CPIQA: Climate Paper Image Question Answering Dataset for Retrieval-Augmented Generation with Context-based Query Expansion

Anonymous ACL submission

Abstract

Misinformation about climate science is a serious challenge for our society. This paper introduces CPIQA (Climate Paper Image Question-Answering), a new question-answer dataset featuring 4,551 full-text open-source academic papers in the area of climate science with 54,612 GPT-4o generated question-answer pairs. CPIQA contains four question types (numeric, figure-based, non-figure-based, reasoning), each generated using three user roles (expert, non-expert, climate sceptic). CPIQA is multimodal, incorporating information from figures and graphs with GPT-4o descriptive annotations. We describe Context-RAG, a novel method for RAG prompt decomposition and augmentation involving extracting distinct contexts for the question. Evaluation results for Context-RAG on the benchmark SPIQA dataset outperforms the previous best state of the art model in two out of three test cases. For our CPIQA dataset, Context-RAG outperforms our standard RAG baseline on all five base LLMs we tested, showing our novel contextual decomposition method can generalize to any LLM architecture. Expert evaluation of our best performing model (GPT-4o with Context-RAG) by climate science experts highlights strengths in precision and provenance tracking, particularly for figure-based and reasoning questions.

1 Introduction

Misinformation about climate science continues to pose a challenge for our society. This poses a serious challenge for public understanding, policymaking and even experts (Lewandowsky, 2020). At the same time, large language models (LLMs) have become powerful tools for information retrieval and evidence synthesis, but they are also highly prone to hallucination—generating incorrect or fabricated facts, references, and claims. Given the high stakes of climate communication, there is a pressing need for a reliable question-answering (QA) system that grounds responses in authoritative scientific sources.

In this work, we introduce CPIQA, a new dataset for climate science QA that incorporates both text and visual data from academic papers. CPIQA consists of 4,551 papers from twelve sources, with extracted figures and their descriptions used as additional evidence in question-answering. The dataset supports three role variations and four question categories designed to reflect different types of real-world climate questions.

Building on CPIQA, we develop a retrieval-augmented generation (RAG)-based chatbot for climate QA. Our system follows a two-stage retrieval process: it first retrieves full papers based on the user's query, then extracts relevant text chunks from the most relevant papers. This approach improves both chunk similarity and cross-relevance of chunks. Further, we introduce Context-RAG, a novel prompting method that enhances retrieval by decomposing a given question into distinct contextual variations before searching for relevant documents. Rather than relying on a single query, our method anticipates different ways the question might be framed—such as a scientific explanation, a policy-related perspective, or a public concern—allowing for more diverse and targeted retrieval. This ensures that retrieved documents are not biased toward a single interpretation of the question.

To evaluate the effectiveness of our method, we test it on SPIQA, a dataset for scientific QA in the computer science domain, in addition to CPIQA. This allows us to assess how well our QA pipeline

generalizes beyond climate science. Finally, we validate the system's outputs through qualitative climate scientist expert evaluation, ensuring that responses are accurate, relevant, concise and aligned with scientific consensus.

By combining structured retrieval with expert-informed question generation, this work contributes a robust, transparent approach to climate QA, helping to bridge the gap between AI-generated answers and reliable scientific communication.

More specifically, our contributions in this work include the following:

- A new multimodal QA dataset resource (CPIQA dataset) for the NLP community based on 4,551 academic climate research paper documents. This dataset is large, annotated with 54,612 question-answer pairs generated by GPT-4o and includes text summaries of all images, graphs and figures within the full text documents. Questions are broken down into figure-based, numeric-based, non-numeric, and reasoning-based types to allow for a finer-grained evaluation of QA performance than most existing QA datasets allow. Our dataset and code is open source and available at GITHUB_URI_REDACTED_PREPRINT and ZENODO_URI_REDACTED_PREPRINT.

- Description of a novel context-based query expansion method for RAG, comprehensively evaluated on both the benchmark SPIQA dataset and our new CPIQA dataset. Context-based query expansion provides a 7.2% improvement over baseline RAG methods across various question types and roles. We include a detailed breakdown of performance across different question types which future researchers can benchmark their models against.

2 Related Work

2.1 Scientific QA Datasets

Table 1 sets out notable QA datasets that have been designed to support scientific domains such as climate science.

