communications earth & environment

Article

A Nature Portfolio journal

https://doi.org/10.1038/s43247-026-03878-1

A domain-adapted retrieval-augmented framework for transparent Earth system assessment

Özge Kart Tokmak ✉, Levke Caesar, Josef Ludescher & Boris Sakschewski

Rapid growth and increasing fragmentation of Earth-system science literature pose a major barrier to timely, evidence-based assessment of global environmental change and planetary boundary conditions. Here we present a domain-adapted retrieval-augmented generation framework for Earth-system research, in which domain adaptation is achieved through curated scholarly corpora and a structured evaluation design, supporting transparent and reproducible deployment. The system continuously curates literature from major scholarly databases and integrates it with locally hosted, open-weight language models to enable source-level traceability, long-term accessibility, and independence from external commercial services. When queried, it retrieves relevant scientific sources and generates responses grounded directly in those documents, reducing screening effort and addressing reliability concerns associated with general-purpose language models. Quantitative evaluation shows high contextual faithfulness (0.96), strong answer relevance (0.88), and robust correctness (0.81). Our findings indicate that domain-adapted retrieval-augmented generation can strengthen the scientific basis for Earth-system assessment by enabling transparent, continuously updated, source-attributed synthesis across rapidly expanding literature.

Global environmental change ranks among the most pressing challenges of this century[1]. Understanding its drivers, interactions, and consequences requires a holistic Earth-system perspective that integrates work from atmospheric science, ecology, hydrology, biogeochemistry, and human-environment research[2,3]. This scientific landscape is vast, interdisciplinary, and rapidly evolving[4]. The volume of peer-reviewed publications grows at a pace that makes staying current increasingly difficult for researchers, assessment teams, and policymakers[4]. As a result, maintaining an up-to-date, coherent synthesis of knowledge across Earth-system domains has become a major bottleneck in global-change research.

Within this broader context, the Planetary Boundaries Framework offers a prominent and well-structured example of an integrated Earth-system assessment. First introduced by Rockström et al. (2009)[5] and subsequently refined by Steffen et al. (2015)[2] and Richardson et al. (2023)[6], the framework provides a scientific basis for defining a "safe operating space" for humanity by identifying biophysical thresholds across nine interconnected domains, including climate change, biosphere integrity, freshwater use, land system change, ocean acidification, ozone depletion, atmospheric aerosol loading, novel entities and biogeochemical flows. Its influence extends across sustainability science, guiding global assessments, policy debates, and modelling initiatives. The framework also illustrates broader challenges facing Earth-system assessment and integration[5]. Literature relevant to each boundary is highly interdisciplinary and continues to expand rapidly[4,7], increasingly challenging manual review processes. The heterogeneity of data types, methodologies, and disciplinary languages further complicates systematic and transparent synthesis. This motivates the development of approaches capable of keeping pace with rapidly expanding scientific domains and providing verifiable, up-to-date syntheses for assessment processes and decision-relevant analysis. This work is motivated by the need for transparent, continuously updated tools that can support Planetary Boundaries assessments (like the annual Planetary Health Check report). As an initial demonstration of this approach, we focus on the Climate Change boundary, which provides both a rich literature base and a well-defined conceptual structure for evaluating retrieval-augmented question-answering performance. More broadly, this setting reflects a common challenge in Earth-system science: integrating rapidly expanding and heterogeneous knowledge into coherent, assessment-relevant synthesis.

Recent studies have applied natural language processing (NLP) methods to assist with large-scale literature analysis in climate and sustainability domains[8–10]. Meanwhile, the emergence of large language models

Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 6012 03, D-14412 Potsdam, Germany.
✉e-mail: oezgeka@pik-potsdam.de

(LLMs) has created new opportunities for automating aspects of scientific synthesis[11–13], yet their direct use in specialised scientific contexts is constrained by well-documented issues[14], including factual unreliability and lack of domain grounding. Retrieval-Augmented Generation (RAG) aims to address some of these limitations by conditioning model outputs on external source documents, and new evaluation approaches, such as the Retrieval-Augmented Generation Assessment (RAGAs) framework, seek to quantify how effectively such systems integrate retrieved information[15–17].

In this paper, we propose a domain-adapted question-answering (QA) system designed to support Earth system and planetary boundaries research through automated, evidence-grounded synthesis of peer-reviewed literature supplemented by OpenAlex-indexed preprints to capture emerging research, with transparent source attribution. Here, "domain-adapted" refers to adaptation through curated Earth-system corpora, domain-expert question design, and evaluation criteria reflecting scientific accuracy and sufficiency within Earth-system research. The system continuously aggregates scientific literature from major scholarly databases and integrates these sources into a structured, searchable knowledge base. A RAG architecture combines semantic retrieval with generative reasoning, enabling users to query the system in natural language and receive transparent, source-grounded answers. The operational QA system is implemented using locally hosted, open-weight language models, enabling transparent configuration and reproducible deployment independent of external commercial APIs.

Several recent studies have demonstrated the feasibility of retrieval-augmented large language models in climate and sustainability contexts. For example, ChatClimate grounds GPT-4 responses in the IPCC AR6 reports to enhance reliability in climate-related question answering[8], while ChatReport applies LLM-based retrieval to corporate sustainability reports with traceable outputs and expert-in-the-loop development[9]. Similarly, ChatNetZero employs RAG with additional reference and anti-hallucination modules to improve factual accuracy in net-zero policy analysis[18]. These systems focus on application-oriented implementations within defined document collections and task settings. Despite recent progress in domain-specific RAG systems, there remains comparatively limited work focusing on reproducible and continuously updateable retrieval frameworks designed for the highly interdisciplinary and rapidly evolving literature of Earth-system science.

In contrast, the present study addresses the methodological design and evaluation of RAG systems for scientific knowledge synthesis in Earth-system science. Specifically, it introduces (i) a cross-database scholarly corpus that integrates publications from multiple sources with automated, periodic ingestion, enabling continuous corpus evolution beyond static collections; (ii) an explicit evaluation framework that separates retrieval and generation performance under controlled answerability assumptions; and (iii) a combined assessment strategy that integrates automated grounding metrics with domain-expert evaluation of scientific sufficiency.

The contribution, therefore, lies not only in applying RAG to Earth-system science but in providing a structured, reproducible framework for constructing, evaluating, and deploying RAG-based systems over evolving scientific literature. This framework is designed to support traceable and updatable knowledge synthesis in Earth-system research, with potential applications in large-scale assessments and interdisciplinary analysis.

To evaluate the system's performance, we constructed a domain-specific dataset and applied the RAGAs framework to assess retrieval quality and the grounding of generated answers. The results indicate high contextual faithfulness (0.96), strong answer relevance (0.88), and robust answer correctness (0.81), showing that the retrieval augmentation improves grounding and factual alignment of generated responses within the controlled evaluation setting. Complementing these quantitative results, an expert evaluation found the answers broadly accurate but often in need of additional contextualisation, quantitative detail, or methodological nuance. This highlights a distinction between paragraph-grounded correctness and the broader expectations of scientific sufficiency in expert assessment. In the remainder of this paper, we describe the system design, present quantitative evaluations of its retrieval-augmented performance, and discuss its potential to strengthen evidence synthesis for Earth-system assessment and planetary boundaries research.

