	1.95	93.38	12.25	21.65	1.40	4.65	3.45	56.75	210,928	9,304	220,232	26,381
Expert-only suite	17.70	0.60	97.59	7.35	12.60	0.00	0.00	0.00	60.88	283,502	11,988	295,490	24,192
General + expert suite	14.05	1.55	94.43	10.40	15.40	1.40	3.05	3.45	63.00	186,216	8,875	195,092	31,437
Atmospheric Python suite	14.60	1.35	94.78	10.35	18.40	1.45	4.10	2.85	60.12	191,208	7,593	198,802	25,344
Atmospheric Fortran suite	14.55	1.85	93.68	11.25	17.10	1.55	3.35	3.30	57.88	195,439	9,823	205,262	31,078
(c) Context retention													
Aggressive compression	24.60	1.95	93.88	12.50	34.90	5.40	6.25	4.00	51.62	187,492	9,545	197,038	10,765
Short retained context	20.20	2.25	93.64	13.75	26.80	2.70	6.40	5.10	56.38	198,062	11,765	209,827	14,682
Long retained context	13.35	1.85	92.77	11.80	16.95	1.20	3.70	4.45	63.38	218,347	9,189	227,536	48,998
Full retained context	12.15	1.30	93.84	11.50	18.00	1.05	3.90	3.40	58.25	180,517	7,182	187,700	80,819
(d) Interaction budget													
12-turn budget	9.95	0.30	97.36	8.50	13.05	1.15	2.70	1.30	49.50	101,826	4,437	106,263	23,002
24-turn budget	12.90	1.65	93.11	10.80	16.90	1.35	4.20	3.30	52.88	153,261	8,143	161,404	32,346
96-turn budget	13.60	1.30	94.12	11.90	18.45	1.60	3.85	4.55	60.12	173,925	8,873	182,798	27,763
Unbounded turn budget	15.20	1.80	94.22	12.30	21.40	1.60	5.15	4.35	64.50	211,751	9,429	221,180	30,131

remain suggestive because their intervals cross zero. On a diagnostic manifest, a few packages whose structure favors one interface can still move the mean substantially.

D.6.2 Trajectory and resource effects

Oracle guidance shortens search without necessarily reducing cost. The evidence map lowers rounds and file reads, but the injected structure itself consumes context and takes time to interpret. Distractors produce the reverse pattern: the agent performs more actions and records more evidence references with almost no quality loss. Obvious clutter is therefore mostly an efficiency tax. A compact evidence description poses a different problem. The agent must decide whether to trust the summary or reopen the underlying source, and the latter is often necessary to recover metadata, aggregation choices, or conflicts. Search length and evidential control are related, but they are not interchangeable.

D.6.3 Capability-conditioned intervention effects

The effective oracles help most with spatiotemporal reconstruction: both exceed 90, compared with 64.40 at baseline. File localization directly removes uncertainty about which records contain the event window and region. Ranking and decision improve less because the oracle does not say how competing quantities should be weighted. Opaque tool names damage the same capability most strongly. Reconstructing a window rarely depends on one call; it requires a sequence of filtering, reading, aggregation, and comparison. Losing the semantic role of any operation makes the resulting timeline harder to recover and harder to audit.

Table 13: **Intervention trajectories.** Panels match Table 7(b); values are per-task means.

	Execution			Evidence and compute				Latency	Tokens			
Intervention	Rounds	Tool calls	Success (%)	Reads	Files	Evidence refs.	Python	Seconds	Input	Output	Total	Cached
(a) Answer-free information oracles												
Relevant-file oracle	15.30	14.30	92.17	5.15	11.30	16.10	4.15	217.26	181,939	5,572	187,512	19,507
Evidence-map oracle	11.10	10.10	93.28	3.10	8.55	11.10	2.75	436.56	290,368	5,050	295,418	62,950
Tool-subset oracle	17.80	16.80	93.18	6.60	13.35	21.60	3.90	212.74	205,252	5,628	210,880	27,738
Planning oracle	16.95	15.95	91.49	5.45	12.65	18.30	4.05	185.18	196,489	5,817	202,306	25,190
(b) Tool-interface presentation												
Hierarchical tool catalog	16.60	15.60	91.97	5.75	12.45	17.75	3.85	197.77	182,350	6,000	188,351	20,416
Descriptive tool names	18.25	17.25	93.42	5.40	12.25	19.65	3.65	150.14	202,390	5,462	207,853	28,749
Dynamic tool descriptions	17.55	16.55	92.12	4.00	12.85	19.95	3.65	182.42	199,673	5,036	204,709	23,053
Opaque tool names	17.90	16.90	92.44	7.00	10.85	16.50	4.30	234.10	215,650	6,294	221,943	23,027
(c) Evidence perturbations												
Equivalent unit change	16.45	15.45	90.70	5.20	13.25	19.25	3.90	176.16	184,679	5,153	189,832	19,904
Added distractor files	19.60	18.60	89.97	6.20	13.10	22.50	4.15	162.54	220,098	5,658	225,756	25,792
Missing evidence	17.40	16.40	89.12	5.65	12.05	21.20	3.55	181.47	198,173	5,219	203,392	19,546
Truncated evidence	16.10	15.10	91.65	4.25	12.75	22.10	3.55	265.79	184,244	5,827	190,071	16,781
Conflicting evidence	16.50	15.50	92.04	5.65	12.55	19.70	3.25	155.59	182,815	5,875	188,689	18,624
(d) Optional meteorological access												
Compact meteorology suite	17.95	16.95	94.90	5.85	12.75	20.60	4.05	223.20	201,057	5,822	206,879	27,571
Cue-routed meteorology suite	15.65	14.65	92.50	5.65	11.50	17.85	3.15	238.22	170,431	5,875	176,306	25,882

