Question Answering in Climate Adaptation for Agriculture: Model Development and Evaluation with Expert Feedback

Vincent Nguyen and **Sarvnaz Karimi** and **Willow Hallgren** and **Mahesh Prakash**
CSIRO, Australia
`firstname.lastname@csiro.au`

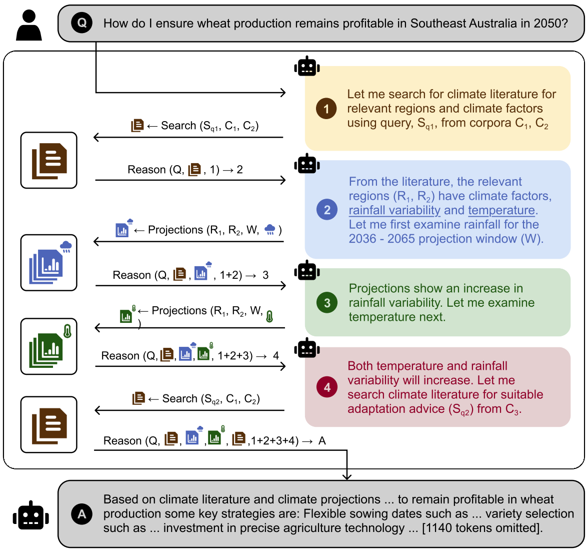
Abstract

The generative capabilities of the large language models (LLMs) are deployed for domain-specific question answering systems. However, their ability to answer climate adaptation questions remains unclear. In particular, can they be used by agronomists and climate scientists to answer questions on the best climate adaptation strategies? Answering questions in this domain requires knowledge of climate data and its uncertainties, and the ability to link them to the broader climate literature while accommodating the unique constraints of users and experts. We investigate the generative and evaluative capabilities of several state-of-the-art LLMs, open-source and proprietary, on climate adaptation for agriculture questions posed by *domain experts* using evaluation criteria designed by the experts.

We propose an iterative exploration framework that enables LLMs to dynamically aggregate information from heterogeneous sources, such as text from climate literature and structured tabular climate data from climate model projections and historical observations. Our experiments demonstrate that LLMs can aggregate heterogeneous data to (1) answer questions, but at a trade-off between presentation quality and epistemological accuracy; and, (2) evaluate answers, but are not as competent at identifying high-quality answers and erroneous information compared to domain experts.

1 Introduction

Domain-specific question answering (QA) has been studied for decades (Green et al., 1961; Woods, 1973). With the increase in generative capabilities of large language models, generative QA has advanced considerably in various domains such as the medical (Singhal et al., 2023), finance (Li et al., 2023), education (Wen et al., 2024) and climate mitigation (Vaghefi et al., 2023). Yet a crucial domain left under-explored is *climate adaptation*—the process of adjusting to current and projected climate change impacts to safeguard people and ecosystems (Runhaar et al., 2018; Lee et al., 2023).


Q How do I ensure wheat production remains profitable in Southeast Australia in 2050?

1 Let me search for climate literature for relevant regions and climate factors using query, S_{q1}, from corpora C_1, C_2

2 From the literature, the relevant regions (R_1, R_2) have climate factors, rainfall variability and temperature. Let me first examine rainfall for the 2036 - 2065 projection window (W).

3 Projections show an increase in rainfall variability. Let me examine temperature next.

4 Both temperature and rainfall variability will increase. Let me search climate literature for suitable adaptation advice (S_{q2}) from C_3.

A Based on climate literature and climate projections ... to remain profitable in wheat production some key strategies are: Flexible sowing dates such as ... variety selection such as ... investment in precise agriculture technology ... [1140 tokens omitted].
← Search (S_{q1}, C_1, C_2)
Reason (Q, , 1) → 2
← Projections (R_1, R_2, W,)
Reason (Q, , , 1+2) → 3
← Projections (R_1, R_2, W,)
Reason (Q, , , , 1+2+3) → 4
← Search (S_{q2}, C_1, C_2)
Reason (Q, , , , ,1+2+3+4) → A

Figure 1: Our framework for generating responses to expert climate questions using scientific literature, reports, climate projections, and observational data.

Climate adaptation, however, poses unique challenges when developing question answering systems. It requires integrating three different categories of information to help shape decision-making and adaptation strategies: (1) climate projections (Sansom et al., 2021; Jagannathan et al., 2023); (2) scientific literature (De La Calzada et al., 2024) and climate adaptation reports (Lemos and Rood, 2010); and, (3) location-specific historical data (Guo et al., 2015; Keenan, 2015). For instance, consider the question "*How do I ensure wheat production remains profitable in Southeast Australia in 2050?*" Answering this requires not only expert knowledge of the relevant climate factors (e.g., rainfall variability and average temperature) but also contemporary knowledge of viable adaptation strategies specific to wheat production drawn from specialised literature.

To advance the climate adaptation QA, we pro-

Findings of the Association for Computational Linguistics: ACL 2025, pages 7045–7075
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

pose a novel dual-purpose framework that enables exploring and aggregating large quantities of climate information from various sources, allowing us to probe the generative and evaluative capabilities of current LLMs. This framework can be adapted to other domains that require domain expertise and different information types and modalities. Our contributions are:

1. A domain-specific framework that allows state-of-the-art (SOTA) models to explore, refine and aggregate climate data and climate literature for QA and evaluation.
2. The first study of its kind on climate adaptation for agriculture QA and evaluation, investigating the capabilities of current SOTA models using our novel expert-created annotation criteria containing presentational and epistemological standards from research and industry. In this study, we investigate the models' baseline domain knowledge, their proficiency in using climate data and climate literature, common errors and hallucinations, and their alignment with human expert evaluations.
3. An analysis of the difficulties in climate adaptation QA encountered by both experts and LLMs.
4. A novel alignment method for domain-specific automatic QA evaluation that improves automatic evaluation by integrating expert feedback directly in few-shot learning exemplars.

2 Methodology

Domain-specific QA often requires specialised knowledge and climate adaptation QA is no exception. Climate adaptation QA requires integrating heterogeneous information sources—including large-scale tabular climate projections, nation-specific historical data, and diverse climate literature such as scientific literature, adaptation reports, and books—to help shape decision-making and adaptation strategies (Lemos and Rood, 2010; Keenan, 2015). We propose a two-step approach to measure the capabilities of LLMs in climate adaptation QA by: (1) constructing the datasets and tools to access heterogeneous climate resources; and, (2) creating a framework for probing the climate adaptation QA and automatic evaluation capabilities of LLMs.

Climate Literature	# Documents	# Chunks
Climate Science	2.36M	27.1M
Environmental Science	144k	163k
Regional Literature	28	1293

Table 1: Statistics on indexed Climate Literature (CL). The chunk size is 512 tokens.

2.1 Constructing Datasets and Resources

Climate Adaptation Questions Human experts created a set of 50 climate adaptation questions that capture some of their information needs and require domain knowledge to answer.

Climate Literature Climate adaptation QA requires domain knowledge regarding noteworthy regional or commodity climate factors, relevant and up-to-date adaptation strategies, and general climate science knowledge. To satisfy this requirement, we curated a corpus that covers literature in general climate science, environmental science, and regional adaptation (Table 1). The literature is sourced from Semantic Scholar, Elsevier, and expert-sourced reports (see Appendix A.1 for details). We created a hybrid retrieval index containing an inverted index and a vector database with their respective embeddings (Nguyen et al., 2022). Following previous work (Nguyen et al., 2024), retrieval from the index uses a hybrid scorer, a linear combination between the BM25 (Robertson et al., 1995) and embedding cosine similarity between question and document embeddings.

Climate Data Climate data, such as climate projections, are typically sourced from large-scale global studies such as CMIP (World Climate Research Programme, 2025), whereas historical observations are restricted to countries. To satisfy both requirements, we source climate data from MyClimateView (Webb et al., 2023), which provides API access to both downscaled CMIP 5 (Taylor et al., 2012) and observational data. We create climate data access tools that connect to all 89 climate API endpoints (see Appendix A.6) from MyClimateView. Most climate-related questions specify locations by name (e.g., Young Town) or amorphously (e.g., Southwest Asia), whereas Climate data is generally organised by coordinates (latitude and longitude). Therefore, we create a tool that resolves locations to coordinates. This allows access to historical data and projections for climate factors such as temperature, rainfall, and soil moisture.

2.2 Generation and Evaluation Framework

One major challenge in using climate data is ambiguity on which climate factor or metric is relevant to a given question (Jagannathan et al., 2023). Furthermore, climate data is very large in volume. For example, CMIP 5 can reach up to 119 TB (Navarro-Racines et al., 2020). Therefore, it is essential to measure the effectiveness of LLMs in exploring and identifying the correct climate data for a given question. Similarly, for climate literature, a subset of the relevant literature is needed to answer adaptation questions. Another challenge is that LLMs may lack the domain knowledge to accurately find relevant information for adaptation questions. An LLM may need to integrate the missing expert knowledge before being able to determine what climate information or literature is needed. Therefore, LLMs must dynamically reason, explore, and identify key climate factors from large datasets while retrieving relevant literature for missing domain knowledge. To achieve this, we propose a framework for iterative exploration of climate data and literature, called *Generation and Evaluation Framework or GEF*, which allows unconstrained exploration of climate knowledge.

Specifically, we prompt LLMs using a function-calling template to: (1) decide the appropriate climate data sources[1] or climate literature corpora for a given question via a reasoning step; (2) perform structured generation of **one** function parameters such as literature search parameters (query, corpora selection, and the number of documents to retrieve) or climate access tool parameters (climate factor, projection window, location, and other specific climate factor parameters), discarding any additional generated parameters for other functions; (3) observe the output of the resulting data from the selected tool called with the generated parameters; and, (4) decide whether to continue exploring additional literature or climate data via a reasoning step (back to step 1) or to aggregate available information to answer the question (see Figure 1 for a general overview). This exploration framework forces the LLM to consider one function call at a time (discarding additional calls) and observe the returned information before continuing; the LLM must integrate the current knowledge it has received to refine its search. Using this framework, we probe LLM models by generating answers to the expert questions (see prompt in Appendix B.1).

Domain Expert Metrics To measure the quality of generated answers, a set of bespoke criteria was created by our in-house domain experts to reflect the *presentation* and *epistemological* (Bulian et al., 2024) expectations of experts in the climate adaptation field. These metrics are defined below, with the full definitions listed in Appendix A.4:

- **Context:** The response provides context to the question and answer.
- **Structure:** The response is formatted well and can be easily read.
- **Use of Language:** The language is appropriate and consistent with industry expert terminology.
- **Comprehensiveness:** The response is complete, the depth of knowledge is appropriate, and of proper length.
- **Specificity:** Provides information specific to the commodity and/or location.
- **Use of Citations**: Citations are appropriate, ample, and do not impede readability.
- **Scientific Accuracy:** The system is faithful to the given source material, and there were no apparent hallucinations.

For each criterion, three sub-questions, rated with *yes*, *no*, or *N/A*, are used. The score for the criterion is calculated by counting the *yes* responses to the sub-questions and ranges from 0 to 3.

2.3 Aligning LLMs with Domain Experts

Expert evaluation within climate science NLP can be challenging and often has low agreement (Leippold and Diggelmann, 2020), given the task difficulty (Varini et al., 2020; Mallick et al., 2025), domain language ambiguity (Varini et al., 2020; Bulian et al., 2024), and the time required to validate answers (Chattopadhyay et al., 2020). Therefore, given the time constraints of experts and low agreement, automatic evaluation with external resources may be a feasible alternative. Using the GEF framework, we probe LLMs' baseline zero-shot and few-shot evaluation capabilities by providing the annotation guidelines and question-generated answer pairs to LLMs (prompt in Appendix B.2). For few-shot, we randomly sampled five exemplars[2],

---

[1] Accessing through the MyClimateView API endpoints.