Results and discussion

Experimental configuration

To evaluate the initial performance of the RAG-based prototype, climate change was selected as a use case among the nine planetary boundaries. The retrieval and generation performance have been evaluated separately using the evaluation framework described in the Methods section. Four open-weight large language models were used in the experiments, using their Ollama model identifiers: *gpt-oss:20b*[19], *deepseek-r1:70b*[20], *llama4:16x17b*[21], and *gemma3:27b*[22], corresponding to models with approximately 20B, 70B, 109B (reported for the 16x17B architecture), and 27B parameters, respectively. Models were run using the default Ollama quantization settings (MXFP4 for *gpt-oss:20b* and Q4_K_M for the remaining models). The experiments were run in a local client-server setup: the web application, vector database, retrieval, and evaluation components were executed on a local application server, while language-model inference was served separately through Ollama on a dedicated machine equipped with two NVIDIA V100 GPUs with 32 GB VRAM each and approximately 64 GB system memory. The models were tested in two modes: (1) retrieval-augmented generation (RAG), using the system's semantic search mechanism to provide grounding evidence, and (2) direct generation without retrieval.

To ensure the models operate as deterministically as possible, they were invoked via Ollama, setting the temperature parameter to 0.00. An exception was made for Gpt-oss, for which the temperature was set to 0.05 with a seed parameter set to 42, as using a temperature of 0.00 caused it to return only short phrase-level answers rather than complete sentences. Performance was quantified using RAGAs metrics, which enable separate assessment of retrieval quality, grounding fidelity, relevance, and factual accuracy.

Retrieval and generation performance

Retrieval performance. We first evaluated the retrieval component using the context precision and context recall metrics from the RAGAs framework. The retriever was configured to retrieve four chunks per query. Across all 50 question-answer pairs, the system achieved a mean context precision of 1.00 and a context recall of 0.98. Under the controlled single-document answerability assumption described in Methods, these values indicate consistent ranking of the relevant chunks containing the reference claims among the top retrieved results. Minor variation in recall was observed for a small subset of questions where not all reference claims were supported within the top retrieved chunks. In these cases, relevant claims were distributed across chunk boundaries or ranked beyond the fixed retrieval window.

Generation performance. RAGAs generation metrics were used to assess how effectively each model used the retrieved evidence. Table 1 reports the mean scores and associated standard deviations for the three metrics (faithfulness, answer relevancy, and answer correctness), while Fig. 1 presents the corresponding mean scores together with 95% bootstrap confidence intervals (based on 5000 resamples). The analysis reveals variability in performance across the four evaluated LLMs: Gemma-3, Gpt-oss, Deepseek-r1, and Llama-4.

Faithfulness measures the degree to which generated answers are grounded in the provided context. Gemma-3 achieved the highest mean score (0.96 ± 0.08), showing both strong adherence and low variance. Deepseek-r1 and Llama-4 performed similarly (0.91 ± 0.17 and 0.91 ± 0.19, respectively), while Gpt-oss showed the lowest and most variable faithfulness score (0.86 ± 0.22).

Answer relevancy measures how directly a generated response addresses the user's question. GPT-oss achieved the highest relevancy (0.88 ± 0.23), followed by Deepseek-r1 (0.86 ± 0.26). Llama-4 yielded the lowest score (0.77 ± 0.18). Some queries received a relevancy score of 0 (reducing the overall averages for all four models) because the judge model

classifies responses as "noncommittal" when they contain words like "*likely*", "*uncertain*" or "*possibly*". However, these terms often originate from the retrieved scientific text rather than from model uncertainty. This behaviour reflects limitations of the judge model in distinguishing contextual scientific uncertainty from genuine noncommittal responses, rather than weaknesses in the underlying generation component.

Answer correctness reflects factual accuracy and semantic similarity of the generated response to the reference answers. Gemma-3 again obtained the highest mean score (0.81 ± 0.13), followed by Gpt-oss (0.77 ± 0.18), Deepseek-r1 (0.75 ± 0.15) and Llama-4 (0.72 ± 0.18).

Overall, these findings indicate that model size (GPT-oss: 20B, Deepseek-r1: 70B, Llama-4: 109B, Gemma-3: 27B) does not appear to translate into higher performance in this retrieval-augmented setting. Gemma-3 provided the strongest balance between faithfulness and factual accuracy, whereas Llama-4 benefited least from retrieval and more often relied on internal knowledge, leading to unsupported extrapolations.

Impact of retrieval augmentation

To contextualise the effect of retrieval within the proposed system, we compared each model's performance in the full RAG configuration with a direct-generation setting (i.e., generation without retrieved context). Table 2 summarises the differences across answer relevancy and answer correctness.

Table 1 | RAGAs generation scores across multiple language models

Model	Faithfulness	Answer Relevancy	Answer Correctness
Gemma-3	0.96 ± 0.08	0.84 ± 0.28	0.81 ± 0.13
Gpt-oss	0.86 ± 0.22	0.88 ± 0.23	0.77 ± 0.18
Deepseek-r1	0.91 ± 0.17	0.86 ± 0.26	0.75 ± 0.15
Llama-4	0.91 ± 0.19	0.77 ± 0.18	0.72 ± 0.18

Mean ± standard deviation for three RAGAs metrics (faithfulness, answer relevancy, and answer correctness) computed over 50 climate-related question-answer pairs. Faithfulness reflects the degree to which generated answers are grounded in the retrieved evidence; answer relevancy measures alignment with the user's query; and answer correctness captures factual accuracy and semantic similarity relative to reference answers (see Methods). The results illustrate substantial differences across models, with Gemma-3 achieving the strongest overall balance of grounding and accuracy.

Across all four models, retrieval was associated with higher answer correctness, although the magnitude of improvement varied by model architecture. Gemma-3 demonstrated the largest overall gain in answer correctness (+0.19), indicating that retrieval gives it substantial factual support, while Deepseek-r1 benefited most in relevancy (+0.22), suggesting that retrieval helps it stay more focused on user intent. Llama-4 showed smaller improvements in both metrics.

As illustrated in Fig. 2, retrieval augmentation had limited influence on answer relevancy for most models, as responses often appeared contextually appropriate even in the absence of retrieved evidence. However, answer correctness generally increased when retrieval was enabled. This pattern suggests that language models may produce responses that sound plausible even when supporting evidence is absent, while retrieval grounding introduces domain-specific information that can improve factual alignment with the literature. Taken together, these observations illustrate how retrieval grounding influences response quality within the evaluated system configuration and align with previously reported benefits of retrieval-augmented approaches.

Expert evaluation results

Two domain experts independently evaluated 24 retrieval-augmented responses generated by the Gemma-3 model, selected for expert review based on its strong balance between faithfulness and answer correctness, in automated evaluations. The 24 questions were randomly sampled from the full evaluation set. Responses were assessed using the *Scientific Accuracy, Completeness & Scientific Sufficiency* and *Alignment with the Source Paper* criteria described in Methods. The individual expert assessments for the sampled responses are provided in the Supplementary Table 1 and can be linked to the question-answer dataset using the question IDs.

Both raters gave relatively high scores on the *Scientific Accuracy* scale (Rater A mean = 4.46, SD = 0.93; Rater B mean = 4.12, SD = 1.19), indicating that they broadly perceived the answers as scientifically correct and consistent with the underlying literature. Scores for *Completeness & Scientific Sufficiency* were moderate (Rater A mean = 3.29, SD = 0.86; Rater B mean = 3.33, SD = 1.40), reflecting that, while responses captured key mechanisms, additional contextual detail, quantitative information, or methodological nuance was often needed. Scores for *Alignment with the Source Paper* were also high (Rater A mean = 4.12, SD = 0.99; Rater B mean = 3.88, SD = 1.33), indicating that responses remained broadly faithful to the findings and claims of the referenced literature.