D.7 Professional meteorological environments

The professional-environment study separates method availability from execution. Compliance records whether the required interactions occurred; call success records whether they completed without adapter, argument, or runtime failure.

D.7.1 Compliance, execution reliability, and implemented scope

Nearly every run attempts the required specialist interaction, so willingness to call the tool is not the bottleneck. Execution is. Specialist-call success ranges from 20.29% for PyET to 97.56% for R climate, and high-success environments usually need fewer repair rounds and smaller contexts. The ranking should not be read as a comparison of languages. It reflects the overlap between the audited operations and the data actually present in the packages. R climate, Julia, and CDO often accept the available time series or grids directly; other adapters require variables, record lengths, or sampling structures that many event packages do not contain.

Applicability and reproducibility. Environment names refer to the audited adapters in Table 14, not to every feature of the upstream packages. The outcome columns make the distinction explicit: R climate has the highest call success, while Julia has the highest Core. A successful call proves only that code ran. It does not establish that the estimator fits the record, that the sampling geometry supports the comparison, or that the returned value answers the question. Those checks sit between execution and scientific use, and the gap between execution reliability and final quality shows why they cannot be collapsed into one metric.

Each adapter therefore records normalized arguments, status, output, and errors. It also declares the operations it implements and validates package-scoped inputs before execution. This makes a justified failure reproducible: another reviewer can see that a method was rejected because the record was too short or a required variable was absent. Merely rerunning a command is weaker. Scientific reproducibility requires reconstructing why the operation was admissible and how its output entered the claim.

Table 14: **Scientific-environment outcomes and execution.** Final Answer, Process, and Core scores are reported with per-task compliance, successful specialist calls, resource use, and implemented operations.

		Outcome			Execution		Resources		
Environment	Operations	Answer	Process	Core	Comply (%)	Success (%)	Rounds	Tokens	Latency (s)
(a) Python scientific libraries									
MetPy	rain summary; profiles; kinematics	68.69	70.38	69.53	100	38.83	25.00	311,043	255.60
xclim	heat/rain extremes; wet/dry spells	67.77	69.62	68.70	95	38.78	24.50	299,074	268.44
climate-indices	SPI; SPEI; PET	64.50	66.75	65.62	95	22.60	30.55	374,199	306.69
PyET	Penman–Monteith; Hargreaves	65.34	67.38	66.36	95	20.29	32.30	387,917	315.96
thermofeel	UTCI; heat index; wind chill	66.78	68.75	67.76	100	26.85	28.55	346,924	357.88
pyextremes	block maxima; POT; return levels	59.02	60.00	59.51	100	34.19	28.05	363,839	289.16
xskillscore	RMSE; correlation; Brier; CRPS	63.37	64.12	63.75	95	57.14	23.70	297,724	231.47
(b) Compiled and domain-language environments									
Julia	summary; integration; sensitivity	**71.77**	70.25	**71.01**	100	91.11	21.50	256,852	204.47
R climate	summary; SPEI; GEV; trend	68.55	**71.38**	69.96	100	**97.56**	21.70	259,984	190.20
CDO	temporal and field aggregates	66.91	69.50	68.20	100	81.63	22.30	268,906	197.90
NCO	inspect; weighted mean; derive field	67.39	65.00	66.19	100	35.59	26.65	321,773	230.95
NCL	summary only	65.03	65.00	65.01	95	40.62	25.70	308,559	270.94
Fortran	rain; kinematics; moisture transport	64.46	64.25	64.35	95	43.48	25.70	301,236	222.60

E PROMPTS AND EXECUTION CONTRACTS

The prompts are grouped by role: package-scoped solving and review, answer–process evaluation, and benchmark construction. Placeholders mark values supplied by the harness. The solver sees no demonstration, hidden reference, or recommended file. Appendix G presents representative runs.

Runtime visibility. The solver sees the public question, event-package root, exposed tool schemas, and package-relative observations. Ground truth, reference solutions, rubrics, computation programs, judge records, and other model answers remain hidden. The trace retains both successful and failed tool calls. Depending on the condition, the controller may request two format repairs, run a bounded evidence review, require a specialist operation, or reserve the final call for a schema-valid submission. None of these transitions adds scientific evidence. Appendix C.2 records their limits.