A significant number of existing QA datasets come from the biomedical and computer science domains, reflecting the heavy use of document-based QA in these fields. While these datasets offer strong benchmarks for scientific QA, they are typically unimodal, focusing exclusively on textual information. Multimodal datasets—those incorporating both text and figures—are far less common, with SPIQA (Pramanick et al., 2024) being the most comprehensive multimodal dataset designed for scientific applications.

Among multimodal datasets, FigureQA (Kahou et al., 2017) is a notable example, containing question-answer pairs for synthetic graphs, figures, and tables. However, it lacks contextual information from accompanying text, making it unsuitable for tasks that require a deeper understanding of scientific literature.

Compared to biomedical and computer science domains, climate science QA datasets are less common. One of the most relevant efforts is ClimaQA (Manivannan et al., 2024), which includes both a 502 question "gold" dataset curated by experts and a larger LLM-generated 3000 question "silver" dataset. ClimaQA is unique in that it supports three types of questions: multiple-choice, cloze-style, and free-form, allowing for a broader range of QA applications. Our CIPQA is significantly larger with 54,612 questions, and unlike ClimaQA which relies on textbook sources our dataset relies on academic paper sources making it suitable for research-driven climate QA.

2.2 Climate Science LLMs

Recent efforts have been made to fine-tune LLMs specifically for climate-related tasks such as fact-grounded QA, ambiguous policy analysis, and misinformation debunking. One such example is ClimateBERT (Webersinke et al., 2022), a model trained on climate-focused text sources to improve NLP performance in this domain. ChatClimate (Vaghefi et al., 2023) grounds GPT-4 responses in IPCC AR6 reports, showing that retrieval significantly improves accuracy. Hallucinations are identified, however, when queries extend beyond the IPCC's coverage. ChatNetZero (Hsu et al., 2024) applies a similar approach to net-zero policies, retrieving structured data on corporate and governmental pledges. While this helps ground responses, the model struggles with policy ambiguity.

Beyond policy analysis, LLMs are being explored for misinformation debunking. Generative Debunking of Climate Misinformation (Zanartu et al., 2024) introduces claim classification and fallacy detection, structuring responses using a fact-myth-fallacy-fact framework. While this improves

Dataset	Question generation	Num QA pairs	Num documents	Paper Source	Domain	Question basis	
						Full text	Figs & tabs
FigureQA	Schema based	1.8M	140k	Synthetic	General	N	Y
BioAsq	Human experts	3.2K	–	PubMed	Biomedical	N	N
PubMedQA	Human experts	1K	120K abstracts	PubMed	Biomedical	Y	N
BioASQ-QA	Human experts	4.7K	–	PubMed	Biomedical	N	N
ArgSciChat	Human experts	41 dialogues	20 papers	arXiv	NLP	Y	N
ScienceQA	Human experts	21K	-	School curriculum	General	Y	Y
QASPER	Human experts	5K	1.5K papers	S2ORC	NLP	N	N
QASA	Human experts	1.8K	112 papers	S2ORC	AI/ML	Y	N
SPIQA	LLMs + Human experts	270K	25.5K papers	arXiv	Computer Science	Y	Y
ClimaQA-Gold	Human Experts	502	23	Textbooks	Climate Science	Y	N
ClimaQA-Silver	LLMs	3000	23	Textbooks	Climate Science	Y	N
CPIQA (ours)	LLMs	55.8k	4650 papers	core.ac.uk	Climate Science	Y	Y

Table 1: Comparison of relevant QA datasets over scientific literature: (Kahou et al., 2017), (Tsatsaronis et al., 2015), (Jin et al., 2019), (Krithara et al., 2023), (Ruggeri et al., 2023), (Lu et al., 2022), (Dasigi et al., 2021), (Lee et al., 2023), (Manivannan et al., 2024) (2)

coherence, LLMs sometimes fail to select the most relevant counterarguments, leading to misdirected rebuttals.

My Climate Advisor (Nguyen et al., 2024) targets the specific domain climate adaptation in agriculture, retrieving information from peer-reviewed research, grey literature, and climate projection data. It tailors responses to regional climate risks, offering actionable insights for farmers. A key contribution is its expert-driven evaluation framework, which assesses responses across seven domain-specific criteria. Initial results highlight gaps in retrieval precision and the difficulty of adapting to evolving climate knowledge.

2.3 Retrieval-Augmented Generation

Effective retrieval-augmented generation (RAG) depends on retrieval quality, query formulation, and model alignment with retrieved knowledge. Traditional RAG pipelines perform a single retrieval step, which can fail when initial queries are too vague or incomplete (He et al., 2024). Recent research has explored iterative retrieval, query reformulation, and domain-specific adaptations to improve response accuracy.

