My Climate CoPilot:
A Question Answering System for Climate Adaptation in Agriculture

Vincent Nguyen[1] and **Willow Hallgren**[2] and **Ashley Harkin**[3]
Mahesh Prakash[1] and **Sarvnaz Karimi**[1]
[1]CSIRO, Data61, Australia
[2]CSIRO, Agriculture and Food, Australia
{firstname.lastname}@csiro.au
[3]Bureau of Meteorology, Australia
{ashley.harkin}@bom.gov.au

Abstract

Accurately answering climate science questions requires scientific literature and climate data. Interpreting climate literature and data, however, presents inherent challenges such as determining relevant climate factors and drivers, interpreting uncertainties in the science and data, and dealing with the sheer volume of data. MY CLIMATE COPILOT is a platform that assists a range of potential users, such as farmer advisors, to mitigate and adapt to projected climate changes by providing answers to questions that are grounded in evidence. It emphasises transparency, user privacy, low-resource use, and provides automatic evaluation. It also strives for scientific robustness and accountability. Fifty domain experts carefully evaluated every aspect of MY CLIMATE COPILOT and based on their interactions and feedback, the system continues to evolve.

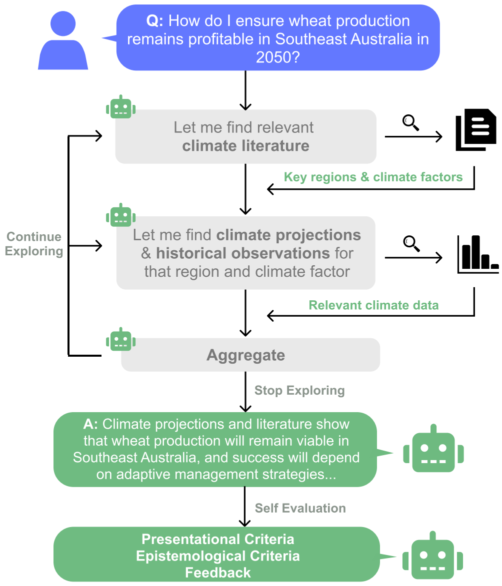

Figure 1: MY CLIMATE COPILOT retrieves, filters, and combines relevant climate literature and climate data to answer climate expert questions for climate adaptation.

1 Introduction

Contemporary information-seeking and knowledge discovery in climate science requires access to and understanding of an increasing amount of climate data (Sansom et al., 2021; Jagannathan et al., 2023) and scientific literature (De La Calzada et al., 2024; Lemos and Rood, 2010). Given the pressing concern of climate change, systems that cater to the needs of individuals dealing with climate risk, such as farm advisors, become increasingly important. *Climate adaptation*—a sub-domain in climate science that aims to safeguard against projected climate impacts for people and ecosystems (Runhaar et al., 2018; Lee et al., 2023)—is our focus.

On the user side, we consider climate adaptation experts who advise farmers (e.g., agronomists), who seek information on adaptation practices relevant to a specific commodity and location. Their clients, farmers, need this information to adapt to future climate impacts to maintain financial and food security, leading to an improvement in their climate resilience. However, even for experts, finding this information is time-consuming. This is due to the particular challenges in using climate science for adaptation purposes, such as the uncertainties in climate change projections, the need for locally relevant climate information, and the sheer scale of the data—the amount of literature doubles every 8 years (Haunschild et al., 2016; Khojasteh et al., 2024) and terabytes of climate data created every 5 years (World Climate Research Programme, 2025).

To assist with information seeking for climate adaptation, we design MY CLIMATE COPILOT (MYCC), an LLM-based question answering system (Figure 1). MYCC answers agriculturally-relevant climate impact and adaptation questions by exploring relevant climate data and climate literature while providing the users with intermediary reasoning traces, as well as all the data found at that point for transparency. It also provides self-evaluation using criteria developed by experts to

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 62–70
July 27 - August 1, 2025

assist users in evaluating the response. Overall, the main features of our system are:

Expert-guided MyCC is continually evolving based on consultation with domain experts and evaluation studies.

Accessible Users can engage in multi-turn conversations to facilitate complicated requests or clarify important concepts. Self-evaluation with expert criteria also provides less climate-literate individuals with context to judge responses.

Transparent MyCC is designed to be highly transparent. All planning, data, or tools used by the model are clearly shown to the end user.