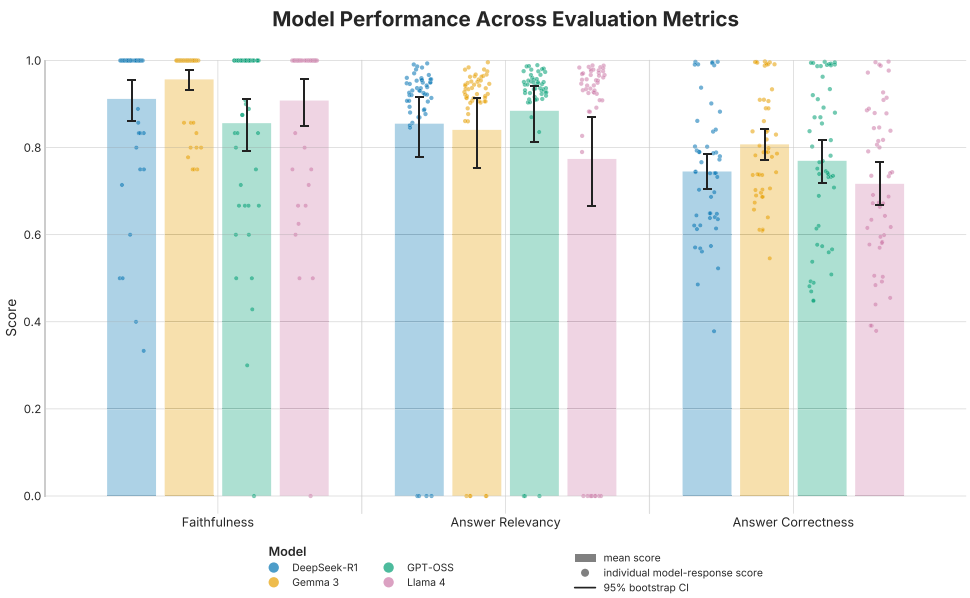

Fig. 1 | Performance of four language models across the RAGAs evaluation metrics. Bars show mean scores, points show individual model-response scores, and error bars show 95% bootstrap confidence intervals of the mean based on 5,000 resamples. Results are shown for three RAGAS metrics: faithfulness (grounding in retrieved context), answer relevancy (alignment with the user query), and answer correctness (factual accuracy and semantic similarity). For each model-metric combination, n = 50, corresponding to one model response to each of 50 climate-related queries. Gemma-3 achieves the strongest overall balance between grounding and factual accuracy, while Llama-4 shows the smallest gains from retrieval. Despite substantial variation in parameter size (Gpt-oss: 20B, Deepseek-r1: 70B, Llama-4: 109B, Gemma-3: 27B), the results show no clear positive correlation between model size and performance.

Table 2 | Comparison of retrieval-augmented and direct LLM performance

Model	Answer Relevancy (RAG)	Answer Relevancy (Direct)	Answer Correctness (RAG)	Answer Correctness (Direct)
Gemma-3	0.84	0.81	0.81	0.62
Gpt-oss	0.88	0.88	0.77	0.60
Deepseek-r1	0.85	0.63	0.75	0.58
Llama-4	0.77	0.73	0.72	0.63

Mean answer relevancy and answer correctness scores for four open-weight models in the RAG and direct-generation settings. Retrieval improves correctness across all models, with varying effects on relevancy.

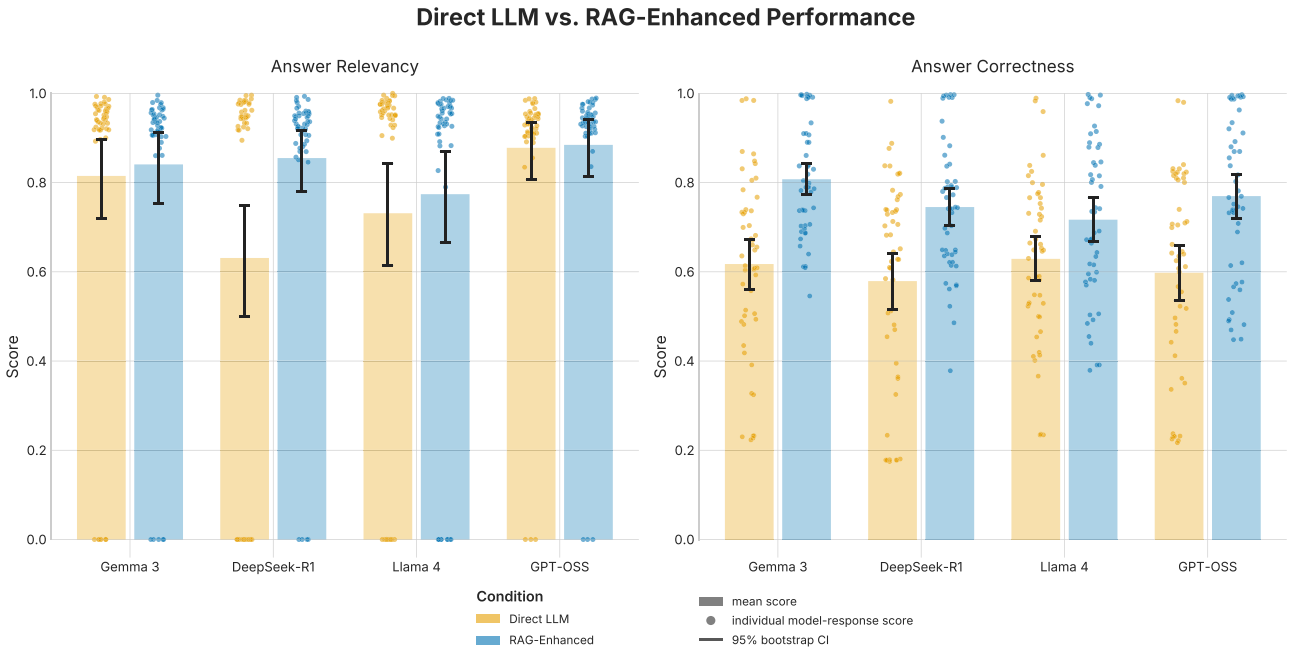

Fig. 2 | Effect of retrieval on answer relevancy and correctness across models. Bars show mean scores, points show individual model-response scores, and error bars show 95% bootstrap confidence intervals of the mean based on 5,000 resamples. Results are shown for direct generation and RAG-supported generation across two RAGAS metrics: answer relevancy and answer correctness. For each model generation setting-metric combination, $n = 50$, corresponding to one response to each of 50 climate-related queries. Retrieval has limited influence on answer relevancy for most models but substantially increases answer correctness, indicating that grounding responses in retrieved domain literature improves factual accuracy.

These results highlight a divergence between automated evaluation metrics, which capture retrieval grounding and contextual correctness, and expert assessments, which more critically evaluate scientific sufficiency, completeness, alignment with the source literature, and broader domain consistency. Agreement rates were 0.875 for *Scientific Accuracy* and *Alignment with the Source Paper*, and 0.708 for *Completeness & Scientific Sufficiency*.

