My Climate Advisor:
An Application of NLP in Climate Adaptation for Agriculture

Vincent Nguyen[1] **Sarvnaz Karimi**[1] **Willow Hallgren**[2] **Ashley Harkin**[3] **Mahesh Prakash**[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

Climate adaptation for agriculture necessitates tools that equip farmers and farm advisors with relevant and trustworthy knowledge to help make them more resilient to climate change. We introduce *My Climate Advisor*, a question-answering (QA) prototype that synthesises information from different data sources, such as peer-reviewed scientific literature and high-quality, industry-relevant grey literature to generate answers, with references, to a given user's question. Our prototype uses open-source generative models for data privacy and intellectual property protection, and retrieval augmented generation for answer generation, grounding and provenance. While there are standard evaluation metrics for QA systems, no existing evaluation framework suits our LLM-based QA application in the climate adaptation domain. We design an evaluation framework with seven metrics based on the requirements of the domain experts to judge the generated answers from 12 different LLM-based models. Our initial evaluations through a user study via domain experts show promising usability results.

1 Introduction

Climate change impacts are seen across the globe in many different ways, from an increase in temperatures to an increase in the frequency of natural disasters. According to the United Nations Framework Convention on Climate Change (Bodansky, 1993), climate change adaptations are increasingly necessary to adjust and respond to the impacts of climate change. These can include technological developments (Smithers and Blay-Palmer, 2001), behavioral changes (Lenzholzer et al., 2020), early warning systems for extreme events (de Perez et al., 2022), and improved risk management (Massetti and Mendelsohn, 2018). In agriculture, climate adaptation means improving farmers' capacity to deal with climate change. This adaptation can include the development and use of tools to increase their knowledge of climate change and methods to make them more resilient (Cradock-Henry et al., 2020). Our study contributes to the goal of making such knowledge accessible. Specifically, our contributions are two-fold: (1) To make the evolving knowledge of climate change and adaptation practices accessible, we have developed a question-answering tool called *My Climate Advisor* (MCA). It is a prototype online service for farmers and farm advisors to gain easier access to information from scientific literature, grey literature and reports, as well as future climate projection data. Given a farmer or farm advisor's question, it responds with information synthesised from the literature alongside references for further reading; and, (2) We propose a novel framework for evaluating such a system, with seven different evaluation criteria, which we share through an annotation guideline together with our initial experimental results. Note that the domain experts carefully design these criteria.[1]

The tool will integrate with *My Climate View*'s API, allowing access to climate historical and projection data within a 100-year window for a breadth of Representative Concentration Pathway (RCP) emission scenarios (Van Vuuren et al., 2011).

2 Background and Related work

We provide a background on climate adaptation, related tools and research in the climate change-agriculture space below.

Climate Adaptation Climate adaptation is described as an adjustment in a social, economic or ecological setting in response to actual or expected climate change (Armstrong et al., 2015). In agriculture, farmers need to adjust their practices to improve resilience to variations in temper-

[1]This tool is to be made public, however, it is currently (June 2024) private while further developments and testing are underway.

ature, precipitation patterns and extreme weather events (Bate et al., 2019). Farmers may need to implement new technologies, crop cultivars and management techniques to ensure food security or economic security in a sustainable manner (Fosu-Mensah et al., 2012). To help farmers adapt to climate change, a goal of My Climate Advisor is to produce location and commodity-relevant, up-to-date management advice from the literature.

My Climate View My Climate View (Webb et al., 2023)[2] is a service that provides climate projections for commodities and regions within Australia. The service is backed by climate indices constructed by climate and commodity experts and climate information from the Australian Bureau of Meteorology. The service is being continually updated with a continuing user engagement initiative. We obtain the data specific to Australian climate through this service.

NLP for Climate Science Machine learning in the climate science domain has been prevalent for years. Many efforts have been dedicated to climate modelling (Dueben and Bauer, 2018; Bittner et al., 2023), disaster prediction (Haggag et al., 2021; Keum et al., 2020), climate change in finance and commerce (Nguyen et al., 2021), climate forecasting (Nguyen et al., 2023) and to inform policy change (Milojevic-Dupont and Creutzig, 2021). However, natural language processing (NLP) for climate science is under-explored.

NLP techniques have been utilised as an analysis tool to provide an overview of climate sentiment on social media, (Prasse et al., 2023; Pupneja et al., 2023) for events such as the Conference of the Parties on Climate Change, (Pupneja et al., 2023) or government policies (Greenwell and Johnson, 2023). Aside from analysis, NLP techniques helped with the monitoring of climate technology innovation (Toetzke et al., 2023), strategies for Environmental, Social and Governance (ESG) investment decision-making (Visalli et al., 2023) and the filtering of literature related to adaptation or mitigation strategies for climate-change-related health problems (Berrang-Ford et al., 2021).

Annotated datasets are crucial for evaluating NLP models. The existing datasets include stance detection for climate change mitigation on social media (Vaid et al., 2022), and global warming in the news (Luo et al., 2020), claim verification for climate change (Leippold and Diggelmann, 2020) and question-answering for both carbon disclosure and climate risk disclosure (Spokoyny et al., 2023). Climate-aware or Green Machine Learning has become more relevant over the years (Cowls et al., 2023). This is also reflected in the NLP community, in the form of Green NLP intending to reduce carbon emissions in the training process of NLP models by re-using pretrained models (Wolf et al., 2020) or in the disclosing or tracking of carbon emissions from NLP models (Strubell et al., 2019; Hershcovich et al., 2022).

A common approach in NLP is to pre-train foundation models with a language model objective for downstream tasks (Devlin et al., 2019). These models have been used in the form of Transformer (Vaswani et al., 2017) encoder-based models such as ClimateBERT (Bingler et al., 2022), which was pretrained on climate-related news articles, research abstracts, and corporate climate reports using domain-adaptive pre-training (Gururangan et al., 2020), and CliMedBERT (Jalalzadeh Fard et al., 2022) which proposed pre-training on climate science literature (Berrang-Ford et al., 2021), climate-policy documents and IPCC reports. However, such approaches using masked language modeling (Devlin et al., 2019) are becoming less prevalent in the question-answering space.

Instead, recently, there has been a shift in the NLP community in adopting Large Language Models (LLMs) pretrained on an autoregressive language modelling task (Brown et al., 2020) and fine-tuned with instructions and human preference labels (Ouyang et al., 2022). These have been used in a chatbot question-answering context (Vaghefi et al., 2023) to provide climate-related information from a combination of Intergovernmental Panel on Climate Change (IPCC) reports and internal LLM knowledge.

However, absent from the literature is NLP for climate change-related agriculture or climate adaptation management advice for agriculture. To the best of our knowledge, we present the first study that collates agriculture-related literature to answer climate change-related agricultural questions and provides climate risk management options to farmers and farm advisors.

