

CAIRNS: Balancing Readability and Scientific Accuracy in Climate Adaptation Question Answering

Liangji Kong
liangji.kong@student.unsw.edu.au
University of New South Wales
Sydney, Australia

Aditya Joshi
aditya.joshi@unsw.edu.au
University of New South Wales
Sydney, Australia

Sarvnaz Karimi
sarvnaz.karimi@csiro.au
CSIRO's Data61
Sydney, Australia

Abstract

Climate adaptation strategies are proposed in response to climate change. They are practised in agriculture to sustain food production. These strategies can be found in unstructured data (for example, scientific literature from the Elsevier website) or structured (heterogeneous climate data via government APIs). We present Climate Adaptation question-answering with Improved Readability and Noted Sources (CAIRNS), a framework that enables experts—farmer advisors—to obtain credible preliminary answers from complex evidence sources from the web. It enhances readability and citation reliability through a structured ScholarGuide prompt and achieves robust evaluation via a consistency-weighted hybrid evaluator that leverages inter-model agreement with experts. Together, these components enable readable, verifiable, and domain-grounded question-answering without fine-tuning or reinforcement learning. Using a previously reported dataset of expert-curated question-answers, we show that CAIRNS outperforms the baselines on most of the metrics. Our thorough ablation study confirms the results on all metrics. To validate our LLM-based evaluation, we also report an analysis of correlations against human judgment.

CCS Concepts

• Computing methodologies → Natural language processing.

Keywords

climate science, large language models, question-answering, natural language processing

ACM Reference Format:

Liangji Kong, Aditya Joshi, and Sarvnaz Karimi. 2025. CAIRNS: Balancing Readability and Scientific Accuracy in Climate Adaptation Question Answering. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3701716.3715501>

1 Introduction

Climate adaptation deals with supporting people and ecosystems in the wake of projected climate impacts [8]. In agriculture, the knowledge of adaptation strategies is written in the scientific literature

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW Companion '25, Sydney, NSW, Australia

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1331-6/25/04
<https://doi.org/10.1145/3701716.3715501>

and industry reports. One way to make this knowledge accessible is through Question-Answering (QA) tools, which could utilise large language models (LLMs) with retrieval-augmented generation (RAG) from online data sources [6]. A previously reported RAG system for climate adaptation QA in agriculture identifies three critical gaps: (1) localisation of queries; (2) integration of evidence from multiple sources, such as climate data APIs, scientific literature or industry reports; and (3) evaluation of these systems with experts [4]. Moreover, these systems often face a trade-off between factual accuracy and human readability because scientifically rigorous answers tend to be less accessible, whereas fluent narratives risk oversimplifying or distorting evidence. In this paper, we address the three gaps and propose a trade-off mechanism called Climate Adaptation question-answering with Improved Readability and Noted Sources (CAIRNS), a *location-aware and evidence-traceable*, a QA framework for agricultural climate adaptation. This work is aligned with the United Nations Sustainable Development Goal (SDG) 13, *Climate Action*, aiming to balance these competing goals by combining structured prompting for readability with strict evidence constraints for factual integrity. CAIRNS first identifies and normalizes location entities, then applies location-weighted retrieval to align regional relevance. For evaluation, CAIRNS adopts a hybrid framework that combines a seven-dimensional evaluation rubric with faithfulness metrics. Using an expert-curated dataset of 50 climate adaptation question-answers that reflect real information needs and requiring domain knowledge to answer, CAIRNS reports an improved specificity and verifiability, but also achieves a better trade-off between location prompting and retrieval precision, as compared with baselines and its ablations. Our contributions are: (1) a unified framework integrating online research papers and climate data that results in state-of-the-art average results; (2) ScholarGuide prompt, a structured prompting technique balancing readability and accuracy; and (3) a human-LLM correlation analysis to solidify the claims.

2 Related Work

Retrieval-augmented generation (RAG) has significantly improved large language models on knowledge-intensive tasks [2]. However, most existing systems focus on generic or news-based domains, with limited optimization for high-stakes applications such as agricultural climate adaptation. Recent domain-specific QA studies [4, 6, 7] incorporate document retrieval and climate data, but still rely on static retrieval and generic prompting, limiting region-specific and verifiable answers. Geography-aware QA has been explored in other domains [1, 3], but these methods are not tailored to climate reasoning. Moreover, existing studies still treat climate data and literature as separate sources, lacking a unified evidence