Qualitative feedback provided further insight into these patterns. A recurring theme across the qualitative feedback was that scientific evaluation is inherently open-ended: experts can always identify additional processes, caveats, uncertainties, or alternative studies that could complement a given answer. As a result, an answer may be accurate yet still feel incomplete from an expert perspective. Several responses were judged as factually correct but missing key qualifiers (such as baselines, scenario dependence, uncertainty ranges) or relying too heavily on a single source without incorporating relevant contrasting findings from the broader literature. Experts also emphasised the importance of scientific precision. In some cases, hedging language ("may", "could", "likely") present in the source text was reduced or omitted in the generated responses, which was interpreted as a loss of precision. This highlights a tension between producing concise, declarative answers and preserving the qualified statements that are central to scientific communication. Interestingly, evaluation patterns suggested that greater domain expertise corresponded to more critical assessments: the more deeply an expert understood a topic, the more gaps, missing contextualisation, or definitional ambiguities they identified. This reflects the broader challenge that even human literature synthesis struggles with, namely reconciling differences in baselines, definitions, spatial and temporal scales, and methodological assumptions across studies.

Finally, several comments indicated that while the grounding mechanism reliably prevents hallucinations, not all experts prioritised groundedness in the same way as LLM evaluation frameworks. Experts tended to assess responses against their broader domain knowledge, which may extend beyond the retrieved text. This reveals an important distinction: grounding ensures reliability, but narrow grounding to a single paper can produce scientifically correct yet less synthetic answers, whereas experts often expect responses that integrate multiple competing findings, uncertainties, and literatures.

These findings provide indicative insights into expert perceptions of answer quality and do not constitute definitive performance estimates.

Conclusion

This study provides a domain-adapted Retrieval-Augmented Generation (RAG) framework for Earth system and planetary boundary research, supporting automated, evidence-based synthesis of scientific literature. The framework integrates automated cross-database literature ingestion and a structured evaluation design, alongside open-weight, locally hosted

language models to enable transparent and reproducible deployment. By combining semantic retrieval with generative reasoning, the system produces context-grounded, verifiable answers that support up-to-date understanding across multiple Earth-system domains.

Evaluation using the RAGAs framework indicates that the prototype reliably retrieves relevant context and generates grounded, scientifically coherent answers. High context precision and recall observed under the controlled single-document answerability setting reflect consistent chunk-level ranking performance. Together, these findings support the feasibility of implementing a fully local, reproducible RAG pipeline for knowledge-intensive Earth-system research. Although the evaluation was designed as a controlled prototype-scale benchmark, the results demonstrate the practical potential of such workflows for continuously evolving Earth-system literature and establish a foundation for larger-scale scientific evaluation and deployment.

While automated evaluation enables systematic comparison across models, expert assessment remains essential for understanding performance in scientific workflows. Domain specialists judged the responses as broadly accurate, while frequently identifying the need for additional contextualisation, quantitative specification, or methodological nuance. This reflects the distinction between paragraph-grounded retrieval synthesis and the broader cross-study interpretation typically performed by domain experts. In this sense, the system functions as a transparent, evidence-grounded synthesis aid that supports literature navigation and structured summarisation within research workflows.

Building on these observations, expert feedback indicates that, although responses may be accurate relative to a single source, scientific expectations often extend beyond individual studies to broader contextualisation across the literature. This highlights a key challenge for retrieval-augmented systems: bridging the gap between source-grounded evaluation and the integrative synthesis required for scientific assessment. This divergence between automated metrics and expert judgement further highlights the need for evaluation approaches that combine grounding-based metrics with domain-informed assessment of scientific sufficiency and cross-study synthesis.

Addressing this gap, an important next step is the development of benchmarks that explicitly evaluate multi-document evidence aggregation in Earth-system science. Constructing such resources will require substantial expert curation to formulate questions requiring cross-study reasoning, reconcile inconsistencies across studies, and annotate distributed evidence. Establishing community-standard benchmarks therefore remains a non-trivial but valuable direction for advancing retrieval-augmented systems in environmental research.

Future work should extend the framework across planetary boundaries and publication periods, and strengthen multi-document synthesis capabilities including explicit linkage between synthesized claims and supporting evidence.

This work sits within a growing ecosystem of domain-adapted LLM tools developed for climate and sustainability applications, including systems for climate-related question answering[8,10], corporate sustainability reporting[9], net-zero policy assessment[18], and agricultural climate adaptation advice[23,24]. These efforts highlight the value of domain-grounded retrieval and transparent sourcing in improving reliability. The present system extends this line of research to the broader, highly interdisciplinary domain of Earth system and planetary boundaries science, offering an approach that emphasises traceability, local control, continuous corpus evolution, and long-term reproducibility.

Overall, the results provide a foundation for developing an independently verifiable, extensible AI-assisted workflow that supports researchers and assessment teams in navigating rapidly evolving environmental knowledge while making explicit both the strengths and the current limitations of retrieval-augmented approaches in scientific assessment contexts. By enabling transparent, continuously updated synthesis across rapidly expanding Earth-system literature, this framework can support more systematic and reproducible planetary boundary assessments and Earth-system diagnostics.

Methods

The proposed system is designed as a question-answering tool with two principal components. The data collection component gathers scientific literature related to the concept of planetary boundaries. The Retrieval-Augmented Generation provides semantic search and contextualised responses to user queries. These components are detailed below. Figure 3 shows the complete pipeline of the proposed system, including the data collection process and the Retrieval-Augmented Generation framework used for question answering.

Data collection

Relevant literature on planetary boundaries is retrieved, integrated, and stored to support the Retrieval-Augmented Generation (RAG) system. The pipeline accommodates both automated and manual ingestion modes.

Automated paper ingestion. For each planetary boundary, a domain-specific Boolean search query is constructed and executed through the official APIs of Web of Science, Scopus, and OpenAlex. Other literature databases can also be included due to the flexible design. The retrieved results are deduplicated using DOI and metadata fields and stored in the database. When accessible, open-access or institutionally available papers are downloaded. If full-text access is unavailable, abstracts are retained in their place.

Because OpenAlex indexes both peer-reviewed publications and preprints, emerging research outputs may also be included in the corpus. The inclusion of preprints supports early visibility of evolving scientific discussions in rapidly developing domains. All retrieved sources are stored with explicit metadata and source attribution, enabling users to distinguish between peer-reviewed and non-peer-reviewed materials.

The automated fetching process runs periodically (e.g. quarterly) to ensure continuous integration of newly published literature and to maintain an up-to-date database. Each record is stored with its associated metadata, including title, authors, publication year, DOI, keywords and source.

Manual paper ingestion. In addition to automated retrieval, the system allows for manual addition of publications. Local PDF papers can be directly imported and processed. The structured full-text content and metadata are extracted and stored in the same unified database as the automatically collected records, ensuring consistency across data sources.

Text processing and embedding. PDF documents are parsed, and full texts are extracted by the Grobid[25] parser. Full texts or, if not available, abstracts are chunked using LangChain's[26] Recursive Character Text Splitter. It splits texts into small, overlapping chunks by recursively dividing them at natural boundaries (e.g. paragraphs and sentences), preserving context while staying within the given character limit. In this work, each document is divided into overlapping text chunks with a limit chunk size of 800 characters and an overlap of 100 characters. This configuration was chosen to balance retrieval granularity and semantic coherence in scientific writing. Smaller chunks increase ranking specificity but risk splitting causal arguments across segments, whereas larger chunks preserve context at the expense of retrieval precision. The overlap helps maintain continuity across adjacent segments.

For each chunk, vector representations (embeddings) are generated using the BAAI/bge-base-en-v1.5[27] model. This model was selected due to its competitive performance among open-weight English embedding models on the Massive Text Embedding Benchmark (MTEB) leaderboard[28], as well as its suitability for local deployment. The choice aligns with our emphasis on reproducibility and independence from proprietary APIs. The embeddings are stored in the PGVector[29] vector database, which enables efficient similarity search through cosine similarity.