[2] Number of exemplars was chosen using a small, separate validation set.

provided that they were not from the same question, for in-context learning (ICL). Given the often low agreement in the domain, we also propose a novel few-shot learning method that can align (Liu et al., 2023b; Hwang et al., 2023) LLMs towards expert evaluation preference by directly embedding expert feedback (ExF) within the exemplars. This feedback approach differs from prior approaches such as ICL with explanations (Ye and Durrett, 2022) or chain-of-thought (Wei et al., 2022). Generally, an explanation would contain factual reasoning (similar to a chain of thought) or rationale for the annotations provided. On the other hand, feedback differs from explanations, as the expert's feedback may not explain the annotation scores they chose but instead proposes how a response can be improved. Given the time constraints of experts, an added benefit of using feedback over explanations is that they are easier to obtain as it requires a lower mental load.

We use Kendall's Tau (Kendall, 1938) as our primary metric for correlation between automatic and expert evaluations, as well as expert annotator agreement. However, while Kendall's Tau can consider the strength of association between annotations, it assigns an equal weighting to each ranking position for annotations. Therefore, we use a secondary metric, Ranked Biased Ordering (RBO) (Webber et al., 2010), which considers both the association between annotations and their positions. We use the same hyperparameters from the original work, which attribute 86% of the weight to the top-10 ranked items.

3 Experimental Setup

Embeddings To encode document and query embeddings, we used Stella 1.5b v5[3], which was observed to perform best on a validation set of human judgments of query-document pairs (more details in Appendix A.3).

LLMs We evaluated three open and closed-weight SOTA LLMs[4]—Claude Sonnet 3.5 v2 (2024-10-22), OpenAI GPT-4o (2024-02-15), and Qwen 2.5-72b Instruct (Qwen 72b for short)—for both answer generation and automatic evaluation (Appendix A.2 for more details). Using a greedy sampling approach, we generated responses to 50 climate adaptation questions under two conditions: with our GEF framework and without, as a baseline. In total, 300 question-answer pairs were generated.

Human expert annotation Thirteen climate experts annotated these question-generated answer pairs using the seven criteria outlined in Section 2.2. When deemed necessary, experts also provided feedback to highlight specific problems in responses. Different experts re-annotated a fifth of the annotations to measure agreement. For half of the responses generated with GEF, experts annotated the correctness of the LLM's use of the climate data access and literature search tools. We publicly release the annotations from the experts through CSIRO's data access portal: `https://doi.org/10.25919/x5wq-n705`.

Annotator Details All annotators were experts in climate science, including agronomists and climate change scientists. Ethics approval was given by CSIRO's ethics committee, which covered all user studies, annotation, and feedback collection. All annotators were volunteers and could choose to withdraw from the study at any point. A brief one-hour introductory session was held to provide context for the annotation process. Initially, 26 climate experts filled out an expression of interest form. However, only thirteen performed the annotations as the others withdrew from the study. Annotations were anonymised during evaluation and error analysis. Overall, the annotation process took four months from ethics approval, expression of interest, annotation briefing and guidelines, and annotation time. To ensure anonymity, annotators were given individual sets of annotations via a unique link. Annotations were collated, and annotator details were not retained.

Automatic evaluation We investigated the climate adaptation QA evaluation performance of the LLMs using: (1) zero-shot and few-shot learning with and without expert feedback; and, (2) with GEF and without as our baseline.

4 Results and Analysis

RQ1: How well-presented and epistemologically accurate are LLM-generated answers? Our results indicate that the experimented LLMs were better at presentational aspects (context, structure, language) than epistemological ones (citations, specificity, scientific accuracy) (Table 2); this finding aligns with prior research in climate NLP (Bulian

---

[3] `https://huggingface.co/NovaSearch/stella_en_1.5B_v5` (Accessed: 12/20/2024)

[4] SOTA in function calling and structured generation.

	Context	Structure	Language	Compreh-ensiveness	Specificity	Citations	Accuracy	AVG
Qwen 72b	1.864	2.525	**2.966**	2.339	1.881	0.000	1.441	1.860
+GEF	1.690 ↓	2.034* ↓	2.707* ↓	1.828* ↓	1.707 ↓	**1.069*** ↑	1.483 ↑	1.788 ↓
GPT-4o	1.839	2.411	2.946	1.946	1.732	0.018	1.286	1.740
+GEF	1.689 ↓	2.066* ↓	2.852 ↓	1.803 ↓	1.721 ↓	0.672* ↑	1.410 ↑	1.745 ↑
Sonnet 3.5	1.600	**2.750**	2.883	2.117	2.033	0.000	1.233	1.802
+GEF	**2.063*** ↑	2.190* ↓	2.794 ↓	**2.349*** ↑	**2.063** ↑	0.778* ↑	**1.587*** ↑	**1.975*** ↑

Table 2: Climate-Augmented Question Answering evaluation using human experts. Overall, the experts agreed that the use of climate data and literature improves the quality of responses, with large increases in citations and scientific accuracy. However, this came at the expense of structure and language. A perfect score would be 3.0 for a given criterion. ↑ indicates improvement over the baseline while ↓ indicates a reduction. * indicates a statistically significant difference ($p < 0.05$; sign test) from the baseline model. We present an **unweighted** average (AVG) for ease of reading, but note that the metrics should not be considered equally important.

et al., 2024; Nguyen et al., 2024). Although including external resources via GEF generally improved epistemological qualities, it came at the expense of presentation quality. A common theme is that the structure of responses worsened as climate data was difficult to represent as text without hindering readability. Two of the three models improved with GEF, while Qwen performed worse. Overall, Sonnet 3.5 with GEF performed best, improving over the baseline in most criteria aside from structure and language.

Furthermore, of the 300 QA pairs annotated by experts, only two were given perfect scores (an example is shown in Appendix Figure 6), and eight of which were one-off from a perfect score. This highlights that current LLMs are far from being able to provide accurate, well-presented answers reliably. Given the scepticism and confusion surrounding climate science (Oreskes and Conway, 2011), a higher standard must be applied here. The trade-off between epistemological and presentation quality can be detrimental in real-world settings as presentation quality is just as important as the epistemological quality (Hsu et al., 2024; Zanartu et al., 2024). From these findings, we conclude that the current LLMs cannot yet provide answers to climate adaptation questions at the standards set by experts, even with the inclusion of external resources. Future work should aim to improve the epistemology without a significant loss in presentation quality.

RQ2: How does proficiency in using climate data and literature affect generated answer quality? We find that different models have varying capabilities in utilising external resources (Table 3), and that proficiency in external data exploration and generated answer quality is correlated (lower correctness in Table 3 correlates to lower scores in Table 2). For instance, Qwen+GEF had the highest climate data (CD) use with moderate literature (CL) use, leading to higher scores in citations and accuracy. However, most of these were used incorrectly (68% overall correctness) as the generated responses contained irrelevant information, lowering specificity and comprehensiveness and leading to a lower score than the baseline.

Model	Avg. CD Use	Avg. CL Use	Overall Correctness
Qwen 72b	4.04 (65%)	0.44 (92%)	0.68
GPT-4o	2.86 (78%)	0.08 (100%)	0.79
Sonnet 3.5	2.76 (92%)	1.16 (100%)	0.96

Table 3: Climate data (CD) and climate literature (CL) usage per question for Climate Question Answering. A subset of the generated responses is manually inspected for CD+CL exploration correctness. In brackets is the percentage of correctly explored climate data or climate literature. For example, out of all the climate data tools used Sonnet, 92% of those were considered correct use.

A similar trend occurred with GPT-4o, where incorrectly used CD (78% overall correctness) also lowered specificity. Furthermore, as CL usage was also low, we found that GPT-4o missed the appropriate literature needed to answer questions, reducing the context and comprehensiveness scores. However, given that the overall correctness was higher than Qwen, the improvements in epistemological criteria outweighed performance drops from the answer presentation.

Sonnet 3.5 was the only model that saw overall significant ($p < 0.05$) performance gains, as it used both CD and CL with high correctness (0.96). We

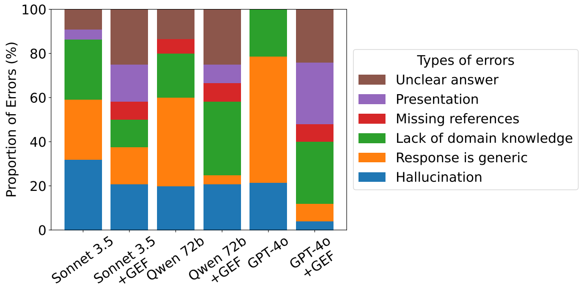


Figure 2: Percentage of error types made by baseline and GEF-based methods.

observe that CL use was higher than other models, leading to improvements in the context and comprehensiveness criteria because relevant domain knowledge could be incorporated into the response. Additionally, correct CD use led to significant improvements in accuracy and specificity. All models had significantly ($p < 0.05$; sign test) lowered structure scores when using GEF. From expert feedback, we find that this primarily results from (1) the difficulty of representing climate data textually, which often makes responses challenging to read and follow (see Appendix Figure 7); and, (2) the position of summaries varied within the response, even from the same model, and often they were after the references or within the references.

RQ3: What are common errors that LLMs make in climate adaptation QA? From expert feedback (Figure 2), we find that the baseline methods tend to (1) be too generic to be useful; for example, Sonnet 3.5 stated the temperature would be warmer for a region but did not say by how much or how it would affect specific commodities; (2) have hallucinations; for example, GPT-4o provided references which did not exist, while Sonnet 3.5 hallucinated almond pollination temperature thresholds; and, (3) lack the domain knowledge to describe appropriate and relevant climate risks. For example, Sonnet 3.5 stated that temperature was the only factor to consider for apple sunburn risk, but sunlight intensity and canopy density were also important to consider. On the other hand, we find that GEF methods tend to draw scepticism from experts, resulting in 43% more critical feedback than baseline methods. Common errors with LLMs using GEF we find are:

Unclear or incomplete answers When given access to external resources, LLMs over-rely on climate data for general knowledge questions, which leads to irrelevant answers (lowering specificity and context). For example, the question "How accurate are climate projections", GPT-4o responded with examples of climate projections for different regions it had explored, but did not answer the question. Moreover, complex jargon such as 'air drainage', specific carbon emission scenarios (RCP), or climate factors was used without providing the appropriate context or reasoning. Many responses also did not provide reference points for climate factors, leading to incomplete answers. For example, Sonnet 3.5 claimed that the number of hot days was expected to be 25-30 in the 2050 projection window but did not provide the current year for context.

Missing references As responses became more specific, models provided highly specific advice without citations; this contrasts with the baselines, where the responses were too generic to warrant citation.

Presentation Presentation quality tended to worsen (GPT-4o was most impacted) as the structure of responses became more challenging to read due to inconsistency in referencing format or climate data presentation.

Lack of domain knowledge Although the use of domain knowledge improves with GEF, there are numerous misinterpretations of climate data. For example, all models represented individual years as representative averages of climate projection windows—a practice which is frowned upon in climate science (Nissan et al., 2019). We also observe that LLMs tend to be more confident, making claims without considering the potential ramifications. One such example was when Qwen 72b recommended large-scale intensive farming systems to improve climate resilience, which was faithful to the source documents but did not consider potential water or welfare challenges.