CoRAG (Chain-of-Retrieval Augmented Generation) (Wang et al., 2025) introduces stepwise retrieval reasoning, allowing the model to dynamically reformulate queries based on retrieved evidence, significantly improving multi-hop QA. Similarly, RICHES (Retrieval Interlaced with Sequence Generation) (Jain et al., 2024) integrates retrieval within the decoding process, eliminating the need for a separate retriever module. This improves response fluency but can introduce hallucinations if retrieval is inconsistent.

Ensuring alignment between retrieved knowledge and generated responses is another key challenge. CoV-RAG (Chain-of-Verification RAG) (He et al., 2024) introduces a verification step that evaluates and refines retrieved documents before answer generation, reducing retrieval errors and hallucinations. RAGAR (RAG-Augmented Reasoning) (Khaliq et al., 2024) extends this further with hierarchical retrieval techniques (CoRAG and ToRAG - Tree-of-RAG) that decompose complex claims into sub-questions, retrieving evidence iteratively for fact-checking in multimodal political discourse.

Beyond reasoning techniques, RAG-Studio (Mao et al., 2024) focuses on domain-specific adaptation, addressing a major limitation of general-purpose RAG models. It introduces a self-alignment framework, where the retriever and generator co-train on synthetic domain-specific data, improving retrieval precision and factual grounding without requiring manually labeled examples. This approach outperforms traditional RAG fine-tuning in specialized domains such as law, finance, and biomedicine.

Our Context-RAG approach is motivated by previous work on multi-step query reformulation, but extending it to novelly focus on extracting distinct contexts in which the question can be re-framed to provide more diverse and user role-targeted retrieval.

3 Methods

3.1 CPIQA Dataset

To develop CPIQA, we curated a dataset of climate-related academic papers, integrating both textual and visual information for the RAG QA task.

We sourced papers from relevant open source climate science journals, identified by climate science expert recommendations. Using CrossRef, we retrieved the DOIs of all available articles from these journals published between 2020 and 2024. We sourced full-text PDFs from CORE.ac.uk (Knoth et al., 2023), an open-access repository of academic publications.

For each document, we extracted full text using *pymupdf4llm*, introducing a filter for documents with significant chunks of missing text. Figures and captions were extracted using *pdffigures 2.0* (Clark and Divvala, 2016), aligning with the CPIQA approach. We use GPT-4o (OpenAI et al., 2024) to generate detailed figure retrieval-friendly descriptions based on the extracted figure type, caption and raw image file. This allows for text-only embeddings to be used in a RAG setting, although image-caption pairs are included in the release.

We generated question-answer pairs by presenting GPT-4o with the full text and figure descriptions. We utilise role-based prompting, generating questions for the general public, climate experts and climate sceptics. Additionally, we generate multiple question types to encourage a wide breadth of questions. Full prompt variations can be found in B

3.2 Question-Answering Architecture

Our baseline two-stage RAG pipeline follows a standard retrieval approach, designed for comparability with SPIQA and evaluation of source attribution. The retriever embeds the user query, and retrieves relevant full text documents. These are used as a filter for the second stage, where the same query is used to retrieve chunks and figures from the filtered documents, maintaining continuity between chunks if required. Retrieved chunks and figure descriptions are inserted into a prompt template alongside the question, from which the LLM generates the answer.

We use *NovaSearch/stella_en_1.5B_v5* (Zhang et al., 2024) as our embedding model due to it being the highest ranked on the MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2022) for the retrieval task with a minimum tokens of at least 100k+, which is a requirement for embedding the majority of documents in CPIQA. In cases where the document is longer than the max-tokens, we chunk the document, maximising token count.

3.3 Context-Based Query Expansion

Context-RAG first seeks to understand the context and intent behind the question. Instead of simply asking, "What do we need to know to answer this question?", our approach reframes it as, "What is the context of this question?" or "Why is this question being asked?". This decomposition enables retrieval that is broader, more targeted, and better aligned with the underlying information need.

The LLM breaks the input question into three distinct contextual perspectives, each represented as a descriptive paragraph, ensuring that retrieval is not biased toward a single interpretation. These are used as part of stage one - retrieval of full text documents. Further, we use the same LLM to break down each context into a set of domain-specific key terms that are up to a sentence in length. This gives finer granularity in the second stage of retrieval.