Privacy Preserving To maintain data privacy, we use private APIs when interacting with proprietary models and only collect data submitted by the user.

2 Related Work

Some climate-related models in the NLP field use Transformer-based (Vaswani et al., 2017) or encoder-based (Devlin et al., 2019) models. For example, ClimateBERT (Bingler et al., 2022) is pre-trained with climate news, research abstracts and climate reports; or CliMedBERT (Jalalzadeh Fard et al., 2022) proposes pre-training on climate science literature (Berrang-Ford et al., 2021), Intergovernmental Panel on Climate Change (IPCC) reports and climate policy documents.

Other systems utilise Large Language Models (LLMs). ChatClimate (Vaghefi et al., 2023) answers general climate change questions using information from IPCC reports via retrieval augmented generation (RAG) or direct prompting (internal LLM knowledge). Similarly, ChatNetZero (Hsu et al., 2024) answers questions relating to broad net-zero domain knowledge such as terminology to articulate net-zero commitments using RAG over an expert-curated corpus. ClimatePolicyRadar (Juhasz et al., 2024) answers questions about individual climate law and policy documents while providing insights into website design. ClimSight (Koldunov and Jung, 2024) provides insights for agriculture, urban planning, disaster management, and policy development using a combination of RAG from climate literature and climate data based on provided coordinates.

ClimSight is the closest to our work; however, it is not suited for an expert audience as it focuses on multiple objectives, is limited to one location, a single conversation turn, and uses one climate projection model. In retrospect, our system is highly specialised and designed for experts by experts. It helps them find information from specialised corpora and climate data, allowing them to provide management advice for climate adaptation needs from multiple locations and multiple climate projection models while providing a multi-turn interface for follow-up questions.

3 Resources and Datasets

Climate Data Climate data often includes *observations* and *projections*. Historical observations are generally sourced from national databases. Climate projections, on the other hand, are sourced from large-scale studies. The Coupled Model Intercomparison Project (CMIP) (World Climate Research Programme, 2025) provides future climate projections on a global scale. My Climate CoPilot includes both these data sources by integration with My Climate View (Webb et al., 2023), a platform that provides projections and observations for a given location and commodity from downscaled CMIP 5 (Taylor et al., 2012) and national observational data for Australia. From the My Climate View API, we create tools that wrap API access to 89 different climates.

Climate Literature Our system combines the data from My Climate View APIs and provides climate adaptation advice for climate expert questions. Such advice must be relevant to regional or commodity climate factors and needs to be up-to-date. To meet these criteria, we develop a literature corpus with two levels of granularity: (1) international literature, which encompasses the agriculture and general climate literature from across the globe; and, (2) regional literature, collected from expert-gathered grey literature, industry reports, and climate indices derived from scientific research. We store these corpora in a hybrid retrieval index that combines an inverted index with a vector database. Following our 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.

International Literature It is filtered from the S2ORC corpus and the top journals from Elsevier. For S2ORC, we filter 2.36M documents from 7.3M based on the document's 'fields of study' facet

(Agricultural and Food Sciences and Environmental Science). From Elsevier, we collect the top 100 agriculture journals ranked by impact factor, totalling 246k documents. We then remove documents missing the following facets: title, abstract, DOI, or body text, leading to 144k documents.

Regional Literature Early feedback from climate scientists (Nguyen et al., 2024) indicates that international literature is often irrelevant when answering questions related to Australia. Therefore, climate experts curated 29 grey literature articles that are highly specific to key growing locations and their respective commodities and climate factors within Australia. This regional literature can be used to tailor responses to the regional context of the question.

4 My Climate CoPilot

Our system, My Climate CoPilot, is an evolving dialogue-based platform that provides evidence-grounded answers to questions by climate or agriculture experts on climate adaptation management.

A typical question-answering dialogue with My Climate CoPilot involves five steps: (1) iterative planning; (2) dynamic tool selection and data exploration; (3) response generation; (4) self-evaluation; and, (5) multi-turn user feedback or edits. Our evaluation studies with experts (Nguyen et al., 2025) showed the importance of transparency. That is, experts want to see all the data that goes into the LLM and the processes behind the scenes. As such, all of the steps are shown to the user.