Hallucination Although the number of extrinsic hallucinations (Dziri et al., 2021) lowered when using GEF, intrinsic hallucinations (Cao et al., 2022) became more common. For example: (1) Sonnet 3.5 interpreted literature mentioned a 79% reduction in *exports* as a 79% reduction in *production*; (2) Qwen 72b stated a reduction in cold days for lambing which contradicted the input climate data and literature; and, (3) GPT-4o used *maximum* temperature as the climate factor when accessing cli-

	Context	Structure	Language	Compreh-ensiveness	Specificity	Citations	Accuracy	AVG $\tau \uparrow$	AVG $rbo \uparrow$
Expert	**0.217**	0.192	**0.303**	-0.054	0.063	0.307	-0.114	0.131	**0.430**
LLM					**Zero-shot**				
Qwen 72b	0.126	0.163	-	0.125	0.122	0.511	0.080	0.193	0.235
+GEF	0.114	0.240	-	0.073	0.153	0.513	0.080	0.205	0.312
GPT-4o	0.043	0.323	-	0.195	0.106	0.525	0.094	0.217	0.268
+GEF	0.100	0.367	-	0.128	0.125	0.536	0.025	0.215	0.246
Sonnet 3.5	0.178	0.352	-	0.172	0.063	0.513	0.118	0.235	0.247
+GEF	0.113	0.370	-	0.237	0.082	0.460	0.052	0.224	0.249
				Few-shot					
Qwen 72b	0.066	0.099	-	0.052	0.155	0.512	0.095	0.163	0.254
+GEF	0.107	0.188	-	-0.030	0.152	0.523	0.025	0.161	0.282
+ExF	0.094	0.132	-	0.111	0.208	0.518	0.013	0.180	0.277
+GEF+ExF	0.123	0.215	-	0.038	0.185	0.534	0.044	0.190	0.324
GPT-4o	0.040	0.126	-	0.183	0.209	**0.563**	-0.003	0.186	0.204
+GEF	0.077	0.286	-	0.161	0.184	0.536	0.061	0.217	0.268
+ExF	0.022	0.154	-	0.203	**0.229**	0.527	-0.021	0.186	0.235
+GEF+ExF	0.062	0.274	-	0.179	**0.229**	0.536	0.055	0.223	0.295
Sonnet 3.5	0.092	0.374	-	0.252	0.169	0.535	0.148	0.262	0.235
+GEF	0.057	0.386	-	0.276	0.197	0.509	**0.191**	0.269	0.248
+ExF	0.110	**0.395**	-	**0.287**	0.157	0.551	0.084	0.264	0.266
+GEF+ExF	0.121	0.364	-	0.252	0.185	0.539	0.181	**0.274**	0.273

Table 4: Comparison between human annotators and LLM annotators using Kendall's Tau (τ) and rank-biased overlap (rbo), detailed results with rbo are shown in Table 7. The language criterion agreement could not be measured as each method predicted yes for each sub-question, meaning the answers could not be ranked. We experimented with different sampling methods in Table 5 to remedy this. Expert feedback (ExF) is used in the few-shot setting but not in zero-shot, as this is part of the ICL examples.

mate data but reported the *average* temperature in the response. Furthermore, extrinsic hallucinations, although lowered, were still present. For example, experts found an attribution error (Yue et al., 2023) where Sonnet 3.5 cited literature about apple sunburn *impacts* and used this to make claims about sunburn *management practices* not present in the literature.

RQ4: How effective are experts at evaluating climate QA responses? From Table 4, we find that inner-annotator agreement between human experts was generally weak, $\tau = 0.131$, highlighting the difficulty of annotation within the domain. By contrast, rbo was much higher, indicating that human experts tended to agree on high-quality responses. When reversing the ordering, we observe that $rbo = 0.301$, meaning human experts also agree on lower-quality examples. We can conclude that experts can decide on the high and low-quality responses but have difficulty evaluating medium-quality ones. For individual criteria, experts tended to agree better on presentational criteria than epistemological ones. Interestingly, agreement on language was not as high as expected, even though the average scores of generated answers were close to 3.0; there may be a few but very sparse disagreements. Comprehensiveness ($\tau = -0.054$) and accuracy ($\tau = -0.114$) were particularly divisive as these were subjective and expertise-based. Overall, this finding is not surprising as established literature in climate science NLP shows agreement is generally poor (Leippold and Diggelmann, 2020; Zanartu et al., 2024) but does establish climate adaptation QA evaluation as a difficult task.

RQ5: How effective are LLMs at evaluating climate QA responses compared to experts? From Table 4, we observe that zero-shot evaluation methods had better agreement with experts than one another regarding avg τ. Using GEF did not substantially change the correlation with experts across all models. This can be partly explained by the fact that climate data and literature exploration for GPT-4o and Qwen 72b are much lower when compared to generation (Figure 3). Therefore, there should not be much difference between the methods. On the other hand, Sonnet remains

	Context	Structure	Language	Comprehensiveness	Specificity	Citations	Accuracy	AVG $\tau \uparrow$	AVG $rbo \uparrow$
Question	**0.122**	0.358	-	**0.312**	**0.235**	**0.448**	0.117	0.265	0.270
Model	0.073	**0.406**	-	0.296	0.127	0.491	0.157	0.258	0.245
Random	0.121	0.364	-	**0.252**	0.185	**0.539**	**0.181**	**0.274**	**0.273**

Table 5: Effect of different sampling strategies on the Sonnet 3.5 model with GEF and ExF. Question sampling means we use the other five annotations of the question as exemplars (each question had six annotations), while for model sampling, we sample exemplars from the same model. Even under different sampling strategies, the language criteria agreement is still undefined.

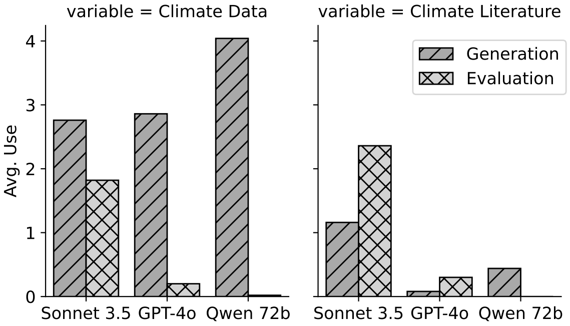


Figure 3: Average climate data and scientific literature usage in zero-shot LLMs.

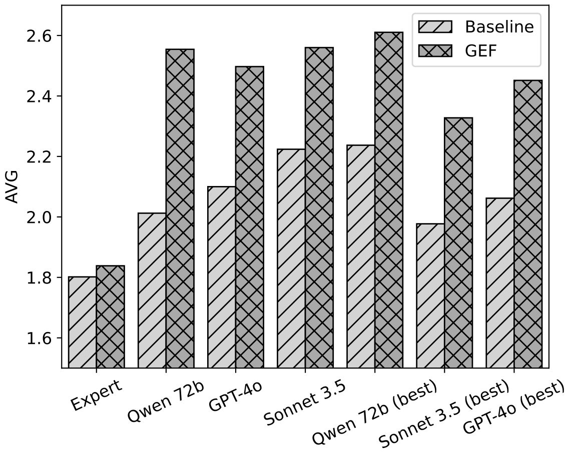


Figure 4: The AVG represents the average across all seven metrics for GEF-based methods (Qwen 72b+GEF, GPT-4o+GEF, Sonnet 3.5+GEF) and baseline methods (Qwen 72b, GPT-4o, Sonnet 3.5). Best indicates the method with the highest correlation with experts from Table 4 for that model (e.g., Qwen 72b (best) is Qwen 7b + GEF (Zero Shot)).

consistent with external data exploration, which likely implies that GEF does not directly improve alignment with experts.

This does change when few-shot learning is used, as using GEF generally leads to improved correlation. Adding expert feedback (ExF) in exemplars also improved alignment, although external data exploration was an overall richer signal. Interestingly, few-shot Qwen 72b and GPT-4o performed worse than their zero-shot counterparts, which can likely be attributed to problems with reasoning over long sequences, given that Sonnet 3.5 improved with few-shot learning. Another explanation is that the number of exemplars was not well-tuned for the models, even though we had tuned this parameter with a validation set. GPT-4o improved over its zero-shot baseline with GEF and ExF as it explored more climate data and literature than Qwen. The best automatic evaluation method was Sonnet+GEF+ExF, which consistently aligned with humans across criteria.

Generally, expert agreement tended to be higher for presentational criteria such as context, language, and structure. At the same time, automatic measures had a higher agreement on epistemological qualities such as specificity, citations, and specificity. Interestingly, no automatic method could correctly predict the language criteria, as all methods predicted perfect scores for each QA pair. A possible reason may be that the random selection of the few-shot exemplars did not contain samples with varied language annotation scores. To rule this out, we conducted an experiment where we changed the sampling strategies used to select the few-shot exemplars (Table 5) that would contain samples of varied language annotations. However, even with different sampling strategies, the language criteria still could not be predicted.

Given that all automatic evaluation methods had higher agreement than human annotators, it can be inferred that LLMs are somewhat an average of expert preference; that is, LLMs can correlate with different experts even though expert agreement is low. However, experts were still better at picking out high-quality examples as *rbo* was always higher than automatic methods. Importantly, given the correlation between correct external data use and improved performance (RQ2), experts are sensitive

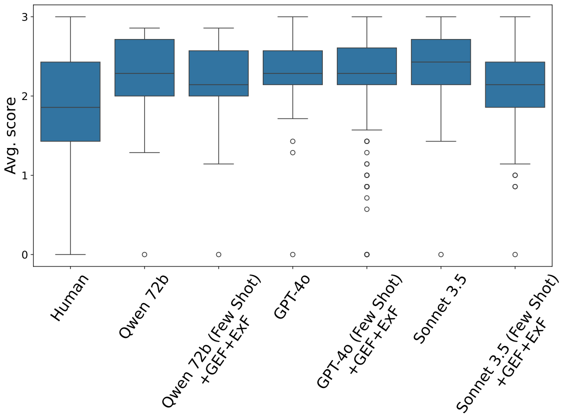


Figure 5: Distribution of scores from human experts, and baseline LLM evaluations and their few-shot counterparts with GEF and ExF counterparts.

to responses that appear high-quality (answers that cited literature and climate data for evidence) but use erroneous data. LLMs do not have this capability, evidenced by Figure 4 (and Appendix 8), where LLMs consistently reported GEF models well above the baselines. Experts place GEF-based methods slightly higher than the baselines, whereas LLMs place GEF-based models 0.5 points higher on average. While overestimation shifts with few-shot learning and expert feedback (Figure 5), LLMs still overestimate the performance of GEF-based methods over baselines and are not yet reasonable evaluators for climate adaptation QA.

5 Related Work

Climate Science in NLP In the climate NLP space, there has been a shift from pre-trained language models like BERT (Prasse et al., 2023; Pupneja et al., 2023; Greenwell and Johnson, 2023; Toetzke et al., 2023; Visalli et al., 2023) towards LLMs. Arabic mini-ClimateGPT (Mullappilly et al., 2023) creates an English instruction fine-tuning dataset from ChatGPT before translating it to Arabic and fine-tuning a Vicuna (Zheng et al., 2023) model. The similarly named ClimateGPT (Thulke et al., 2024) introduced domain-specific foundation models for climate science, which have been trained from scratch on the Llama2 (Touvron et al., 2023) architecture and instruction fine-tuned using domain-specific datasets. ClimateGPT was evaluated on climate science-specific benchmarks (Spokoyny et al., 2023). For climate science question answering, ChatClimate (Vaghefi et al., 2023) integrated information from IPCC reports or internal LLM knowledge to provide traceable answers to general climate science questions. Climate science question answering is similar to evidence-based QA Schimanski et al. (2024), which corroborates evidence from various sources to provide accurate answers.

Evaluation with LLMs As question answering moves from extractive to generative, using standard gold standard evaluation (Kamalloo et al., 2023) is increasingly more difficult. Given that LLMs are now the de facto standard for QA, it is important to evaluate their responses due to their propensity for hallucination (Chang et al., 2024). A growing trend is to use the LLMs themselves as evaluators (Lin and Chen, 2023; Chiang and Lee, 2023; Zheng et al., 2023) in a reference-free setting such as G-EVAL (Liu et al., 2023a), C-Eval for Chinese texts (Huang et al., 2023) and L-Eval for long-form text (An et al., 2024). These methods show that LLMs can correlate well with human annotators over traditional lexical metrics (Lin, 2004; Banerjee and Lavie, 2005), but there is a potential bias towards LLM-generated text and positional bias (Wang et al., 2024).

6 Conclusions

We systematically analyse the climate adaptation question answering capabilities of LLMs and find that their baseline answer generation capabilities fall below the expectations of climate experts. Although grounding LLMs with external resources, such as climate data and climate literature, improves epistemological accuracy, it degrades answer presentation and introduces new sources of errors, such as intrinsic hallucination and misinterpreting data. Additionally, our study shows that LLMs cannot yet serve as reasonable evaluators of climate adaptation responses, achieving moderate agreement with experts. Even when equipped with external data, LLMs lacked the ability to properly validate erroneous answers and identify higher-quality examples while being too lenient in scoring compared to experts. Overall, our work measures the capabilities of current LLMs in answering climate adaptation questions and the challenges of working with external resources in domain-specific question answering. It also provides a foundation for the field of climate adaptation question answering, which could be expanded to other similar scientific areas.