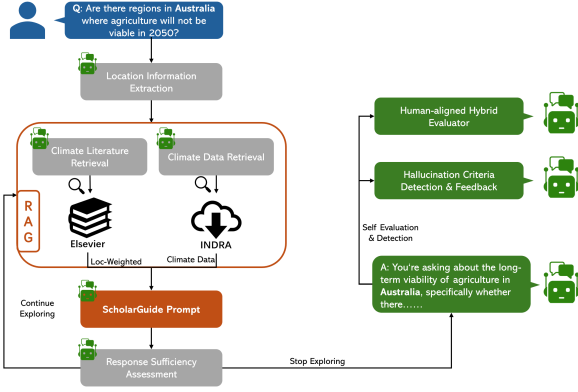


Figure 1: Overview of the CAIRNS framework.

chain [4, 6]. On the evaluation side, Nguyen et al. [6] proposed a seven-dimensional human rubric and examined consistency between LLM and expert assessments, establishing a foundation for human-aligned evaluation in this domain. Building on their rubric, CAIRNS introduces faithfulness metrics and a κ -weighted hybrid evaluator to align human and automated scoring for scalable domain QA, while its structured prompting component directly tackles the persistent trade-off between readability and factual accuracy often overlooked in prior RAG-based systems.

3 Methodology

CAIRNS comprises four stages: (1) location-weighted retrieval on climate literature from the Elsevier website and data collection from a climate data API; (2) answer generation with ScholarGuide Prompt; (3) answer verification; and (4) hybrid evaluation. These stages (Figure 1) are described as follows.

To retrieve the relevant documents, we design a hybrid retriever that combines sparse lexical matching (Okapi BM25) and dense semantic retrieval (BAAI/bge-base-en-v1.5 [9]) through a weighted scoring scheme. Hybrid retrieval alone is insufficient because location cues strongly affect relevance in climate adaptation QA. Therefore, we augment the base hybrid retrieval score with a location similarity term to balance semantic and spatial relevance:

$$\text{HybridScore}' = (1 - \beta) \cdot \text{HybridScore} + \beta \cdot \text{LocSim}, \quad \beta \in [0, 1]. \quad (1)$$

The location similarity (LocSim) measures overlap between the question’s location entities and those in each document. After retrieval, a two-stage *action-parameter* reasoning process selects the appropriate climate-data API and generates parameters. The structured climate information returned by the API is converted into citation-ready text using verbalisation (for example, “the average rainfall in Picton is 835 mm”), and merged with web-sourced literature references to form a unified, verifiable evidence chain.

CAIRNS then follows a ReAct-style reasoning-retrieval loop [10]. It retrieves evidence, generates an initial answer, and applies a sufficiency checker. If information gaps are found, the system generates new sub-questions and re-enters the retrieval loop to supplement the information until the answer is deemed sufficient. At the heart of answer generation in CAIRNS is the **ScholarGuide Prompt**

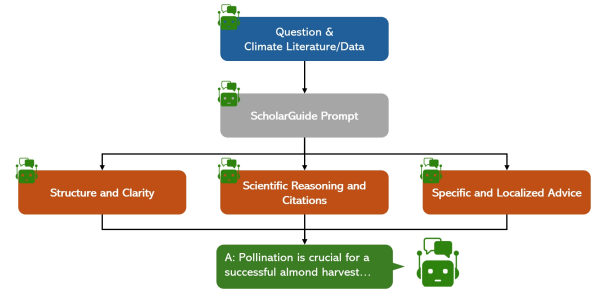


Figure 2: ScholarGuide Prompt: structured reasoning for readable and verifiable answers.

(Figure 2). This prompt is central to balancing readability and factuality. For climate adaptation QA, the ScholarGuide prompt, which is a single structured prompt with guidance, explicitly informs the LLM of the grading rubric to optimise the answers. The prompt frames the LLM as an “academic writing coach” and adopts a three-way mechanism that factors (a) structure and clarity, (b) scientific reasoning and citations, and (c) specific and localized advice.

Hybrid Evaluation Framework. We adopt the seven-dimensional human rubric of Nguyen et al. [6] and augment it with faithfulness metrics. Firstly, we compute a faithfulness score by linearly combining an automatic metric (e.g., RAGAS) and an LLM-based assessor:

$$S_{\text{faithful}} = \alpha S_{\text{RAGAS}} + (1 - \alpha) S_{\text{LLM}}, \quad \alpha \in [0, 1]. \quad (2)$$

The weighting coefficient α is validated on four external datasets to balance automatic and LLM assessments.