[2] https://myclimateview.com.au/

3 Methods

My Climate Advisor is currently designed as a question-answering tool[3] with several components and data sources. We detail our data collection method and corpora used for the Retrieval Augmented Generation (RAG) and the retrieval algorithm to search over the corpora. For generation, we detail the Large Language Model (LLM) used in the study and the decoding algorithms and hyperparameters used for answer generation.

3.1 Data Collection and Indexing

Climate adaptation information needs to be trustworthy and relevant. We therefore gather information from reputable sources such as peer-reviewed published agriculture literature, books, expert-curated documents and high-quality industry grey literature.

For peer-reviewed agriculture literature, we gather articles from the S2ORC corpus (Lo et al., 2020), snapshot on 2023-11-03. The initial size of the corpus was 12.4 million articles. We filter the corpus using the 'fields of study' facet provided by semantic scholar (Kinney et al., 2023). Documents matching the fields of study 'Agricultural and Food Sciences' and 'Environmental Science' are retained, resulting in 1.88 million documents. We remove documents without body text or a Digital Object Identifier (DOI), leaving a final set of 1.36 million articles. We use this corpus for general-purpose agriculture-related questions in our first index.

From this corpus, we filter the documents found in the top 100 agriculture journals ranked by impact score (13,400 documents). However, not all journals could be found within S2ORC. We supplement the rest from the Elsevier [4] snapshot 2023-11-03, leading to a total of 126,000 articles. We use this corpus for more precise climate adaptation advice, forming our second index.

For our third index, we use an expert-curated document set containing information on agricultural commodities, climate adaptation and growing conditions for the Australian climate. We augment it with information from books and industry reports containing information on climate risk and adaptation methods relevant to the Australian climate.

[3]The restrictions on the inputs and outputs for users will require a thorough investigation. See Appendix D for more details.

[4]https://www.elsevier.com/en-au/about

Corpus	# Documents	# Chunks (C=400)	Size (GB)
S2ORC	1.36M	30.6M	124
Top Journals	126K	221K	8.3
Grey Literature	28	1513	0.008

Table 1: Corpus statistics.

This corpus is highly specialised; as such, it is the smallest of the three indexes, with 28 documents.

For indexing, we chunk all documents using a semantic chunking parser[5] to 400 tokens, roughly the size of a paragraph, and ensure we split at sensible sentence boundaries. For each chunk, we use a sentence encoder (Reimers and Gurevych, 2019), JinaBERT (Günther et al., 2023), to produce contextual embeddings which are then normalised and byte quantised. Further details on the statistics of the datasets can be found in Table 1.

3.2 Generative Models

Causal LLMs provide a conditional probability distribution over an output vocabulary, V, given an input sequence, $S = (w_1, ..., w_2)$ or preceding context (Jurafsky and Martin, 2009):

$$P(w_n | w_1, ..., w_{n-1}), w \in V. \tag{1}$$

To select the word to decode from the probability distribution at each autoregressive timestep, t, we use maximum likelihood (greedy decoding) to enable reproducibility and reduce hallucinations from pseudo-randomness (Ippolito et al., 2019; Peng et al., 2023):

$$\hat{w}_t = \arg\max_{w \in V} P(w | \boldsymbol{w}_{<t}). \tag{2}$$

When LLMs are fine-tuned with instructions (Chung et al., 2024), they can generate responses given a prompt S_p as an assistant rather than behaving as a text completion language model (Ouyang et al., 2022).

We use an open-source LLM, in this case, Llama 3-8b (Touvron et al., 2023), which has been instruction fine-tuned. Using an open source allows control over the privatisation of the user's data, compliance with API agreements, use of scientific literature and most importantly, reliability, which cannot be achieved with proprietary mixture-of-expert models as they are non-deterministic (Hayes et al., 2024). Open source allows access to the

[5]https://crates.io/crates/text-splitter, (Accessed: 15/5/24)

weights, which can be beneficial for precise safeguarding with control vectors (Zou et al., 2023). Furthermore, although they have more representation power, proprietary models tend to be more resource-heavy, contributing to climate change (Rillig et al., 2023).

3.3 Retrieval Augmented Generation

We use retrieval augmented generation (RAG) to generate answers using scientific document snippets as context. Using RAG emphasises the provenance of scientific literature as the LLM can be instructed via system prompt to provide the DOI of any relevant document snippets used to generate the answer. We also provide these references in our user interface for further transparency.

It also uses an API from My Climate View (Webb et al., 2023) for location and commodity-specific information, such as noteworthy climate factors[6].

We use Naive RAG (Gao et al., 2023) to synthesise information from an inverted index with a Hierarchical Navigable Small World (HNSW) vector store. For retrieval, we use a hybrid scoring to capture orthogonal signals from keyword matching and semantic similarity (Wang et al., 2021; Nguyen et al., 2022). The hybrid score, S, is a function of an exact-matching (lexical overlap) and soft-matching (vector embeddings)[7] of tokens component. The hybrid scorer is used to rank the query $q \in Q$ and document $d \in D$ pairs as follows,

$$S(q,d) = \beta(\alpha \sum_{t \in q \cap d} f(t) + (1-\alpha)\frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|}), \quad (3)$$

where $f(t)$ is a function of term, which uses document-level or term-level statistics to produce a score given an exact match between the query and document terms, the vector representations, or embedding representations, $\vec{x} = Enc(x), x \in (q,d)$, is given by a universal embedding model, Enc. A soft-match can be computed using cosine similarity between the vector representations. The hyperparameter α is a weighted linear combination of the exact-matching and soft-matching components. Finally, the entire score is multiplied by an index-specific weight, β, which denotes the importance of the index/corpus. We set $\beta = 1$ and $\alpha = 0.02$ in our experiments. The matching components can be interchanged with any model; currently, we use BM25 (Robertson et al., 1994) for our exact-matching component and Jina BERT (Günther et al., 2023) for soft-matching.

[6]API access was not used for the evaluation experiments.

[7]We use the terminology from (Gao et al., 2021).

4 Experiments

To understand how our tool performs, we benchmark it against other existing and proprietary methods. With consultation of climate risk and adaptation experts, we created 15 questions about Australian climate adaptation (Appendix **??**), which we used to generate responses. These questions range from general climate change and adaptation questions to more difficult commodity and region-specific questions.

4.1 Evaluation

Evaluating the capabilities of abstractive QA systems using standardised benchmarks remains challenging due to problems such as data contamination (Sainz et al., 2023), hallucination (Li et al., 2023) and sycophancy (Sharma et al., 2023). Automatic metrics for abstractive question answering such as BERT-score, METEOR, and ROUGE suffer from lexical insensitivity and negation errors, which distort the semantics of text (Saadany and Orasan, 2021) and have bias towards machine-written text (Caglayan et al., 2020) leading to a low alignment with human annotators (Liu et al., 2023).