Limitations

Our study is not without limitations, such as (1) the use of single-turn/single-response question answering; (2) the evaluation of answers and tool parameters and not the climate data/scientific literature; (3) having two or fewer annotations per sample rather than having three experts deciding on majority labels; (4) a limited number of questions and answers; (5) the use of three LLMs for generating and evaluation; (6) study design; (7) lack of more advanced prompting strategies; (8) unable to measure the language evaluation capability of LLMs; and, (9) the metrics were not weighted. While we could not explore these limitations further in this paper, we highlight these as a potential starting point for researchers who want to explore this domain further and also provide the rationale below.

A major limiting factor was that the climate experts were volunteers, so asking for more of their time remained impractical. Further, many volunteers withdrew from the study due to the lack of time, lowering the annotation workload that could be reasonably given to the remaining experts. Therefore, a multi-turn QA setup or climate data/scientific literature evaluation would exceed this workload (Limitations 1 and 2). The limited annotation budget also lowered the number of questions that could be made, as these require careful consideration from experts and also the number of answers that could be evaluated for each question (Limitations 3 and 4).

Although stronger LLMs existed, we opted to use three LLMs (Limitation 5) for generation and evaluation for the following reasons: (1) many of the stronger LLMs were not capable of function calling and structured generation (e.g., OpenAI o1); (2) the cost of running these models would be prohibitively expensive in terms of monetary costs for proprietary models and computational costs for open-source models, lowering the number of tested models; and, (3) hardware limitations, we did not have resources to run very large open-source models (e.g., Mistral-Large 123b and Llama3 405b).

The study design introduced distrust from some climate scientists as the responses became highly specific. The study was designed to be single-blind; participants were not told which method or model they annotated. However, scepticism increased with GEF-based approaches because annotators could not see the data that models used to generate the answer (Limitation 6). Although the overall findings seem unaffected, as experts could still determine low-quality and high-quality responses, this could lower expert agreement. Applying more advanced prompting techniques (Limitation 7) would likely improve the quality of the responses and lower scepticism at the risk of making it easy to distinguish our system from the baselines.

We find that automatic LLM evaluators predicted perfect or close to perfect scores for the language criterion in responses, irrespective of method (Limitation 8). This limited the ability of our agreement metrics to capture the LLM's evaluation performance (the responses could not be ranked as they were all tied) to the language criterion, so we chose to exclude these from our automatic evaluation analysis. We did try to rectify this in Appendix 5, but to no avail. We hypothesise that this problem is likely a result of the annotation criteria for language being simplistic, as baseline results for language are almost a perfect score for most methods. Although this is a negative result, we highlight it as an interesting insight for researchers working in automatic LLM evaluation.

Finally, we did not investigate the importance weightings of the metrics and instead opted for looking at trends based on the category of metric (presentation or epistemological) or individual metrics (Limitation 9). This is because the importance of the metrics is dependent on the user. For example, a scientist may value presentation less than a layperson. This importance weighting and the inclusion of an uncertainty metric (Bulian et al., 2024), which could be an important dimension, given the overconfidence of LLMs providing responses, may serve as a foundation for future work.

Ethics Statement

The authors obtained ethics approval from CSIRO (Climate Services for Agriculture (Phase 2) (131/24)). This covered all the user studies and annotation processes involving human experts and collecting their feedback.

All LLM endpoints used in this research are part of an enterprise agreement where any data exchanged remains private and is not retained.

Acknowledgments

Dr Willow Hallgren has a PhD in Climate Change and Adaptation in Agriculture and served as an expert, driving the data curation and annotation for this research. We would like to thank our colleague, Ashley Harkin (The Bureau of Meteorology), for her expert knowledge guiding this work. We also would like to acknowledge others in the Climate Services for Agriculture project at CSIRO, Australia who provided support, helped recruit domain experts and provided their invaluable expertise: Alison Binney, Bianca Warnock, Stephanie Dickson, Bella Robinson, Yick Hung Cheung, and Melissa Spinocchia.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14388–14411, Bangkok, Thailand.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauss. 2024. Assessing large language models on climate information. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 4884–4935.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3340–3354, Dublin, Ireland.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Ashesh Chattopadhyay, Pedram Hassanzadeh, and Saba Pasha. 2020. Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific reports*, 10(1):1317.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Nick Craswell. 2009. *Mean Reciprocal Rank*, pages 1703–1703. Springer, Boston, MA, Boston, MA.

Natalia De La Calzada, Théo Alves Da Costa, Annabelle Blangero, and Nicolas Chesneau. 2024. ClimateQ&A: Bridging the gap between climate scientists and the general public. In *Proceedings of the Tackling Climate Change with Machine Learning Workshop at ICLR*, Wien, Austria.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Western Joint IRE-AIEE-ACM Computer Conference*, page 219–224, Los Angeles, CA.

Matthew P Greenwell and Thomas F Johnson. 2023. Is it all talk: Do politicians that promote environmental messages on social media actually vote-in environmental policy? *Energy, Ecology and Environment*, 8(1):17–27.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2023. Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents. *arXiv e-prints*, page arXiv:2310.19923.

Hua-Dong Guo, Li Zhang, and Lan-Wei Zhu. 2015. Earth observation big data for climate change research. *Advances in Climate Change Research*, 6(2):108–117.

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*, pages 82–92, Bangkok, Thailand.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 62991–63010, New Orleans, LA.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore.

Kripa Jagannathan, Tapan B Pathak, and David Doll. 2023. Are long-term climate projections useful for on-farm adaptation decisions? *Frontiers in Climate*, 4:1005104.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada.

Rodney J Keenan. 2015. Climate change impacts and adaptation in forest management: a review. *Annals of forest science*, 72:145–167.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, et al. 2023. Ipcc, 2023: Climate change 2023: Synthesis report, summary for policymakers. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change. *Intergovernmental Panel on Climate Change (IPCC)*.

Markus Leippold and Thomas Diggelmann. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *Proceedings of the NeurIPS 2020 Workshop: Tackling Climate Change with Machine Learning*, Virtual.

Maria Carmen Lemos and Richard B Rood. 2010. Climate projections and their impact on policy and practice. *Wiley interdisciplinary reviews: climate change*, 1(5):670–682.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the 4th ACM International Conference on AI in Finance*, page 374–382, Brooklyn, NY.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI*, pages 47–58, Toronto, Canada.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. In *Proceedings of the NeurIPS 2023 Workshop: Socially Responsible Language Modelling Research (SoLaR)*, New Orleans, LA.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. 2025. Understanding the impact of climate change on critical infrastructure through nlp analysis of scientific literature. *Sustainable and Resilient Infrastructure*, 10(1):22–39.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia.

Sahal Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Anwer, Salman Khan, and Fahad Khan. 2023. Arabic mini-ClimateGPT : A climate change and sustainability tailored Arabic LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136, Singapore.

Carlos Navarro-Racines, Jaime Tarapues, Philip Thornton, Andy Jarvis, and Julian Ramirez-Villegas. 2020. High-resolution and bias-corrected CMIP5 projections for climate change impact assessments. *Scientific data*, 7(1):7.

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*, pages 27–45, Bangkok, Thailand.

Vincent Nguyen, Maciej Rybinski, Sarvnaz Karimi, and Zhenchang Xing. 2022. Search like an expert: Reducing expertise disparity using a hybrid neural index for COVID-19 queries. *Journal of Biomedical Informatics*, 127:104005.

Hannah Nissan, Lisa Goddard, Erin Coughlan de Perez, John Furlow, Walter Baethgen, Madeleine C Thomson, and Simon J Mason. 2019. On the use and misuse of climate change projections in international development. *Wiley Interdisciplinary Reviews: Climate Change*, 10(3):e579.

Naomi Oreskes and Erik Conway. 2011. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press.

Katharina Prasse, Steffen Jung, Isaac Bravo, Stefanie Walter, and Margret Keuper. 2023. Towards understanding climate change perceptions: A social media dataset. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, New Orleans, LA.

Yashaswi Pupneja, Yuesong Zou, Sacha Levy, and Shenyang Huang. 2023. Understanding opinions towards climate change on social media. In *Proceedings of the NeurIPS 2023 Workshop: Tackling Climate Change with Machine Learning*, New Orleans, LA.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *NIST Special Publication*, 109:109.

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*, 18:1201–1210.

Philip G Sansom, David B Stephenson, and Thomas J Bracegirdle. 2021. On constraining projections of future climate using observations and simulations from multiple climate models. *Journal of the American Statistical Association*, 116(534):546–557.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1913–1931, Bangkok, Thailand.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards Answering Climate Questionnaires from Unstructured Climate Reports. *arXiv e-prints*, page arXiv:2301.04253.

Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. 2012. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change. *arXiv e-prints*, page arXiv:2401.09646.

Malte Toetzke, Benedict Probst, and Stefan Feuerriegel. 2023. Leveraging large language models to monitor climate technology innovation. *Environmental Research Letters*, 18(9):091004.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, page arXiv:2302.13971.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. 2023. Chatclimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4(1):480.

Francesco S. Varini, Jordan L. Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. In *Proceedings of the NeurIPS 2020 Workshop: Tackling Climate Change with Machine Learning*, Virtual.

Francesco Visalli, Antonio Patrizio, Antonio Lanza, Prospero Papaleo, Anupam Nautiyal, Mariella Pupo, Umberto Scilinguo, Ermelinda Oro, and Massimo Ruffolo. 2023. ESG data collection with adaptive AI. In *Proceedings of the 25th International Conference on Enterprise Information Systems*, pages 468–475.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9440–9450, Bangkok, Thailand.

Leanne Webb, Carly Tozer, Lynette Bettio, Rebecca Darbyshire, Bella Robinson, Aysha Fleming, Sigrid Tijs, Roger Bodman, Mahesh Prakash, et al. 2023. Climate services for agriculture: Tools for informing decisions relating to climate change and climate variability in the wine industry. *Australian Journal of Grape and Wine Research*, 2023.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, New Orleans, LA.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. AI for Education (AI4EDU): Advancing Personalized Education with LLM and Adaptive Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 6743–6744, Barcelona, Spain.

William A. Woods. 1973. Progress in natural language understanding: An application to lunar geology. In *Proceedings of the National Computer Conference and Exposition*, page 441–450, New York, NY.

World Climate Research Programme. 2025. Coupled Model Intercomparison Project (CMIP). Accessed: 2025-02-10.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv e-prints*, page arXiv:2412.15115.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *The 2023 Empirical Methods in Natural Language Processing Conference*, pages 4615–4635, Singapore.

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*, pages 46–62, Bangkok, Thailand.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *The 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 46595–46623, New Orleans, LA.

A Appendix

A.1 Climate Literature Corpus Collection Details

Climate literature was collected from the S2ORC (Lo et al., 2020) corpus from Semantic Scholar, high-impact climate adaptation literature from Elsevier, and expert-sourced reports. For S2ORC, documents (spanning the years 1802 to 2024) were filtered from the larger corpus based on the field of study tag (Agricultural Science, Environmental Science). For Elsevier, we selected the top 100 high-impact journals (spanning the years 1982 to 2024) based on expert guidance. Using Elsevier's API[5] we downloaded the open-access articles from these top journals. For reports, experts sourced relevant gray literature (spanning years 2009 to 2023) for the key growing regions and commodities.

A.2 Additional Experimental Details

Open-source generation We ran the open-source model, Qwen 2.5-72b (Yang et al., 2024), at 16-bit precision with Flash Attention (v2) using the llama.cpp[6] library on three H100 Nvidia 80GB GPUs. We used the maximum prompt context window of 128,000 tokens, which required roughly 200GB of GPU VRAM. Each conversation turn for generation took 1 minute, with roughly five turns per question. This approximately amounts to 4 hours for each setting in the question answering task (8 hours total) and 25 hours for each setting in the evaluation task (150 hours total).

Proprietary model generation For Claude, we used Bedrock's Converse API, and for OpenAI (through Azure and Qwen (hosted via Ollama) we used OpenAI's chat completion API. We supplied tools and their descriptions to the system prompt and used the official APIs to detect the function calls and rationale text from the model. We treated each function call as a separate conversation turn by supplying the tool result as a user turn (one at a time), even if the model made several calls in one turn. Then the LLM would either continue to call another function with reasoning text or stop and answer the user's question. We also added a limit of 15 conversation turns before requiring the model to answer the question.