To further align LLM-based scoring with expert judgments, we aggregate multiple evaluators using κ -based agreement weights. Note that these weights apply only to the seven rubric dimensions. For each evaluator i , Cohen’s κ_i defines its normalized weight:

$$w_i = \frac{\max(\kappa_i, 0)}{\sum_j \max(\kappa_j, 0)}. \quad (3)$$

The final hybrid score is the weighted sum

$$S_{\text{final}} = \sum_i w_i S_i, \quad (4)$$

where S_i are component scores (including S_{faithful} and other automated sub-scores). This κ -weighted fusion preserves interpretability while privileging evaluators that empirically align with experts.

4 Experiment Setup

Datasets and Tasks. We evaluate all models on a state-of-the-art, 50-question dataset provided in [6]. The questions are written by domain experts and cover diverse regions and topics. The retrieval corpus consists of a set of 13000 publicly available climate-related papers from Elsevier. We automatically extract and normalize geographic information (*country/adm1*), where *country* denotes the nation and *adm1* corresponds to major cities (e.g., *Australia/Sydney*), as metadata to support location-weighted retrieval. Climate data

are from INDRA API ¹ (e.g., precipitation, temperature, and crop-yield indicators), transformed into structured text snippets for use during generation.

Baselines and Models. We use Gemini-2.0-Flash in LLM-only baselines without external knowledge and only provide the original question. We experiment with multiple ablations of CAIRNS: (1) **Domain RAG (literature-only)**: retrieval from the literature subset only; (2) **Domain RAG (literature+data)**: retrieval augmented with both literature and climate data based on two-stage API reasoning. The base model in the case of CAIRNS and all its ablations is Gemini-2.0-Flash. All models share identical corpora, prompt structures, and evaluation settings for a fair evaluation.

Evaluation Protocol. We build the Hybrid Evaluator using the *My Climate CoPilot User Study Annotations* dataset [5]. Six commercial LLMs from three vendors are compared against human annotations, and Cohen’s κ scores are computed to measure agreement. These κ values determine reliability weights $W_{m,d}$ for each model m and evaluation dimension d . This calibrated Hybrid Evaluator is then used to score the outputs from our various CAIRNS configurations. We evaluate both zero-shot and few-shot settings, using five annotated examples per setting, also from [5]. The evaluator scores across seven dimensions: Context, Structure, Language, Comprehensiveness, Specificity, Citations, and Accuracy (range: 0-3, as in past work), and computes two additional faithfulness metrics (range: 0-1): Citation Rate and Faithfulness Rate.

5 Results

Table 1 shows that CAIRNS provides the best average scores across all experiments, with key trends described as follows.

Impact of ScholarGuide Prompt (SGP). The ScholarGuide Prompt improves readability-related dimensions, including contextual richness, structural coherence, and language fluency. Meanwhile, its citation constraint effectively corrects inconsistencies across cited sources. Although this constraint slightly decreases the overall citation rate, it improves faithfulness, as indicated by a higher FaithRate compared to the ‘-SGP’ ablation. By enforcing unified reference numbering and grounding each factual statement to explicit evidence, SGP enables the model to generate clearer, more verifiable, and location-specific answers. This pattern remains stable across both zero-shot and few-shot settings, demonstrating that structured prompting is a key driver of CAIRNS’s performance.

Impact of Multi-turn ReAct (MTR). The iterative reasoning loop allows the model to refine its intermediate outputs by repeatedly aligning retrieved literature and climate data, leading to more complete and context-aware answers, as reflected by higher comprehensiveness scores. While MTR’s contribution is modest relative to SGP, it complements the structured generation process and improves robustness on multi-source or data-intensive questions.

Location-Weighted Retrieval (LWR). Ablating the LWR component leads to a slight drop in specificity, while other metrics remain largely unchanged. This indicates that LWR plays a positive role in enabling the model to generate region-sensitive answers, particularly in questions where geographic focus is essential.

¹<https://research.csiro.au/indra/>

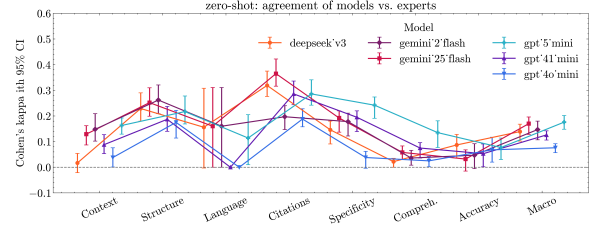


Figure 3: Zero-shot agreement between LLM evaluators and human experts across seven dimensions based on Cohen’s κ with 95% confidence interval (CI).

