EARTHLITBENCH: BENCHMARKING LITERATURE-GROUNDED SCIENTIFIC INQUIRY IN EARTH SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific inquiry in Earth system science demands synthesizing fragmented, cross-study empirical findings across diverse environmental spheres and temporal-spatial scales. Evaluating LLM support for this work therefore calls for assessing evidence selection and scientific justification together. However, existing benchmarks predominantly focus on single-document factoid QA, leaving a critical blind spot in evaluating multi-hop reasoning and evidence synthesis across extensive literature. To bridge this gap, We introduce EarthLitBench, a benchmark of 822 instances across ten tasks organized into four capability groups: problem formulation and literature discovery, study and evidence understanding, scientific reasoning, and knowledge synthesis. Grounding our benchmark in authoritative scientific reviews, we construct traceable knowledge graphs that capture complex scientific relationships, thereby guiding rigorous question generation and evidence verification. To ensure fair evaluation, source reviews, construction graphs, and reference solutions are strictly withheld from candidate models, while evidence retrieval is constrained within an authorized reference corpus. Furthermore, we establish the RERA evaluation protocol, which couples automated source verification with reference-guided LLM judges across four complementary dimensions: Retrieval, Evidence fidelity, Reasoning fidelity, and Answer fidelity. Experiments evaluate nine LLM-only models and compare four of them with fixed retrieval. Models perform relatively well on problem decomposition but have greater difficulty with scientific reasoning, even though explicit references are given. These findings highlight a gap between accessing scientific literature and conducting scientific inquiry, motivating stronger integration of evidence and scientific interpretation. We will make our benchmark publicly available.

1 INTRODUCTION

Scientific inquiry in Earth system science demands synthesizing fragmented empirical evidence across interacting Earth spheres, observational regimes, and numerical models (Steffen et al., 2020). For instance, explaining the mechanistic diversity of the El Niño–Southern Oscillation (ENSO) requires researchers to reconcile conflicting proxy reconstructions with climate model projections, isolate opposing ocean–atmosphere feedbacks (e.g., the Bjerknes positive feedback versus thermodynamic damping), and determine whether teleconnection impacts hold under shifts in background ocean states (Jin et al., 2006). In such investigations, retrieving a relevant study does not establish evidential support for a causal hypothesis, nor does generating a plausible conclusion verify scientific reasoning validity. Evaluating AI scientific assistance therefore requires examining whether models can construct fully justified inferences grounded in primary literature, rather than merely recalling isolated facts or retrieving topically relevant documents.

Scientific knowledge question answering (Rein et al., 2024; Skarlinski et al., 2024), literature retrieval(Ajith et al., 2024), and research-assistance tasks(Skarlinski et al., 2024) address complementary aspects of this challenge. For literature-grounded inquiry, however, finding a relevant source does not establish that its evidence supports a particular conclusion, and a correct conclusion does not establish that the supporting reasoning is valid. Evaluation must connect the requested scientific operation to the selected sources, located evidence, and explicit justification. This motivates a benchmark that examines these requirements together across problem formulation, study understanding, scientific reasoning, and synthesis.

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.

In bridge this gap, we introduce EARTHLITBENCH, a diagnostic benchmark comprising approximately 822 instances across ten tasks structured into four core inquiry dimensions: problem formulation and literature discovery, study and evidence understanding, scientific reasoning, and knowledge synthesis. Each task requires models to provide selected sources, located evidence, an explicit justification, and a task-specific output. Crucially, to expose specific capability bottlenecks without compounding error cascades across multi-step execution workflows, each task is evaluated as an independent scientific operation under uniform evidentiary standards.

Constructing traceable evaluation targets without costly manual labeling or hallucination, we establish a review-grounded benchmark factory. Expert domain reviews naturally synthesize disparate primary studies, resolving methodological conflicts and outlining mechanistic frontiers. Our factory extracts these relationships into source-linked knowledge graphs, generates candidate tasks from structured subgraphs, and enforces multi-stage verification against both cited primary texts and expert-defined scientific constraints. At evaluation time, synthesizing reviews and graphs are withheld, restricting models to a frozen authorized reference corpus. We formulate the *RERA* protocol to evaluate performance across four interdependent dimensions—***Reference selection, Evidence fidelity, Reasoning fidelity, and Answer fidelity***—unifying deterministic bibliographic validation with reference-grounded LLM judging across unaided LLMs, fixed-retrieval pipelines, and research agents.

The experiments reveal an uneven capability profile: models perform relatively well on problem decomposition but have greater difficulty with cross-study comparison, mechanism reasoning, and transfer across conditions. Fixed retrieval improves problem formulation and literature discovery across all four paired systems, whereas gains on scientific reasoning tasks are smaller and less consistent. These observations highlight a gap between access to scientific information and its interpretation. They motivate evaluation that examines how models connect evidence to scientific judgments, alongside whether they can identify useful sources, without attributing remaining failures solely to either retrieval or model capability.

Our contributions are threefold, summarized as following:

- We introduce EarthLitBench, consist of ten tasks, to evaluate literature-grounded scientific inquiry through task-specific outputs and evidence requirements.
- We develop a review-grounded benchmark construction procedure including explicit scientific relationships with source tracing and verification of candidate benchmark data.
- We define the RERA evaluation protocol and provide controlled comparisons revealing differences across inquiry tasks and the uneven benefits of fixed retrieval.

2 RELATED WORK

2.1 LITERATURE-GROUNDED RESEARCH EVALUATION AND SCIENTIFIC INQUIRY

Scientific benchmarks assess different components of research competence. ScienceQA (Lu et al., 2022) evaluates multimodal science question answering, while SciBench (Wang et al., 2024) emphasizes advanced scientific problem solving. Literature-grounded evaluation makes evidence selection part of the task: SciFact (Wadden et al., 2020) couples claim verification with supporting or refuting abstracts, and SciFact-Open (Wadden et al., 2022) extends verification to a larger retrieval setting. Together, these lines move evaluation beyond answer correctness toward the evidence supporting an answer. Scientific inquiry additionally requires deciding whether studies are comparable, explaining disagreements, and determining where a conclusion applies (Xu et al., 2025; Yifei et al., 2025). These judgments depend on study conditions and relationships among findings, which claim verification and problem solving alone do not characterize. EarthLitBench therefore evaluates task-specific scientific operations alongside source selection, located evidence, and explicit justification, spanning question formulation, evidence understanding, reasoning, and synthesis.