By shifting retrieval focus from the question itself to its underlying context, we hypothesize that Context-RAG improves recall, diversity, and factual grounding, ensuring that responses draw from a broader and more relevant evidence base. Further, this prompt structure can be applied prior to any other prompt decomposition or expansion method so should be seen as a complimentary method.

4 Results

We evaluate our proposed Context-RAG method against the standard two-step RAG approach across two datasets: SPIQA, a benchmark for scientific paper image question answering, and CPIQA, our newly introduced dataset for climate science. Performance is measured using BERTScore-F1 across multiple test cases and language models.

4.1 Context-RAG

Table 2 demonstrates the two-step RAG approach has a 7% lower BERTScore-F1 compared to the best open source models tested, and our Context-RAG a 3% lower score. Given our change in SPIQA problem formulation, from a one-step QA task where the relevant source paper is provided to a two step QA task where the source paper must be retrieved, this lower performance was expected. In the SPIQA dataset test-A contains LLM-generated

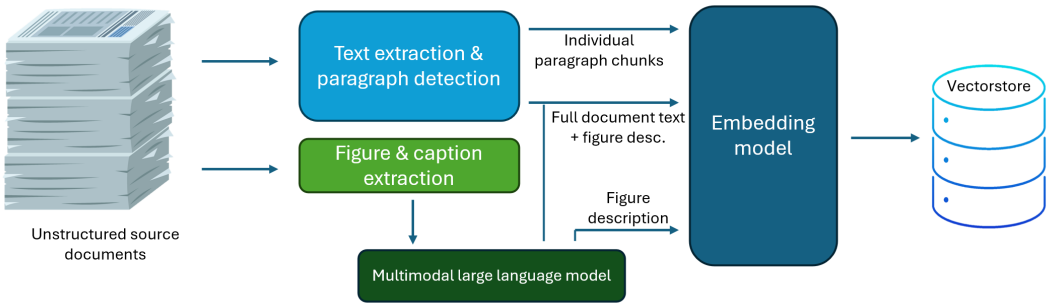

Figure 1: Generic pipeline used to construct CPIQA dataset & set up vectorstore prior to retrieval task

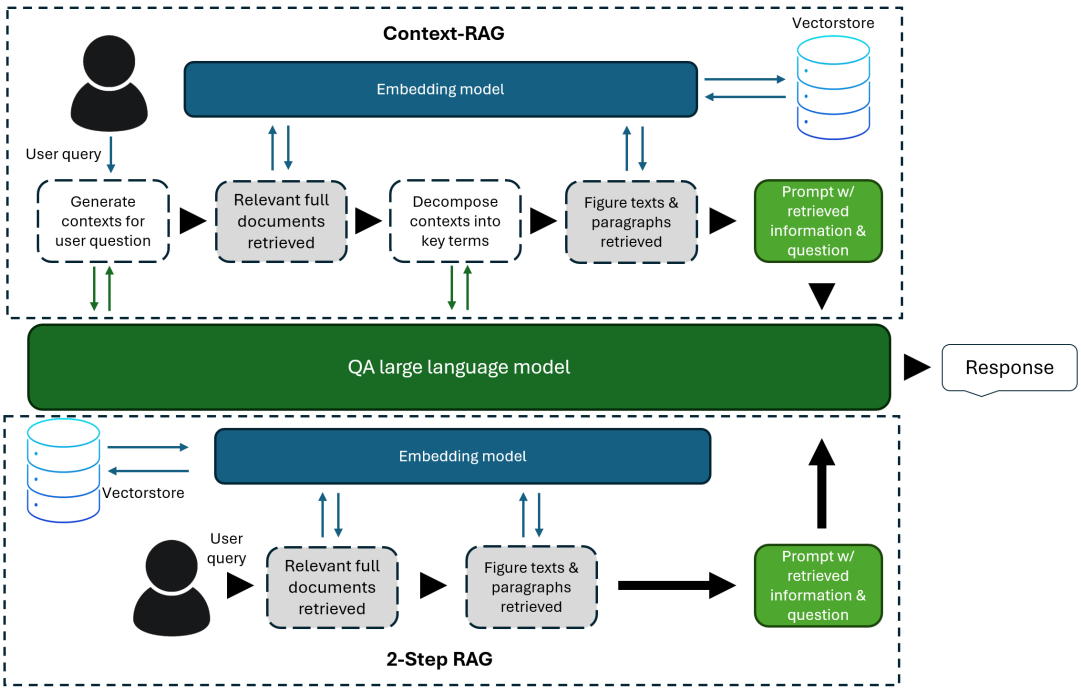

Figure 2: Architecture diagram demonstrating distinction between two-step RAG and Context-RAG