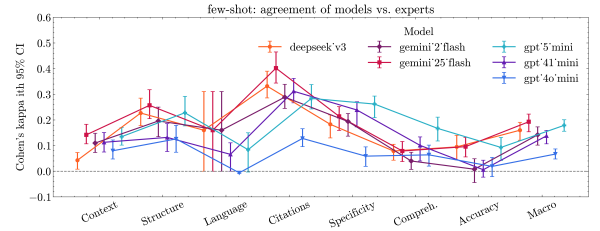


Figure 4: Few-shot agreement between LLM Evaluators and Human Experts across Seven Dimensions (Cohen’s κ with 95% CI).

Effect of Data-augmented Retrieval. We investigated the role of structured climate data in retrieval-augmented generation. Although we hypothesized that integrating structured climate data would enhance the factual support and interpretability of generated answers, the results suggest that this capability remains at an early stage. As shown in Table 1, the DATA_RAG (literature+data) pipeline underperforms the RAG pipeline (literature-only) across most metrics, with the decline most pronounced in accuracy. This indicates that the current model is still limited in its ability to leverage structured climate data to substantiate or refine textual reasoning.

6 Evaluating the LLM-based Evaluator

It is vital that the validity of LLM-based evaluation be scrutinized. To evaluate the evaluation framework itself, we first calculated the agreement between multiple LLM-based evaluators and human experts across all assessment dimensions using Cohen’s κ . This reflects the feasibility and trustworthiness of automated evaluation on each dimension. As illustrated in Figures 3 and 4, the consistency between LLM evaluators and experts varies notably across dimensions. For citations and structure, there is relatively high alignment, while specificity and comprehensiveness remain challenging, showing broader confidence intervals and lower κ values. These differences reveal that individual evaluators demonstrate varying stability and reliability across different aspects of answer quality. As a result, we used the κ values of each evaluator and dimension as weights to design a domain-specific Hybrid Evaluator tailored for agricultural climate adaptation QA. This evaluator integrates the strengths of multiple LLM-based assessors without requiring any fine-tuning or reinforcement learning, resulting in the highest overall consistency and credibility of automated evaluation observed in our study.

Table 1: Comparative evaluation of CAIRNS and baseline pipelines across seven dimensions and faithfulness metrics. MTR = Multi-Turn ReAct. SGP = ScholarGuide Prompt. CD = Climate Data. LWR = Location-Weighted Retrieval. Bold values indicate the best score in each dimension. For ablation tests, - indicates that the component of CAIRNS that has been removed, and + indicates the external component that has been added. N/A: Not Applicable.

Method	Context $\uparrow$ (0-3)	Structure $\uparrow$ (0-3)	Language $\uparrow$ (0-3)	Citations $\uparrow$ (0-3)	Specificity $\uparrow$ (0-3)	Compreh. $\uparrow$ (0-3)	Accuracy $\uparrow$ (0-3)	AVG $\uparrow$ (0-3)	CitRate $\uparrow$ (0-1)	FaithRate $\uparrow$ (0-1)
Baselines										
Gemini (Zero-shot)	2.84±0.01	2.0±0.01	3.0±0.0	0.42±0.06	1.04±0.06	2.60±0.21	1.84±0.05	1.96±0.024	N/A	N/A
Gemini (Few-shot)	2.86±0.05	1.98±0.01	3.0±0.0	0.36±0.06	1.02±0.05	2.56±0.18	1.84±0.08	1.95±0.021	N/A	N/A
CAIRNS										
CAIRNS (Zero-shot)	3.0±0.0	3.0±0.0	3.0±0.0	2.94±0.0	1.86±0.02	2.50±0.13	2.78±0.07	2.72±0.02	0.528	0.814
CAIRNS (Few-shot)	3.0±0.0	3.0±0.0	3.0±0.0	2.94±0.0	1.88±0.11	2.24±0.18	2.80±0.06	2.69±0.03	0.528	0.814
Ablations of CAIRNS (Zero-shot)										
-SGP -MTR	1.66±0.04	2.44±0.02	3.0±0.0	2.86±0.01	1.52±0.10	1.98±0.09	2.82±0.04	2.33±0.02	0.733	0.689
-SGP	1.78±0.07	2.44±0.05	3.0±0.0	2.94±0.02	1.60±0.03	2.18±0.11	2.86±0.07	2.39±0.02	0.740	0.799
-SGP -MTR +CD	1.50±0.03	2.36±0.02	2.96±0.01	2.82±0.02	1.58±0.07	2.02±0.21	2.48±0.06	2.23±0.03	0.754	0.688
-SGP +CD	1.64±0.05	2.44±0.05	3.0±0.0	2.90±0.06	1.54±0.06	2.14±0.12	2.64±0.05	2.32±0.03	0.778	0.832
-LWR	3.0±0.0	3.0±0.0	3.0±0.0	2.92±0.03	1.78±0.02	2.30±0.27	2.82±0.06	2.69±0.04	0.511	0.804
-MTR	3.0±0.0	3.0±0.0	3.0±0.0	2.88±0.04	1.82±0.03	2.34±0.15	2.84±0.07	2.69±0.02	0.490	0.781
Ablations of CAIRNS (Few-Shot)										
-SGP -MTR	1.64±0.06	2.42±0.04	3.0±0.0	2.88±0.03	1.54±0.04	1.96±0.11	2.84±0.05	2.33±0.02	0.733	0.689
-SGP	1.78±0.05	2.44±0.01	3.0±0.0	2.96±0.01	1.60±0.09	2.00±0.14	2.84±0.05	2.37±0.02	0.740	0.799
-SGP -MTR +CD	1.50±0.05	2.34±0.05	2.96±0.01	2.82±0.04	1.54±0.08	2.16±0.09	2.48±0.07	2.25±0.02	0.754	0.688
-SGP +CD	1.64±0.06	2.46±0.04	3.0±0.0	2.94±0.05	1.54±0.06	2.22±0.07	2.60±0.05	2.34±0.01	0.778	0.832
-LWR	3.0±0.0	3.0±0.0	3.0±0.0	2.92±0.04	1.80±0.05	2.40±0.13	2.82±0.08	2.69±0.02	0.511	0.804
-MTR	3.0±0.0	3.0±0.0	3.0±0.0	2.86±0.02	1.84±0.11	2.20±0.13	2.84±0.06	2.68±0.02	0.490	0.781