We, therefore, rely on two experts, a climate scientist and an agronomist, to evaluate the system responses of our system (with and without RAG) and proprietary methods: GPT-3.5, GPT-4, Gemini, Claude, Mistral and the 70B variant in a single-blind study. For all models, including ours, we use the default settings aside from temperature, which we manually set to 0. Specifically for the Llama models, we use the defaults from the llama.cpp library[8]. The Llama 3 models used in the experiments are all the instruct-tuned variants from Meta's official repository. However, for Mistral (Jiang et al., 2023), we use a variant that is instruction fine-tuned with OpenHermes 2.5 (Teknium, 2023) and preference aligns using direct preference optimisation (DPO) (Rafailov et al., 2024) with Argilla's DPO mix (Argilla, 2024).

Given that the Llama family models do not provide a default system prompt, we use a customised system prompt depending on whether or not RAG was used. Details of these prompts can be found in

[8]https://github.com/ggerganov/llama.cpp, (Accessed: 15/5/24)

Appendix C.

The expert annotators curated the following set of 15 questions for the Australian climate to which each system generated responses:

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?

We used maximum likelihood decoding for each model by setting the temperature to zero. The annotators were given the generated responses without knowing the model used to generate the response. They were the literature alongside references for further reading; and asked to evaluate the 15 question-response pairs according to the following annotation criteria and the Likert scale (Likert, 1932):

1. Context: Does the LLM provide enough background information to understand its response?
   - 1.1. Attempts to give some broader context to explain the issue.
   - 1.2. Provides an introductory paragraph to introduce the topic.
   - 1.3. Provides a summary paragraph at the end.
2. Readability: Is the response of the LLM easy to read?
   - 2.1. Overall, the response is well-structured and easy to read.
   - 2.2. Headings and subheadings are well structured and logical and with appropriate categories.
   - 2.3. Used dot points appropriately.
3. Language: Does the LLM use fluent industry terminology?
   - 3.1. Phrasing is appropriate (easy to read, fluent) and not awkward or incorrect.
   - 3.2. Correct use of grammar.
   - 3.3. Consistent with the language used within the industry.
4. Provenance: Does the LLM provide relevant citations to its answers?
   - 4.1. Citations are used appropriately with respect to the context.
   - 4.2. The number of citations used is appropriate (not too few, not too many, regarding what we might expect for the topic).
5. Specificity: Is the information in the response relevant? For instance, to location, time and commodity in question?
   - 5.1. Gives information that is specific to a commodity.
   - 5.2. Gives information specific to the location/region in question, where applicable.
   - 5.3. Where there is no information specific to a location, the LLM admits this (and, preferably, gives information for the appropriate broader region).
6. Comprehensiveness: Does the LLM respond with a complete answer?

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

7. Scientific accuracy: Is the information correct, given the source material?

    7.1. The citations used accurately cite their source material.

    7.2. The cited source material provides high-quality, reliable scientific information.

    7.3. No obvious hallucinations.

We then normalise each annotator's scores before combining them. This allows us to capture the overall ranking preference of the systems rather than an absolute scoring. The raw unnormalised scores can be found in Appendix Table 3 and 4.

5 Results and Analysis

In the literature, we often see that proprietary generalist models perform better than open-source models (Zhao et al., 2023; Chiang et al., 2024). However, we found no clear distinction between proprietary and open-source models (Table 2). The GPT-4 model responses were preferred most across all metrics except accuracy and citation. However, when inspecting the raw scores, the open-source models, Llama and Mistral, are either tied or were marginally worse than GPT-4. This is encouraging as in our application, given the privacy of our data, we cannot use proprietary models.[9]

In line with prior work, we found that model scale was generally indicative of model performance (Hoffmann et al., 2022; Caballero et al., 2023); the Llama3 70b variant outperformed its 8b and 7b variants, for the Claude family, Opus outperformed Haiku, Gemini 1.5 outperformed 1.0 and GPT-4 outperformed GPT-3.5.

Agreement Inner-annotator agreement using Kendalls's Tau (Kendall, 1938) led to 0.319 (moderate) agreement and an overlap of 41.5%. Although the annotators mutually drafted the evaluation criteria, *scientific accuracy* was a source of significant disagreement (Table 1). One annotator penalised responses that were not self-contained; that is, the response must contain scientifically robust sources to back up any claims. The other annotator used their knowledge to determine the scientific validity

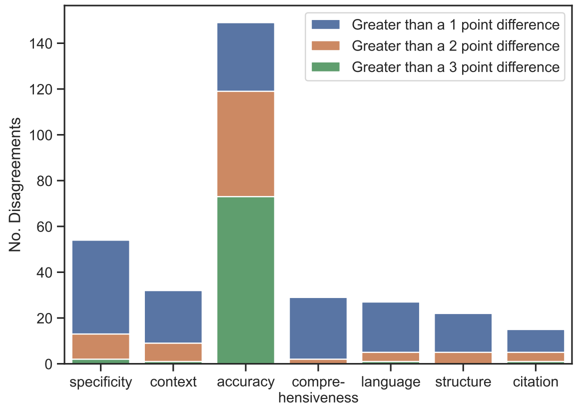

Figure 1: The number of disagreements between annotators for each criterion for the annotation task. A disagreement is defined as when the annotators give different annotations to one another.

of the claims. Noting that verification of climate-related claims has been established as a low agreement task (Leippold and Diggelmann, 2020).

Another source of disagreement was with specificity however, upon inspection, many of these disagreements were within one point and can be attributed to human error or bias. We can further back this claim by looking at the sentiment of scores. When the labels are binarised, scores higher than 2 become positive, and scores 2 or less become negative. In this binary setting, Kendall's Tau agreement is 0.488 (moderate), with an overlap of 76.6%, which can be interpreted as the annotator's overall sentiments of responses being closely aligned. When removing accuracy annotations from this calculation, strong agreement is reached at 0.635 with an overlap of 85.4%, highlighting that the annotator's sentiments are closely aligned.

System Preference Both annotators preferred GPT-4 with Llama-3 70B faring well also. The initial results indicated that the most scientifically accurate model is Claude Opus (one annotator). Both annotators agreed that ChatGPT (GPT-3.5 turbo) was the worst model. This is noteworthy given that it is currently the most popular public-facing chat model. When analyzing the combined raw distribution of scores (Figure 3), we note that the highest performing question-response pair was from the llama-variants, Llama 3 8b + RAG and Mistral 7b + RAG, to questions 6 and 15 respectively from each annotator (see Appendix B). These responses were not only scientifically accurate but were stylistically similar to the responses from GPT-4, where a list of dot points is given, a summary and refer-

[9] Raw scores are in Appendix Table 3 & 4.