PROMPT RUNTIME **Package-scoped research suite**

SOLVER `EarthVerse interactive solver`

```
SYSTEM
You are an autonomous extreme-event benchmark solver. You must solve the task
by planning, selecting local files, executing tools, reading observations, and
then producing a structured final answer. Use only local event-package evidence
returned by tools. Do not use web knowledge. Never read computed_gt.json,
solution_en.md, review.json, or compute_gt.py. Prefer deterministic Python
calculations when the task asks for metrics. You must make real tool calls
before finalizing. Every assistant message must follow one of the two response
contracts below.

TASK MESSAGE
Task id: <<TASK_ID>>
Event id: <<EVENT_ID>>
Question path: <<QUESTION_PATH>>
Event package path: <<PACKAGE_ROOT>>

Question:
<<PUBLIC_QUESTION>>

Evidence boundary:
Strict package-only mode: only the current question and event package evidence
are visible.

Tools:
<<REGISTERED_TOOL_CATALOGUE>>
```

PROMPT RUNTIME **Package-scoped research suite** *continued*

```
TOOL ACTION CONTRACT
Action        tool_call
Tool          one registered tool name
Arguments     only the arguments required by that tool
Reason        why this call is needed for the current analysis

FINAL ANSWER CONTRACT
Action        final_answer
Answer        the exact structure requested by the public question
Confidence    one numeric confidence value
Evidence      for each claim: package-relative source file and support type
              (direct metric, report claim, metadata, image context, table
              value, derived metric, or unsupported boundary)
Research      a brief ordered summary of the reasoning
Limitations   important unsupported or unresolved elements

Start by inspecting the package inventory, then choose only the files needed
for this event-specific disaster analysis.
```

REVIEW **Final evidence review**

```
Begin a final evidence-review phase before submitting the answer. Treat the
prior answer as a draft: re-check its evidence provenance, formulas, units,
intermediate values, competing physical explanations, and every required
output field against the package. Use additional tools when verification is
useful. Return either one tool action or a revised final answer; do not merely
endorse the draft. You have at most <<MAX_REVIEW_TURNS>> review turns.
```

PROMPT EVALUATION **Joint answer and process judge**

JUDGE **Answer and research-process evaluation**

```
SYSTEM
You are a strict but fair evaluator for a local extreme-event benchmark. Your
job is to judge answer correctness and task-rubric quality. Use partial 0-100
scores and avoid old deterministic exact matching.

CALIBRATION CORE
Evaluate this model response for an extreme-event evidence-and-reasoning
benchmark. Make one combined evaluation and produce two core 0-100 scores.

- answer_correctness_score is the direct partial-credit comparison between the
  model's final answer and the ground truth. Treat the answer as a set of units:
  required fields, labels, decisions, values, units, orderings, alternatives,
  stage boundaries, and required conclusions. Missing required fields are
  incorrect; correct prose cannot replace a missing structured field.
- llm_rubric_score must come from the task's own 20-point rubric. Compare the
  answer and visible trace against every criterion, assign earned points out of
  20, and convert the result to percent.
- Do not invent extra credit. A plausible narrative, many tool calls, or reading
  the right files is not by itself a correct solution process. Evidence,
  intermediate values, transformations, and conclusions must agree.
- Keep the two scores distinct. A correct-looking final answer cannot substitute
  for missing process evidence, and an attempted method cannot receive full
  rubric credit when its result is wrong.
- Use only the supplied package evidence, ground truth, task solution, rubric,
  final answer, and model trace. Do not rely on web knowledge.

Runtime payload:
<<TASK_METADATA>>
<<QUESTION>>
<<GROUND_TRUTH>>
<<SOLUTION_AND_20_POINT_RUBRIC>>
<<MODEL_FINAL_ANSWER>>
<<MODEL_TOOL_TRACE>>
<<MODEL_RUN_NOTES>>
```

PROMPT EVALUATION **Joint answer and process judge** *continued*

```
EVALUATION RECORD
Scores        answer correctness and rubric-based process quality
Answer units  each required unit and whether it was satisfied
Rubric ledger criterion-level points earned out of 20
Diagnosis     verdict, critical errors, missing elements, unsupported claims
Rationale     one sentence explaining the score pair
```

PROMPT CONSTRUCTION **Benchmark construction suite**

QUALITY Task construction acceptance review

```
ROLE
You are the final quality reviewer for one extreme-event benchmark task.

INPUTS
- The public question (question_en.md)
- The expert solution (solution_en.md)
- The executable ground-truth program (compute_gt.py)
- The generated scoring contract (computed_gt.json)
- The optional review record (review.json)

TASK
Verify that the five artifacts describe one event-specific scientific target.
Do not approve a rewritten public question when the solution or GT still scores
an older package-inspection objective.

ACCEPTANCE CHECKS
1. The public question reads like a real disaster-analysis assignment.
2. It contains no local path, visible CSX identifier, package inventory,
   available-material wording, AOI, quality-gate, evidence-boundary, or benchmark
   phrasing, and it does not reveal the useful files.
3. The solution contains Final Answer, Key Computations, Reasoning Path,
   Computed Interpretation, and Scoring Rubric.
4. computed_gt.json declares schema_version 1.0, task and event identifiers,
   task type, evaluation mode, primary GT, key values, and a top-level rubric.
5. Deterministic tasks contain compact machine-checkable targets. Open expert
   tasks contain concrete judge anchors rather than a fabricated exact essay.
6. The rubric has 5-8 criteria whose points sum exactly to 20.
7. Former distractors are rejected through numeric tests, inequalities, process
   timing, or competing-mechanism checks rather than option prose.
8. compute_gt.py runs successfully and regenerates the scoring contract.

OUTPUT
Return changed or reviewed file paths, event and hazard type, task type,
evaluation mode, acceptance status for every check, and any remaining ambiguity,
missing-data risk, temporal/spatial mismatch, or unsupported overclaim.
```