2.2 EARTH SCIENCE BENCHMARKS FOR LARGE LANGUAGE MODELS

Earth science benchmarks extend evaluation from domain knowledge to exploratory and interactive tasks. ClimaQA (Manivannan et al., 2025) uses climatology textbooks for question answering, with

Table 1: Comparison of Earth Science Scientific-Capability Benchmarks. Materials, evaluation focus, and response formats.

Benchmark	Materials / Source	Evaluation Focus	Response Format
ClimaQA	Climatology textbooks	Domain knowledge and understanding	QA, multiple-choice & cloze responses
EarthSE	Earth science literature	Knowledge application and exploration	QA responses and multi-turn dialogue
EarthVerse	Earth-system events data	Scenario-based Data Usage	Agent responses supported by scenario evidence
EarthLitBench	Reviews with references	Scientific Inquiry	Evidence grounded scientific inquiry

expert-validated and automatically generated subsets. EarthSE (Xu et al., 2026) draws on Earth science literature to assess knowledge application and exploration through QA and multi-turn dialogue. EarthVerse (Cui et al., 2026) evaluates agents in Earth-system event scenarios using multiple information sources, including remote-sensing and ground-based data. These settings address complementary aspects of scientific capability: disciplinary knowledge, literature-based exploration, and situated evidence use (Table 1).

EarthLitBench focuses on a specific need within this landscape: assessing how systems use published evidence to justify scientific judgments across studies and conditions. Its ten tasks distinguish evidence discovery from interpretation, reconciliation, mechanism reasoning, transfer, and synthesis. Shared output requirements and task-specific criteria make both the evidence basis and the resulting judgment assessable; paired LLM-only and fixed-retrieval conditions examine how access to retrieved material changes performance. This complements domain QA and interactive exploration with an explicit evaluation of literature-grounded scientific inquiry.

3 EARTHLITBENCH

3.1 INQUIRY TASKS AND BENCHMARK INTERFACE

EarthLitBench formalizes literature-grounded scientific inquiry through ten tasks organized into four progressive capability groups (Table A1): problem formulation and literature discovery (T01–T02) establish inquiry scope and evidence requirements; study and evidence understanding (T03–T05) extract and verify findings; scientific reasoning (T06–T08) resolve cross-study comparisons, causal mechanisms, and condition transfer; and knowledge synthesis (T09–T10) derive collective conclusions and justified future directions. These tasks are evaluated independently without mandatory execution pipelines or cumulative difficulty assumptions. Each instance is indexed by a single primary judgment and scoring target, receiving self-contained context and constraints via public inputs without leaking upstream gold solutions.

To enforce rigorous scientific accountability, candidate systems interact with a standardized task interface under controlled evidence access. While necessary attachments (e.g., regional figures or tables) are provided with the prompt, external literature access is governed by the evaluation track (closed-book, fixed retrieval), with underlying survey graphs and gold references strictly withheld. Across all conditions, candidate systems must return a standardized four-part response: explicit source citations, localized evidence pointers (specific passages, panels, or cells), a deductive scientific justification, and the requested deliverable. For inferential and analytical tasks (T01 and T03–T10), deliverables constitute structured accounts such as question decompositions, cross-study reconciliations, or time-bounded gap analyses; for literature discovery (T02), the target is an eligibility-screened inventory mapping studies to evidence needs with explicit coverage bounds.

Reference solutions establish verifiable evidential anchors and scoring rubrics without presuming exhaustive uniqueness. Because legitimate alternative scientific pathways exist across Earth system inquiries, candidate submissions are not penalized for formulating alternative question decompositions or sound scientific justifications, provided their supporting premises are verified within the reference corpus. Crucially, scoring is grounded strictly on the explicit, literature-anchored scien-

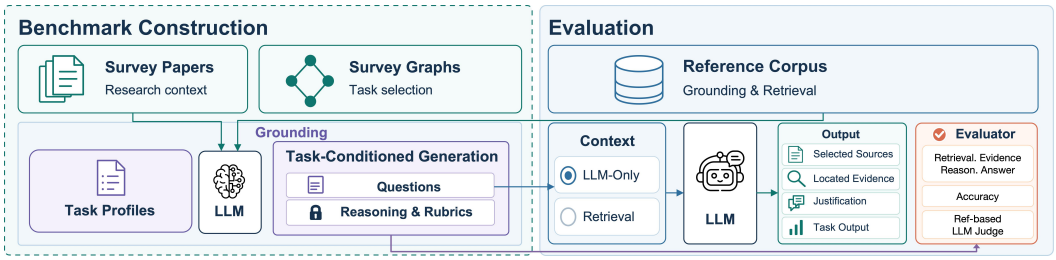


Figure 1: Overview of Benchmark construction and evaluation. Survey papers provide hidden research organization and problem context, while source-linked survey graphs support task selection and provenance. Authorized reference literature supplies grounding evidence during construction and the permitted corpus during evaluation. Candidates receive questions or only the declared evidence access; each task returns retrived sources, located evidence, an explicit scientific justification, and a task-specific output. Hidden reference solutions and construction supervision are evaluator-only. RERA combines source verification with reference-based judging and allows valid alternative answers and explanations.

tific justifications submitted by the candidate, rather than speculative completions or ungrounded internal model reasoning traces.

3.2 Review-Grounded Construction and Verification

We construct benchmark instances from review articles and their cited literature, using a source-traceable graph to organize task-conditioned contents. Figure 2 summarizes three stages of benchmark construction.

Source-traceable graph construction. MinerU parses each review PDF into Structured texts and image assets. Heading hierarchy and associate figures and tables with their identifiers, captions are preserved. Body text is split into token-bounded chunks with overlap confined to the same section. The title, abstract, and keywords guide selection of a domain profile from a shared Earth science schema dictionary, constraining extraction types and relations. An LLM jointly extracts entities, claims, and relations from body chunks; media descriptions and bibliography entries are processed parallelly. Every extracted scientific node must have verifiable support in the source review; claims and relations additionally undergo checks of attribution, meaning, direction, and conditions. Unsupported records and alignment failures are rejected. For uncertain entities or complex chunks, re-extraction are allowed to ensure result correctness.