Iterative planning MyCC was originally designed as a RAG system with query rewriting (Nguyen et al., 2024), however, since questions in climate science involve multi-step reasoning over heterogeneous sources, we moved to an agentic framework. It uses an LLM to determine the user's intent and what climate APIs or climate literature are needed. In a traditional agent framework, the planning stage typically creates a single task plan illustrating actions and tools needed to complete the user's request at the beginning. In the climate adaptation domain, this approach would not work because relevant parameters such as climate factors, growing regions, and commodities might not be known before searching the literature and are required to interact with the climate data endpoints. For example, if a user asks *"What can I grow in South Western Australia in 2050?"*, the LLM agent must: (1) determine the coordinates of the location of interest; (2) find relevant climate factors and drivers from the literature; (3) use this information to filter climate data relevant to specific commodities and growing regions; and, (4) search the literature again based on trends in the climate data.

Overall, this means that the correct tools cannot be known beforehand or planned in a single step, and therefore, planning should be continual and influenced by the past trajectory of choices.

Dynamic tool selection and data exploration Many climate adaptation questions require resolving the precise spatial coordinates of a location and climate factors to obtain tabular climate data (Jagannathan et al., 2023). Furthermore, given the size of the climate data and climate literature (terabytes), it is not feasible to explore them in their entirety for a given user request. We, therefore, formulate the task of climate data and climate literature exploration as follows.

Given a user question, q, we use an LLM agent to select the appropriate tool $c \in C^d$, where d represents the cardinality of tools and generates its reasoning, $r \in S^{V_d}$, where V_d is the vocabulary size of the LLM, for selecting the tool c. This is a multi-step process, such that at each time step t, the current tool selection and reasoning are influenced by the past trajectory and conversation history such that $(c^t, r^t) = LLM((c^1, r^1), ..., (c^{t-1}, r^{t-1}))$. This process continues until the LLM agent decides to terminate exploration and collate an answer.

Self-evaluation Our prior work (Nguyen et al., 2025), using few-shot and human feedback alignment, found that LLMs could match expert-level performance for climate science response evaluation. After response generation, we prompt the LLM, in a new and separate conversation, to do an evaluation of the response across seven *presentational* and *epistemological* dimensions (Bulian et al., 2024) designed by experts (Nguyen et al., 2024). Each dimension has a checklist of three sub-criteria, which, when summed, can be used as the overall score.

1. **Context**
   - 1.a Attempts to give some broader context to explain the issue
   - 1.b Provides an introductory paragraph to introduce the topic

1.c Provides a summary paragraph at the end

2. **Structure**

   2.a Overall response is well structured, easy to read

   2.b Headings and subheadings are well structured and logical, and with appropriate categories

   2.c Dot points are used appropriately

3. **Use of Language**

   3.a Phrasing is appropriate (easy to read, fluent) and not awkward or incorrect

   3.b Correct use of grammar

   3.c Consistent with language used within the industry

4. **Use of Citations (where used)**

   4.a Citations are used appropriately

   4.b The number of citations used is appropriate

   4.c Citations are easy to read

5. **Specificity**

   5.a Gives information which is specific to a commodity, if appropriate

   5.b Gives information which is specific to the location/region in question, where applicable

   5.c Where there is no information specific to a location, the system admits this

6. **Comprehensiveness**

   6.a The system's response is comprehensive and does not just give a partial, incomplete answer

   6.b Shows depth of knowledge or understanding regarding the topic

   6.c Answers beyond the question's scope to provide context

7. **Scientific Accuracy**

   7.a Is the information scientifically robust? Answer to the best of your knowledge

   7.b Does the response meet scientific expectations? (consider own knowledge or through supported literature)

   7.c Does the response have any errors? Answer to the best of your knowledge

User feedback and edits To enhance MYCC through user feedback, we collect the feedback post-generation in two forms: (1) user preference; and, (2) edits. *User preference* comes in three categories: positive, neutral, and negative. Aside from assessing expert sentiment, the positive and negative categories are used as signals for human alignment via reinforcement learning (Ouyang et al., 2022) for preliminary testing with open-source models. Experts can *edit the responses* to suit their needs; these edits are collected for supervised tuning of downstream open-source models.