Model	Evaluation Criteria							
	Context	Structure	Language	Specificity	Comprehensiveness	Accuracy	Citation	Avg. Score
GPT 4-Turbo	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**	1.05	0.00	**2.00**
Llama 3 70b	1.83	1.83	1.68	1.96	1.61	1.05	0.16	1.85
Claude 3 Opus	1.52	1.56	1.57	0.83	1.52	**1.69**	0.00	1.69
Llama 3 8b + RAG (Ours)	1.15	0.94	1.29	0.84	1.11	1.04	**2.00**	1.54
Gemini 1.5 Pro	1.40	1.50	1.57	1.44	1.65	0.92	0.00	1.54
Llama 3 8b	1.59	1.44	1.51	1.60	1.29	0.64	0.04	1.46
Mistral 7b + RAG	1.39	0.89	1.20	0.73	0.93	0.90	1.65	1.39
Claude 3 Haiku	1.20	1.44	1.30	1.01	1.30	0.82	0.00	1.23
Mistral 7b	1.34	1.11	1.34	1.06	0.94	0.61	0.48	1.15
Llama 3 70b + RAG	0.94	0.72	0.94	0.64	0.70	0.80	1.94	1.08
Gemini 1.0 Pro	0.00	0.39	0.23	1.17	1.02	0.31	0.00	0.54
GPT 3.5-Turbo	0.20	0.00	0.36	0.00	0.00	0.00	0.08	0.00

Table 2: Responses generated by 12 models were annotated for climate adaptation-related questions based on seven criteria (scores of 0 to 4). The values in the tables are from the normalised sum of two annotators. The models are ranked by average score.

ences at the end. Therefore, we find that there is potential for our tool to outperform GPT-4 once aligned with this style of response. Both annotators agreed on the worst performing question-response pair, where Gemini 1.0-pro responded to question 3 with a hallucinated *Apples do not get sunburned* response. An initial hypothesis could be that the model was trained with incorrect data. However, this did not occur with Gemini 1.5-pro, assumed to be trained with similar data, where the model responded with the correct strategies to prevent sunburn risk.

Regarding individual scores, the first annotator (Table 3) generally preferred the non-RAG models due to the stylistic issues mentioned earlier. In contrast, the second annotator (Table 4) preferred the RAG models due to their scientific accuracy and provenance.

Question difficulty A hypothesis that can be reasonably drawn is that LLMs should struggle with questions that are more specific to locations, commodities and time periods. However, we did not see this trend within our annotation. Instead, from Figure 2, we see that questions requiring more reasoning tended to be more difficult (questions 3, 8, 11) for the LLMs over questions more knowledge-recalled oriented (questions 5, 9, 15). In particular, question 8 was difficult as many models responded by telling the user to check the weather forecasts rather than a concrete response. The GPT-4 fared the worst for question 13; although the response was stylistically well-received, it uses generic terminology that is not in line with the industry standard, opting for the term *growth* over the more accurate *vegetative growth* or *tuber bulking*. GPT-4 also had a problem with question 8, where it explained what climate projections were but did not elaborate on their accuracy.

Some questions were underspecified to test the applicability to the Australian climate, such as question 12. Surprisingly, only four models failed to recognise that Young was a town in New South Wales, Australia. Claude's Opus model performed the worst on this question, providing a generic response about its inability to access climate projection data and, therefore, unable to answer the question. A similar answer was provided by Claude Haiku, but the model still provided an answer after its generic response. Mistral 7b and Claude Haiku had a similar issue but with question 7 and question 11, respectively, where they provided a generic response about being unable to predict weather patterns. The RAG models underperformed for specific questions for which the counterpart model did not. A detailed results table for each question and model pair can be found in the Appendix: Table 5.

Ablation on RAG Our ablation analysis reveals that our in-house RAG models were more scientifically accurate than their counterparts. However, this was at the expense of the other metrics, such as readability and background information context. We suspect the model might be using terminology based on the academic context and omitting context as there is an assumption that the user has read the retrieved literature. Furthermore, annotators mentioned that the models included references within their responses, making them longer and more challenging to read. However, including references allows users to read further and verify information. Although our method is scientifically robust, it may

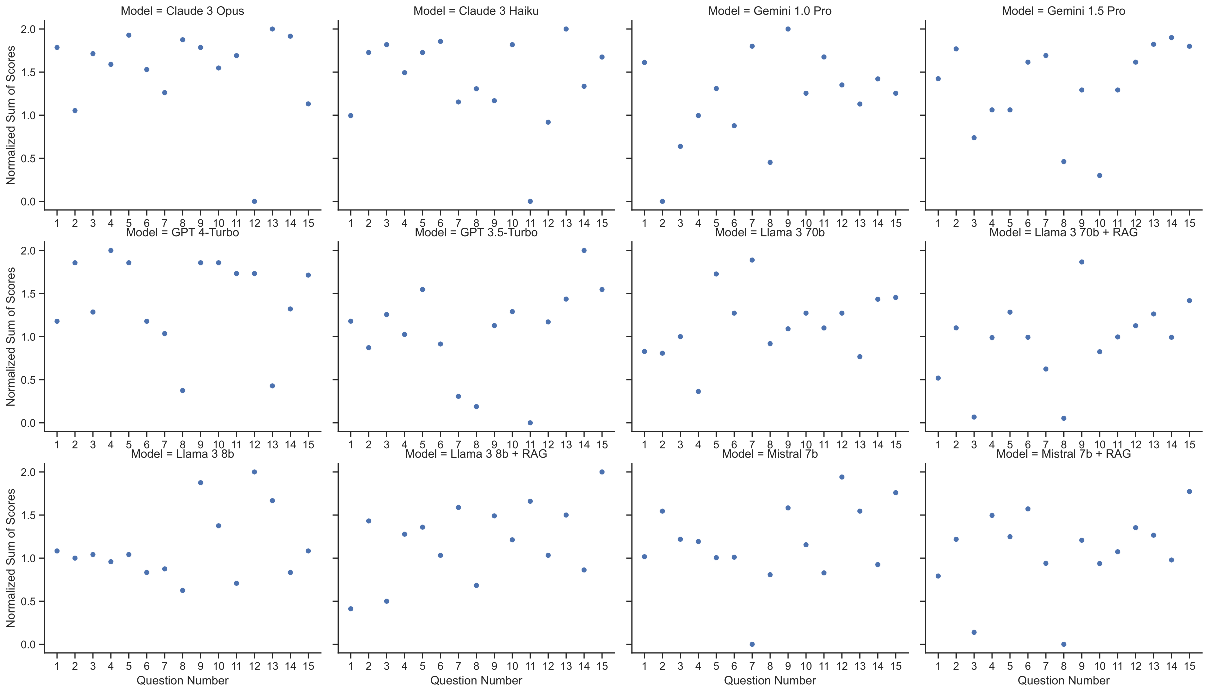

Figure 2: The normalised sum of the two annotator's scores for each response generated by 12 models for each of the 15 questions. Each sub-graph contains the normalised score sum of a particular model plotted against the question number.

not align with the user, who prefers their responses to be structured in a particular way. Fine-tuning the model to include its references at the end of the answer is needed as part of future work.