Evidence-grounded Task Selection Each task specifies eligible seed-node types, relation patterns where applicable, and evidence requirements. We search each review graph for matching seeds and collect the nodes and edges guided by task profile. Mechanism tasks select directed process relations and their conditions; comparison tasks gather study-specific claims, methods, and scopes. Selection preserves relevant limitations and counter evidence, and comparison sets need not form a connected subgraph. Source links recover the selected records' section context and associated media. For claims and mechanism steps, inline citations in their supporting passages are resolved to bibliography entries, then checked against the cited studies for actual support. Citation matching alone does not establish support. The selected graph, contextual passages, and verified reference evidence form the generation context. Missing or ambiguous evidence triggers review or further evidence collection; only seeds with sufficient candidate-accessible evidence proceed to generation.

Data generation and screening. Given the task profile, sampled nodes and edges, source context, and verified reference evidence, an LLM generates a question, reference reasoning, reference answer (QRA), and rubric. Reference reasoning links evidence-backed premises, scientific operations, conclusions, and limitations. Screening uses the sampled graph together with the original evidence to check source fidelity, answerability from candidate-accessible materials, and whether the question requires the selected scientific structure. Complexity reflects necessary relations, operations, and independent studies, rather than raw graph size or repeated citations. For deduplication, node–edge overlap within reviews, shared evidence across reviews, and question similarity flag candidate pairs.

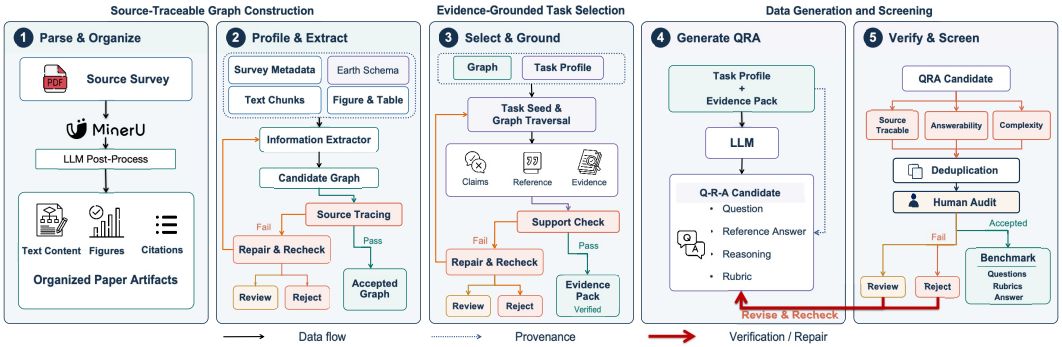


Figure 2: Five-stage survey-to-benchmark construction pipeline. (1) Parse& Organize preserves section, media, and citation context. (2) Profile & Extract uses a metadata-selected schema profile and an LLM to generate a per-paper graph, but candidate graph records enter the accepted graph only after a backward source-span and evidence check; bounded repair must be rechecked, and unresolved cases are reviewed or rejected. (3) Select & Ground builds a task seed and evidence pack by checking citation scope and actual primary-source support. (4) Generate QRA produces a question, reference answer, evidence-linked reasoning, and rubric. (5) Verify & Screen checks grounding, answerability, required scientific operations, human-audit cases, and duplicates before benchmark assembly.

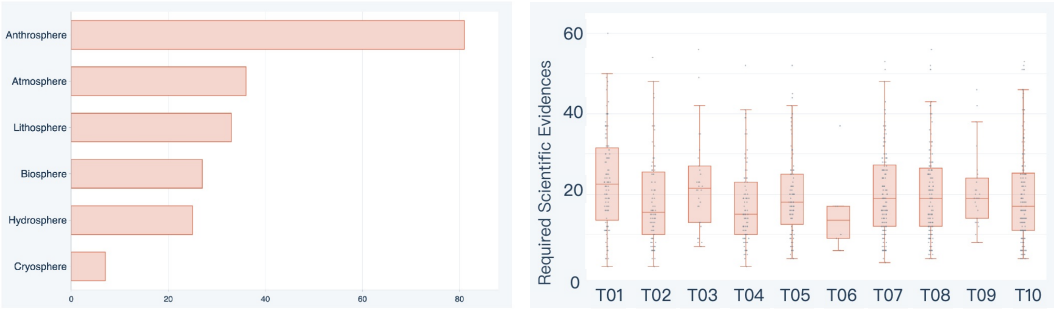


Figure 3: Distribution of semantic and complexity of EarthLitBench. *Left*: distribution of sphere tags of source surveys papers, showing the scientific broadness in earth science. *Right*: distribution of required scientific evidences in rubric of each instance of each tasks, demonstrating the complexity of EarthLitBench.

We then compare evidence groups, target operations, and answer meaning to distinguish paraphrases from distinct tasks using the same sources. Items failing verification are rejected or revised once and reverified; unresolved cases remain under review. Human audits supplement automated checks and final quality assessment. The sampled graph and its evidence mappings remain hidden construction records.

3.3 Dataset Composition

We collected 112 surveys published on top-tier academic journals since 2025, and constructed a benchmark dataset consisting of 822 instances. For the diversity of topic covered by EarthLitBench, we group the collected survey by sphere tags of corresponding research topic, and sphere tags are annotated by domain experts. Surveys may cover several spheres as many researches are cross-sphere. Distribution of survey topics are shown in Fig. 3 (a). Meanwhile, we demonstrate complexity of EarthLitBench by the required scientific concepts of each instance. Fig. 3 (b) shows the distribution the evidences required by data for each tasks. Most tasks have a mean evidence number about 20, and it shows that difficulty measured by evidence number is quite similar across tasks.

Table 2: RERA Evaluation Dimensions. Task-specific criteria specialize a common 0–4 scale.

Dimension	What Is Assessed	Scoring Method
Retrieval (Reference Selection)	Validity, eligibility, relevance, and coverage of the selected sources relative to required evidence needs	Source verification, 5-Grade Ref-based LLM judge
Evidence Fidelity	Whether candidate-located sources faithfully support stated facts, attribution, quantities, and conditions	5-Grade Ref-based LLM judge
Reasoning Fidelity	Whether scientific operations and inferences validly connect premises to judgments	5-Grade Ref-based LLM judge
Answer Fidelity	Correctness, completeness, and scope of the requested conclusion or output	5-Grade Ref-based LLM judge

4 Evaluation Protocol

4.1 Evaluation Setting and Evidence Access

We conduct evaluation under two evidence-access conditions over authorized reference corpus. LLM-only (pure LLM) receives the question, task instructions, scope and time constraints, and necessary attachments, without corpus-search or browsing access. LLM + RAG (Lewis et al., 2021) additionally receives material selected by a shared pipeline and cannot issue follow-up searches. All conditions use the same task requirements and scoring targets; source reviews, construction graphs, sampled subgraphs, reference solutions, and hidden assessment mappings remain unavailable in evaluation.