7 Conclusions and Future Work

CAIRNS is a question-answering framework for agricultural climate adaptation that balances readability, accuracy, and citation transparency. It integrates structured prompting (SGP), multi-turn reasoning, and a consistency-weighted Hybrid Evaluator. Experimental results demonstrate that SGP substantially improves structural clarity and citation reliability without sacrificing factual precision, reflecting a meaningful step toward aligning QA outputs with domain researchers' communication preferences. In contrast, data-augmented retrieval remains less effective at the current stage. By leveraging inter-model agreement as a weighting signal, our Hybrid Evaluator achieves the highest credibility among automated evaluations without fine-tuning or reinforcement learning. The evaluator leverages inter-model agreement as a weighting signal and can be directly applied to low-resource or expert-scarce settings for question-answering. In the future, we plan to enhance location-aware reasoning by incorporating fine-grained geographic information, such as latitude and longitude, to enable document transferability across similar climatic regions and to generalize literature-based insights. In parallel, we aim to establish a larger human-aligned evaluation benchmark that spans diverse question types and subdomains, serving both as a calibration resource for the Hybrid Evaluator and as a fine-tuning corpus for future domain-specific models, promoting reproducible and trustworthy QA research in climate adaptation.

References

- [1] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 513–523.
- [2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [3] Haonan Li, Martin Tomko, and Timothy Baldwin. 2023. Location Aware Modular Biencoder for Tourism Question Answering. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*. 95–109.
- [4] Vincent Nguyen, Willow Hallgren, Ashley Harkin, Mahesh Prakash, and Sarvnaz Karimi. 2025. My Climate CoPilot: A Question Answering System for Climate Adaptation in Agriculture. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics, Association for Computational Linguistics, Vienna, Austria. <http://hdl.handle.net/102.100.100/706402?index=1> csiro:EP2025-1223.
- [5] Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2025. My Climate Copilot User Study Annotations. CSIRO. Data Collection. doi:10.25919/x5wq-n705
- [6] Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, and Mahesh Prakash. 2025. Question Answering in Climate Adaptation for Agriculture: Model Development and Evaluation with Expert Feedback. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics, Association for Computational Linguistics, Vienna, Austria. <http://hdl.handle.net/102.100.100/706403?index=1> csiro:EP2025-0636.
- [7] David Osei Opoku, Ming Sheng, and Yong Zhang. 2025. DO-RAG: A Domain-Specific QA Framework Using Knowledge Graph-Enhanced Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.17058* (2025).
- [8] Hens Runhaar, Bettina Wilk, Åsa Persson, Caroline Uittenbroek, and Christine Wamsler. 2018. Mainstreaming climate adaptation: taking stock about “what works” from empirical research worldwide. *Regional environmental change* 18, 4 (2018), 1201–1210.
- [9] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. *arXiv:2309.07597* [cs.CL]
- [10] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Received 17 November 2025; revised xxx; accepted xxx