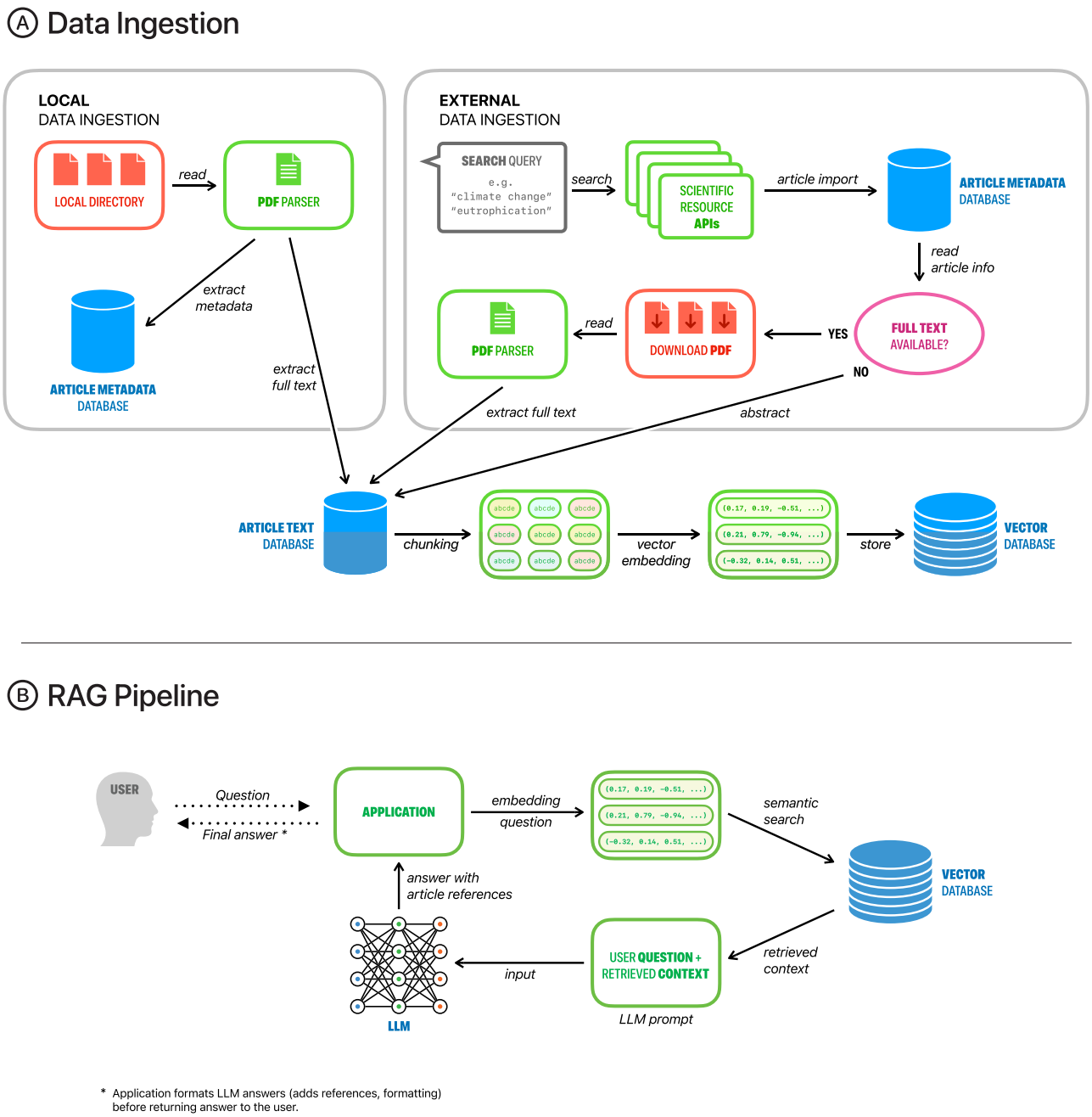

Fig. 3 | Overall pipeline for building a scientific literature knowledge base on Earth systems and planetary boundaries and for operating a question-answering system grounded in that knowledge base. A Shows how external sources and local PDFs are ingested, parsed, chunked, embedded, and stored in a vector database. **B** Illustrates how a user's question is converted into an embedding, matched to relevant text segments through semantic search, and combined with the retrieved context to produce an informed, referenced answer.

Retrieval-augmented generation (RAG)

Retrieval-Augmented Generation is an architecture introduced by Lewis et al. (2020)[15] that helps large language models produce more correct and reliable answers by using information taken from an outside text collection. In this way, the LLM's responses are based on real and verifiable sources instead of only its internal knowledge. In this study, the RAG pipeline is implemented using Ollama, an open-source framework that facilitates the deployment and management of local LLMs. Using local LLMs provides full control over model selection. This approach allows the system to run models entirely on local infrastructure, removing dependencies on external APIs.

The proposed RAG-based system enables researchers to interact with the literature database through a user interface. When a user submits a query, it is converted into an embedding vector using the same embedding model employed for document embedding mentioned above. The system retrieves the k-most semantically similar document chunks from the vector database by a similarity search. These retrieved chunks are provided to the language model as contextual input along with the user's question. A structured prompt is formed that instructs the LLM to generate an answer grounded in the provided evidence. The generated response, together with the relevant retrieved passages, is presented to the user. The interface highlights the text segments used by the LLM to formulate its response and provides direct links to the source publications, as shown in Fig. 4, allowing users to verify the answer and explore additional context.

The system's performance was evaluated using a dataset compiled for this study and multiple RAGAs metrics as well as through expert assessments described in the following sections.