Models may identify sources through titles, authors, years, or DOIs rather than internal corpus IDs. A valid selected source need not occur in the fixed results or search log, but must belong to the authorized corpus and meet the item's scope and time requirements. Public corpus eligibility rules do not supply a list of relevant sources. Search tools are unavailable to all LLMs, so that they cannot expand the corpus through open-web browsing. Source-specified questions retain their necessary attachments in every condition.

The evaluator receives the original response, task rubric, verified reference solution and evidence, and assessment checklist. It verifies submitted source identities against the frozen manifest and inspects the evidence actually identified by the candidate, including valid alternatives absent from the reference solution. It does not supply missing candidate evidence or reconstruct an omitted justification. Reviews and graphs support construction audits, not a requirement to reproduce hidden wording or structure. Scoring does not require proof that a candidate accessed a correctly identified source. We use Qwen3.8-27B as judge model with references.

The main results cover LLM-only and LLM with RAG. Paired comparisons use shared model versions and identical eligible questions, with attachments, modality permissions, and output budgets controlled. Differences measure system performance under different evidence access, not isolated reasoning ability. Experiments mainly covers both open-source and close-source LLMs, and results are presented in Tab. 3.

4.2 RERA Scoring and Aggregation

The final scientific responses are evaluated by proposed RARE protocol. All four dimensions apply to T01–T10 under each protocol-defined condition (Table 2).

Rubrics and score definitions. Shared dimension anchors are specialized by task rubrics and reviewed item checklists of necessary content, evidence needs, and acceptable alternatives. Each dimension receives one holistic integer score per item: 0, absent or wholly invalid; 1, limited validity with a central failure; 2, partially correct with substantive gaps; 3, substantially correct with minor deficiencies; and 4, all required criteria met. Checklists do not change scales or weights or introduce requirements absent from the public question. A wholly absent or wholly invalid dimension receives

0; partial omissions follow the 1–3 anchors according to their scientific consequences. Appendix B defines the anchors and task-specific targets.

Source selection and evidence use. Retrieval judges the adequacy of the final source selection after identity, corpus-membership, time, and duplicate checks. Retrieved documents earn no automatic credit; valid alternative sources can receive full credit without matching the construction bibliography. Evidence separately assesses whether the submitted facts and locations are supported. Reasoning assesses the explicit scientific operations, and Answer assesses the requested product. Dimensions can share errors but require separate reasons; a correct answer does not compensate within another dimension for missing evidence or justification. Citation count, and explanation length are not scoring targets.

Evaluation procedure. A fixed reference-based judge uses verified sources, the original response, task rubric, item checklist, and reference materials to return four scores and source-linked reasons. Identity ambiguity and valid alternatives refer to LLM judge model, not automatic hallucination labels. Precision/Recall/F1 are auxiliary diagnostics on controlled subsets with adequate relevance annotation.

Task-level aggregation. Let $r_{mi}^{\text{source}}, e_{mi}, r_{mi}^{\text{reasoning}}, a_{mi} \in \{0, 1, 2, 3, 4\}$ be the four raw grades for model m on item i. The same rule applies to every task and condition:

$$s_{mi} = 100 * \frac{1}{16}\left(r_{mi}^{\text{source}} + e_{mi} + r_{mi}^{\text{reasoning}} + a_{mi}\right), \qquad S_{m,t} = \frac{1}{N_t}\sum_{i \in I_t} s_{mi}.$$

Here I_t is the predeclared eligible item set within the condition and modality track, and $N_t = |I_t|$. Required-output failures remain in the denominator; evaluator failures require resolution rather than candidate penalties. Equal weighting is a reporting convention, not evidence that ordinal anchors are interval-scaled or dimensions independent. The composite is compensatory, not accuracy or an all-criteria pass rate. Table 3 reports task mean score.

Reliability. The evaluation protocol fixes rubrics, judge configuration, and sensitivity analyses on development data before held-out evaluation. Human audits assess dimension-level agreement and task-level adequacy, with uncertainty accounting for shared source reviews. The audit procedure is specified in Appendix B; reliability results remain to be reported.

5 Experiments and Analysis

5.1 Systems and Experimental Setup

Construction diagnostics and validation. For knowledge graph construction and benchmark data generation, we use a local-deployed Qwen3.8-27B model (Qwen Team, 2026) for its multimodal capacity and efficiency. To guarantee the dataset quality generated, we introduce several review and screening mechanism in data flow. This part shows the effectiveness of screening procedure in 2 core steps, graph construction and benchmark generation. We manually selected 10 surveys and analyzed patterns of rejected entities and items in the graph construction. There are 33.6% for self-loop, 42.25% for fake relations, 17.6% for schema mismatch, which are three major mistakes.

Systems and conditions. We compare nine LLM-only models and four LLM + fixed retrieval systems (Table 3). The fixed-retrieval subset consists of DeepSeek-V4-Pro, Kimi K2.6, GLM-5.3, and GPT-5.6 Sol; all four also have LLM-only baselines. Paired comparisons use the same model version and eligible questions across all ten tasks. Both settings retain the question, instructions, and necessary attachments; only fixed retrieval adds shared retrieved evidence. Construction and judging models are documented separately, including role overlap. Search-tool evaluation is outside the reported experiments.

Retrieval augmented configuration. A shared pipeline uses only the question and public task instructions to query a frozen corpus and index. Query formation, ranking, evidence selection, and context limits are fixed across candidate models; results are reused for each model on that question. Content retains document identities and source locations without hidden relevance labels or reference-solution mappings. Retrieved results are attached to model input following naive RAG

Table 3: Main Results across Inquiry Tasks. Task-level RERA scores on a 0–100 scale (higher is better) for all evaluated systems. Score is the mean of the four normalized dimension grades, averaged first within items and then within a task (Section 4.2). Highest scores of each task are bolded.