5 Implementation Details

5.1 MY CLIMATE COPILOT Development

The MY CLIMATE COPILOT is composed of three layers with the Rust programming language: (1) a self-made backend library that handles prompting and interacting with LLMs; (2) a middleware server that handles communication between the client and various APIs such as the My Climate View APIs, Elasticsearch, and the Python interpreter; and, (3) the frontend website (web assembly) or application (rust native) which experts can ask their climate adaptation questions on (see Figure 2).

5.2 List of available tools to the model

For climate data access tools, we created a tool that correlates to one of the 89 endpoints on My Climate View[1]. For search tools, we created a custom scorer that allows access to the Elasticsearch instance by inputting a query, corpus of interest, and number of documents to search.

5.3 Hybrid Index Implementation

Our hybrid index was implemented with Elasticsearch[2], which allowed the construction of an inverted index and a vector store for hybrid scoring. To create the embeddings for the vector database, we experimented with a variety of embedding models and evaluated them against a set of human judgments produced by experts. The judgments were created for 15 questions, with 20 documents per query, where the documents were retrieved using a hybrid scorer with BM25 (Robertson et al., 1995) and JinaBERT (Günther et al., 2023) from our past work (Nguyen et al., 2024). Two climate experts annotated the document-query pairs for relevance.

[1]`https://dev.indraweb.io/`

[2]`https://www.elastic.co/elasticsearch` Last Accessed: 1/12/2025

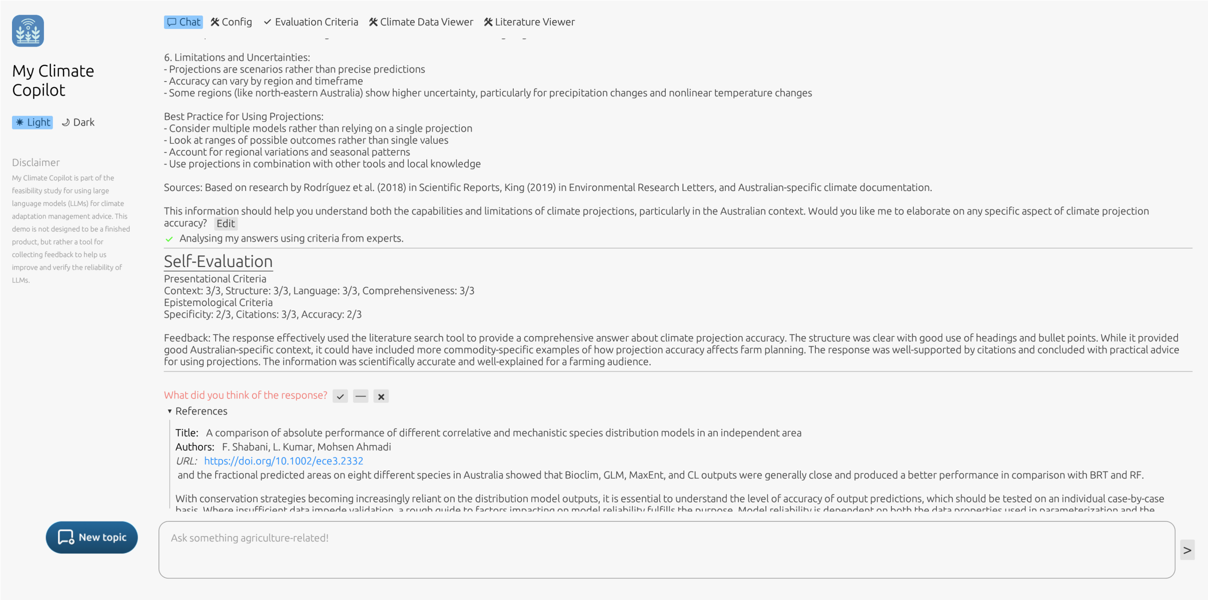

Figure 2: Frontend of the MY CLIMATE COPILOT system. The interface is designed to be transparent, containing references to literature and tabs to access the raw climate data and literature in more detail. Responses from MY CLIMATE COPILOT include a self-evaluation step to help users critique responses.

Embedding Model Selection Human judgments from the previous step were used to empirically evaluate the top models from the MTEB benchmark (Muennighoff et al., 2023) (See Table 1)). Using nDCG10 (Craswell, 2009) as the primary metric, `Stella 1.5b v5`[3] scored the highest and was chosen as our embedding model.