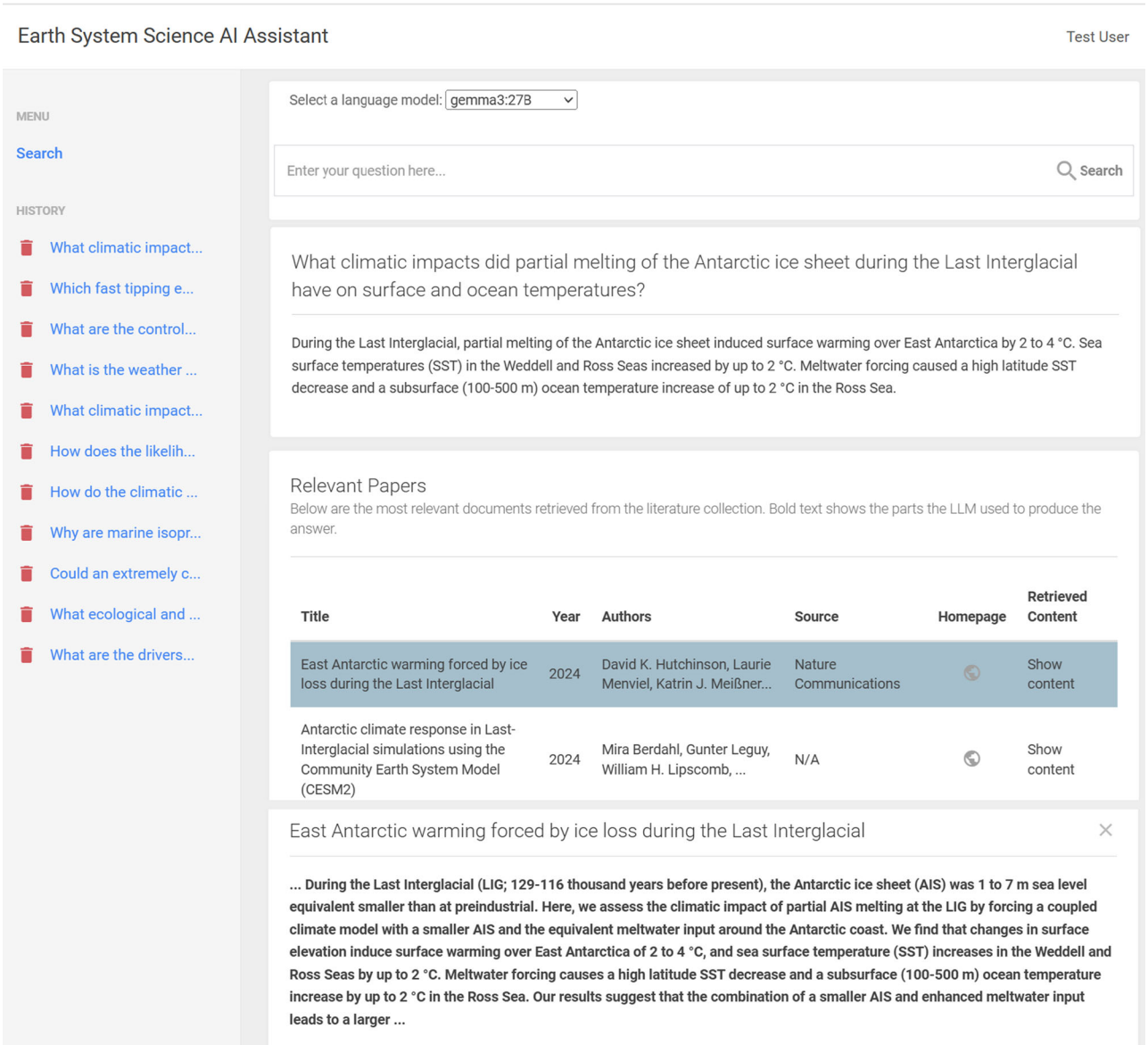

Fig. 4 | User interface of the system. After the user submits a question, the system retrieves the most relevant text segments and uses them to produce an answer. The final output presents both the answer and the supporting references, with the specific parts of the source texts that informed the response highlighted in bold so that users can directly see the underlying evidence. The text excerpt at the bottom of the figure is from ref. 45.

Evaluation framework

Evaluation corpus. To assess overall system performance, we constructed a corpus of 12,834 climate-related scholarly documents from Web of Science, Scopus, and OpenAlex for the year 2024, with the Boolean search query provided in Supplementary Note 1, formatted appropriately for each source. For 3634 of these documents, full texts were retrieved, and abstracts were retained for the rest. All documents were indexed for semantic retrieval.

Evaluation dataset construction. A set of 50 open-ended question-answer pairs was compiled by domain experts using only the retrieved texts of randomly selected peer-reviewed climate research articles. The questions were derived from paragraphs in these papers identified through the predefined Boolean query described above, ensuring that the evaluation reflects topics represented in the assembled corpus. They span major thematic areas including physical climate processes, mitigation measures, impacts, adaptation, and feedback mechanisms. Because each question-answer pair required expert formulation and verification against the source article, the resulting dataset comprises 50 questions that enable controlled assessment of retrieval and generation performance within this experimental framework. The question set therefore serves as a structured evaluation dataset and does not constitute a comprehensive benchmark for ranking language models.

Our evaluation adopts a controlled single-document answerability assumption, whereby each question is constructed to be answerable from a single research article. Single-document answerability is an established evaluation regime in reading-comprehension benchmarks such as SQuAD[30], where questions are grounded in a provided passage and answers are anchored within that same text. In the scientific domain, QASPER similarly evaluates question answering over individual research papers, with answers supported by explicit evidence spans within a single article[31]. Likewise, LitQA2 constructs scientific questions whose answers appear in the body of a specific paper, intentionally constraining answer locality[32]. Similarly, QASA collects questions generated while reading individual research articles, with answers grounded in evidential rationales extracted from the same paper[33].

Consistent with these paradigms, our questions are formulated from selected paragraphs to enable controlled assessment of retrieval

performance at the chunk level. As the automated evaluation relies on reference answers for each question, this setup ensures that the relevant evidence supporting the answer is present within the indexed corpus. Retrieval is performed over the full indexed corpus without access to document identifiers, such that successful performance corresponds to correctly ranking the chunk(s) containing the relevant evidence among all indexed documents.

Automated evaluation. A RAG pipeline consists of two components: a retriever component that retrieves additional context from an external database for the LLM to answer the query, and a generator component that generates an answer based on a prompt augmented with the retrieved information. When evaluating a RAG pipeline, both components should be evaluated separately and quantitatively.

RAGAs (Retrieval-Augmented Generation Assessment) is a framework for evaluation of RAG pipelines, providing a set of metrics that assess how well the retriever selects relevant context chunks, how effectively the LLM uses those chunks, and the overall answer quality. The evaluation is performed automatically by a judge LLM, without any human scoring or manual assessment[17]. RAGAs metrics were computed using the RAGAs Python package, version 0.4.2. In this study, GPT-4o Mini served solely as an external evaluation model within the RAGAs framework, with the temperature parameter fixed at 0.00 to encourage deterministic behaviour. However, because language models remain inherently non-deterministic, small variations in scoring may occur across repeated runs. The deployed QA system itself relies exclusively on locally hosted, open-weight language models; GPT-4o Mini was not part of the operational pipeline.

Five RAGAs metrics that together capture the key components of RAG performance have been computed: *context precision* and *context recall* measure how effectively the retriever selects relevant information; *faithfulness*, *answer relevancy* and *answer correctness* assess the model's ability to generate responses that are grounded, relevant, and factually accurate. Higher scores indicate higher overall RAG performance.

Context retrieval metrics. *Context precision* evaluates how well the retriever ranks the most relevant pieces of information at the top of the retrieved results for a given query. It shows whether the retrieved context primarily consists of relevant chunks and whether these appear early in the ranking. It is computed as the average of precision@k values for each chunk in the retrieved context, where precision@k represents the ratio of relevant chunks among the top *k-ranked* chunks[34].

Context recall assesses how completely the retriever captures the relevant information for a given query. It measures the proportion of relevant documents or chunks that were successfully retrieved and emphasises the system's ability to avoid missing important content. It is calculated as the ratio of the number of claims in the reference answer supported by the retrieved context over the total number of claims in the reference answer.

Under the single-document answerability assumption described above, successful retrieval corresponds to correctly ranking the chunk or set of chunks containing the claims reflected in the reference answer. In this setting, context precision evaluates whether those relevant chunks appear among the highest-ranked retrieved results, while context recall measures the proportion of reference claims supported by the retrieved context. Because each reference answer is derived from a source paragraph that is segmented into retrievable chunks, retrieval of the relevant chunk(s) yields near-maximal recall by design. In this setting, high mean precision and recall values reflect consistent chunk-level ranking performance within this controlled evaluation setting and should be interpreted in light of the single-document answerability assumption.

Generation metrics. *Faithfulness* assesses how factually consistent a generated response is with the information contained in the retrieved context. A response is considered faithful when all of its claims can be directly supported by the retrieved context. To calculate this metric, three steps are followed: all the factual claims made in the generated response are identified. Whether each claim can be inferred or substantiated by the retrieved context is verified. Finally, the ratio of the number of claims in the generated response supported by the retrieved context over the total number of claims in the generated response is calculated.