Model	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10
Qwen3.8-27B (Qwen Team, 2026)	48	38	31	25	27	39	34	29	35	52
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	47	59	44	28	32	37	41	33	40	59
Kimi K2.6 (AI, 2026)	55	42	46	31	45	40	49	35	41	55
GLM-5.3 (GLM-5-Team, 2026)	72	68	47	37	43	44	46	40	49	62
Gemini 3.7 Flash Google (2026)	66	73	51	46	52	48	42	39	47	65
Claude Opus 4.8 (Anthropic, 2026)	58	81	58	52	58	53	48	45	53	69
Claude Opus 5.5 Anthropic (2026)	71	**89**	**65**	**57**	69	59	51	**58**	**62**	70
GPT-5.6 Sol OpenAI (2026a)	70	65	53	49	62	56	55	44	51	65
GPT-6 Astra OpenAI (2026b)	**79**	75	62	52	**71**	**64**	**58**	53	57	**72**
DeepSeek-V4-Pro + RAG	70	69	48	31	32	41	40	35	46	65
Kimi K2.6 + RAG	66	79	**53**	36	49	47	55	38	46	63
GLM-5.3 + RAG	**77**	85	50	41	44	47	51	**46**	**57**	**70**
GPT-5.6 Sol + RAG	75	**87**	**53**	**49**	**62**	**56**	**55**	44	51	65

setting, and models cannot issue follow-up searches. They select sources and construct their responses, including T02's organized evidence set. A valid authorized source recalled independently is eligible even if absent from the returned list. Final-source scores therefore assess system outputs, not the fixed retriever alone.

Execution controls. Paired settings use the same task instructions, output requirements, decoding settings, and output budget, apart from evidence-access instructions. Context budgets accommodate shared results without model-specific truncation. Text and visual tracks use pre-declared modality-compatible item sets. Freeze corpus/index versions, prompts, repetitions, and parameters before testing. Disable browsing and undeclared provider-side tools in LLM-only and fixed retrieval.

5.2 Scientific Inquiry Performance and Retrieval Effects

Scientific reasoning consistently lags behind conceptual organization. Every evaluated LLM-only model scores higher on question decomposition (T01) than on comparison, mechanism, and transfer reasoning (T06–T08). This consistent ordering reveals a shared performance gap between organizing a scientific inquiry and developing an evidence-grounded scientific account. Qualitative inspection supports this distinction: models often produce coherent problem structures but omit the scientific relations, applicability conditions, or supporting evidence needed to justify their conclusions. Their ability to structure an inquiry therefore does not reliably extend to carrying out the evidence-dependent operations required to resolve it. Within the evaluated tasks, scientific reasoning emerges as a recurring limitation across models, rather than a weakness confined to a particular system.

Proprietary models show a larger advantage on reasoning-intensive tasks. The evaluated proprietary models outperform the open-weight models overall, with an average score gap of [A] points. This difference is task-dependent: the gap increases from [B] points on question decomposition and literature selection (T01–T02) to [C] points on comparison, mechanism, and transfer reasoning (T06–T08). The larger separation on reasoning-intensive tasks shows that the model-group difference is not a uniform shift in performance across the benchmark. Instead, it is more pronounced when responses must connect findings, justify scientific relations, and preserve applicability conditions. Scientific reasoning thus represents a particularly important dimension of performance differentiation between the evaluated model groups. This comparison characterizes the tested systems; it does not establish weight availability as the cause of the observed gap.

Retrieval improves organization and selection more than scientific reasoning. For the same four models and eligible items, fixed retrieval improves T01 scores by 5–23 points and T02 scores by 10–37 points, whereas changes on T06–T08 range from -1 to +7 points. Notably, the smallest

reported T02 gain exceeds the largest reported gain on any of the three reasoning tasks. Supplying retrieved context therefore yields substantially different benefits across scientific operations: it improves question decomposition and literature selection more consistently than comparison, mechanism, and transfer reasoning. This pattern is consistent with a persistent limitation in model-side evidence integration and scientific inference, rather than a difficulty resolved by additional context alone. Incomplete retrieval coverage remains an alternative explanation, so these results do not isolate intrinsic reasoning ability. Nevertheless, they show that retrieval augmentation, as implemented here, is insufficient to produce comparable improvements in scientific reasoning.

Open questions are hard to grade critically. As a representative task of scientific reasoning and knowledge synthesize, T10 seems to be not as hard as other tasks. Inspection suggests that some corresponding future-work passages in the source reviews provide relatively limited evidential justification as the task is a open question and inevitable. High scores may therefore reflect the production of plausible extensions without equally strong evidence that the proposed directions address underdeveloped areas of the literature. This possibility should not be conflated with demonstrated evaluator leniency, nor should valid alternative proposals be penalized merely for differing from the reference. A stronger claim about research-gap identification requires checking whether proposed directions are tied to documented limitations, unresolved findings, or unmet evidence needs.

6 Conclusion

We introduced EarthLitBench, a benchmark for literature-grounded scientific inquiry in Earth science, comprising ten tasks, a review-grounded construction procedure, and the four-dimensional RERA evaluation protocol. The benchmark connects problem formulation and literature discovery with evidence understanding, scientific reasoning, and knowledge synthesis under controlled evidence access. Our experiments identify a gap between organizing scientific information and developing complete, evidence-supported judgments; retrieval helps address the former more consistently than the latter. EarthLitBench provides a basis for evaluating progress toward scientific inquiry systems that can both locate relevant literature and reason faithfully with its evidence.

7 Limitation

Our survey-grounded construction approach has three main limitations. First, constructing a dataset with corpus for publicly available benchmarks is constrained by access rights and redistribution permissions. Closed-access literature for which redistribution rights cannot be obtained cannot be included in the public version, resulting in the exclusion of some data due to a lack of evidentiary support. Second, the disciplinary coverage and temporal scope are determined by the survey papers—and their cited references—that are available and selected within the chosen publication period. Consequently, the benchmark dataset may suffer from under-representation regarding disciplines with fewer surveys during that period and emerging topics not yet summarized in review articles. These constraints limit the representativeness of the benchmark dataset and make it difficult to generalize research findings to disciplines and literature falling outside its scope.

AI Use Statement

We used generative AI tools to assist with software engineering and code implementation, and with manuscript drafting, revision, and language polishing. LLMs also form part of the research methodology: they assist with extracting and organizing literature-grounded knowledge and generating benchmark questions, reference justifications, and reference answers (Section 3.2), and serve as reference-based judges in the evaluation protocol (Section 4.2). These methodological uses are distinct from writing assistance. The study itself evaluates the literature-grounded scientific inquiry capabilities of LLMs in Earth science under the evidence-access conditions specified in §4.1.

LLM generation or judging alone does not establish scientific validity. The construction and evaluation protocols specify source-grounding checks, verification procedures, and human audits (Appendices D and F). The authors are responsible for reviewing AI-assisted work and take full responsibility for the final manuscript, code, benchmark artifacts, and reported results, including their scientific claims and interpretations.

