Evaluating Retrieval Augmented Generation to Communicate UK Climate Change Information

Arjun Biswas[1], **Hatim Chahout**[2], **Tristan Pigram**[2], **Hang Dong**[1]
Hywel T.P. Williams[1],**Fai Fung**[2], **Hailun Xie**[1]
[1]University of Exeter, Exeter, United Kingdom
[2]Met Office, Exeter, United Kingdom
{ab1574, h.t.p.williams, h.dong2, h.x.xie}@exeter.ac.uk,
{hatim.chahout, fai.fung, tristan.pigram}@metoffice.gov.uk

Abstract

There is a huge demand for information about climate change across all sectors as societies seek to mitigate and adapt to its impacts. However, the volume and complexity of climate information, which takes many formats including numerical, text, and tabular data, can make good information hard to access. Here we use Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) to create an AI agent that provides accurate and complete information from the United Kingdom Climate Projections 2018 (UKCP18) data archive. To overcome the problematic hallucinations associated with LLMs, four phases of experiments were performed to optimize different components of our RAG framework, combining various recent retrieval strategies. Performance was evaluated using three statistical metrics (faithfulness, relevance, coverage) as well as human evaluation by subject matter experts. Results show that the best model significantly outperforms a generic LLM (GPT-3.5) and has high-quality outputs with positive ratings by human experts. The UKCP Chatbot developed here will enable access at scale to the UKCP18 climate archives, offering an important case study of using RAG-based LLM systems to communicate climate information.

1 Introduction

Climate services are data, information, and knowledge provided to support decision-making about climate change (Global Framework for Climate Services, 2025). In the UK, the national government funded the UK Met Office to produce the UK Climate Projections (UKCP) (Lowe et al., 2018), a compilation of high-quality climate models, outputs, and analyses that help organizations prepare and adapt to climate change. Similar efforts are underway in other countries (e.g. Climate Change in Australia (CSIRO and Bureau of Meteorology, 2015) and CH2018 in Switzerland (Fischer et al., 2022)). The audiences for such climate services can be very large and diverse; for example, the UKCP data portal has over 11,000 registered users and is widely used in national government policy (Department for Environment, Food and Rural Affairs, 2024) and environmental regulations (Environment Agency, 2024), as well as business adaptation planning (Anglian Water, 2020) and best practice guidelines for local governments preparing for climate change (ADEPT, 2019). One major challenge is tailoring such services to specific and local user contexts. There are too few human experts to serve the complex climate information needs of such a large and diverse set of users. Generative AI tools offer a potential solution, allowing a user to extract bespoke climate information tailored to their own local context, through simple natural language interfaces. However, it is very important that such tools provide high-quality information; poor quality or incorrect information could cause harm by worsening climate-related decision-making.

In this study, we present the development of an LLM-based climate service that is intended to help deliver UKCP climate information. The UKCP archive contains a wide variety of complex scientific content (Met Office, 2025). A helpdesk is provided and human experts assist the UKCP user community in navigating the complex UKCP archive, offering user guidance and scientific documentation to improve access and utilization. Our tool is conceptualized as an automated support tool that can respond to typical UKCP helpdesk queries with accurate and trustworthy information. If deployed, this will reduce pressure on human experts and allow a greater number of UKCP users to be served. Here we describe our development of this tool in the form of a chatbot that uses Retrieval Augmented Generation (RAG). We evaluate a number of different information chunking, retrieval, ranking, and query expansion strategies,

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 126–141

creating and testing 14 different RAG pipelines. Performance is evaluated using a range of automated metrics (including a novel *coverage* metric) and human evaluation of outputs by subject matter experts (SME) in climate science. Results show that our RAG-based chatbot communicates accurate and relevant information from the UKCP archive, avoiding hallucinations or deviation from the content in the UKCP archive, and outperforming a non-specialized LLM-based chatbot. Overall positive ratings by human experts are achieved for our best RAG system (S2BH-CHR-MQG5).