The most surprising observation was that the Llama3 70b RAG variant under-performed. In particular, the questions that the retriever failed to find relevant impacted the models the most. In particular, as Llama3 70b is more aligned with instruction-following, it suffered the most performance drop as it refused to answer questions where the answer cannot be found in the documents. This was seen in question 3, where the documents referred to sunburn as *sunscald* and did not contain relevant information related to sunburn risk. A similar occurrence happened with question 8, where the retriever found information about the number of days over 40 degrees in Adelaide (South Australia), but the models were either too aligned with instruction-following (Llama3 70b) or misinterpreted the locations (Mistral 7b + RAG). Overall, we observe that the relevance of retrieved documents impacted the RAG models. However, smaller models were less inclined to follow instructions and answered using their internal knowledge rather than our documents and scored higher.

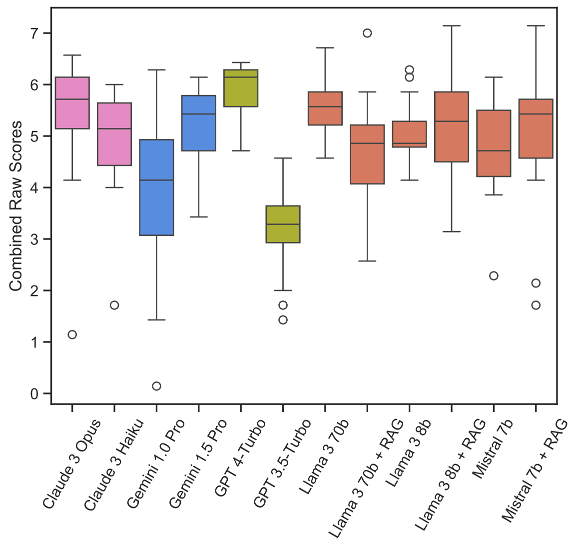

Figure 3: The raw sum of two annotators for the 12 models. Model families are grouped by colour.

6 Conclusions

My Climate Advisor is a question-answering tool designed to provide trustworthy climate change risk and adaptation information for farmers and their advisors. Our tool is created on an in-house Llama 3 with RAG, which synthesises information from peer-reviewed scientific literature and trustworthy grey literature. With the assistance of domain experts, we design an evaluation framework that outlines criteria designed to differentiate LLM-generated answers to a set of questions. While our initial evaluations show a gap between our tool and the leading proprietary systems, the outcome is still encouraging. Our analysis shows that our tool is on par for scientific accuracy while providing provenance for explainability.

Our system can be fine-tuned for further improvements in the near future. Note that due to privacy concerns and the financial and environmental costs of proprietary LLMs, we are limited to open-source models. We will refine the prompting strategy to synthesise climate adaptation information better without sacrificing readability. Finally, we plan to expand the input to multimodal data, including numerical data and graphs, for more accurate representations of climate data including climate projections.

7 Limitations

Some limitations include the lack of prompt engineering for each model. We used the default settings, aside from the temperature setting. However, we believe this is a fair comparison using the default settings. Our tool is also limited in comparison to proprietary offerings, but given that it will be continually updated and supported, we believe that our tool will eventually surpass proprietary offerings while reaping the benefits of using open-source models such as mitigating privacy concerns, protecting intellectual property, integration with control vectors and reducing carbon emissions. Finally, although the annotation guidelines were created jointly by the experts when it came to annotation, there were some interpretations of the criteria. We tried to overcome this limitation by normalising the scores and considering the ranks of the models rather than the raw scores. Despite these limitations, the findings of this study should inform similar studies on the capabilities of proprietary models and open-source LLMs for answering questions in the climate change adaptation domain.

8 Ethical Concerns

We use open-source LLMs to ensure user data privacy and intellectual property protection. We do not use cookies or any tracking mechanism for the users interacting with the My Climate Advisor tool. Given the climate impact of LLMs, it is critical to use power-efficient hardware alongside local LLMs where environmental impacts can be minimised.

References

Argilla. 2024. Argilla dpo mix: A small, high-quality dpo dataset. Accessed on 10.04.2024.

J. Armstrong, R. Wilby, and R. J. Nicholls. 2015. Climate change adaptation frameworks: an evaluation of plans for coastal Suffolk, UK. *Natural Hazards and Earth System Sciences*, 15(11):2511–2524.

Bate Godwill Bate, Jude Ndzifon Kimengsi, and Sani Gur Amawa. 2019. Determinants and policy implications of farmers' climate adaptation choices in rural cameroon. *Sustainability*, 11(7).

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.

Matthias Bittner, Sanaa Hobeichi, Muhammad Zawish, Samo Diatta, Remigious Ozioko, Sharon Xu, and Axel Jantsch. 2023. An LSTM-based downscaling framework for australian precipitation projections. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*.

Daniel Bodansky. 1993. The united nations framework convention on climate change: a commentary. *Yale Journal of International Law*, 18:451.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2023. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of* the 28th International Conference on Computational Linguistics, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv e-prints*, page arXiv:2403.04132.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2023. The ai gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI& Society*, 38(1):283–307.

Nicholas A Cradock-Henry, Paula Blackett, Madeline Hall, Paul Johnstone, Edmar Teixeira, and Anita Wreford. 2020. Climate adaptation pathways for agriculture: Insights from a participatory process. *Environmental Science & Policy*, 107:66–79.

Erin Coughlan de Perez, Laura Harrison, Kristoffer Berse, Evan Easton-Calabria, Joalane Marunye, Makoala Marake, Sonia Binte Murshed, Erlich-Honest Zauisomue, et al. 2022. Adapting to climate change through anticipatory action: The potential use of weather-based early warnings. *Weather and Climate Extremes*, 38:100508.

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.

Peter D Dueben and Peter Bauer. 2018. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009.

Benedicta Y Fosu-Mensah, Paul LG Vlek, and Dilys Sefakor MacCarthy. 2012. Farmers' perception and adaptation to climate change: a case study of sekyedumase district in ghana. *Environment, Development and Sustainability*, 14:495–505.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv e-prints*, page arXiv:2312.10997.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online.

May Haggag, Ahmad S Siam, Wael El-Dakhakhni, Paulin Coulibaly, and Elkafi Hassini. 2021. A deep learning model for predicting climate-induced disasters. *Natural Hazards*, 107:1009–1034.

Jamie Hayes, Ilia Shumailov, and Itay Yona. 2024. Buffer Overflow in Mixture of Experts. *arXiv e-prints*, page arXiv:2402.05526.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy.

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*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, page arXiv:2310.06825.

D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

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

Ho Jun Keum, Kun Yeun Han, and Hyun Il Kim. 2020. Real-time flood disaster prediction system by applying machine learning technique. *KSCE Journal of Civil Engineering*, 24(9):2835–2848.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *arXiv e-prints*, page arXiv:2301.10140.

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

Sanda Lenzholzer, Gerrit-Jan Carsjens, Robert D Brown, Silvia Tavares, Jennifer Vanos, YouJoung Kim, and Kanghyun Lee. 2020. Awareness of urban climate adaptation strategies–an international overview. *Urban climate*, 34:100705.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140):55.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. 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.