Answer relevancy evaluates how well a model's response aligns with the user's original question. A response is considered relevant when it directly addresses the user's question and stays focused on the intended topic. Therefore, this metric indicates how well the model understands and responds to the user's query, independent of factual accuracy. To compute this metric, the following steps are performed: the judge LLM generates a set of questions (three questions by default) based on the generated response by RAG. The cosine similarity between the embedding of the user's question and the embedding of each generated question is computed. These cosine similarity values are averaged to obtain the final score.

Answer correctness determines the correctness of the generated response with respect to the reference answer as a weighted sum of factual correctness and semantic similarity. The weights are 0.75 and 0.25 by default, respectively. Semantic similarity measures how alike the generated response and the reference answer are in meaning. The similarity score is computed to quantify the semantic similarity between the embedding of the generated response and the embedding of the reference answer[35]. Factual correctness evaluates whether the generated response contains factually consistent information. It evaluates the consistency of factual statements in the generated response with those in the reference answer. This assessment is based on the following concepts:

- *True Positive (TP):* Facts or statements that appear in both the reference answer and the generated response.
- *False Positive (FP):* Facts or statements that appear in the generated response but are not found in the reference answer.
- *False Negative (FN):* Facts or statements that are present in the reference answer but missing from the generated response.

Factual correctness is quantified by calculating the F1 Score, which is formulated as:

$$\mathbf{F1Score} = |\mathrm{TP}|/(|\mathrm{TP}| + 0.5 \times (|\mathrm{FP}| + |\mathrm{FN}|))$$

Expert evaluation. To complement the automated RAGAs metrics, we conducted an expert evaluation focusing on the scientific reliability of retrieval-augmented responses. A randomly sampled subset of 24 questions from the full set of 50 was evaluated for the model that achieved the strongest overall performance under automated evaluation, reflecting the need for domain-specific expertise in this analysis. As such, this assessment provides an exploratory qualitative evaluation of scientific accuracy and sufficiency. Two domain experts independently scored each response along three dimensions using a 1–5 scale: *Scientific Accuracy* (factual correctness relative to retrieved papers and domain knowledge), *Completeness & Scientific Sufficiency* (inclusion of key mechanisms, processes, or estimates required for correct scientific interpretation), and *Alignment with the source paper* (consistency with the findings and claims of the referenced literature) The detailed scoring rubric, including scale anchors and evaluation criteria, is documented in the Supplementary Note 2. Ratings were conducted independently by the two experts without consultation. Agreement between expert scores was assessed using the proportion of responses with an absolute difference of ≤1 point. In this evaluation setting, this serves as a descriptive measure of consistency.

Tools and technologies

The prototype has been developed in Python using the Django framework[36]. The LangChain[26] library was employed to implement the RAG pipeline, while open-weight language models were accessed through a locally hosted Ollama[37] server. The entire application operates within a Docker[38] based architecture, consisting of independent containers for the application,

PostgreSQL[39] database, the Grobid PDF parser[25], and the Ollama server. Containerization ensures reproducibility, scalability, and portability across computing environments.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Background

Natural language processing (NLP) for climate and environmental science

NLP techniques have been employed in climate and environment-related text analysis, including financial disclosure analysis[40,41], corporate climate disclosure analysis[42], detecting environmental claims[43], question answering for climate adaptation in agriculture[23] and detecting sustainable development goals in environmental reports[44].

Question answering (QA)

Recent work explored LLM-based question-answering systems for climate and sustainability domains, which inform the design of the proposed approach. Four related QA systems are briefly reviewed below.

ChatClimate (Vaghefi et al., 2023)[8] is a domain-adapted conversational AI system designed to enhance the reliability of large language models (LLMs) in climate-related question-answering tasks. It is built upon GPT-4. The system integrates access to the Intergovernmental Panel on Climate Change's Sixth Assessment Report (IPCC AR6), which is an authoritative scientific source in the domain, to mitigate issues of hallucination and outdated knowledge that are common in general-purpose LLMs. ChatClimate is evaluated in three configurations: standalone GPT-4, ChatClimate relying solely on IPCC reports, and a Hybrid model combining GPT-4's internal knowledge with targeted retrieval from IPCC sources. Expert evaluations demonstrate that the hybrid model delivers more accurate and contextually grounded responses and highlight the efficacy of augmenting LLMs with domain-specific and up-to-date external data.

ChatNetZero (Hsu et al., 2024)[18] is an LLM-based chatbot for supporting net-zero emissions research and planning by analyzing net-zero related text documents and serving as a question-answering platform for climate policy-specific information. It is a RAG-based system that incorporates specialized anti-hallucination and reference modules to ground its responses in verified, domain-specific sources. Comparative evaluations against leading LLMs, including GPT-4, Gemini, Coral, and ChatClimate, show that ChatNetZero gives more factually accurate answers when benchmarked against source documents. In addition to quantitative validation, the expert surveys they conducted revealed that ChatNetZero produces more accurate outputs; however, it sometimes falls short in providing the contextual richness and response length preferred by users.

CHATREPORT (Ni et al., 2023)[9] is an LLM-based system designed to automate the analysis of corporate sustainability reports, which are traditionally too voluminous and complex for efficient human review. In light of increasing concerns about corporate transparency in the climate change context, such automation tools are gaining interest. Challenges in scalable report analysis are the hallucination issues of LLMs and the difficulty of integrating domain expertise into the AI development process. CHATREPORT addresses these challenges with two strategies: ensuring traceability of generated content to mitigate hallucination, and incorporating domain experts directly into the iterative prompt engineering and system development loop. They evaluated the system on over 1000 sustainability reports and demonstrated the system's potential for enabling scalable, reproducible, and transparent report analysis.

My Climate Advisor (Nguyen et al., 2024)[24] is a domain-specific QA system developed to support climate adaptation efforts in agriculture by delivering trustworthy, relevant information to farmers and agricultural advisors. The system uses an in-house LLaMA 3 model with RAG to synthesize scientific publications and high-quality grey literature to provide grounded answers with clear resources. Initial user studies show that My Climate Advisor does not yet surpass leading proprietary systems in all dimensions; however, it performs comparably in terms of scientific accuracy and excels in explainability through transparent sourcing.

Each of the above systems contributes insights to our proposed QA system for Earth system and planetary boundaries science. They highlight the importance of domain-specific knowledge integration. Building on these advances, the proposed system aggregates the latest planetary boundaries literature and employs a similar RAG methodology to ensure that answers are both up-to-date and verifiable.

Data availability

The question-and-answer dataset generated for this study is publicly available at https://github.com/ozgeozge/PB_Assistant/tree/main/paper/data. The public release includes the large majority of records and is limited to those derived from openly licensed source publications. The numerical source data underlying Figs. 1 and 2 are available at https://doi.org/10.6084/m9.figshare.32939171.

Code availability

The code used to implement the RAG system is available at: https://github.com/ozgeozge/PB_Assistant.