2 Background and Related Work

Climate science and projections about future climate change are typically presented as complex datasets, scientific reports, articles, and other technical content. Currently, human climate scientists are needed to interpret this information for non-experts (Intergovernmental Panel On Climate Change (IPCC), 2023). While generative AI might help increase access to climate information at scale, effective decision-making around climate entails accurate translation of complex concepts and provision of trustworthy information. AI tools for communicating climate science must prioritize output accuracy and scientific quality.

Generally, LLMs are prone to hallucinations while having strong generative capability – providing responses that appear grammatically correct, fluent, and authentic, but actually deviate from source inputs (faithfulness) and/or fail on factual accuracy (factualness), offering outdated or incorrect information (Ji et al., 2023; Xu et al., 2024). Answers may also be incomplete, generic, or vague. This has led to methods that provide LLMs with additional domain-specific information to improve performance in applications requiring precise answers (Wu et al., 2023; Peng et al., 2023). Two popular approaches include domain-specific training of LLMs and RAG. Below we summarise studies that use these approaches in the domain of climate change.

Earlier studies adapt encoder-only, discriminative LLMs like BERT (Devlin et al., 2019) for climate communication tasks. ClimateBERT (Webersinke, 2022) was trained on approximately 2 million paragraphs of climate-related information, including reports, scientific paper abstracts, and news articles. The training process included general pre-training, followed by domain-specific training on climate information and then downstream training for specific tasks like classification, sentiment analysis, and fact-checking. Another example is ClimateBERT-NetZero (Schimanski et al., 2023) which fine-tuned BERT to detect whether a text contains a net zero or reduction target, and thus support subsequent data analyses.

Until most recently, generative LLMs are applied to convert climate information. ClimateGPT (Thulke et al., 2024) is a foundation model trained on a large corpus of climate-related texts. Training of ClimateGPT involved pre-training and instruction fine-tuning. As reported (Thulke et al., 2024), the pre-trained model outputs are domain-specific but suffer from hallucinations and cannot provide detailed information. ClimateGPT was then expanded by integrating a simple hierarchical RAG system, leading to improved performance. Also, training LLMs is energy-consuming and cannot easily adapt to new information, e.g., for climate projection.

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) is a process of incorporating information from external databases to increase the answer accuracy of LLMs in domain-specific applications. It works by extracting relevant information (retrieval), processing the retrieved information with other external sources to create a structured prompt (augmentation), and summarising the combined information using an LLM (generation). The study (Fore et al., 2024) shows that RAG helps to improve the factual metrics of answers using in-context learning, which effectively mitigates conflicting information from the training set, for question answering with climate-related claims.

ChatClimate used a RAG-based system to communicate climate information (Vaghefi et al., 2023). This RAG system extracted the top-n pieces of relevant information for a given query from the IPCC Report. More recently, ChatNetZero (Hsu et al., 2024) is a RAG-based chatbot targeting the net zero domain. Our RAG-based system further explores a variety of chunking, retrieval and query rewriting strategies to enhance the RAG process. We focus on the dynamic, future climate projection data, instead of the current climate reports.

Robust evaluation of answer quality and information retrieval strategies is vital to ensure RAG systems' correctness and trustworthiness, as they are highly sensitive to noisy or irrelevant con-

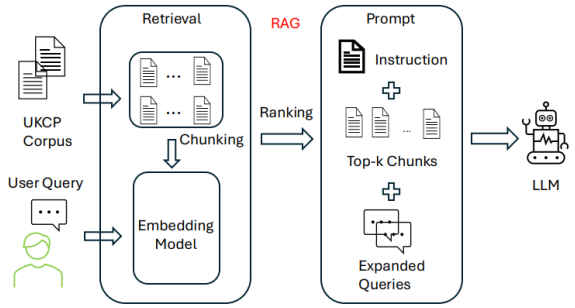


Figure 1: Overall framework of the UKCP Chatbot.

texts (Shi et al., 2023). A notable RAG-focused evaluation tool is RAGAS (Es et al., 2024), but it does not capture all dimensions of trustworthiness and accuracy. In the climate-focused applications above, human evaluators are used to judge answer correctness (Thulke et al., 2024), or automated evaluation is performed using benchmark datasets (Webersinke, 2022). These approaches become difficult in specialized domains, such as climate change and projection, where non-experts may not spot mistakes and there are no domain-specific benchmarks. In this work, we provide a phased automated evaluation with new datasets and human evaluation design with climate experts.