REFERENCES

Moonshot AI. Kimi k2.6: Advancing open-source coding and agent swarm capabilities. https://huggingface.co/moonshotai/Kimi-K2.6, April 2026.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Anthropic. Claude (large language model). https://www.anthropic.com, 2026. Version: Claude 5.5 Opus, Accessed: 2026-09-26.

Zhiqing Cui, Xinxiang Yin, Yihong Tang, Xinglang Zhang, Yuanzhe Hu, Siru Zhong, Weidong Tang, Yuxuan Liang, Weijia Li, Ming Jin, Shirui Pan, Yuhao Kang, Dingyi Zhuang, and Jinhua Zhao. Earthverse: Benchmarking scientific agents across dynamic earth systems and natural hazards, 2026. URL https://arxiv.org/abs/2608.23525.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

GLM-5-Team. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google. Gemini 3.7 flash. Web release, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash.

Fei-Fei Jin, Seon Tae Kim, and Luis Bejarano. A coupled-stability index for enso. *Geophysical Research Letters*, 33(23), 2006. doi: https://doi.org/10.1029/2006GL027221. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006GL027221.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. ClimaQA: An automated evaluation framework for climate question answering models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=goFpCuJalN.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026a. Accessed: 2026-09-26.

OpenAI. Introducing GPT-6 Sol and Luna. https://openai.com/index/introducing-gpt-6-sol-and-luna/, 2026b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Will Steffen, Katherine Richardson, Johan Rockström, Sarah E. Cornell, Ingo Fetzer, Elena M. Bennett, Reinette Biggs, Stephen R. Carpenter, Wim de Vries, Cynthia A. de Wit, Carl Folke, Dieter Gerten, Jens Heinke, Georgina M. Mace, Linn M. Persson, Veerabhadran Ramanathan, Belinda Reyers, and Sverker Sörlin. The emergence and evolution of earth system science. *Nature Reviews Earth & Environment*, 1:54–63, 2020. doi: 10.1038/s43017-019-0005-6.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. Scifact-open: Towards open-domain scientific claim verification, 2022. URL https://arxiv.org/abs/2210.13777.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning*, 2024.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry, 2025. URL https://arxiv.org/abs/2507.16280.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jyYE06FL8G.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics, 2025. URL https://arxiv.org/abs/2509.00496.

A TASKS AND INTERFACES

A.1 OVERVIEW OF TASK LIST

Details of task definition and corresponding scientific operation and expect output are presented in Tab. A1.

Table A1: Inquiry Tasks and Benchmark Interface. Four connected capability groups, distinguished by their primary scientific operations and outputs.

Capability Group	**Task ID / Name**	**Primary Scientific Operation**	**Primary Output**
Problem Formulation and Literature Discovery	T01: Existing research and scientific question decomposition	Decompose objects, variables, subquestions, scope, and dependencies	Structured question decomposition and evidence needs
Problem Formulation and Literature Discovery	T02: Literature discovery and screening	Identify eligible studies and organize sources around explicit evidence needs	Included sources, located screening reasons, study–need mappings, and coverage limits
Study and Evidence Understanding	T03: Study design, data, and method understanding	Explain how a study was conducted and what its assumptions permit	Structured study account with assumptions and scope
Study and Evidence Understanding	T04: Quantitative result and figure/table interpretation	Interpret quantities, baselines, trends, units, and uncertainty	Located result interpretation and any necessary calculations
Study and Evidence Understanding	T05: Claim–evidence and citation-role verification	Judge whether specified evidence supports, qualifies, contradicts, or cannot support a claim	Claim–evidence verdict with citation-role justification
Scientific Reasoning	T06: Cross-study comparability and disagreement reconciliation	Assess comparability and explain agreement or disagreement under study-specific conditions	Comparability assessment and qualified reconciliation
Scientific Reasoning	T07: Mechanism, causal, and feedback reasoning	Explain supported process pathways, feedback, and causal limits	Evidence-linked mechanism account with conditions and alternatives
Scientific Reasoning	T08: Scale, region, time, and scenario transfer	Assess a finding against explicitly changed target conditions	Transfer judgment with assumptions and failure boundaries
Knowledge Synthesis	T09: Multi-evidence synthesis and uncertainty calibration	Weigh independent evidence, counterevidence, and study quality	Collective conclusion with calibrated uncertainty and scope
Knowledge Synthesis	T10: Research gap and future work identification	Identify justified limitations of current knowledge within a fixed literature time frame	Evidence-grounded gaps, significance, and corresponding research directions

A.2 CONNECTED INQUIRY, INDEPENDENTLY EVALUATED INSTANCES

The four groups connect question and evidence organization (T01–T02), study and evidence understanding (T03–T05), scientific reasoning (T06–T08), and knowledge synthesis (T09–T10). A useful conceptual progression is:

Question and evidence needs $\rightarrow$ eligible studies $\rightarrow$ understood and checked findings $\rightarrow$ comparisons, mechanisms, and transfer judgments $\rightarrow$ collective conclusions and research directions.

This is a dependency map, not a prescribed sequence. Evidence checking can motivate further screening; a synthesis may use comparisons without requiring a mechanism or transfer analysis. An instance is labeled by its principal requested operation and evaluated independently. Its public inputs must specify the necessary problem context without exposing upstream gold or requiring comple-

tion of another benchmark item. Shared sources do not by themselves make two tasks duplicates; changing a task label without changing the required operation does not make two questions distinct.

The cards below describe each task's scientific context, not additional evidence entitlement. LLM-only receives the public question and necessary attachments; fixed retrieval adds shared results, and the protocol-defined search-tool setting permits queries to the same frozen corpus. T02 is eligible in all three. Distinguish corpus eligibility, actual retrieved material, and explicit attachments. Tag each item as open-selection or source-specified before evaluation; specified-source items assess identification/selection rather than autonomous discovery. Do not add unnecessary citation demands to make a source-specified task appear harder.

A.3 TASK CARDS

T01: Existing research and scientific question decomposition

- **Goal and input:** Organize a supplied research question or existing research account within stated objects, scope, and time constraints.
- **Operation and output:** Identify subquestions, variables, methods where relevant, dependencies, and evidence needs; return a structured decomposition with located support and an explanation of those relationships.
- **Links and boundary:** Establishes criteria that can guide T02 and organizes later analysis. It does not require proposing novel directions (T10). A new research agenda without decomposition does not fulfill T01; questions requiring unseen review wording are ineligible.