RUBRIC Event-specific rubric construction

```
ROLE
You are constructing the hidden scoring contract for one benchmark task.

TASK
Create a 20-point rubric that can be applied without guessing the author's
intent. Use 5-8 small, independently judgeable criteria. Every criterion must
contain name, points, description, and partial_credit. The point values must sum
exactly to 20.

REQUIRED COVERAGE
- 3-4 points: final conclusion, numeric target, threshold state, proof result,
  ranking, or concise canonical stance.
- 3-5 points: required quantitative extraction or calculation, including units,
  tolerances, and windows where relevant.
- 2-4 points: formula, index, process-window, causal-chain, or score-ledger logic.
- 2-4 points: fusion of reports, time series, rasters, imagery, exposure,
```

PROMPT CONSTRUCTION **Benchmark construction suite** *continued*

```
  infrastructure, or trajectory evidence into checkable anchors.
- 1-3 points: a compact event-specific consequence that follows from the
  computed evidence.
- 1-3 points: rejection of plausible competing mechanisms or overclaims.
- 1-2 points: complete, concise output in the requested structure.

HARD RULES
Do not use vague criteria such as "good reasoning" or "uses evidence well."
Do not award points for choosing an option after a converted task has become a
numeric or structured analysis. A criterion that depends on an incorrect
scientific result cannot receive full process credit. For llm_judge tasks,
include canonical_stance, required_claims, required_metrics, forbidden_claims,
and acceptable_variants in primary_gt. For hybrid tasks, preserve both the
deterministic core and the reasoning anchors.

RUBRIC RECORD
Total points   exactly 20
Criteria       5-8 independently judgeable items
Each item      event-specific name; points; full-credit requirement;
               partial-credit and error rule
```

SOURCE **Event authenticity and source audit**

```
ROLE
You are auditing the real-event anchor of one event package.

QUESTIONS
1. Did the event occur? Require at least one traceable source confirming the
   event name, time window, or principal affected region.
2. Is the source reliable? Prefer government agencies, international
   organizations, peer-reviewed work, and authoritative scientific institutions;
   then disaster catalogues and major news. Encyclopedias and generic search
   pages are supporting evidence only.
3. Does the source actually correspond to this event? Check event name, time,
   place, and hazard type. A source about the same hazard family but a different
   time or place is not a strong match.

OUTPUT FIELDS
- existence_verdict: confirmed | likely | weak | not_confirmed
- temporal_match: exact | near | broad | mismatch | unknown
- spatial_match: exact | regional | broad | mismatch | unknown
- hazard_match: exact | compound_related | broad_family | mismatch | unknown
- source_reliability: high | medium | low | mixed
- release_decision: keep | keep_with_minor_fix | needs_stronger_source |
  replace_event
- concise evidence and repair notes

DECISION RULES
Use keep when the event is confirmed, at least two of temporal/spatial/hazard
matching are exact or regional, and the preferred evidence is authoritative.
Use keep_with_minor_fix when the event is real but URL, naming, or scale needs a
small correction. Use needs_stronger_source when current support is generic,
secondary, or overly broad. Use replace_event when the event cannot be confirmed
or the cited evidence clearly belongs to another event.
```

F REPRESENTATIVE CASE SET

A worked answer contract. One task examines the 2021 Pacific Northwest heat wave. The solver must distinguish a direct hourly record from a lower-resolution aggregate, identify the matching heat window, and compute heat and humidity statistics on the same temporal support. The arithmetic is elementary; the scientific work lies in choosing compatible observations and carrying that choice into the conclusion.

The six cards below index remote-sensing reconciliation, urban pluvial causality, evidence substitution, multi-agent quantitative decomposition, integrated heat research, and a partial scientific failure. Together they cover five system families without repeating the full records that follow.

REPRESENTATIVE CASE SET

FRONTIER HOSTED GPT-5.6 Sol CSX-189_Q1 Core 99.50
Uljin wildfire consistency. Report claims, wind and precipitation, Sentinel-2 dNBR, annual change, smoke transport, and bounded impact claims. TRACE Sixteen successful calls, including a failed semantic search recovered by a filename query and a final provenance check.

GENERAL HOSTED Tencent Hy3 CSX-059_Q2 Core 100.00
NYC Ida pluvial overload. An hourly rainfall record inside a two-day rain shield, population load, surface change, and a controlled future perturbation. TRACE Ten evidence calls followed by one deterministic calculation of every ratio, normalization term, and scenario score.