Received: 10 December 2025; Accepted: 23 July 2026;
Published online: 10 August 2026

References

1. Calvin, K. et al. *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (Eds.)]*. https://www.ipcc.ch/report/ar6/syr/ (IPCC, 2023).
2. Steffen, W. et al. Planetary boundaries: guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
3. Lenton, T. M. et al. Climate tipping points—too risky to bet against. *Nature* **575**, 592–595 (2019).
4. Bornmann, L. & Mutz, R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **66**, 2215–2222 (2015).
5. Rockström, J. A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
6. Richardson, K. Earth beyond six of nine planetary boundaries. *Sci. Adv*. **9**, eadh2458(2023).
7. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
8. Vaghefi, S. A. et al. ChatClimate: grounding conversational AI in climate science. *Commun. Earth Environ.* **4**, 480 (2023).
9. Ni, J. et al. CHATREPORT: democratizing sustainability disclosure analysis through LLM-based tools. in *Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Feng, Y. & Lefever, E.) 21–51 (Association for Computational Linguistics, 2023).
10. Thulke, D. et al. ClimateGPT: towards AI synthesizing interdisciplinary research on climate change. Preprint at. https://doi.org/10.48550/arXiv.2401.09646 (2024).
11. Brown, T. B. et al. Language models are few-shot learners. in *Proc. 34th International Conference on Neural Information Processing Systems* 1877–1901 (Curran Associates Inc., 2020).
12. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at. https://doi.org/10.48550/arXiv.2302.13971 (2023).
13. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Proc. 36th International Conference on Neural Information Processing Systems* 27730–27744 (Curran Associates, Inc., 2022).

14. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
15. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
16. Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. in *Proc. 39th International Conference on Machine Learning*, 2206–2240 (PMLR, 2022).
17. Es, S., James, J., Espinosa Anke, L. & Schockaert, S. RAGAs: automated evaluation of retrieval augmented generation. in *Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (eds Aletras, N. & De Clercq, O.) 150–158 (Association for Computational Linguistics, 2024).
18. Hsu, A. et al. Evaluating ChatNetZero, an LLM-Chatbot to Demystify Climate Pledges. *Proc. 1st Workshop on Natural Language Processing Meets Climate Change*, 82–92 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.climatenlp-1.6.
19. OpenAI. et al. *gpt-oss-120b & gpt-oss-20b Model Card*. Preprint at. https://doi.org/10.48550/arXiv.2508.10925 (2025).
20. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
21. Unmatched Performance and Efficiency | Llama 4. *Unmatched Performance and Efficiency | Llama 4*. https://www.llama.com/models/llama-4/.
22. Team, G. et al. Gemma 3 Technical report. Preprint at. https://doi.org/10.48550/arXiv.2503.19786 (2025).
23. Nguyen, V. et al. My Climate CoPilot: A Question Answering System for Climate Adaptation in Agriculture. *Proc. 63rd Annual Meeting of the Association for Computational Linguistics*, 62–70 (Association for Computational Linguistics, 2025). https://doi.org/10.18653/v1/2025.acl-demo.7.
24. Nguyen, V. et al. My Climate Advisor: An Application of NLP in Climate Adaptation for Agriculture. *Proc. 1st Workshop on Natural Language Processing Meets Climate Change*, 27–45 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.climatenlp-1.3.
25. GROBID contributors. *GROBID*. GitHub. https://github.com/grobidOrg/grobid (2008).
26. Chase, H. *LangChain [Computer software]*. GitHub. https://github.com/langchain-ai/langchain (2022).
27. Beijing Academy of Artificial Intelligence (BAAI). *BGE v1 & v1.5. BGE Documentation*. https://bge-model.com/bge/bge_v1_v1.5.html (2024).
28. MTEB. *MTEB Leaderboard*. Hugging Face Spaces. https://huggingface.co/spaces/mteb/leaderboard (2022).
29. pgvector. *GitHub*. https://github.com/pgvector.
30. Rajpurkar, P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proc. conference on empirical methods in natural language processing*, 2383–2392 (Association for Computational Linguistics, 2016). https://doi.org/10.18653/v1/D16-1264.
31. Dasigi, P. et al. A dataset of information-seeking questions and answers anchored in research papers. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds. Toutanova, K. et al.), 4599–4610 (Association for Computational Linguistics, 2021).
32. Skarlinski, M. D. et al. Language agents achieve superhuman synthesis of scientific knowledge. Preprint at. https://doi.org/10.48550/arXiv.2409.13740 (2024).
33. Lee, Y. et al. QASA: advanced question answering on scientific articles. in *Proc. 40th International Conference on Machine Learning*, 19036–19052 (PMLR, 2023).
34. Ragas. *Context Precision*. Ragas Documentation. https://docs.ragas.io/en/v0.4.2/concepts/metrics/available_metrics/context_precision/ (2025).
35. Ragas. *Semantic Similarity*. Ragas Documentation. https://docs.ragas.io/en/v0.4.2/concepts/metrics/available_metrics/semantic_similarity/ (2025).
36. Django. *Django Project* https://www.djangoproject.com/.
37. Ollama. https://ollama.com.
38. Docker Inc. *Docker: Accelerated Container Application Development*. https://www.docker.com/.
39. Group, P. G. D. *PostgreSQL*. https://www.postgresql.org/ (2025).
40. Bingler, J. A., Kraus, M., Leippold, M. & Webersinke, N. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *J. Bank. Financ.* **164**, 107191 (2024).
41. Colesanti Senni, C., Vaghefi, S., Schimanski, T., Manekar, T. & Leippold, M. *Using AI to Assess the Decision-Usefulness of Corporates' Nature-related Disclosures*. SSRN Scholarly Paper at. https://doi.org/10.2139/ssrn.4860331 (2024).
42. Schimanski, T. et al. ClimRetrieve: A Benchmarking Dataset for Information Retrieval from Corporate Climate Disclosures. *Proc. Conference on Empirical Methods in Natural Language Processing*, 17509–17524 (Association for Computational Linguistics, 2024).
43. Saha, D., et al. EnClaim: a style augmented transformer architecture for environmental claim detection. in *Proc. 1st Workshop on Natural Language Processing Meets Climate Change*, 123–132 (Association for Computational Linguistics, 2024).
44. Garigliotti, D. SDG target detection in environmental reports using retrieval-augmented generation with LLMs. in *Proc. 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)* (eds Stammbach, D et al.) 241–250 (Association for Computational Linguistics, 2024).
45. Hutchinson, D. K., Menviel, L., Meissner, K. J., & Hogg, A. M. East Antarctic warming forced by ice loss during the Last Interglacial. *Nat. Commun.* **15**, 1026 (2024).

Acknowledgements

We would like to extend our gratitude to the researchers Jonas Kaiser and Lotta Bergfeld at the Potsdam Institute for Climate Impact Research (PIK) for their contributions to the development of the question-answer dataset.

Author contributions

Ö.K.T. and B.S. jointly contributed to the conceptualization of the study. Ö.K.T. designed the study, developed the methodology, conducted the analyses, and drafted the initial manuscript. L.C. carried out the human expert evaluation, contributed to validation, and authored the section reporting the expert evaluation results. J.L. contributed domain expertise, conducted expert evaluation, and contributed to the interpretation of results. B.S. provided supervision and contributed input to the visualizations. Ö.K.T., L.C., J.L. and B.S. jointly reviewed and edited the manuscript and approved the final version.

Funding

Ö.K.T, L.C., J.L. and B.S. are part of the Planetary Boundaries Science Lab's research effort at the Potsdam Institute of Climate Impact Research (PIK) and are funded by the Grantham Foundation for the Protection of the Environment and additional anonymous donors. Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-026-03878-1.

Correspondence and requests for materials should be addressed to Özge Kart Tokmak.

Peer review information *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Joao Duarte and Nandita Basu. A peer review file is available.