T02: Literature discovery and screening

- **Goal and input:** Establish an evidence set for a scoped question and any supplied subquestions, using explicit eligibility criteria and evidence needs within the frozen authorized corpus.
- **Operation and output:** Return included sources, located eligibility facts and screening reasons, study–evidence-need mappings, and bounded coverage limitations. Use the same final-output interface in all conditions; no synthetic scientific conclusion or screening of unseen candidates is required. Exhaustive inclusion/exclusion of a supplied list belongs to a separately reported auxiliary screening test.
- **Links and boundary:** Connects T01's evidence needs to the studies examined in T03–T09. Eligibility is not mutual comparability (T06), a scientific support verdict (T05), or a collective conclusion (T09). A relevant study with different conditions is not excluded merely because direct comparison is difficult. Missing eligibility evidence triggers item review rather than an invented reference decision.

T03: Study design, data, and method understanding

- **Goal and input:** Explain how a specified primary study investigates its question, using identified study materials and the requested design or method focus.
- **Operation and output:** Relate data, variables, methods, assumptions, and scope to the study objective; return a coherent study account with source locations and stated limitations.
- **Links and boundary:** Supplies context for T04 result interpretation, T05 verification, and T06 comparability. It is not principally a numerical-result exercise (T04) or a verdict on whether a particular claim is supported (T05). The required method information must be available in authorized materials.

T04: Quantitative result and figure/table interpretation

- **Goal and input:** Interpret specified quantitative text, figures, or tables with accessible units, labels, baselines, and uncertainty information.

- **Operation and output:** Read and relate quantities, perform necessary calculations or normalization, and explain supported trends or contrasts; identify exact passages, panels, or cells.
- **Links and boundary:** Provides result-level evidence for T05–T09. A within-study result contrast can be T04; explaining cross-study disagreement is T06. Observed trends alone do not justify causal mechanisms (T07). Unreadable essential media cannot support an eligible item; hidden survey figures cannot be silently assumed visible.

T05: Claim–evidence and citation-role verification

- **Goal and input:** Evaluate a specified claim against identified cited material, including the claim's conditions and the citation context needed for interpretation.
- **Operation and output:** Check attribution, scope, and evidential strength; return a supported, qualified, contradicted, or insufficient-support verdict with located justification.
- **Links and boundary:** Uses T03–T04 understanding and supplies checked premises for later reasoning. Unlike T04, its target is the support relation, not only the reported result. Unlike T06, it need not reconcile multiple studies. Citation occurrence alone is not evidence of support.

T06: Cross-study comparability and disagreement reconciliation

- **Goal and input:** Compare specified findings from at least two studies addressing a related question, with relevant methods and scope information.
- **Operation and output:** Align definitions, baselines, scales, and conditions; distinguish genuine disagreement from conditional differences or non-comparability. Return a comparison and appropriately qualified reconciliation.
- **Links and boundary:** Draws on T03–T05 and can inform T07–T09. T02 establishes eligibility, not comparability. T06 may legitimately conclude that disagreement cannot be resolved; T09 instead asks what the wider evidence collectively supports. A forced reconciliation without sufficient evidence is not a valid completion.

T07: Mechanism, causal, and feedback reasoning

- **Goal and input:** Explain a specified process or finding using literature evidence about drivers, intermediate processes, outcomes, and relevant conditions.
- **Operation and output:** Construct supported pathways or feedback accounts, distinguish association from causal evidence, and discuss warranted alternatives and limits.
- **Links and boundary:** May use verified claims and cross-study comparisons and support T08 or T09. Repeating an observed trend is insufficient. Explaining a mechanism does not establish that it transfers to a new setting (T08); unsupported causal steps cannot be supplied to complete an attractive pathway.

T08: Scale, region, time, and scenario transfer

- **Goal and input:** Assess a finding or mechanism from a specified source setting against an explicitly different target setting.
- **Operation and output:** Compare the relevant conditions, test transfer assumptions, and identify supported, conditional, or unsupported extrapolations with failure boundaries.
- **Links and boundary:** Uses study scope, comparisons, and mechanisms from T03–T07 and constrains T09 synthesis. Without explicit source and target conditions, a generic limitations question is not T08. Identifying a transfer limitation does not by itself constitute the research agenda required by T10.

T09: Multi-evidence synthesis and uncertainty calibration

- **Goal and input:** Develop a collective judgment for a scoped question using multiple studies and their relevant conditions, quality, and counterevidence.
- **Operation and output:** Weigh independent evidence, agreement, disagreement, and limitations; return an integrated conclusion with calibrated uncertainty and an explicit applicability range.
- **Links and boundary:** Integrates understanding and reasoning from T03–T08 and can motivate T10. It requires more than a sequence of summaries or a single pairwise reconciliation (T06). Future-work suggestions do not substitute for a current conclusion. Repeated reports of one study are not independent corroboration.

T10: Research gap and future work identification

- **Goal and input:** Identify underdeveloped directions for a topic within an explicit literature time frame, using accessible evidence about knowledge, methods, coverage, and unresolved questions.
- **Operation and output:** Distinguish substantive limitations from merely missing retrieved material; explain each gap's scientific significance and connect it to a justified direction for further work.
- **Links and boundary:** Builds on the limitations exposed by T03–T09. It is not decomposition of an already supplied question (T01) or only an uncertainty summary (T09). A historical limitation is not automatically unresolved at the cutoff, and an absent search hit does not establish field-wide absence. Supported alternative directions are valid; reproducing the hidden review's list is not required.

B Evaluation validity

B.1 Formal applicability by task and condition

Every eligible item has four required dimensions. Missing content is a scoring failure, not N/A; lacking retrieval access is not an exemption. Modality-compatible item sets are declared before comparison. Open-selection and source-specified items are tagged separately: when the question already names its sources or supplies the necessary material, Retrieval assesses identification/selection, not autonomous discovery. Report these strata and their sample sizes; do not impose unnecessary additional references or hide essential attachments to manufacture a discovery requirement.

B.2 Task-specific judging targets

Retrieval coverage concerns the suitability of selected documents for the evidence needs stated in the question, not whether a candidate's specific factual claims are entailed; the latter belongs to Evidence. Reasoning concerns the submitted scientific operations, including short calculations, design explanations, and screening decisions. Answer concerns the final task product, not similarity to reference wording. The dimensions need not be statistically independent: one scientific error can affect more than one, but each affected score needs its own justification.