OPEN-WEIGHT LARGE Qwen3-235B CSX-006_Q2 Core 93.93
Heat-source substitution audit. Variable-level admissibility of an ERA5-Land aggregate substituted for local hourly apparent-heat and warm-night data. TRACE Eight calls covering inventory, source comparison, schema checks, corrected heat metrics, and numerical source impact.

MULTI-AGENT FRAMEWORK GeoMMAgent CSX-063_Q1 Core 83.50
Zhengzhou threshold ledger. Reported burst and multiday rainfall, product ordering, exposure thresholds, and a conjunctive flood diagnosis. TRACE Four coordinator roles and nine package-tool calls for retrieval, calculation, scientific synthesis, and schema assembly.

RESEARCH FRAMEWORK OpenResearcher CSX-018_Q2 Core 86.92
Sahel heat-response model. Hourly heat, wet bulb, warm nights, exposure, services, roads, surface stability, two indices, and a +1 C counterfactual. TRACE Twenty-one calls, including one failed heat script, an explicit repair, and independent checks of peaks, duration, density, scores, and deltas.

GENERAL HOSTED GPT-4o CSX-312_Q2 Core 41.25
Partial scientific failure. The model identifies the coupled Kilauea mechanism but reports lava effusion rate instead of the required summit-collapse-to-flow volume ratio. TRACE Nineteen model rounds and nine tool calls: seven succeed, two guessed paths fail, and the final answer earns Answer 40.00 and Rubric 42.50.

G Detailed research trajectories

Appendix F identifies the six selected systems and tasks. This section expands each card into its scientific target, evidence sequence, calculations, and decision. The Uljin case retains the complete task and action order; the remaining records focus on the steps that determine the result.

G.1 Full case: Uljin report-to-burn consistency

QUESTION CSX-189_Q1

SCIENTIFIC TARGET
Uljin wildfire report-to-burn consistency ledger

A technical review team is checking whether the March 4–13, 2022 Uljin wildfire is numerically consistent with a dry, wind–assisted burn–scar diagnosis rather than a late–rain–control or image–baseline alternative.

EVIDENCE SCOPE
Use only package–local evidence. Do not use point–weather products for this task; build the ledger from the event report, package gridded precipitation and wind summaries, Sentinel–2 dNBR, and annual embedding–change context.

QUESTION CSX-189_Q1 *continued*

REQUIRED COMPUTATION
Compute these ledger values:

– report_dry_wind_flag: 1 if the report links the event to strong winds and dry weather, otherwise 0.
– smoke_transport_flag: 1 if the report states that smoke moved toward southern Japan, otherwise 0.
– era5_mean_wind_speed_mps: vector speed from mean 10 m u and v components.
– era5_wind_bearing_to_deg: direction toward which the mean wind vector points, degrees clockwise from north.
– precip_mean_spread_mm: spread between ERA5–Land, GPM IMERG, and CHIRPS event precipitation means.
– dnbr_mean: mean dNBR over the event burn–change window.
– dnbr_max_to_mean_ratio: dNBR maximum divided by dNBR mean.
– dnbr_to_annual_change_mean_ratio: dNBR mean divided by annual 1–minus–cosine embedding–change mean.
– reported_charred_area_km2: reported hectares converted to square kilometers.

MECHANISM CHECKS Apply these tests:
– report_dry_wind_mechanism_test: pass if report_dry_wind_flag = 1.
– smoke_transport_context_test: pass if smoke_transport_flag = 1.
– positive_heterogeneous_burn_test: pass if dnbr_mean > 0 and dnbr_max_to_mean_ratio >= 2.0.
– late_rain_control_test: pass only if rainfall is the dominant control; otherwise fail.
– nuclear_damage_overclaim_test: pass if the answer avoids inferring nuclear–plant damage or radiation release.

OUTPUT
Evidence ledger
– Report/dry–wind flag and smoke–transport flag
– Mean ERA5 wind speed and bearing
– Precipitation–mean spread across the three gridded products
– Mean dNBR, dNBR maximum–to–mean ratio, and dNBR/annual–change ratio
– Reported charred area in square kilometres

Decision checks
– Report mechanism; smoke context; positive heterogeneous burn
– Late–rain control; bounded nuclear–impact claim

INTERPRETATION
– One or two sentences tied directly to the computed ledger

CASE Research record: CSX-189_Q1

COMPLETE ACTION ORDER; OBSERVATION BODIES ABRIDGED

1	`list_package_files`	**Discover** Enumerated all 39 package files.
2	`read_text_file`	**Anchor** Confirmed Uljin, South Korea, 4–13 March 2022.
3	`search_package_files`	**Mechanism** Located report language linking strong winds and dry weather.
4	`search_package_files`	**Transport** Located the statement that smoke moved toward southern Japan.
5	`search_package_files`	**Impact quantity** Recovered nearly 17,000 charred hectares.
6	`read_text_file`	**ERA5-Land** Read mean precipitation and mean 10 m wind components.
7	`read_text_file`	**GPM IMERG** Read event precipitation mean of 2.884 mm.