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.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online.

Emanuele Massetti and Robert Mendelsohn. 2018. Measuring climate adaptation: Methods and evidence. *Review of Environmental Economics and Policy*.

Nikola Milojevic-Dupont and Felix Creutzig. 2021. Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society*, 64:102526.

Quyen Nguyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. 2021. Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach. *Energy Economics*, 95:105129.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. 2023. Climax: A foundation model for weather and climate. In *Proceedings of the International Conference on Machine Learning*, volume 202, pages 25904–25938.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore.

Katharina Prasse, Steffen Jung, Isaac B 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*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *TREC*, Gaithersburg, MD.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online. INCOMA Ltd.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. *arXiv e-prints*, page arXiv:2310.13548.

John Smithers and Alison Blay-Palmer. 2001. Technology innovation as a strategy for climate adaptation in agriculture. *Applied Geography*, 21(2):175–197.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. Accessed on 10.04.2024.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Last accessed: 15.05.2024.

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.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland.

Detlef P Van Vuuren, Jae Edmonds, Mikiko Kainuma, Keywan Riahi, Allison Thomson, Kathy Hibbard, George C Hurtt, Tom Kram, Volker Krey, Jean-Francois Lamarque, et al. 2011. The representative concentration pathways: an overview. *Climatic change*, 109:5–31.

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

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.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the International Conference on the Theory of Information Retrieval*, page 317–324, New York, NY, USA.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv e-prints*, page arXiv:2303.18223.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv e-prints*, page arXiv:2310.01405.

A Interfaces

A.1 My Climate Advisor interface

We present the user interface of our tool, My Climate Advisor, in Figure 10. The tool is currently in the early stages of development. The interface's main use is to collect feedback from users to improve the retrieval and generation capabilities of the system.

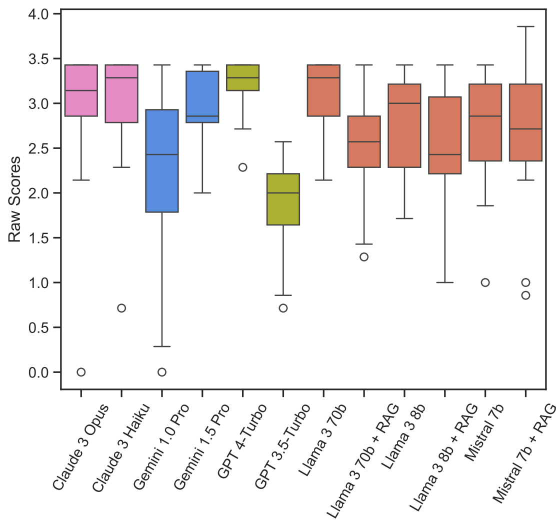

Figure 4: First annotator's average scores. Model families are grouped together by color.

A.2 Annotation interface

Each annotation was tasked with annotating 180 samples in a single-blind study. We use the Label Studio library and interface (Tkachenko et al., 2020-2022) hosted locally. Each annotator was allowed to choose when to do their annotations and which annotations to start from.

B Additional experimental results

The individual scores from the annotators are also included for completeness. Table 3 & 4 show the individual raw scores of each annotator, which were combined and normalised to produce Table 2.

We also include boxplots to show the variance of each method across the questions in Figures 4 & 5, which were combined to produce Figure 3.

The average scores of individual questions and corresponding models are given in Table 5, which provides additional information on Figure 2.

C Additional experimental details: Prompts

We provide additional details on the prompts used in our study for the open-source variants. As these models do not have a default system prompt, we included two styles of system prompts: one that used RAG and one that did not. For the Llama3 models, we used a custom prompt (Appendix Figure 6) for RAG and another prompt (Appendix Figure 7) otherwise. For the Mistral model, we used a similar prompt (Appendix Figure 8) for RAG and a standard prompt (Appendix Figure 9) otherwise.

Model	Evaluation Criteria							
	Context	Structure	Language	Specificity	Comprehensiveness	Accuracy	Citation	Avg. Score
GPT 4-Turbo	**3.90**	**3.70**	**3.70**	**3.70**	**3.70**	**3.80**	0.00	**3.20**
Llama 3 70b	3.70	3.60	3.40	**3.70**	3.40	**3.80**	0.00	3.10
Gemini 1.5 Pro	3.40	3.30	3.50	3.40	3.40	3.70	0.00	3.00
Claude 3 Opus	3.50	3.50	3.50	3.10	3.30	3.40	0.00	2.90
Claude 3 Haiku	3.60	3.60	3.50	3.30	3.10	3.30	0.00	2.90
Llama 3 8b	3.40	3.10	3.30	3.50	2.90	3.30	0.00	2.80
Mistral 7b	3.10	2.90	3.20	3.30	2.70	3.30	0.27	2.70
Mistral 7b + RAG	3.20	2.70	2.90	2.80	2.50	3.20	**1.10**	2.60
Llama 3 8b + RAG	2.60	2.50	2.90	2.90	2.70	3.30	**1.10**	2.60
Llama 3 70b + RAG	2.70	2.40	2.80	2.80	2.20	3.10	**1.10**	2.40
Gemini 1.0 Pro	1.70	2.50	2.70	3.20	2.70	2.90	0.00	2.30
GPT 3.5-Turbo	1.90	1.90	2.40	2.80	1.60	2.50	0.00	1.90

Table 3: First annotator's average scores. In the first column, the models are sorted based on average scores. Bold numbers indicate the highest in the column.

Model	Evaluation Criteria							
	Context	Structure	Language	Specificity	Comprehensiveness	Accuracy	Citation	Avg. Score
GPT 4-Turbo	**3.70**	**3.80**	**4.00**	**3.30**	**3.50**	0.13	0.00	**2.60**
Llama 3 8b + RAG	3.00	3.10	3.90	2.70	2.50	1.10	**1.70**	**2.60**
Claude 3 Opus	2.90	3.20	3.70	2.20	2.80	**2.60**	0.00	2.50
Llama 3 70b	3.50	3.60	3.90	3.20	2.90	0.13	0.27	2.50
Mistral 7b + RAG	2.90	2.80	3.80	2.70	2.30	0.93	1.10	2.40
Llama 3 8b	3.20	3.40	3.80	2.90	2.70	0.07	0.07	2.30
Gemini 1.5 Pro	2.70	3.30	3.70	2.80	3.00	0.00	0.00	2.20
Llama 3 70b + RAG	2.30	2.80	3.60	2.50	2.10	0.87	1.60	2.20
Mistral 7b	2.90	3.00	3.70	2.20	2.10	0.00	0.40	2.00
Claude 3 Haiku	1.90	2.90	3.40	2.10	2.50	0.53	0.00	1.90
Gemini 1.0 Pro	1.00	2.10	2.90	2.70	2.30	0.00	0.00	1.60
GPT 3.5-Turbo	1.30	2.00	3.30	1.10	1.10	0.00	0.13	1.30

Table 4: Second annotator's average scores. In the first column, the models are sorted based on average scores. Bold numbers indicate the highest in the column.