Model ID	**nDCG@10**
NovaSearch/stella_en_1.5B_v5	**0.769**
Alibaba-NLP/gte-Qwen2-7B-instruct	0.763
Salesforce/SFR-Embedding-2_R	0.756
NovaSearch/stella_en_400M_v5	0.700
jinaai/jina-embeddings-v2-base-en	0.662
nvidia/NV-Embed-v2	0.403

Table 1: Embedding model selection. 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`.

Hybrid Index Documents from the climate literature corpora were chunked to 512 tokens. The chunks and their embeddings were indexed in the inverted index with the following metadata: title, authors, DOI, journal/venue, and year. At indexing time, no prompt was used to create the chunked document embeddings; however, at query time, the following was used: `Instruct: Given a web search query, retrieve relevant passages that answer the query. Query: {query}`.

5.4 Backbone LLM Selection

During the development of MYCC, we trialled and evaluated several proprietary and open-source models (Nguyen et al., 2024). Our latest study showed that Claude Sonnet 3.5 had strong generation capabilities for our application (Table 2 for evaluation and Figure 4 for tool use breakdown).

5.5 User Evaluation

While MYCC is not yet publicly available, we recruited experts who helped to critically evaluate and test the system. These experts—agronomists and climate scientists—volunteered from the umbrella research program where *My Climate View* is developed. To date, we have tested MYCC with over 50 different domain experts, which has been, in turn, used to improve the overall system.

For our latest testing phase, we held a one-hour introductory session to provide context and guidance on using the system. After testing the systems, we interviewed two of the experts for one hour about their experiences with MYCC and feedback.

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

Task	Climate Adaptation QA Avg. Score (↑)	Self-Evaluation τ (↑)
Qwen 72b	1.788	0.205
GPT-4o	1.745	0.223
Sonnet 3.5	**1.975**	**0.274**

Table 2: Evaluation of the question answering (QA) and self-evaluation capabilities of various open-source and proprietary LLMs. QA is evaluated by experts using the seven criteria for presentational and epistemological dimensions (Nguyen et al., 2024) and reported here as an average. Self-evaluation is calculated using Kendall's Tau (Kendall, 1938) against expert evaluation, which measures the similarity between the annotation sets.

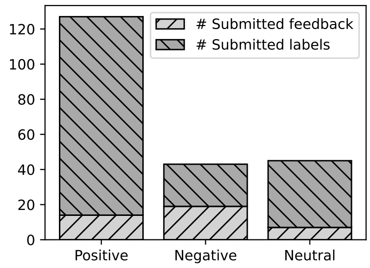

Figure 3: Number of submitted preferences and feedback from climate experts.

6 Evaluation and Analysis

Self reported ratings Climate experts submitted ratings for MyCC (Figure 3), which is used to gauge the sentiment of the expert towards the response. Overall, the experts were positive towards MyCC (64% of all labels, or 82% of positive & negative labels), the neutral category, and negative labels had similar counts. We can safely assume that the baseline capabilities of the systems are reasonable, but there is further room for improvement.

Qualitative feedback Although many of the self-reported labels were positive, the users typically provided optional feedback only for negative sentiment (Figure 3); a similar finding is reported in the financial domain (Colmekcioglu et al., 2022). Positive feedback appreciated the accuracy of responses was high. However, they also highlighted problems such as a lack of relevance to location (14%); these are cases where the system retrieved and used international literature for region-specific questions or minor presentation details (21%) such as citation format or summary location.

The negative feedback from experts emphasised similar points, such as presentational characteristics (52%), awkward answer phrasing (41%) or the referencing style (11%) (some experts did not like explicit references within the text), and regional relevance (15%) (Australia mainly has pasture-based dairy, whereas the US or EU have housed dairy which the response assumed). For the neutral label, all written feedback focused on answer phrasing (earlier iterations) and presentation.

We incorporated the feedback into My Climate CoPilot by creating a location disambiguation tool that converts location names to coordinates and tools for the LLM to select specific literature corpora using a query, corpora of interest, and the number of documents to retrieve.