Test Case	Best open-weight baseline (Pramanick et al., 2024)	2 step RAG	Context-RAG
test-A	61.61	57.54	63.28
test-B	47.21	53.22	53.32
test-C	48.45	32.27	34.20
Overall	54.57	47.85	51.31

Table 2: Comparison of our standard two-step RAG and Context-RAG methods on the SPIQA dataset, using *Llama-3.3-70B-Instruct*, compared to baseline results: *LLaVA-1.5-7B* (Liu et al., 2023) for test-A, test-B and *InstructBLIP-7B* (Dai et al., 2023) for test-C. Baseline results experimental setup provides source paper, whereas our setup retrieves from the entire dataset. *bert-base-uncased* is used as the evaluation model for BERT-score (Zhang* et al., 2020).

QAs whilst test-B and test-C have human-written QAs. For two-step RAG we see a 6% improvement for test-B. With Context-RAG, we see an improvement of 4% over two-step RAG, outpeforming the best open source models in test-A by 2% and test-B by 6% showing the potential for our Context-RAG method.

4.2 Climate Question-Answering

A summary of our CPIQA dataset can be found in table 3. We define a train/test/validation split to improve comparability to figure work that may use this data.

Split	Paper count	Question count	Figure count
Train	4255	51060	38325
Validation	99	1188	903
Test	197	2364	1816

Table 3: Summary of CPIQA dataset size incl. number of documents, questions and figures

LLM	2 step RAG	Context-RAG
GPT-4o	67.18	69.10
Gemini 2.0-flash	62.22	64.21
Llama-3.3-70B-Instruct	64.38	65.35
DeepSeek-R1-Distill-Qwen-32B	64.79	65.47
Gemma-2-27b-it	62.32	62.05

Table 4: Comparison of our standard two-step RAG and Context-RAG methods on our CPIQA dataset. Evaluated using BERT-score F1 using the model *microsoft/deberta-xlarge-mnli* (He et al., 2021)

On CPIQA (table 4), we compare both RAG methods across five LLMs. GPT-4o achieves the highest overall performance, with Context-RAG (69.10) slightly surpassing the two-step approach (67.18). Gemini 2.0-flash follows closely, showing a similar pattern, where retrieval based on generated contexts consistently improves results. Other models, such as *Llama-3.3-70B-Instruct* and *DeepSeek-R1-Distill-Qwen-32B*, show a smaller gap between the two approaches, suggesting that context informed retrieval benefits higher-capacity models more significantly.

Table 5 provides insights into the retrieval effectiveness of two-step RAG vs. Context-RAG when retrieving the specific source paper for GPT-4o. Interestingly, two-step RAG achieves a higher correct retrieval rate (60%) than Context-RAG (39%). However, despite retrieving the correct document less frequently, Context-RAG still yields a higher F1 score (70.96 vs. 68.71) which suggests the enhanced retrieved diversity of Context-RAG is allowing it to generate better overall answers.

4.2.1 Expert Evaluation

We asked academic climate science experts to evaluate our best performing model, GPT-4o, according to the qualitative citeria and scoring guidelines below:

- Answer precision: Degree of errors in the answer (1 - lots of errors, 5 - no errors). Unrelated to the question, consider only the answer independently of the question.
- Answer recall: To what degree does the response answer the question? Consider the relevance to the question (1 - irrelevant to the question, 5 - fully covers the question)
- Answer provenance: Is the answer using information from the source document? (1 = not based on context paper; 5 = fully based on context paper)
- Answer conciseness: Does the answer contain waffle or does it go off on a tangent to the question? (1 = verbose; 5 = concise)

The experts were given the question, generated answer, and full PDF source document. Due to expert availability, a random 6% sample of the test set was evaluated by our experts balanced by question type. Table 6 presents the expert evaluation of GPT-4o with Context-RAG, analyzing performance across different question audiences and types. Context-RAG achieves high conciseness scores across all audiences ($\geq$4.1), indicating its ability to generate succinct responses. Non-figure-based and numeric questions exhibit strong precision and recall, particularly for the climate expert role, where numeric questions achieve 4.1 precision and 4.7 recall. Questions generated using the climate expert role had significantly higher provenance scores, especially for numerical (4.6) questions, suggesting that the experts found the answers well-supported by evidence in the source paper. For the general public and climate sceptic roles, Context-RAG achieves moderate performance across all dimensions. Numeric questions for the climate sceptic role showed 3.7 precision and 4.1 recall, while figure-based and reasoning questions had slightly lower provenance scores (2.4-2.7), indicating some difficulty in tracing sources. For the general public role, provenance remains lowest for reasoning questions (2.4), suggesting challenges in aligning broad responses with domain-expert expectations. Overall, our expert qualitative evaluation results align with the