D Restrictions on User Inputs or Outputs

Given the problems with LLMs with regards to reward hacking and teacher forcing (Zhao et al., 2023) which can lead to hallucination or misinformation. It is prudent to think of the ways that farmers or their advisors will interact with our tool. We denote three possible variants of usage that have to do with the user access or openness to the inputs (questions) and the outputs (LLM responses):

1. Input Open, Output Open: Chat-style interface. Users can freely input questions to produce outputs. This requires the most amount of safeguarding and may be difficult to reliably control in practice.

2. Input Open, Output Closed: The users may submit questions, however, they will be given responses that are embedded within a prefilled frequently asked questions (FAQ). This FAQ will be continually updated with LLM responses but can be checked beforehand.

3. Input Closed, Output Closed: The user cannot control the inputs, and instead is given a response by the LLM based on the information of location and commodity that has been prefilled for a related service.

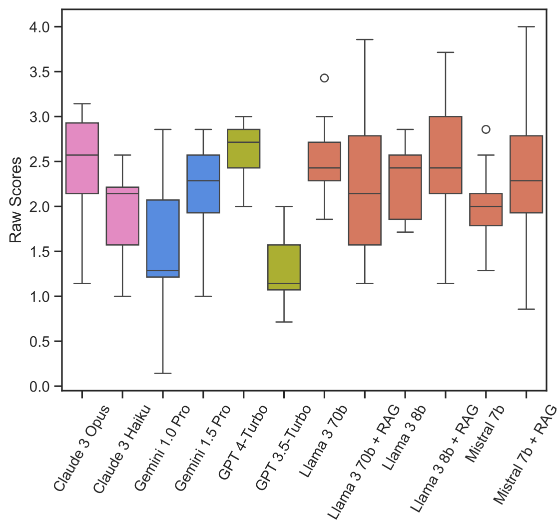

Figure 5: Second annotator's average scores. Model families are grouped together by color.

Llama3 RAG prompt

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant designed to help answer a farmer's agriculture-related questions. Use the following documents to help answer the user's questions.
If you are unsure of your answer, inform the user to check the information with their farm advisor. <|eot_id|><|start_header_id|>user<|end_header_id|>

What are the ideal pollination conditions for growing almonds? <|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 6: Prompt used for Llama3 + RAG.

Llama3 prompt

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant designed to help answer a farmer's agriculture-related questions.
If you are unsure of your answer, inform the user to check the information with their farm advisor. <|eot_id|><|start_header_id|>user<|end_header_id|>

What are the ideal pollination conditions for growing almonds? <|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 7: Prompt used for Llama3.

Mistral RAG prompt

<s><|im_start|>system You are a helpful AI assistant designed to help answer a farmer's agriculture-related questions. If you are unsure of your answer, inform the user to check the information with their farm advisor.<|im_end|> <|im_start|>user What are the ideal pollination conditions for growing almonds?<|im_end|> <|im_start|>assistant

Figure 8: Prompt used for Mistral 7b + RAG.

Mistral prompt

<s><|im_start|>system You are a helpful AI assistant designed to help answer a farmer's agriculture-related questions.
Use the following documents to help answer the user's questions.
If you are unsure of your answer, inform the user to check the information with their farm advisor.<|im_end|> <|im_start|>user What are the ideal pollination conditions for growing almonds?<|im_end|> <|im_start|>assistant

Figure 9: Prompt used for Mistral 7b.

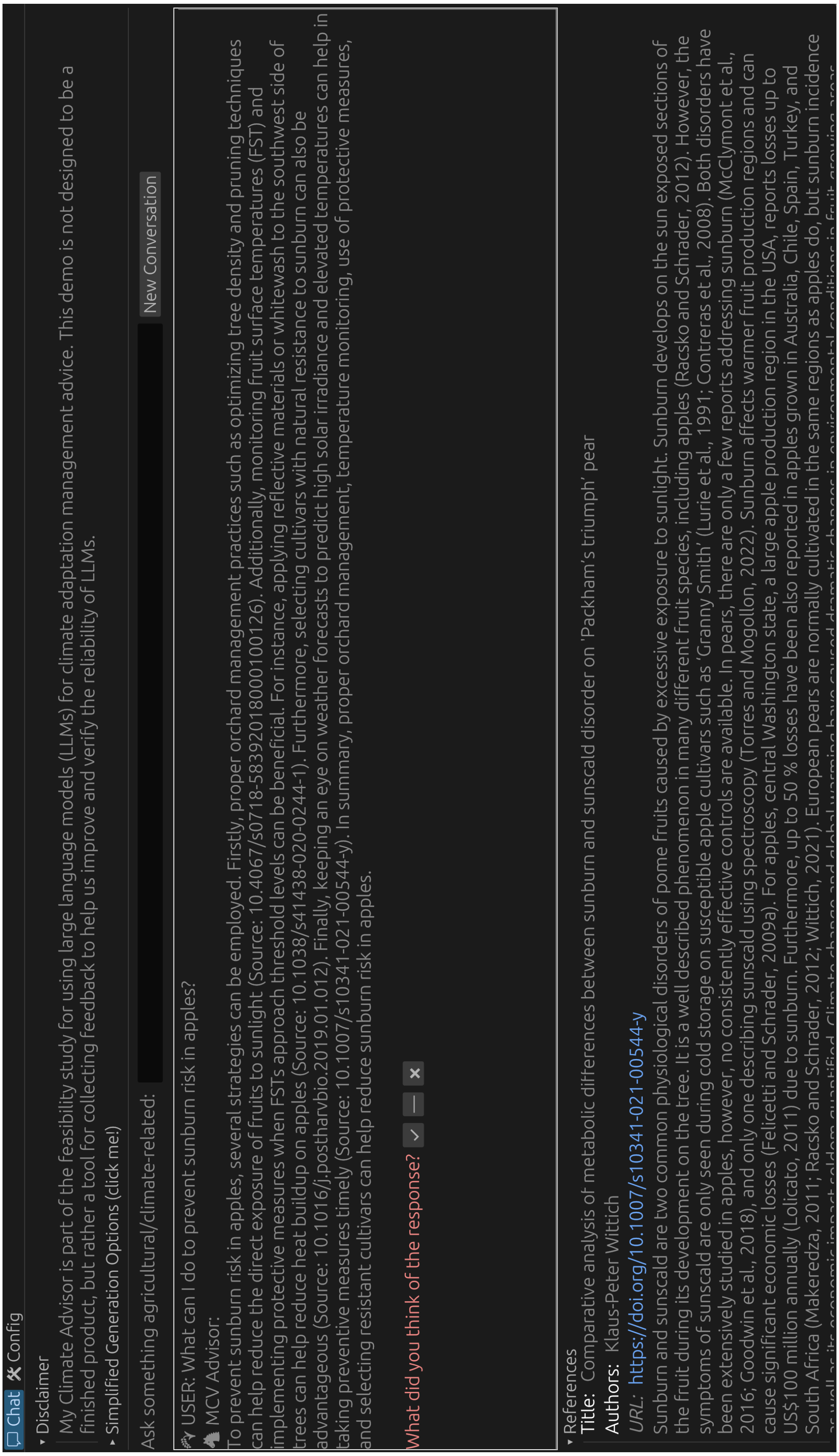

Figure 10: User interface of the prototype My Climate Advisor. The user inputs their question to the LLM, and the response and the references used to generate that response are provided.