Interviews Self-reported labels and qualitative feedback can be limited in understanding the views of the experts. Therefore, we interviewed two experts, each for an hour, after they used MyCC for two weeks. The experts recalled that answers were comprehensive and had highly relevant information at times for questions that were well-structured, such as *"Can you propose heat-tolerant hop varieties that might be used as an adaptation strategy for climate change? What trade-offs might be necessary, such as quality or yield?"* However, questions that were more general or applied to multiple regions such as *"What parts of Australia might become less suitable for wool growing over the next 50 years? What would be the main reasons for any change?"* received answers that were too generic, because the data to answer this question was not readily available. Otherwise, the system could sometimes infer additional context beyond the question; although, in many cases, the additional context missed was irrelevant or incomplete.

6.1 Common Question Themes

To get a sense of the information needs of experts, we analysed the types of questions (with percentage) in the 2180 question-answer pairs:

Agricultural Practices (34%) Questions about best practices for growing specific crops under changing climate conditions (e.g., "How do I grow the best quality avocados?", "What are the ideal pollination conditions for growing apples?").

Climate Change Impact (28%) Questions that were about climate factors such as temperature changes, rainfall patterns, and extreme weather events, and their effects on agriculture. For example, *"How will climate change affect drought*

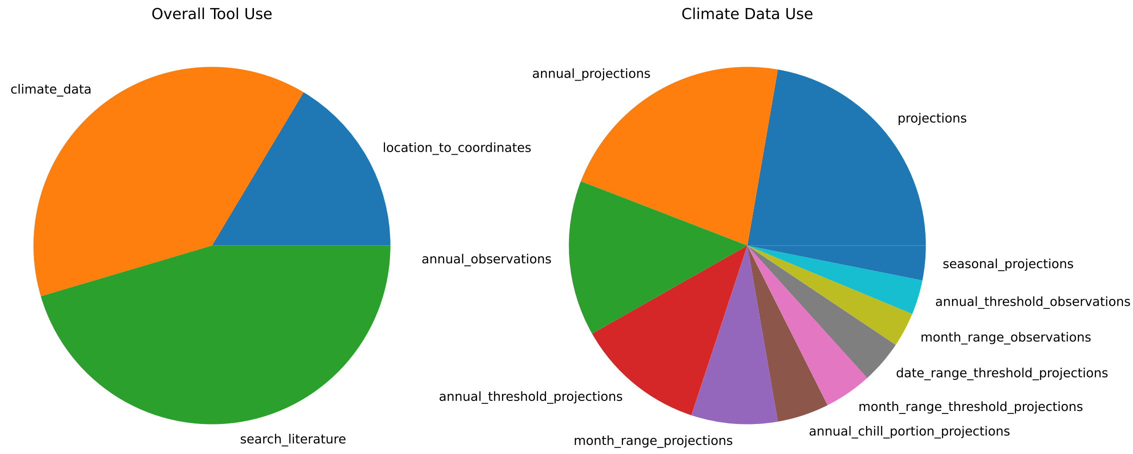

Figure 4: Overall tool distribution (left) and climate data tool use distribution (right) by Claude Sonnet 3.5 during our latest evaluation testing with experts (i.e., the results from Table 2).

occurrence and how will that impact food security in Australia in 2040?".

Crop-Specific Concerns (15%) Questions focused on the impact of climate change on specific crops and how to mitigate those impacts (e.g., "How will heat days impact wine production in 2050?", "What does heat stress during the flowering and grain-filling periods do to a wheat crop?").

Regional Climate Projections (10%) Questions about climate projections for specific regions and their implications for agriculture (e.g., "What is the climate forecast for Melbourne in 30 years?", "How will the weather in Fairfield, NSW change in 2050? What does this mean for crops?").

Adaptation Strategies (8%) Questions that sought advice on how to adapt agricultural practices to cope with climate change (e.g., "What strategies can I use to manage soil moisture at sowing in wheat?", "How can I prepare my dairy for a warmer climate in Tatura?").

Climate Data (5%) Questions focused on understanding and interpreting climate data and projections (e.g., "What is potential evapotranspiration?", "How confident are climate projections?").

7 Future Developments

My Climate CoPilot is continually improving with expert feedback from systematic and formal user studies (Nguyen et al., 2024, 2025). Our evaluation with experts showed that presentational characteristics are highly valued when it comes to question answering. Therefore, we plan to fine-tune open-source models with the data we have collected from the platform to ensure that they align with expert preference but also remain scientifically robust. Another way is to improve provenance by adapting entity linking techniques such as REAL (Shlyk et al., 2024), to link references to where they are used within the generated text and improve transparency. While many generated answers were highly specific and expert-aligned, there were cases where they were too generic when there was insufficient data. That is, the system did not find the correct literature to answer the question or the information did not exist in our corpora. We will further augment the retrieval system with specialised literature, with further developments aiming to support general questions that span globally by integrating with data from CMIP.