Generation costs For proprietary models, we used the APIs with data privacy agreements, namely Azure API[7] for interacting with GPT-4o

---

[5] `https://dev.elsevier.com/` Last accessed: 12/20/2024

[6] https://github.com/ggerganov/llama.cpp (Last Accessed: 1/17/2025)

[7] `https://azure.microsoft.com/en-us/products/ai-services/openai-service`) (Last Accessed: 1/17/2025)

and Amazon Bedrock[8] for Claude Sonnet 3.5 v2. The cost of question answering generation was approximately 58 USD for GPT-4o and 55 USD for Claude Sonnet 3.5. For evaluation, it was 402 USD for GPT-4o and 294 USD for Claude Sonnet 3.5.

A.3 Embedding Model Selection

Human judgments We selected our embedding model using a small set of human judgments (query-document-relevance triples). This set was created by first running a hybrid scorer with BM25 and JinaBERT (Günther et al., 2023) against the climate corpora index using the first fifteen climate expert questions. A climate expert was asked to annotate the top 20 documents from each query for relevance (300 total), creating human judgments.

Model ID	**nDCG@10**
NovaSearch/stella_en_1.5B_v5	**0.7688**
Alibaba-NLP/gte-Qwen2-7B-instruct	0.7630
Salesforce/SFR-Embedding-2_R	0.7562
NovaSearch/stella_en_400M_v5	0.6999
jinaai/jina-embeddings-v2-base-en	0.6621
nvidia/NV-Embed-v2	0.4026

Table 6: Embedding model selection experiments. We experimented against the top five models from the MTEB leaderboard (10-30-2024). URLs for the model can be generated by prepending `https://huggingface.co/` to the model ID. For example `https://huggingface.co/NovaSearch/stella_en_1.5B_v5`.

Model Selection Using these human judgments, we empirically evaluated (Table 6) the top embedding models from MTEB (Muennighoff et al., 2023) and selected the best model, Stella 1.5b v5, using nDCG10 (Craswell, 2009). All documents in our corpus were chunked with a context window of 512, as this was the window size the model was trained with. For the query prompt, we use `Instruct: Given a web search query, retrieve relevant passages that answer the query. Query: {query}`.

A.4 Annotation Criteria

We provide the expanded criteria created and used by climate experts for climate science QA. In addition to the annotation guidelines, annotators were provided with a brief introduction session (1 hour) regarding the annotation process.

Instructions This activity has been designed to gather feedback on each of the responses provided by 3 LLMs (we refer to all as systems). This allows us to understand how the responses/outputs compare to other systems.

It should be noted that all systems selected for testing will be anonymised to avoid bias.

1. **Context**
    - 1.a. Attempts to give some broader context to explain the issue
        - **Rating:** Yes/No
    - 1.b. Provides an introductory paragraph to introduce the topic
        - **Rating:** Yes/No
    - 1.c. Provides a summary paragraph at the end
        - **Rating:** Yes/No
2. **Structure**
    - 2.a. Overall response is well structured, easy to read
        - **Rating:** Yes/No
    - 2.b. Headings and subheadings are well structured and logical, and with appropriate categories
        - **Rating:** Yes/No
    - 2.c. Dot points are used appropriately
        - **Rating:** Yes/No
3. **Use of Language**
    - 3.a. Phrasing is appropriate (easy to read, fluent) and not awkward or incorrect
        - **Rating:** Yes/No
        - **Explanation:**
    - 3.b. Correct use of grammar
        - **Rating:** Yes/No
    - 3.c. Consistent with language used within industry
        - **Rating:** Yes/No
4. **Use of Citations (where used)**
    - 4.a. Citations are used appropriately
        - **Rating:** Yes/No/N/A
        - **Explanation:**
    - 4.b. The number of citations used is appropriate
        - **Rating:** Yes/No/N/A
    - 4.c. Citations are easy to read

---

[8] https://aws.amazon.com/bedrock/ (Last Accessed: 1/17/2025)

    - **Rating:** Yes/No/N/A

5. **Specificity**
    - 5.a. Gives information which is specific to a commodity, if appropriate
        - **Rating:** Yes/No
    - 5.b. Gives information which is specific to the location/region in question, where applicable
        - **Rating:** Yes/No
    - 5.c. Where there is no information specific to a location, the system admits this
        - **Rating:** Yes/No
6. **Comprehensiveness**
    - 6.a. The system's response is comprehensive and does not just give a partial, incomplete answer
        - **Rating:** Yes/No
    - 6.b. Shows depth of knowledge/understanding regarding topic
        - **Rating:** Yes/No
    - 6.c. Answers beyond the question's scope to provide context
        - **Rating:** Yes/No
7. **Scientific Accuracy**
    - 7.a. Is the information scientifically robust? Answer to the best of your knowledge
        - **Rating:** Yes/No/N/A
        - **Explanation:**
    - 7.b. Does the response meet scientific expectations? (consider own knowledge or through supported literature)
        - **Rating:** Yes/No/N/A
    - 7.c. Does the response have any errors? Answer to the best of your knowledge
        - **Rating:** Yes/No/N/A

A.5 Questions

We provide the question set that was used for climate science QA generation below:

1. What are the ideal pollination conditions for growing almonds?
2. What can I do to prevent sunburn risk in apples?
3. What varieties of apples are more tolerant to sunburn?
4. What regions will support growing cotton in 2070?
5. How does the climate in South West Western Australia compare from 1970 to now?
6. What will be the greatest climate risk for growing wheat in the wheatbelt in 2050?
7. Will my rainfall continue to increase in variability in Northern NSW?
8. In north-east SA, how many days will I likely experience over 45 degrees?
9. How accurate are climate projections?
10. What is the difference between a heatwave and a hot day?
11. Will we likely see less cold risk days over the lambing season in central Tasmania?
12. How will climate change impact cherry production in Young?
13. What is the production cycle of potatoes?
14. Are there regions in Australia where agriculture will not be viable in 2050?
15. Will commodity distribution in Australia change under a future climate?
16. What are the 6 main climatic drivers of chickpea growth in Australia?
17. What happens to chickpea crops when there is a deficit of soil moisture at the time of sowing?
18. What is the likelihood of the incidence of extreme heat increasing in the future during the reproductive phase of my chickpea crops (grown in Biloela in central Qld)?
19. What is the extreme heat threshold of Kabuli chickpeas and does it differ from Desi?
20. What is the extreme heat threshold of Kabuli chickpeas and does it differ from Desi? Which type of chickpea is more heat-resistant?
21. What is the critical temperature below which chickpea crops are non-viable?
22. Where in Australia might chickpeas become viable in 2040 where they are not viable now?

23. How cool does the temperature have to be, and for how long, to induce flowering in Avocados? Are these periods of cool weather going to become rarer in the Northern Rivers region by 2030?

24. How cool does the temperature have to be, and for how long, to induce flowering in Avocados? Are these periods of cool weather going to become rarer in the Northern Rivers region by 2040?

25. What is the most heat resistent cultivar of avocado available?

26. What are the main climate change risk factors for chickpeas in Australia?

27. What are the main climate change risk factors for mangos in Australia?

28. What are the main climate change risk factors for apples in Australia?

29. What temperature is considered the threshold for increased risk of sunburn in apples?

30. What are the main ways climate change can impact farmers trying to grow avocados?

31. When growing avocados, what temperature is the threshold for a ‘cool day’ during flower induction?

32. What are the main ways climate change will impact farmers trying to grow field tomatoes?

33. What are the main ways climate change will impact farmers trying to grow chickpeas?

34. What are the main ways climate change will impact farmers trying to grow cherries?

35. What is the threshold for a hot day during the flowering season when growing avocados?

36. In the context of growing avocados in Shepparton VIC, what is meant by a pollination event?

37. Are there existing Australian agricultural regions that could become unviable in the future?

38. What is the date range within which flowering occurs in the northern avocado growing region?

39. If a minimum temperature threshold for wheat is crossed, does this always mean there will be a frost event?

40. What does heat stress during the flowering and grain-filling periods do to a wheat crop?

41. How can wine grapes be managed to reduce the risk of frost?

42. What factors influence soil moisture?

43. Will bananas be able to be grown in Northern NSW in 2050?

44. Will sugarcane still be a viable crop for Northern NSW in 2040?

45. Given their chill requirements, will cherries, almonds and apples still be able to be grown in Australia by 2100?

46. What are the most important climate change risks of for cherry production and how are these likely to change by 2050?

47. How will the increased severity and duration of heatwaves affect fresh field tomato production in the coming decades?

48. What will the impact of increased rainfall extremes be on tomatoes, and especially during the time from flowering to harvest?

49. What factors influence climate risk for temperate perennial tree crops?

50. What are the areas in Australia where wine grapes are currently grown, that might struggle to grow wine grapes in 2050?

What are the main ways climate change can impact farmers trying to grow avocados?

Main Ways Climate Change Can Impact Avocado Farmers in Australia

Climate change can significantly impact avocado farmers in several ways, affecting both the suitability of growing regions and the overall productivity of avocado orchards. Here are the main impacts:

1. **Temperature Changes:**
    - **Increasing Minimum Temperatures:** Higher minimum temperatures during the coldest months can positively affect avocado growth by reducing frost damage and improving pollination conditions. However, excessively high temperatures can also cause heat stress, particularly during critical periods such as pollination and fruit set.
    - **Heat Stress:** Avocados are sensitive to high temperatures, which can lead to reduced yields and poor fruit quality. Heat stress can also affect the tree's overall health and resilience.
2. **Precipitation Patterns:**
    - **Drier Conditions:** Reduced rainfall in some regions can lead to water stress, making it necessary for farmers to invest in irrigation systems or other water management practices. Drought conditions can significantly reduce avocado yields.
    - **Wetter Conditions:** Excessive rainfall can increase the risk of diseases such as root rot and fungal infections, which can be detrimental to avocado trees.
3. **Shift in Suitable Growing Regions:**
    - **Expansion of Suitable Areas:** Climate change may make some previously unsuitable areas more favourable for avocado cultivation due to increasing temperatures. For example, regions that were too cold or dry might become suitable with the right management practices.
    - **Contraction of Suitable Areas:** Conversely, some current growing regions may become less suitable due to changes in temperature and precipitation patterns. This could lead to a decrease in highly suitable (S1) areas by up to 41% under the RCP 8.5 scenario by 2050.
4. **Soil and Land Suitability:**
    - **Soil Moisture:** Changes in soil moisture levels due to altered precipitation patterns can affect root health and nutrient uptake, impacting overall tree growth and fruit production.
    - **Land Degradation:** Increased temperatures and changes in rainfall can exacerbate land degradation issues such as salinity and erosion, further challenging avocado cultivation.
5. **Pest and Disease Management:**
    - **New Pests and Diseases:** Climate change can alter the distribution and prevalence of pests and diseases, introducing new challenges for farmers. For example, warmer temperatures can lead to increased pest populations.
    - **Disease Susceptibility:** Changes in temperature and humidity can affect the susceptibility of avocado trees to various diseases, requiring more intensive management practices.
6. **Economic Impacts:**
    - **Increased Costs:** Farmers may face higher costs for irrigation, disease control, and other management practices necessary to adapt to changing conditions.
    - **Market Dynamics:** Changes in global avocado production due to climate change can affect market prices and demand, impacting the economic viability of avocado farming.