CASE **Research record: CSX-189_Q1** *continued*

8	`read_text_file`	**CHIRPS** Read event precipitation mean of 4.456 mm.
9	`read_text_file`	**Sentinel-2 dNBR** Read mean 0.303255 and maximum 0.708527.
10	`search_package_files`	**Embedding search** Initial semantic query returned no match.
11	`search_package_files`	**Recovery** A filename-oriented query located annual cosine-change statistics.
12	`read_text_file`	**Annual context** Read mean 1-minus-cosine change of 0.013761.
13	`search_package_files`	**Smoke check** Confirmed the report-side smoke passage.
14	`read_text_file`	**Full report** Read the authoritative HTML surrounding all extracted claims.
15	`python_exec`	**Deterministic ledger** Computed wind speed 1.392 m/s, bearing 331.253 degrees, precipitation spread 6.389 mm, dNBR heterogeneity 2.336, dNBR/annual-change ratio 22.037, and 170 km2 charred area.
16	`search_package_files`	**Final provenance check** Re-queried the most easily misquoted smoke-transport claim before finalization.

COMPUTED LEDGER

DRY-WIND FLAG **1** SMOKE FLAG **1** MEAN WIND **1.392043 m/s** BEARING **331.252581 deg**

PRECIPITATION SPREAD **6.389107 mm** MEAN dNBR **0.303255** dNBR MAX/MEAN **2.336407**

dNBR/ANNUAL CHANGE **22.037202** CHARRED AREA **170.0 km^2**

PASS Report mechanism PASS Smoke context PASS Heterogeneous burn

FAIL Late-rain control PASS Bounded nuclear-impact claim

EVIDENCE RECONCILIATION

The burn scar is not accepted on appearance alone. The trace anchors two report claims, reconciles three precipitation products, and keeps event-window dNBR separate from annual embedding change. Step 10 also preserves the failed semantic search and the filename-based recovery that follows.

CASE **NYC Ida: urban pluvial overload**

TASK

SCIENTIFIC TARGET Reconstruct the New York City phase of post-tropical Cyclone Ida as a record hourly burst embedded in a broader two-day rain shield.

EVIDENCE SCOPE Keep station-scale hourly rainfall distinct from the wider gridded rainfall field, then add population exposure and surface-change context.

REQUIRED COMPUTATION Compute the baseline response stress and repeat it under the prescribed future perturbation.

RESEARCH TRAJECTORY: TEN EVIDENCE CALLS AND ONE CALCULATION

1	`list + anchor`	**Window** Enumerated 39 files and fixed the inclusive 1–2 September 2021 event window.
2	`read report`	**Local forcing** Recovered the Central Park hourly record and the conservative 10-inch regional reference.

CASE NYC Ida: urban pluvial overload *continued*

3	`read 3 products`	**Rain shield** Read GPM IMERG, CHIRPS, and ERA5-Land event means and maxima.
4	`read exposure`	**Context** Read 5.013 million exposed people, Sentinel-1 mean change of 0.234 dB, and annual embedding change of 0.022.
5	`python_exec`	**Ledger + scenario** Converted units, reconciled products, computed ratios and normalized terms, and reran the weighted index under the prescribed perturbation.

EVIDENCE, METRICS, AND DECISION

WINDOW **2 days** RECORD HOUR **88.138 mm** REGIONAL REFERENCE **254.000 mm**
GRID PEAK **136.888 mm**

GRID MEAN **53.427 mm** HOUR/PEAK **0.644** REGIONAL/PEAK **1.856** BURST/AREAL DAY **3.299**

POPULATION **5.013 million** RAIN LOAD **267.812 million-person-mm** SURFACE CHANGE **0.337**

DECISION The baseline response-stress index is **72.6**; it rises to **80.1** under the future perturbation (**+7.5**). The local hourly record is not treated as a substitute for the wider gridded rain shield.

CASE Heat-source substitution audit

TASK

SCIENTIFIC TARGET Determine whether a same-window ERA5-Land aggregate can replace the local hourly heat source.

EVIDENCE SCOPE Audit variable availability for apparent-temperature duration and warm-night exposure rather than judging a source by filename.

DECISION RULE Accept the substitution only when the replacement preserves every variable needed for the requested metrics.

RESEARCH TRAJECTORY: EIGHT SUCCESSFUL CALLS

1	`list_package_files`	**Discover** Enumerated 37 files and identified the two competing heat products.
2	`read anchor`	**Window** Fixed 31 March–4 April 2024.
3	`read sources`	**Compare** The local source contains hourly temperature, humidity, apparent temperature, and daily minima; the aggregate lacks the last two required roles.
4	`python_exec`	**Schema audit** Confirmed that apparent temperature is the first missing field and daily minimum temperature is also absent.
5	`python_exec`	**Correct + compare** Recomputed the target heat metrics and measured the peak-temperature error introduced by the substitute.