Furthermore, although our previous studies and expert guidance led us to the implementation of self-evaluation, we have yet to assess the impact of this. We plan to assess expert reception and feedback in a future study.

8 Conclusions

We present My Climate CoPilot, a question-answering system that helps users improve their knowledge of climate change impacts and adaptation in the Australian agricultural sector. Our system helps users find relevant climate data and literature for their climate adaptation needs and provides management advice to reduce climate risk. My Climate CoPilot is transparent, privacy-aware, extensible, and continually evolving under expert guidance.

Limitations

One limitation of MyCC is that questions that require climate data aggregation from multiple locations (e.g., a question asking about climate factors across Australia), may be difficult to answer given the limited context windows of models. A comprehensive evaluation for this is planned and will require expert guidance to validate these difficult questions.

Another limitation is that our current system is specialised for Australian agriculture and climate adaptation by design. However, we plan to support general questions globally by using the international climate literature we have collected and integrating data from CMIP. This was out of scope for our current study, as evaluation for this will be significantly more challenging given the scale and climate variations between countries, which will require international experts.

Ethics Statement

The authors obtained ethics approval from CSIRO for all user studies and annotations, including expert evaluation and collection of expert feedback (Climate Services for Agriculture (Phase 2) (131/24)).

All LLM endpoints used in evaluation or as part of the system have an enterprise agreement, under which any data exchanged remains private and is not retained.

Acknowledgements

Dr Willow Hallgren has a PhD in Atmospheric Science and has worked in the field of Climate Change and Adaptation in Agriculture. Ashley Harkin is an agronomist and farmer advisor. They both were instrumental in providing domain expertise. We would like to sincerely thank all the members of the Climate Services for Agriculture (CSA) project team who contributed their knowledge and expertise in developing and evaluating this system.

References

Lea Berrang-Ford, Anne J Sietsma, Max Callaghan, Jan C Minx, Pauline FD Scheelbeek, Neal R Haddaway, Andy Haines, and Alan D Dangour. 2021. Systematic mapping of global research on climate and health: a machine learning review. *The Lancet Planetary Health*, 5(8):e514–e525.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

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 of *Proceedings of Machine Learning Research*, pages 4884–4935. PMLR.

Nazan Colmekcioglu, Reza Marvi, Pantea Foroudi, and Fevzi Okumus. 2022. Generation, susceptibility, and response regarding negativity: An in-depth analysis on negative online reviews. *Journal of Business Research*, 153:235–250.

Nick Craswell. 2009. *Mean Reciprocal Rank*, pages 1703–1703. Springer US, 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 ICLR*. International Conference on Learning Representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN.

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.

Robin Haunschild, Lutz Bornmann, and Werner Marx. 2016. Climate change research in view of bibliometrics. *PloS one*, 11(7):e0160393.

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

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.

Babak Jalalzadeh Fard, Sadid A. Hasan, and Jesse E. Bell. 2022. Climedbert: A pre-trained language model for climate and health-related text. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*.

Matyas Juhasz, Kalyan Dutia, Henry Franks, Conor Delahunty, Patrick Fawbert Mills, and Harrison Pim. 2024. Responsible Retrieval Augmented Generation for Climate Decision Making from Documents. *arXiv e-prints*, page arXiv:2410.23902.

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

Danial Khojasteh, Milad Haghani, Abbas Shamsipour, Clara C Zwack, William Glamore, Robert J Nicholls, and Matthew H England. 2024. Climate change science is evolving toward adaptation and mitigation solutions. *Wiley Interdisciplinary Reviews: Climate Change*, 15(4):e884.

Nikolay Koldunov and Thomas Jung. 2024. Local climate services for all, courtesy of large language models. *Communications Earth & Environment*, 5(1):13.

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)*.

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.

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. Association for Computational Linguistics.

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

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, and Mahesh Prakash. 2025. Question Answering in Climate Adaptation for Agriculture: Model Development and Evaluation with Expert Feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. REAL: A retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

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