Scientific References:

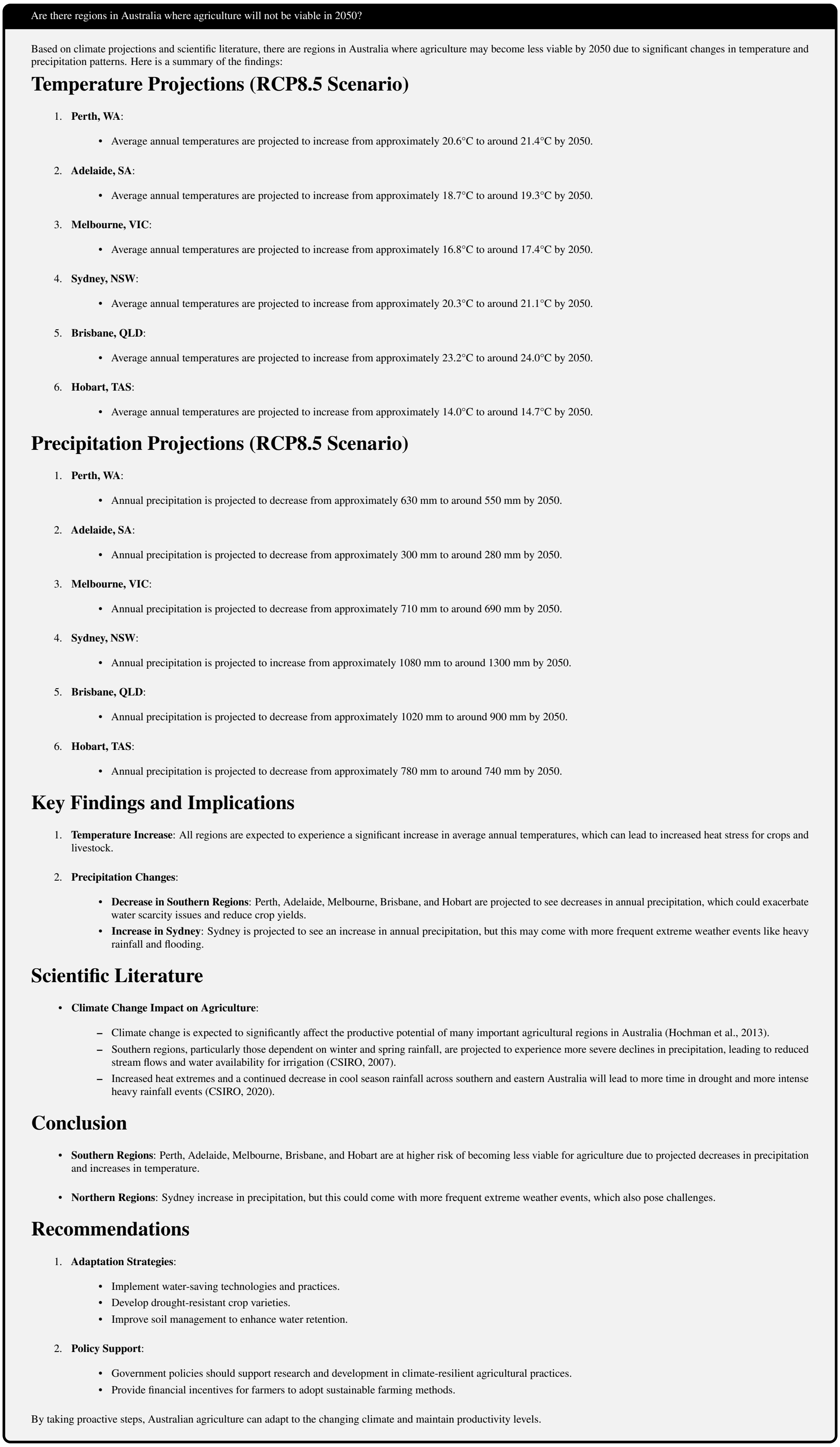
- Grütter, R., Trachsel, T., Laube, P., & Jaisli, I. (2022). Expected global suitability of coffee, cashew and avocado due to climate change. PLoS ONE.

- Melo, O., & Foster, W. (2021). Agricultural and Forestry Land and Labor Use under Long-Term Climate Change in Chile. Atmosphere.

- Chung, S., Rho, H., Lim, C., Jeon, M., Kim, S., Jang, Y., & An, H. (2022). Photosynthetic response and antioxidative activity of 'Hass' avocado cultivar treated with short-term low temperature. Scientific Reports.

These impacts highlight the need for adaptive strategies and management practices to ensure the sustainability of avocado farming in Australia under changing climatic conditions.

Figure 6: An example of a perfect response from Qwen 72b.



Are there regions in Australia where agriculture will not be viable in 2050?

Based on climate projections and scientific literature, there are regions in Australia where agriculture may become less viable by 2050 due to significant changes in temperature and precipitation patterns. Here is a summary of the findings:

Temperature Projections (RCP8.5 Scenario)

1. **Perth, WA**:
    - Average annual temperatures are projected to increase from approximately 20.6°C to around 21.4°C by 2050.
2. **Adelaide, SA**:
    - Average annual temperatures are projected to increase from approximately 18.7°C to around 19.3°C by 2050.
3. **Melbourne, VIC**:
    - Average annual temperatures are projected to increase from approximately 16.8°C to around 17.4°C by 2050.
4. **Sydney, NSW**:
    - Average annual temperatures are projected to increase from approximately 20.3°C to around 21.1°C by 2050.
5. **Brisbane, QLD**:
    - Average annual temperatures are projected to increase from approximately 23.2°C to around 24.0°C by 2050.
6. **Hobart, TAS**:
    - Average annual temperatures are projected to increase from approximately 14.0°C to around 14.7°C by 2050.

Precipitation Projections (RCP8.5 Scenario)

1. **Perth, WA**:
    - Annual precipitation is projected to decrease from approximately 630 mm to around 550 mm by 2050.
2. **Adelaide, SA**:
    - Annual precipitation is projected to decrease from approximately 300 mm to around 280 mm by 2050.
3. **Melbourne, VIC**:
    - Annual precipitation is projected to decrease from approximately 710 mm to around 690 mm by 2050.
4. **Sydney, NSW**:
    - Annual precipitation is projected to increase from approximately 1080 mm to around 1300 mm by 2050.
5. **Brisbane, QLD**:
    - Annual precipitation is projected to decrease from approximately 1020 mm to around 900 mm by 2050.
6. **Hobart, TAS**:
    - Annual precipitation is projected to decrease from approximately 780 mm to around 740 mm by 2050.

Key Findings and Implications

1. **Temperature Increase**: All regions are expected to experience a significant increase in average annual temperatures, which can lead to increased heat stress for crops and livestock.
2. **Precipitation Changes**:
    - **Decrease in Southern Regions**: Perth, Adelaide, Melbourne, Brisbane, and Hobart are projected to see decreases in annual precipitation, which could exacerbate water scarcity issues and reduce crop yields.
    - **Increase in Sydney**: Sydney is projected to see an increase in annual precipitation, but this may come with more frequent extreme weather events like heavy rainfall and flooding.

Scientific Literature

- **Climate Change Impact on Agriculture**:
    - Climate change is expected to significantly affect the productive potential of many important agricultural regions in Australia (Hochman et al., 2013).
    - Southern regions, particularly those dependent on winter and spring rainfall, are projected to experience more severe declines in precipitation, leading to reduced stream flows and water availability for irrigation (CSIRO, 2007).
    - Increased heat extremes and a continued decrease in cool season rainfall across southern and eastern Australia will lead to more time in drought and more intense heavy rainfall events (CSIRO, 2020).

Conclusion

- **Southern Regions**: Perth, Adelaide, Melbourne, Brisbane, and Hobart are at higher risk of becoming less viable for agriculture due to projected decreases in precipitation and increases in temperature.
- **Northern Regions**: Sydney increase in precipitation, but this could come with more frequent extreme weather events, which also pose challenges.

Recommendations

1. **Adaptation Strategies**:
    - Implement water-saving technologies and practices.
    - Develop drought-resistant crop varieties.
    - Improve soil management to enhance water retention.
2. **Policy Support**:
    - Government policies should support research and development in climate-resilient agricultural practices.
    - Provide financial incentives for farmers to adopt sustainable farming methods.

By taking proactive steps, Australian agriculture can adapt to the changing climate and maintain productivity levels.

Figure 7: An example of a poorly structured response from Qwen 72b when GEF is used.

A.6 Toolbox: list of available tools to the model

We detail the climate data access tools (namely tools #1 - #90), as well as the literature search tool (#91). We included a Python tool (#92) that LLMs had access to, however, this was rarely used, if at all.

1. - **Name**: get observations
   - **Description**: Get daily or monthly climate observation data
   - **Parameters**: lon, lat, variable, frequency, startYear, endYear
2. - **Name**: get annual observations
   - **Description**: Get annual climate observation data for a specific location and variable.
   - **Parameters**: lon, lat, variable, start_year, end_year
3. - **Name**: get month range observations
   - **Description**: Get climate observation data for specified month or month range.
   - **Parameters**: lon, lat, variable, month_range, start_year, end_year
4. - **Name**: get seasonal observations
   - **Description**: Get seasonal climate observation data
   - **Parameters**: lon, lat, variable, season, start_year, end_year
5. - **Name**: get date range observations
   - **Description**: Get climate observation data for the specified date range
   - **Parameters**: lon, lat, variable, startYear, endYear, startDate, endDate
6. - **Name**: get month range threshold observations
   - **Description**: Get historical climate threshold data for specified month or month range, e.g., number of January days above 35.
   - **Parameters**: lon, lat, variable, month_range, start_year, end_year, threshold, operator
7. - **Name**: get date range threshold observations
   - **Description**: Get historical climate threshold data for specified date range, e.g., number of days above 35 between March 15 and July 10.
   - **Parameters**: lon, lat, variable, start_year, end_year, start_date, end_date, threshold, operator
8. - **Name**: get annual threshold observations
   - **Description**: Get historical annual climate threshold data, e.g., annual number of days above 35.
   - **Parameters**: lon, lat, variable, start_year, end_year, threshold, operator
9. - **Name**: get seasonal threshold observations
   - **Description**: Get seasonal historical climate threshold data, e.g., number of winter days below 2 degrees.
   - **Parameters**: lon, lat, variable, season, start_year, end_year, threshold, operator
10. - **Name**: get combined threshold observations
    - **Description**: Get historical climate combined threshold data for specified date range, e.g., number of 5 Apr - 25 Aug days tmin < 12 and tmax < 20.
    - **Parameters**: lon, lat, startYear, endYear, startDate, endDate, logicalOp, variable1, threshold1, operator1, variable2, threshold2, operator2
11. - **Name**: get annual combined threshold observations
    - **Description**: Get historical annual climate threshold data for a combination of two variables, e.g., annual number of days where tmin < 12 & tmax < 20.
    - **Parameters**: lon, lat, startYear, endYear, logicalOp, variable1, threshold1, operator1, variable2, threshold2, operator2
12. - **Name**: get month range combined threshold observations
    - **Description**: Get historical climate combined threshold data for specified month or month range, e.g., number of Apr-Aug days tmin < 12 and tmax < 20.