Method	Retrieval result	Retrieval rate %	F1 score
2 step RAG	Correct	60%	68.71
	Incorrect	40%	66.12
Context-RAG	Correct	39%	70.96
	Incorrect	61%	67.97

Table 5: Retrieval rate of the specific source paper for GPT-4o, and its corresponding F1 socre

LLM	Question Audience	Question Type	Precision	Recall	Proven-ance	Concise-ness
GPT-4o Using Context-RAG (Best tested approach)	General public	Figure-based	3.6	3.7	2.8	4.9
		Numeric	2.9	3.6	3.0	4.6
		Non-fig	4.2	4.3	3.0	4.9
		Reasoning	3.4	3.7	2.4	4.7
	Climate sceptic	Figure-based	3.9	3.6	3.3	4.3
		Numeric	3.7	4.1	2.9	4.3
		Non-fig	3.4	3.3	2.4	4.4
		Reasoning	4.0	3.6	2.7	4.3
	Climate expert	Figure-based	3.9	3.6	3.7	4.1
		Numeric	4.1	4.7	4.6	4.8
		Non-fig	3.9	4.3	4.4	4.7
		Reasoning	4.0	4.4	4.4	4.4

Table 6: Expert evaluation of our best approach across roles and evaluation types on a scale of 1-5

trends demonstrated in the BERTscore-F1 results shown in table 7.

5 Discussion

5.1 Context-RAG vs two-step RAG: Retrieval vs Answer Quality

Our results highlight key differences between Context-RAG and the two-step RAG approach in terms of retrieval accuracy and answer quality. As shown in table 5, two-step RAG achieves a higher retrieval rate for the exact source paper (60% vs. 39%), while Context-RAG has a lower rate of exact source matches but produces slightly higher F1 scores in answer generation. This suggests that Context-RAG, despite not always retrieving the original source, provides sufficient and relevant information for generating high-quality answers.

One possible explanation for this is the nature of climate science literature, where overlapping factual content across multiple papers may reduce the importance of retrieving a specific source. Many academic papers cite and build upon each other, meaning that relevant information can often be found in multiple documents. Context-RAG's ability to extract and structure key concepts before retrieval may allow it to synthesize information from related sources, even if the exact original paper is not retrieved. This could explain its relatively strong answer quality despite a lower direct retrieval rate.

This trade-off is further reflected in our broader evaluation metrics. In our Climate QA setting (table 4), Context-RAG yields improved BERT-scores compared to two-step RAG, particularly for more complex questions. This indicates that selecting and structuring context before retrieval may contribute to better alignment with model-generated responses. However, two-step RAG's higher retrieval rate suggests it may be more reliable when strict source matching is a priority.

These findings suggest that retrieval rate alone is not always the best indicator of final answer quality. While two-step RAG more frequently retrieves the intended source, Context-RAG appears to generate answers that are at least as effective, if not more so, in terms of response accuracy.

5.2 Performance Across Different Models

Our evaluation shows that the performance of Context-RAG compared to two-step RAG, whilst always better, varies across models. Larger models, such as GPT-4o and Gemma, show greater improvements in answer quality with Context-RAG, suggesting that their enhanced reasoning capabilities allow them to make better use of retrieved context. For smaller models, the improvements are less pronounced, indicating that they may struggle to leverage retrieved information as effectively.

Notably, context generation can be done in addition to any other prompt augmentation or decomposition technique, though potential impact on performance is not evaluated in this work.

5.3 SPIQA vs CPIQA: Domain-Specific Insights

Comparing SPIQA and CPIQA, we observe distinct trends that highlight domain-specific retrieval challenges. Context-RAG demonstrates consistent improvements over two-step RAG across both datasets, but CPIQA remains more challenging due to domain-specific complexities. Specifically, climate science papers frequently cite each other and share overlapping facts, making it harder for retrieval models to isolate the most relevant document before evidence extraction. This is reflected in CPIQA's lower retrieval accuracy despite the improved context expansion.