EVIDENCE, METRICS, AND DECISION

MAX AIR **44.6 C** MAX APPARENT **43.2 C** HOURS APPARENT $\geq$ 40 C **16** WARM NIGHTS $\geq$ 27 C **5**

AGGREGATE MAX **35.3 C** PEAK DIFFERENCE **9.3 C**

MISSING FIELDS **apparent temperature; daily minimum**

CASE Heat-source substitution audit *continued*

SOURCE SWAP FAILS The aggregate cannot compute apparent-heat duration or warm nights. The corrected label is `sustained_apparent_heat_with_warm_night_exposure`.

CASE Zhengzhou: coordinated threshold ledger

TASK

SCIENTIFIC TARGET Test whether the 17–23 July 2021 Zhengzhou flood combined an exceptional hourly burst, multiday extreme rainfall, and dense urban exposure.

REQUIRED COMPUTATION Build the rainfall-ratio, product-order, population, and infrastructure threshold ledger.

DECISION RULE Assign the final label only when all four numerical thresholds pass.

RESEARCH TRAJECTORY: FOUR ROLES AND NINE PACKAGE-TOOL CALLS

1	`Agent 0: retrieve`	**Evidence** Read the event report, three precipitation products, WorldPop, and OSM with strict provenance.
2	`Agent 1: compute`	**Ledger** Calculated rainfall ratios, product order, and the four threshold flags.
3	`Agent 2: synthesize`	**Mechanism** Joined physical rainfall evidence with population and road exposure.
4	`Agent 3: assemble`	**Output** Built the requested schema without changing the shared evidence boundary.
5	`python_exec`	**Verify** Recovered hour share 0.3272, annual ratio 0.963, and four true threshold flags.

EVIDENCE, METRICS, AND DECISION

THREE-DAY RAIN **617.1 mm** ONE-HOUR RAIN **201.9 mm** HOUR SHARE **0.3272** ANNUAL RATIO **0.963**

MEAN ORDER **GPM > CHIRPS > ERA5-Land** MAX GRID MEAN **460.5514 mm** MAX GRID **484.255 mm**

POPULATION **5.2625 million** HIGHWAYS **618** AMENITIES **50** TUNNELS **2**

ALL FOUR THRESHOLDS PASS Hour share ≥ 0.30, annual ratio ≥ 0.90, population ≥ 5 million, and highway features ≥ 200. Final label: `short_burst_multiday_urban_exposure`.

CASE Sahel heat: response load and warming scenario

TASK

SCIENTIFIC TARGET Model a five-day Sahel heat episode and its response burden.

EVIDENCE SCOPE Combine hourly heat and Stull wet-bulb temperature, warm nights, population, services, roads, gridded context, and surface stability.

REQUIRED COMPUTATION Calculate both response indices at baseline and under a +1 C counterfactual.

RESEARCH TRAJECTORY: TWENTY-ONE CALLS

CASE **Sahel heat: response load and warming scenario** *continued*

1	`list + open`	**Core inputs** Read hourly heat, WorldPop, AOI geometry, and OSM infrastructure.
2	`python: exposure`	**Counts** Computed schools, health facilities, shelters, major roads, AOI area, and population density.
3	`python: heat`	**Failed attempt** The first heat script failed because it assumed the wrong hourly field names.
4	`python: repair`	**Schema recovery** Re-read the actual fields and computed Stull wet-bulb temperature and warm nights.
5	`open products`	**Context** Read ERA5-Land temperature, GPM and CHIRPS precipitation, and Sentinel-2 dNBR.
6	`python checks`	**Heat metrics** Independently verified air and apparent peaks, duration, degree-hours, and density.
7	`python scenario`	**Indices** Rechecked baseline and +1 C scores, threshold counts, and deltas.

HEAT, EXPOSURE, AND SURFACE CONTEXT

MAX AIR **44.6 C** MAX APPARENT **43.2 C** MAX WET BULB **16.3605 C** HOURS $\geq$ 40 C **16** DEGREE-HOURS **24.4** WARM NIGHTS **5**

POPULATION **170,311.46** DENSITY **3.46 km^{-2}** SCHOOLS **6** HEALTH FACILITIES **2** SHELTERS **1** MAJOR ROADS **57**

SURFACE CHANGE **0.024472** SURFACE STABILITY **0.97553**

INDICES AND +1 C SCENARIO

BASE HEAT RESPONSE **77.344912** BASE PRIORITY **86.374648**

FUTURE MAX APPARENT **44.2 C** FUTURE HOURS $\geq$ 40 C **25** FUTURE DEGREE-HOURS **40.4**

FUTURE HEAT RESPONSE **84.444912** FUTURE PRIORITY **93.094648** DELTAS **+7.1; +6.72**

INTERPRETATION The explicit failed computation and repair remain in the record; the final scenario is supported by independently rechecked heat, exposure, and response terms.

CASE **Partial failure: correct mechanism, wrong physical index**

TASK

SCIENTIFIC TARGET Diagnose the 2018 Kilauea eruption as isolated lower-rift, isolated summit, rainfall-triggered, or coupled summit-drainage/lower-rift activity.

MECHANISM CHECK Use the evidence to distinguish the coupled explanation from all three alternatives.

OUTPUT Return one physical index and value, three evidence clues, and an operational implication.