- **Parameters**: lon, lat, monthRange, startYear, endYear, variable1, threshold1, operator1, variable2, threshold2, operator2, logicalOp

13.
- **Name**: get seasonal extreme observations
- **Description**: Get seasonal historical extreme climate data, e.g., hottest day tmax for winter.
- **Parameters**: lon, lat, variable, season, startYear, endYear

14.
- **Name**: get annual extreme observations
- **Description**: Get historical annual extreme climate data (e.g., wettest/hottest/coldest recorded daily values)
- **Parameters**: lon, lat, variable, start_year, end_year

15.
- **Name**: get month range extreme observations
- **Description**: Get historical climate extremes for specified month or month range, e.g., wettest day precip for April to August.
- **Parameters**: lon, lat, variable, month_range, start_year, end_year

16.
- **Name**: get date range extreme observations
- **Description**: Get historical extreme climate data for specified date range, e.g., wettest day rainfall between March 15 and July 10.
- **Parameters**: lon, lat, variable, startYear, endYear, startDate, endDate

17.
- **Name**: get month range ffdi
- **Description**: Get historical FFDI data for selected month or month range.
- **Parameters**: lon, lat, month_range, start_year, end_year, threshold

18.
- **Name**: get ffdi
- **Description**: Get monthly FFDI observation data, e.g., get number of days per month where FFDI > 25
- **Parameters**: lon, lat, frequency, startYear, endYear, threshold

19.
- **Name**: get annual ffdi
- **Description**: Get historical annual FFDI data, e.g., get the number of days per year where FFDI > 25
- **Parameters**: lon, lat, start_year, end_year, threshold

20.
- **Name**: get seasonal ffdi
- **Description**: Get historical seasonal FFDI data, e.g. get the yearly number of summer days where FFDI > 25
- **Parameters**: lon, lat, season, start_year, end_year, threshold

21.
- **Name**: get month range heatwaves
- **Description**: Get historical heatwave data for selected month or month range.
- **Parameters**: lon, lat, monthRange, startYear, endYear, days, tmax, tmin, overlap

22.
- **Name**: get annual heatwaves
- **Description**: Get historical annual heatwave data based on specified parameters.
- **Parameters**: lon, lat, startYear, endYear, days, tmax, tmin, overlap

23.
- **Name**: get date range heatwaves
- **Description**: Get historical heatwave data for selected date range
- **Parameters**: lon, lat, startYear, endYear, startDate, endDate, days, tmax, tmin, overlap

24.
- **Name**: get seasonal heatwaves
- **Description**: Get historical seasonal heatwave data
- **Parameters**: lon, lat, season, startYear, endYear, days, tmax, tmin, overlap

25.
- **Name**: get month range cold exposure observations
- **Description**: Get historical climate cold exposure data for specified month or month range.
- **Parameters**: lon, lat, month_range, start_year, end_year, tmin, precip

26.
- **Name**: get date range cold exposure
- **Description**: Get historical cold exposure data for specified date range.
- **Parameters**: lon, lat, startYear, endYear, startDate, endDate, tmin, precip

27.
- **Name**: get seasonal cold exposure

- **Description**: Get historical seasonal climate cold exposure data
- **Parameters**: lon, lat, season, start_year, end_year, tmin, precip

28.
- **Name**: get annual cold exposure
- **Description**: Get historical annual cold exposure data based on specified parameters.
- **Parameters**: lon, lat, start_year, end_year, tmin, precip

29.
- **Name**: get seasonal soil moisture
- **Description**: Get historical seasonal average soil moisture data
- **Parameters**: lon, lat, season, start_year, end_year

30.
- **Name**: get date range soil moisture
- **Description**: Get historical average soil moisture data for specified date range
- **Parameters**: lon, lat, startYear, endYear, startDate, endDate

31.
- **Name**: get annual soil moisture data
- **Description**: Get historical annual average soil moisture observation data
- **Parameters**: lon, lat, start_year, end_year

32.
- **Name**: get soil moisture
- **Description**: Get daily or monthly soil moisture observation data
- **Parameters**: lon, lat, frequency, startYear, endYear

33.
- **Name**: get month range soil moisture
- **Description**: Get historical average soil moisture data for selected month or month range.
- **Parameters**: lon, lat, month_range, start_year, end_year

34.
- **Name**: get month range evapotranspiration
- **Description**: Get historical potential evapotranspiration data for selected month or month range.
- **Parameters**: lon, lat, month_range, start_year, end_year

35.
- **Name**: get date range evapotranspiration
- **Description**: Get historical potential evapotranspiration data for specified date range
- **Parameters**: lon, lat, startYear, endYear, startDate, endDate

36.
- **Name**: get seasonal evapotranspiration
- **Description**: Get historical seasonal potential evapotranspiration data
- **Parameters**: lon, lat, season, start_year, end_year

37.
- **Name**: get evapotranspiration
- **Description**: Get daily or monthly potential evapotranspiration observation data
- **Parameters**: lon, lat, frequency, startYear, endYear

38.
- **Name**: get annual evapotranspiration data
- **Description**: Get historical annual potential evapotranspiration data
- **Parameters**: lon, lat, start_year, end_year

39.
- **Name**: get annual chill portions
- **Description**: Get historical annual chill portion data
- **Parameters**: lon, lat, variable, startYear, endYear

40.
- **Name**: get growing degree days
- **Description**: Get observed growing degree days for selected date range
- **Parameters**: lon, lat, startYear, endYear, startDate, endDate, baseTemp

41.
- **Name**: get annual growing degree days
- **Description**: Get observed annual growing degree days
- **Parameters**: lon, lat, start_year, end_year, base_temp

42.
- **Name**: get month range growing degree days
- **Description**: Get observed growing degree days for selected month or month range
- **Parameters**: lon, lat, monthRange, startYear, endYear, baseTemp

43.
- **Name**: get seasonal growing degree days

- **Description**: Get observed seasonal growing degree days
- **Parameters**: lon, lat, season, start_year, end_year, base_temp

44.
- **Name**: get seasonal thi data
- **Description**: Get historical seasonal number of days where THI >= threshold
- **Parameters**: lon, lat, index, threshold, season, start_year, end_year

45.
- **Name**: get date range thi data
- **Description**: Get historical number of days where THI >= threshold during specified date range.
- **Parameters**: lon, lat, index, threshold, startYear, endYear, startDate, endDate

46.
- **Name**: get annual thi data
- **Description**: Get historical annual number of days where THI >= threshold
- **Parameters**: lon, lat, index, threshold, start_year, end_year

47.
- **Name**: get month range thi data
- **Description**: Get historical number of days where THI >= threshold during specified month range.
- **Parameters**: lon, lat, index, threshold, month_range, start_year, end_year

48.
- **Name**: get projections
- **Description**: Get climate projections data
- **Parameters**: lon, lat, variable, frequency, years, emission

49.
- **Name**: get month range projections
- **Description**: Get climate projections for selected month or month range.
- **Parameters**: lon, lat, variable, month_range, years, emission

50.
- **Name**: get seasonal projections
- **Description**: Get seasonal climate projections for a given location and parameters.
- **Parameters**: lon, lat, variable, season, years, emission

51.
- **Name**: get annual projections
- **Description**: Get annual climate projections for a given location and variable.
- **Parameters**: lon, lat, variable, years, emission

52.
- **Name**: get date range projections
- **Description**: Get climate projections for selected date range
- **Parameters**: lon, lat, variable, years, emission, start_date, end_date

53.
- **Name**: get date range threshold projections
- **Description**: Get projected climate threshold data for specified date range, e.g., number of days with at least 5mm precip between March 15 and July 10.
- **Parameters**: lon, lat, variable, years, emission, start_date, end_date, threshold, operator

54.
- **Name**: get annual threshold projections
- **Description**: Get annual projected climate threshold data, e.g., annual number of days above 35.
- **Parameters**: lon, lat, variable, years, emission, threshold, operator

55.
- **Name**: get seasonal threshold projections
- **Description**: Get projected climate threshold data for specified season, e.g., number of summer days above 35.
- **Parameters**: lon, lat, variable, season, years, emission, threshold, operator

56.
- **Name**: get month range threshold projections
- **Description**: Get projected climate threshold data for selected month or month range, e.g., number of days below 2 degrees from start of May to end of September.
- **Parameters**: lon, lat, variable, month_range, years, emission, threshold, operator

57.
- **Name**: get combined threshold projections
- **Description**: Get projected climate combined threshold data for selected date range, e.g., number of days with tmin

< 12 and tmax < 20 from 15 Apr to 30 Aug.
- **Parameters**: lon, lat, years, emission, startDate, endDate, logicalOp, variable1, threshold1, operator1, variable2, threshold2, operator2

58.
- **Name**: get month range combined threshold projections
- **Description**: Get projected climate combined threshold data for selected month or month range, e.g., number of days with tmin < 12 and tmax < 20 from Apr to Aug.
- **Parameters**: logicalOp

59.
- **Name**: get annual combined threshold projections
- **Description**: Get annual projected climate combined threshold data for a given location and parameters.
- **Parameters**: lon, lat, years, emission, variable1, threshold1, operator1, variable2, threshold2, operator2

60.
- **Name**: get month range ffdi projections
- **Description**: Get projected FFDI data for selected month or month range.
- **Parameters**: lon, lat, month_range, years, emission, threshold

61.
- **Name**: get ffdi projections
- **Description**: Get monthly FFDI projection data, e.g., get number of days per month where FFDI > 25
- **Parameters**: lon, lat, frequency, years, emission, threshold

62.
- **Name**: get annual ffdi projections
- **Description**: Get projected annual FFDI data, e.g., get the number of days per year where FFDI > 25
- **Parameters**: lon, lat, years, emission, threshold

63.
- **Name**: get seasonal ffdi projections
- **Description**: Get projected seasonal FFDI data, e.g., get the yearly number of summer days where FFDI > 25
- **Parameters**: lon, lat, season, years, emission, threshold

64.
- **Name**: get month range heatwaves projections
- **Description**: Get projected heatwave data for selected month or month range.
- **Parameters**: lon, lat, monthRange, years, emission, days, tmax, tmin, overlap

65.
- **Name**: get annual heatwaves projections
- **Description**: Get projected annual heatwave data based on specified parameters.
- **Parameters**: lon, lat, years, emission, days, tmax, tmin, overlap

66.
- **Name**: get date range heatwaves projections
- **Description**: Get projected heatwave data for selected date range
- **Parameters**: lon, lat, years, emission, startDate, endDate, days, tmax, tmin, overlap

67.
- **Name**: get seasonal heatwaves projections
- **Description**: Get projected seasonal heatwave data
- **Parameters**: lon, lat, season, years, emission, days, tmax, tmin, overlap

68.
- **Name**: get month range cold exposure projections
- **Description**: Get projected climate cold exposure data for specified month or month range.
- **Parameters**: lon, lat, month_range, years, emission, tmin, precip

69.
- **Name**: get date range cold exposure projections
- **Description**: Get projected cold exposure data for specified date range.
- **Parameters**: lon, lat, years, emission, startDate, endDate, tmin, precip

70.
- **Name**: get seasonal cold exposure projections
- **Description**: Get projected seasonal climate cold exposure data
- **Parameters**: lon, lat, season, years, emission, tmin, precip

71.
- **Name**: get annual cold exposure projections

- **Description**: Get projected annual cold exposure data based on specified parameters.
- **Parameters**: lon, lat, years, emission, tmin, precip

72. - **Name**: get seasonal soil moisture projections
    - **Description**: Get projected seasonal average soil moisture data
    - **Parameters**: lon, lat, season, years, emission

73. - **Name**: get date range soil moisture projections
    - **Description**: Get projected average soil moisture data for specified date range
    - **Parameters**: lon, lat, years, emission, startDate, endDate

74. - **Name**: get annual soil moisture projections
    - **Description**: Get projected annual average soil moisture data
    - **Parameters**: lon, lat, years, emission

75. - **Name**: get soil moisture projections
    - **Description**: Get projected daily or monthly soil moisture data
    - **Parameters**: lon, lat, frequency, years, emission

76. - **Name**: get month range soil moisture projections
    - **Description**: Get projected average soil moisture data for selected month or month range.
    - **Parameters**: lon, lat, month_range, years, emission

77. - **Name**: get month range evapotranspiration projections
    - **Description**: Get projected potential evapotranspiration data for selected month or month range.
    - **Parameters**: lon, lat, month_range, years, emission

78. - **Name**: get date range evapotranspiration projections
    - **Description**: Get projected potential evapotranspiration data for specified date range
    - **Parameters**: lon, lat, years, emission, startDate, endDate

79. - **Name**: get seasonal evapotranspiration projections
    - **Description**: Get projected seasonal potential evapotranspiration data
    - **Parameters**: lon, lat, season, years, emission

80. - **Name**: get annual evapotranspiration projections
    - **Description**: Get projected annual potential evapotranspiration data
    - **Parameters**: lon, lat, years, emission

81. - **Name**: get annual chill portion projections
    - **Description**: Get projected annual chill portion data based on provided parameters.
    - **Parameters**: lon, lat, variable, years, emission

82. - **Name**: get annual growing degree days projections
    - **Description**: Get projected annual growing degree days
    - **Parameters**: lon, lat, years, emission, baseTemp

83. - **Name**: get seasonal growing degree days projections
    - **Description**: Get projected seasonal growing degree days
    - **Parameters**: lon, lat, season, years, emission, baseTemp

84. - **Name**: get date range growing degree days projections
    - **Description**: Get projected growing degree days for specified date range
    - **Parameters**: lon, lat, years, emission, startDate, endDate, baseTemp

85. - **Name**: get month range growing degree days projections
    - **Description**: Get projected growing degree days for specified month or month range.
    - **Parameters**: lon, lat, month_range, years, emission, base_temp

86.
- **Name**: get seasonal thi projection data
- **Description**: Get projected seasonal number of days where THI >= threshold
- **Parameters**: lon, lat, index, threshold, season, years, emission

87.
- **Name**: get thi projection data
- **Description**: Get projected number of days where THI >= threshold during specified date range.
- **Parameters**: lon, lat, index, threshold, years, emission, start_date, end_date

88.
- **Name**: get annual thi projection data
- **Description**: Get projected annual number of days where THI >= threshold
- **Parameters**: lon, lat, index, threshold, years, emission

89.
- **Name**: get month range thi projection data
- **Description**: Get projected number of days where THI >= threshold during specified month range.
- **Parameters**: lon, lat, index, threshold, month_range, years, emission

90.
- **Name**: location to coordinates
- **Description**: Get the longitude and latitude geological coordinates of a location in Australia.
- **Parameters**: suburb, state

91.
- **Name**: search literature
- **Description**: Retrieve relevant literature from an Elasticsearch instance containing Australian-specific and non-Australian-specific literature.
- **Parameters**: query, index, size

92.
- **Name**: execute python code
- **Description**: Executes Python code using anaconda3. Data from previous outputs is stored in a cache variable.
- **Parameters**: code

B Climate adaptation question answering prompts

We provide the prompts that we used for our experiments:

B.1 Generation prompts

- Baseline prompt for QA (Figure 8).
- GEF prompt for QA (Figure 10).