The expert evaluation of Context-RAG on CPIQA suggests that provenance and precision are particularly important for climate science experts, as climate-related claims often require precise attribution to datasets, models, or prior research. In contrast, SPIQA, which focuses on interpreting structured results in computer science papers, may place relatively less emphasis on cross-document attribution and more on model reasoning over structured information. These differences suggest that retrieval and reasoning challenges may manifest differently across domains.

5.4 Breakdown by Question Type and Audience

Performance varies across different question types and target audiences, highlighting distinct challenges in retrieval and answer generation. As shown in table 7, numeric and figure-based questions benefit the most from Context-RAG, with consistent improvements across models. This suggests that retrieving structured, contextually relevant information before chunk selection is particularly useful for questions requiring precise data interpretation.

Reasoning-based questions show smaller gains, indicating that retrieval improvements alone may not fully address challenges in multi-step inference. This aligns with previous findings that complex reasoning tasks often depend more on a model's intrinsic capabilities than retrieval alone.

Audience-specific performance trends also reveal key insights. Questions targeted at climate experts generally yield the highest scores, suggesting that expert-level queries align well with retrieved academic content. In contrast, questions posed from a sceptic's perspective score lower, likely due to misalignment between the retrieved scientific literature and the framing of the question. This highlights the difficulty of addressing sceptical viewpoints in a fact-based retrieval system.

6 Conclusion

To support research in climate-focused QA, this paper introduces CPIQA, a dataset built from over 4,551 climate science papers and 54,612 GPT-4o generated question-answer pairs, integrating both text and figure-based question answering. CPIQA incorporates expert-informed question generation and multimodal evidence retrieval, making it a valuable resource for future work in climate AI.

We describe Context-RAG, a novel retrieval-augmented generation (RAG) approach that improves answer quality by structuring retrieval around contextual variations of a question. Unlike traditional RAG methods that directly retrieve documents based on the query, Context-RAG first generates multiple contextual perspectives, retrieves documents accordingly, and then refines retrieval using key domain-specific terms. Our evaluation on CPIQA, a new multimodal climate QA dataset described in this paper, and SPIQA, a scientific paper image QA benchmark dataset, demonstrates that Context-RAG outperforms the standard two-step RAG approach in answer quality, even when exact document retrieval rates are lower.

Our results show that Context-RAG improves performance across various question types and user audiences, particularly for numeric and figure-based questions. Larger models, such as GPT-4o, benefit most from this structured retrieval approach, leveraging contextually relevant evidence for improved reasoning. Furthermore, our expert evaluation of the best-performing model reinforces the effectiveness of Context-RAG in real-world climate science applications.

These findings highlight the importance of evidence-based QA methods. Future directions for this work include the exploration of domain-specific fine-tuning of RAG QA models, a more complete evaluation of the effectiveness of different RAG prompting techniques, and exploring enhancements to Context-RAG that are more explicitly tailored to our four different question types.

7 Limitations

Our GPT-4o generated question-answer pairs are sourced from single source documents, and do not

consider answers that might span multiple documents. Other documents in our dataset may contradict or deviate from the source document and this is an exciting area for future work to explore, as we show with Context-RAG increased performance even when the specific source document was not retrieved.

Our CPIQA dataset has GPT-4o generated QA pairs. Whilst we performed a qualitative climate scientist expert evaluation for our RAG answers in terms of precision, provenance and conciseness, it was not feasible to perform expert analysis of the generated QA pairs themselves due to the size of our dataset and availability of our experts.

In this paper, we only use LLaMa-based models for evaluation on SPIQA due to time constraints. We expect our RAG results will generalize to any base LLM on any scientific paper QA task, but this paper has not explicitly confirmed this and we leave it as an item for future work. We did test CIPQA on five LLMs which strongly suggests our hypothesis for this is correct.

Our RAG experiments were run on eight H100 GPU cards using approximately 60 GPU hours of compute time. The GPT-4o QA pair generation took twelve hours and cost $550.