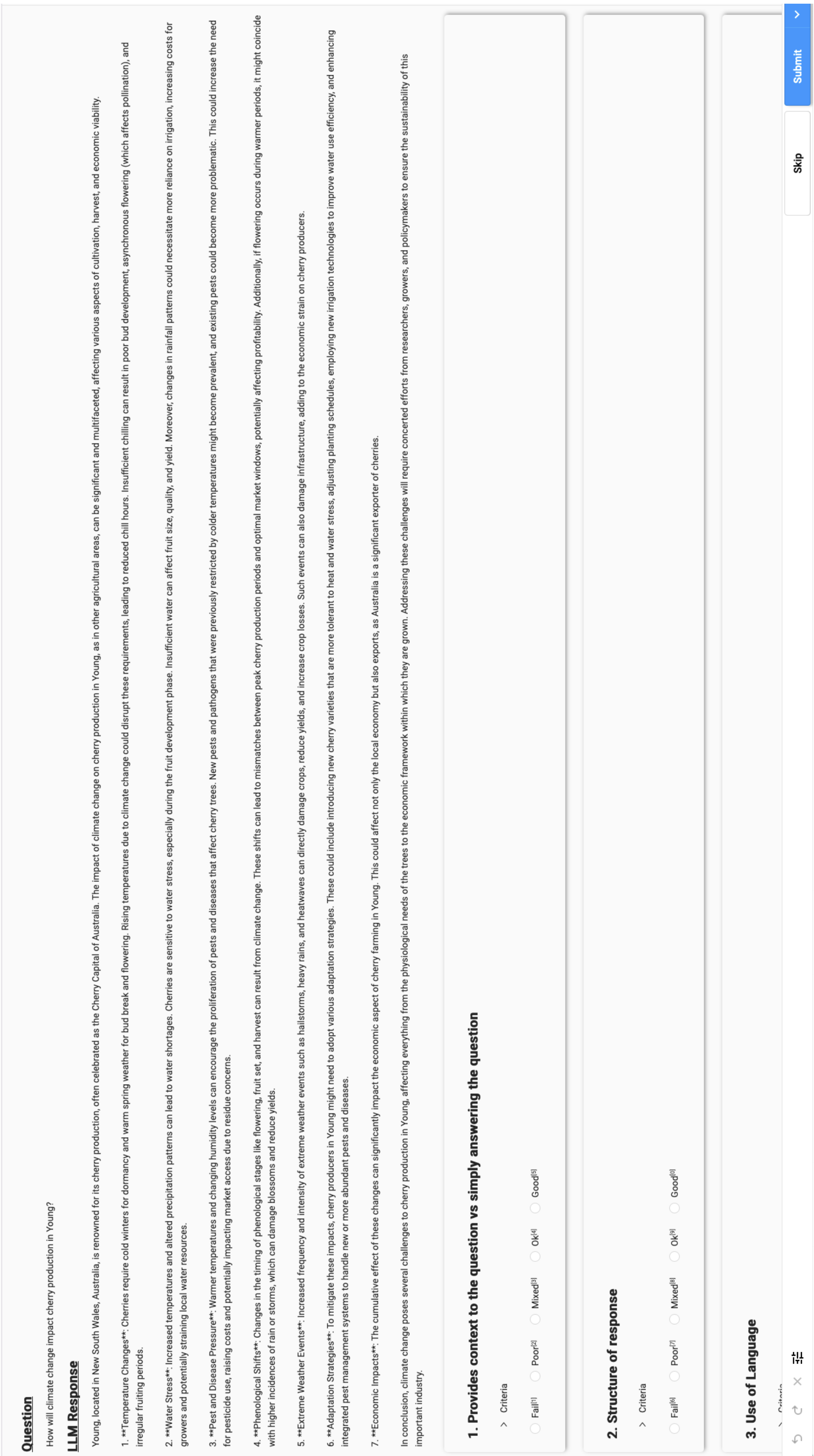

Figure 11: Annotation interface used to grade LLM responses to agriculture questions.

	Claude 3 Opus	Claude 3 Haiku	Gemini 1.0 Pro	GPT 4-Turbo	Mistral 7b + RAG	Llama 3 70b	Gemini 1.5 Pro	Mistral 7b	Llama 3 8b	Llama 3 8b + RAG	Llama 3 70b + RAG	GPT 3.5-Turbo
Q1	1.79	1.00	1.61	1.18	0.79	0.83	1.42	1.02	1.08	0.41	0.52	1.18
Q2	1.05	1.73	0.00	1.86	1.22	0.81	1.77	1.55	1.00	1.43	1.10	0.87
Q3	1.71	1.82	0.64	1.29	0.14	1.00	0.74	1.22	1.04	0.50	0.07	1.26
Q4	1.59	1.49	1.00	2.00	1.50	0.36	1.06	1.19	0.96	1.28	0.99	1.03
Q5	1.93	1.73	1.31	1.86	1.25	1.73	1.06	1.01	1.04	1.36	1.28	1.55
Q6	1.53	1.86	0.88	1.18	1.57	1.27	1.62	1.01	0.83	1.03	0.99	0.91
Q7	1.26	1.15	1.80	1.04	0.94	1.89	1.69	0.00	0.88	1.59	0.62	0.31
Q8	1.88	1.31	0.45	0.38	0.00	0.92	0.46	0.81	0.62	0.68	0.05	0.19
Q9	1.79	1.17	2.00	1.86	1.21	1.09	1.29	1.58	1.88	1.49	1.87	1.13
Q10	1.55	1.82	1.25	1.86	0.94	1.27	0.30	1.16	1.37	1.21	0.82	1.29
Q11	1.69	0.00	1.68	1.73	1.07	1.10	1.29	0.83	0.71	1.66	1.00	0.00
Q12	0.00	0.92	1.35	1.73	1.35	1.27	1.62	1.94	2.00	1.03	1.13	1.17
Q13	2.00	2.00	1.13	0.43	1.27	0.77	1.82	1.55	1.67	1.50	1.26	1.44
Q14	1.92	1.33	1.42	1.32	0.98	1.43	1.90	0.93	0.83	0.86	0.99	2.00
Q15	1.13	1.67	1.25	1.71	1.77	1.45	1.80	1.76	1.08	2.00	1.42	1.55

Table 5: Normalised sum of average scores from both annotators for each question and model.