B.2 Evaluation prompts

- Baseline prompt for QA evaluation (Figure 9).
- GEF prompt for QA evaluation (Figure 11).

Baseline QA Prompt

You are a knowledgeable and helpful farm advisor for Australian users. Your goal is to provide accurate and relevant advice using your internal knowledge.
Instructions:

1. **Contextual Relevance:** Tailor advice to the specific needs and conditions of Australian farming.
2. **Proactive Assistance:** Additional relevant information or resources might benefit the user.
3. **Single-Turn Clarity:** Ensure your response is straightforward and comprehensive, as there will be no user feedback or clarification opportunity. Do not ask follow-up questions. Instead, state your assumptions and answer the user's question, however vague.

Figure 8: Baseline generation prompt for Climate Science QA.

Baseline Evaluation Prompt

You are a farm advisor tasked with evaluating the quality of responses to questions based on a specific set of criteria. Below is a rubric that you should use to assess the responses. For each criterion, provide a rating (Yes/No/N/A) and a brief explanation for your rating.
Criteria:

`<Annotator Criteria>`

Please return the evaluation in the following JSON format:

`<JSON format>`

Figure 9: Baseline prompt for climate science QA evaluation.

GEF QA Generation Prompt

You are a knowledgeable and helpful farm advisor for Australian users. Your goal is to provide accurate and relevant advice using the supplied tools, including climate historical and projection information, Python for calculations, and scientific literature for credible references.
Instructions:

1. **User Location:** If the user provides a region instead of a specific location, choose a location that best represents that region.
2. **Tool Usage:**
   - **Climate Data:** Utilize historical climate data and projections to provide contextually relevant advice.
   - **Python:** This is used to interact with climate data and verify mathematical calculations.
   - **Scientific Literature:** Reference credible scientific literature to ensure accuracy.
3. **Climate Information:** Utilize historical climate data and projections to provide contextually relevant advice.
4. **Mathematical Accuracy:** Verify all calculations using the Python tool.
5. **Scientific Literature:**
   - Reference credible scientific literature to ensure accuracy.
   - Verify all answers against scientific literature to ensure accuracy and credibility.
   - Focus on including only the most relevant literature in your answers to keep them clear and helpful.
   - Include scientific references used in responses in a way that does not impede reading.
6. **Contextual Relevance:** Tailor advice to the specific needs and conditions of Australian farming using climate data and projections.
7. **Proactive Assistance:** Additional relevant information or resources might benefit the user.
8. **Tool Utilization:** Run the necessary tools before responding to the user's query.
9. **Single-Turn Clarity:** Ensure your response is straightforward and comprehensive, as there will be no user feedback or clarification opportunity. Do not ask follow-up questions. Instead, state your assumptions and answer the user's question, however vague.

JSON Schema for Climate Data
The climate data is also stored in the CACHE variable and is accessible via the Python tool and the index provided. The schema of the JSON data is as follows:

```
{
  "data": [
    {
      "yearRange": "string",
      "model": "List[string]",
      "rcp": "string",
      "averageYearValue": [
        {
          "year": "integer",
          "averageValue": "float"
        }
      ]
    }
  ]
}
```

Explanation:

- **data:** An array of objects, each representing a dataset.
  - **yearRange:** A string indicating the range of years (e.g., "2056-2085").
  - **model:** A list of strings representing the models used (e.g., "ACCESS1-0").
  - **rcp:** A string indicating the Representative Concentration Pathway (e.g., "rcp85").
  - **yearValue:** An array of objects, each containing:
    - **year:** An integer representing the year.
    - **value:** A float representing the value for that year.

Figure 10: GEF prompt for Climate Science Question Answering. We did not explicitly include tool descriptions as these were handled by the official libraries from OpenAPI (GPT-4o and Qwen) or Amazon Bedrock (Claude).

GEF Evaluation Prompt

You are a farm advisor tasked with evaluating the quality of responses to climate adaptation questions based on a specific set of criteria. Below is a rubric that you should use to assess the responses. For each criterion, provide a rating (Yes/No/N/A) and a brief explanation for your rating. You must use the provided tools to help assess the validity of scientific claims, climate projections and observations in the responses. The tools provided include climate and projection information, Python for calculations and scientific literature search.

JSON Schema for Climate Data

The climate data is also stored in the CACHE variable and is accessible via the Python tool and the index provided. The schema of the JSON data is as follows:

```
{
  "data": [
    {
      "yearRange": "string",
      "model": "List[string]",
      "rcp": "string",
      "averageYearValue": [
        {
          "year": "integer",
          "averageValue": "float"
        }
      ]
    }
  ]
}
```

Explanation:

- **data:** An array of objects, each representing a dataset.
  - **yearRange:** A string indicating the range of years (e.g., "2056-2085").
  - **model:** A list of strings representing the models used (e.g., "ACCESS1-0").
  - **rcp:** A string indicating the Representative Concentration Pathway (e.g., "rcp85").
  - **yearValue:** An array of objects, each containing:
    - **year:** An integer representing the year.
    - **value:** A float representing the value for that year.

Criteria:

`<Annotator Criteria>`

Here are a few examples of climate scientists' judgments, ensure you use the same format:

```
###Question
What are the main ways climate change will impact farmers trying to grow chickpeas?

###Response
<Generated Response>

###Judgment
{"context_a": 0, "context_b": 0, "context_c": 1,
"structure_a": 1,  "structure_b": 1, "structure_c": 1,
"language_a": 1, "language_b": 1, "language_c": 1, ... }

###Feedback
<When using ExF, feedback is added here>

<4 additional few-shot examples>
```

Please return the evaluation in the following JSON format:

`<JSON format>`

Figure 11: The GEF prompt for climate science answer QA evaluation. Few-shot examples (Question, Response, Judgment) were included in few-shot runs, but not for zero-shot. For few-shot runs, expert feedback is given only for ExF runs. We did not explicitly include tool descriptions as these were handled by the official libraries from OpenAPI (GPT-4o and Qwen) or Amazon Bedrock (Claude).

	Context		**Structure**		**Language**		**Compreh-ensiveness**		**Specificity**		**Citations**		**Accuracy**		**AVG**	
	τ	rbo	τ	rbo	τ	rbo	τ	rbo	τ	rbo	τ	rbo	τ	rbo	τ	rbo
Expert	**0.217**	**0.438**	0.192	**0.926**	**0.303**	**0.910**	0.307	0.165	0.063	0.164	-0.054	0.319	-0.114	0.087	0.131	**0.430**
								Zero-shot								
Qwen 72b	0.126	0.344	0.163	0.547	-	-	0.125	0.057	0.122	0.051	0.511	0.045	0.080	0.013	0.193	0.235
+GEF	0.114	0.291	0.240	0.712	-	-	0.073	0.021	0.153	0.072	0.513	0.046	0.080	0.013	0.205	0.312
GPT-4o	0.043	0.324	0.323	0.706	-	-	0.195	0.013	0.106	0.085	0.525	0.047	0.094	0.012	0.217	0.268
+GEF	0.100	0.298	0.367	0.708	-	-	0.128	0.013	0.125	0.012	0.536	0.046	0.025	0.013	0.215	0.246
Sonnet 3.5	0.178	0.290	0.352	0.713	-	-	0.172	0.014	0.063	0.001	0.513	0.053	0.118	0.094	0.235	0.247
+GEF	0.113	0.178	0.370	0.713	-	-	0.237	0.017	0.082	0.002	0.460	0.053	0.052	0.053	0.224	0.249
								Few-shot								
Qwen 72b	0.066	0.325	0.099	0.716	-	-	0.052	0.114	0.155	0.197	0.512	0.161	0.095	0.014	0.163	0.254
+GEF	0.107	0.320	0.188	0.690	-	-	-0.030	0.123	0.152	0.326	0.523	0.221	0.025	0.013	0.161	0.282
+ExF	0.094	0.341	0.132	0.718	-	-	0.111	**0.213**	0.208	0.212	0.518	0.169	0.013	0.011	0.180	0.277
+GEF+ExF	0.123	0.323	0.215	0.690	-	-	0.038	0.184	0.185	**0.500**	0.534	0.213	0.044	0.034	0.190	0.324
GPT-4o	0.040	0.355	0.126	0.499	-	-	0.183	0.099	0.209	0.092	**0.563**	0.169	-0.003	0.012	0.186	0.204
+GEF	0.077	0.326	0.286	0.702	-	-	0.161	0.094	0.184	0.310	0.536	0.164	0.061	0.012	0.217	0.268
+ExF	0.022	0.359	0.154	0.606	-	-	0.203	0.113	**0.229**	0.104	0.527	0.216	-0.021	0.014	0.186	0.235
+GEF+ExF	0.062	0.331	0.274	0.702	-	-	0.179	0.094	**0.229**	0.456	0.536	0.175	0.055	0.011	0.223	0.295
Sonnet 3.5	0.092	0.108	0.374	0.716	-	-	0.252	0.109	0.169	0.208	0.535	0.235	0.148	0.037	0.262	0.235
+GEF	0.057	0.196	0.386	0.716	-	-	0.276	0.108	0.197	0.202	0.509	0.179	**0.191**	0.086	0.269	0.248
+ExF	0.110	0.048	**0.395**	0.716	-	-	**0.287**	0.119	0.157	0.269	0.551	**0.424**	0.084	0.021	0.264	0.266
+GEF+ExF	0.121	0.218	0.364	0.716	-	-	0.252	0.129	0.185	0.199	0.539	0.249	0.181	**0.128**	**0.274**	0.273

Table 7: Comparison between expert and LLM annotators using Kendall's Tau and rank-biased overlap. Language field could not be measured as each method predicted perfect language for every prediction.

	Context	**Structure**	**Compre-hensiveness**	**Specificity**	**Citations**	**Accuracy**	**AVG**
GPT-4o	1.743	2.317	0.018	1.957	2.148	1.913	1.683
+GEF	2.312 ↑	2.747 ↑	1.773 ↑	1.998 ↑	2.385 ↑	1.925 ↑	2.190 ↑
Qwen 72b	1.977	2.687	0.077	2.030	2.405	1.940	1.853
+GEF	**2.402** ↑	2.753 ↑	2.423 ↑	2.078 ↑	**2.542** ↑	1.943 ↑	2.357 ↑
Sonnet 3.5	1.702	2.647	0.022	2.362	2.550	1.975	1.876
+GEF	2.253 ↑	**2.887** ↑	**1.960** ↑	**2.435** ↑	2.795 ↑	**2.093** ↑	**2.404** ↑

Table 8: Climate-Augmented Question Answering evaluation using LLM annotators. Table 4 gives the mean scores of all methods. Overall, the LLMs gave higher scores to responses and universally agreed that using climate data and literature improves the quality of responses, with large increases in both presentation and epistemological quality; this means that LLMs could be fooled by the inclusion of climate data and climate literature regardless of its correctness. Furthermore, a baseline model has beaten no method with GEF. A perfect score would be 3.0 for a given criteria. ↑ indicates improvement over the baseline while ↓ indicates a reduction.