Acknowledgments

Redacted for anonymous submission

References

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng, Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 4599–4610.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393, Miami, Florida, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

Palak Jain, Livio Baldini Soares, and Tom Kwiatkowski. 2024. From RAG to riches: Retrieval interlaced with sequence generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8887–8904, Miami, Florida, USA. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. 2023. Core: A global aggregation service for open access papers. *Nature Scientific Data*, 10(1):366.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data 2023 10:1*, 10:1–12.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Stephan Lewandowsky. 2020. Climate change disinformation and how to combat it. *Annual Review of Public Health*, 42:1–21.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai Wei Chang, Song Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2024. Climaqa: An automated evaluation framework for climate foundation models.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. RAG-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735, Miami, Florida, USA. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My climate advisor: An application of NLP in climate adaptation for agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro,

Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:7684–7699.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:1–28.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment 2023 4:1*, 4(1):1–13.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

Francisco Zanartu, Yulia Otmakhova, John Cook, and Lea Frermann. 2024. Generative debunking of climate misinformation. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 46–62, Bangkok, Thailand. Association for Computational Linguistics.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Results breakdown

Table 7 breaks down the results of models on CPIQA by prompt variation.

B Prompts

B.1 QA template with context

```
You are an assistant for climate
research question-answering tasks.
Use the following pieces of
retrieved context to answer the
question. If you don't know the
answer, say that you don't know. Use
three sentences maximum and keep
the answer concise.
Retrieved information: {context}
Question: {question}
Answer:
```

B.2 Stage 1 contexts generation template

```
Given a question, describe in detail 3 contexts or domains in which it can be asked, explain the contexts with a paragraph each. Include titles of academic documents that could be used in the context. Give the contexts as 3 paragraphs with no headings.
Question: {question}
Contexts:
```

B.3 Stage 2 keyword generation template

```
Given a question and context about the question, decompose the question and context into a set of relevant long-form query sentences for evidence document retrieval (RAG) that can answer the question. Present each sentence on a newline only with no headings.
Context: {context}
Question: {question}
Decomposed phrases:
```

Large language model	Question Audience	Question Type	2 Step RAG (BERTScore-F1)	Context-RAG (BERTScore-F1)
OpenAI GPT-4o	General public	Numeric	73.67	76.65
		Figure-based	66.40	67.06
		Non-fig	64.25	67.10
		Reasoning	63.41	63.81
	Climate sceptic	Numeric	64.61	65.55
		Figure-based	64.36	66.15
		Non-fig	64.97	66.32
		Reasoning	64.97	66.39
	Climate expert	Numeric	78.48	81.34
		Figure-based	68.62	70.73
		Non-fig	67.69	69.92
		Reasoning	63.97	66.63
Google Gemini 2.0-flash	General public	Numeric	64.11	64.93
		Figure-based	61.81	63.70
		Non-fig	60.64	62.75
		Reasoning	59.28	61.63
	Climate sceptic	Numeric	60.84	62.35
		Figure-based	60.07	61.97
		Non-fig	60.35	62.23
		Reasoning	60.23	62.35
	Climate expert	Numeric	70.02	72.35
		Figure-based	64.66	67.18
		Non-fig	64.76	66.01
		Reasoning	60.04	62.35
Llama-3.3-70B-Instruct	General public	Numeric	63.64	72.11
		Figure-based	64.13	67.48
		Non-fig	63.00	64.81
		Reasoning	62.33	62.05
	Climate sceptic	Numeric	63.33	61.22
		Figure-based	62.93	61.16
		Non-fig	63.28	60.23
		Reasoning	63.14	60.04
	Climate expert	Numeric	70.26	77.59
		Figure-based	66.93	66.66
		Non-fig	66.90	66.10
		Reasoning	63.32	63.89
DeepSeek-R1-Distill-Qwen-32B	General public	Numeric	70.40	67.78
		Figure-based	65.05	65.16
		Non-fig	63.30	66.04
		Reasoning	62.25	61.48
	Climate sceptic	Numeric	62.80	63.56
		Figure-based	62.58	64.07
		Non-fig	63.16	64.28
		Reasoning	63.50	64.45
	Climate expert	Numeric	73.26	74.40
		Figure-based	63.75	64.43
		Non-fig	64.87	65.74
		Reasoning	61.16	63.31
gemma-2-27b-it	General public	Numeric	68.76	67.25
		Figure-based	64.13	63.99
		Non-fig	62.43	62.81
		Reasoning	58.20	58.94
	Climate sceptic	Numeric	60.82	60.74
		Figure-based	59.81	61.22
		Non-fig	61.35	62.24
		Reasoning	60.52	62.00
	Climate expert	Numeric	71.95	64.65
		Figure-based	60.60	63.17
		Non-fig	62.67	63.18
		Reasoning	53.15	54.59

Table 7: Evaluation of models across question types and RAG methods. Questions are divided into *numeric*, *figure bases*, *non-figure based* and *reasoning based*