MODEL ROUNDS **19** TOOL CALLS **9** SUCCESS **7** ERRORS **2** UNIQUE EVIDENCE FILES **4**

OBSERVED RESEARCH SEQUENCE

1	`list_package_files`	**Discover** Enumerated all 34 package files.

CASE **Partial failure: correct mechanism, wrong physical index** *continued*

2	`read USGS report`	**Primary evidence** Opened the authoritative summit-collapse and Lower East Rift Zone report.
3	`search package`	**Weak recovery** Two semantic searches returned no direct match for the requested coupling language.
4	`read anchor`	**Event bounds** Confirmed the 3 May–4 August 2018 event window and Kilauea location.
5	`read guessed paths`	**Tool errors** Two nonexistent guessed files produced the run's only tool errors.
6	`search + reread`	**Final extraction** Recovered downrift propagation and effusion above 100 m^3/s, then finalized without computing the 0.8/0.8 volume ratio.

RETURNED ANSWER

DIAGNOSIS **coupled summit drainage and lower-rift effusion**

PHYSICAL INDEX **Lava Effusion Rate** VALUE **100 m^3/s**

CLUES Downrift magma propagation after summit collapse; lava discharge above 100 m^3/s; coordinated summit and rift monitoring.

REQUIRED ANSWER

DIAGNOSIS **coupled summit drainage and LERZ effusion**

PHYSICAL INDEX **summit-collapse-to-lava-flow volume ratio** VALUE **1.0**

The index follows from approximately **0.8 km^3** of summit collapse and **0.8 km^3** of lava flow. A complete answer also uses partial summit drainage, near-daily collapses, and M_w 4.7–5.4 energy release, then rejects isolated-rift, isolated-summit, and rainfall-triggered alternatives.

ASSESSMENT

ANSWER **40.00** RUBRIC **42.50** CORE **41.25**

PARTIAL FAILURE The model identifies the correct coupled mechanism and gives a useful operational implication, but substitutes a real yet irrelevant rate for the required volume ratio. It also omits the summit evidence and explicit alternative-mechanism rejection. The failure is scientific and evidential, not merely a formatting breakdown.

H Detailed Related Work

Earth observation and scientific multimodality. EarthSE, MSEarth, GeoMMBench, AtmosSci-Bench, GeoBench-VLM, and XLRS-Bench test scientific knowledge, figure understanding, atmospheric reasoning, and remote-sensing interpretation (Xu et al., 2026; Zhao et al., 2026; Xiao et al., 2026; Li et al., 2025; Danish et al., 2025; Wang et al., 2025). Prithvi-EO-2.0 and TerraMind study representations across sensors and downstream tasks, while GEO-Bench-2 finds that no representation dominates every sensing regime (Szwarcman et al., 2026; Jakubik et al., 2025; Simumba et al., 2025). In each case, the observation or dataset is supplied.

Disaster intelligence. CrisisMMD, xBD, and FloodNet cover crisis classification, damage assessment, segmentation, and visual question answering from supplied media (Alam et al., 2018; Gupta et al., 2019; Rahnemoonfar et al., 2021). DORA adds expert-authored emergency tasks, a declared multimodal manifest, typed tools, and replayable reference calls (Wang et al., 2026). It moves closer to an operational workflow, although the relevant observations and layers are still declared in advance.

Tool-using and scientific agents. ReAct interleaves reasoning and action, while Toolformer and ToolLLM study learned API use (Yao et al., 2023; Schick et al., 2023; Qin et al., 2024). AgentBench, GAIA, and OSWorld evaluate multi-step interaction in executable environments (Liu et al., 2024; Mialon et al., 2024; Xie et al., 2024). ScienceAgentBench, SciAgent, and SciAgentGym extend this setting to code, data, scientific functions, and long-horizon orchestration (Chen et al., 2025; Ma et al., 2024; Shen et al., 2026). BLADE and DiscoveryBench further expose multiple valid analysis paths and decompose data-driven discovery into evaluable facets (Gu et al., 2024; Majumder et al., 2025).

EarthVerse couples expert authoring with a scientific tool suite (Figure 3). Experts frame and verify each investigation; numerical, atmospheric, and geospatial tools support executable analysis. Evaluation then scores source choice, intermediate work, and physical interpretation separately from the final answer.

Geospatial and deep-research systems. ThinkGeo evaluates tool execution over remote-sensing imagery; Earth-Agent, OpenEarthAgent, Spatial-Agent, and GeoMMAgent coordinate Earth-observation or geospatial operations (Shabbir et al., 2025; Feng et al., 2026; Shabbir et al., 2026; Bao et al., 2026; Xiao et al., 2026). DeepResearcher, OpenResearcher, and MiroThinker study iterative retrieval and long research trajectories (Zheng et al., 2025; Li et al., 2026; MiroMind Team et al., 2025; Yu et al., 2026). Intern-S1, GeoGPT, and GeohazardGPT test domain-specific pretraining (Bai et al., 2025; Zhang et al., 2024; Ge et al., 2025). EarthVerse evaluates the combined system: whether reasoning and interaction can turn retrieved evidence and tool output into a coherent account of an evolving event.