Rubric hierarchy and scoring unit. Shared anchors define score meaning; task criteria define scientific operations; a reviewed item checklist states necessary content, source needs and locations, numerical tolerances where relevant, and acceptable alternatives. Essential requirements and central-error criteria are fixed before candidate evaluation and must follow from the public question. Generated checklists require verification; they are not accepted rubrics by default. Each item uses the same rubric across models and conditions. Alternative-source judgments and reference corrections are versioned and applied consistently to all affected responses.

A *central* defect invalidates a principal source requirement, claim, inference, or task product identified by the checklist. A *substantive* defect leaves meaningful partial validity but fails a required component; a *minor* defect does not alter the principal judgment or its scope. A central error caps the affected dimension at 1; use 0 if no valid required content remains. A questionable citation is

Table B1: Scientific Targets for Reference-Based Judging. Task criteria specialize the shared anchors without changing scales or weights.

Task	Retrieval (Reference Selection) target	Evidence Fidelity target	Reasoning Fidelity target	Answer Fidelity target
T01	Sources address the required objects, variables, scope, and evidence needs	Sources support the stated research objects, variables, and scope	Decomposition and dependencies are justified by the question and evidence needs	Relevant components are covered without changing the supplied problem
T02	Eligible sources cover the explicit evidence needs and selection criteria	Located study facts accurately establish eligibility or exclusion conditions	Stated criteria are applied consistently to those facts and evidence needs	The evidence set is organized into correct study–need mappings with explicit, bounded coverage qualifications
T03	The specified study and necessary design, data, or method sources are identified	Design, data, method, and assumption descriptions match the study	Explanations correctly relate design choices and assumptions to what the study can establish	The requested study account is accurate, complete, and scoped
T04	Sources contain the target results, figures/tables, and required interpretive context	Values, units, baselines, and uncertainty are correctly attributed and located	Calculations, normalization, and interpretation follow from the reported quantities	The quantitative interpretation is correct and appropriately qualified
T05	The cited work and necessary claim/citation context are identified	Cited material and citation context are accurately represented	The support verdict respects evidential strength, conditions, and attribution	The verdict and citation-role judgment answer the specified claim
T06	Sources cover the compared studies and their relevant methodological or scope differences	Study-specific findings, methods, and conditions are represented faithfully	Comparability checks and explanations of disagreement follow from those differences	The comparison or unresolved-disagreement conclusion is justified and bounded
T07	Sources cover required process links and alternatives requested by the question	Each asserted process link has appropriate support and conditions	Pathway direction, feedback, alternatives, and causal strength are handled validly	The mechanism account covers the requested process without unsupported causal claims
T08	Sources address relevant original and target conditions and requested boundary evidence	Source and target conditions, including relevant counterevidence, are correctly stated	Transfer assumptions and failure boundaries are tested rather than asserted	The applicability judgment answers the specified target setting with limits
T09	Sources cover required evidence types, independent studies, and relevant counterevidence	Evidence is faithfully represented, including relevant counterevidence and study dependence	Weighting and uncertainty reflect study quality, comparability, and independence	The collective conclusion is complete, calibrated, and scoped
T10	Sources cover documented advances and limitations within the stated literature cutoff	Evidence supports a time-bounded limitation, not merely an absent retrieval hit	The gap's significance and proposed direction follow from the documented limitation	Gaps and directions are specific, justified, and responsive to the topic

not automatically a central defect, and unresolved identity is not a confirmed fabrication. Rubric definition and empirical judge reliability are assessed separately.

Dimension boundaries. A correct bibliography can earn Retrieval credit without Evidence credit when the response supplies no located factual support. Source identity resolution does not prove that the model read the source. A correct answer can earn Answer credit without Evidence or Reasoning credit; missing citations do not automatically erase explicitly valid reasoning. Conversely,

Table B2: Shared 0–4 Judge Anchors. Identity and eligibility checks inform Retrieval; all four grades remain task-conditioned judgments.

Score	Retrieval (Reference Selection)	Evidence Fidelity	Reasoning Fidelity	Answer Fidelity
4 — Complete	Valid, eligible, relevant sources adequately cover all required evidence needs; no substantive selection defect remains	All required substantive claims have locatable support; attribution, quantities, and conditions are faithful	All required operations and inferential links are valid, with relevant assumptions and uncertainty handled	All requested conclusions or structured components are correct, complete, and appropriately scoped
3 — Minor deficiencies	The main evidence needs are covered; only minor selection or coverage deficiencies remain	Central claims are supported; only minor attribution or coverage deficiencies remain	The central inference and necessary operations are sound; only minor explanatory gaps remain	The requested judgment is correct and substantially complete, with only minor omissions or imprecision
2 — Partial	Some required needs are covered, but substantive coverage gaps or unsuitable selections remain	Some required claims have faithful support, but substantive coverage or fidelity gaps remain	Some necessary operations are valid, but a substantive step or condition is missing or incorrect	Meaningful parts are correct, but a substantive requested component or qualification is missing or incorrect
1 — Central failure	Only limited relevant sources remain; the principal source base is missing or depends on a fabricated or ineligible source	Limited valid support remains, but a central required claim is unsupported or relies on fabricated, misattributed, or misrepresented evidence	Limited valid reasoning remains, but a central inferential error invalidates the proposed judgment	Limited correct content remains, but the principal conclusion or requested output is incorrect
0 — Absent or invalid	No sources are submitted, or none is valid, eligible, and task-relevant	No required located evidence is supplied, or none provides valid support for the required claims	No required justification is supplied, or no valid task-relevant operation is demonstrated	No required answer is supplied, or no task-relevant answer component is correct

a valid inference from stated premises cannot repair a misquoted result. Score these defects separately rather than propagating a single failure to all four dimensions. Supported non-comparability, insufficient-support, and uncertainty judgments can receive full credit when responsive to the task; a generic refusal is not such a judgment.

Illustrative aggregation checks, not experimental results. Raw Retrieval/Evidence/Reasoning/Answer grades of 3/2/4/3 yield $100 \times 12/16 = 75$ for any task, including T02. Grades of 4/4/4/4 yield 100; 0/0/0/0 yield 0. A correct, explicitly justified answer with no selected sources or located evidence can receive 0/0/4/4, yielding 50, not a certification of evidence-grounded success. Equal scores do not imply identical failure patterns.