3 Methodology

We develop a conversational question-answering system to provide relevant, accurate, and trustworthy information related to the UKCP archive (hereafter referred to as "UKCP Chatbot"). Answers must be based only on the available UKCP archive; in other words, answers should only use UKCP data and not "general knowledge" or other external information. This makes the task a complex testbed for the faithfulness and hallucination of an LLM-based RAG system.

Our structured RAG system integrates different document chunking strategies, retrieval methods, and query expansion into well-defined retrieval pipelines. Multiple proposed pipelines are evaluated to optimise information retrieval, answer relevance and accuracy. A hybrid evaluation approach was used, incorporating both automated and human assessments. To demonstrate the advantage of the RAG approach, we also compared a general-purpose LLM, GPT-3.5-0125[1], used outside the RAG framework and based only on its general knowledge. The same GPT-3.5 LLM was used within several of our RAG system components so gives a fair comparison. User surveys were conducted to understand subjective perceptions of the UKCP Chatbot.

3.1 System Overview

The overall functionality of the chatbot is shown in Figure 1. The user enters a query q as input. The RAG system then parses the query and extracts the most relevant data chunks from the UKCP corpus. The extracted information is encapsulated into a prompt to a LLM to summarise the information and generate an answer. GPT-3.5 was used for its efficiency and cost-effectiveness. Conversation history is recorded to better understand the context of user queries. The components of the RAG system are optimized by comparison of several alternatives in each case; these choices are described below. The system is developed using a JavaScript front-end interface (see Figure 5 in Appendix F) and Python for back-end data manipulation.

3.2 Data Preparation

The UKCP archive contains diverse UK-focused climate data and information for a wide audience including scientific researchers, policymakers, industry professionals, and members of the public. The archive provides climate projections to the year 2100 based on model projections of future climate conditions for a number of greenhouse gas emission scenarios. Information is presented as published literature, observations, and climate model data. Here we focus on documents available from the UKCP archive, which include scientific reports, fact sheets, technical and guidance documentation, stakeholder engagement materials, and case study reports.

The corpus consists of 85 documents in raw PDF format in complex layouts. From this corpus, four segmented datasets were created by using different chunking approaches: *fixed-length*, *paragraph*, *section*, and *summary* methods. Details of data extraction, document segmentation, data cleaning, and data representation are in Appendix A.

3.3 RAG Framework

To develop the optimal RAG pipeline, we divide the RAG methodology into four components: document segmentation (chunking), chunk retrieval,

[1] https://platform.openai.com/docs/models/gpt-3.5-turbo

Table 1: Design options for RAG system components tested by phased evaluation.

Phase	Component Evaluated	Model ID	Component Variant
1	Chunking	F5	Fixed-Length (5-chunk context)
		F10	Fixed-Length (10-chunk context)
		P5	Paragraph (5-chunk context)
		P10	Paragraph (10-chunk context)
2	Retrieval	H20	Hierarchical (20 documents)
		S2B	Small-to-big
		S2BH15	Hierarchical (15 documents) & Small-to-big
		S2BH20	Hierarchical (20 documents) & Small-to-big
3	Ranking	S2BH-LST	Lost-in-the-middle
		S2BH-CHR	Cohere
		S2BH-DVR	Diversity
		S2BH-REC	Reciprocal
4	Query Expansion	S2BH-CHR-MQG3	Multiple Query Generation (3 queries)
		S2BH-CHR-MQG5	Multiple Query Generation (5 queries)

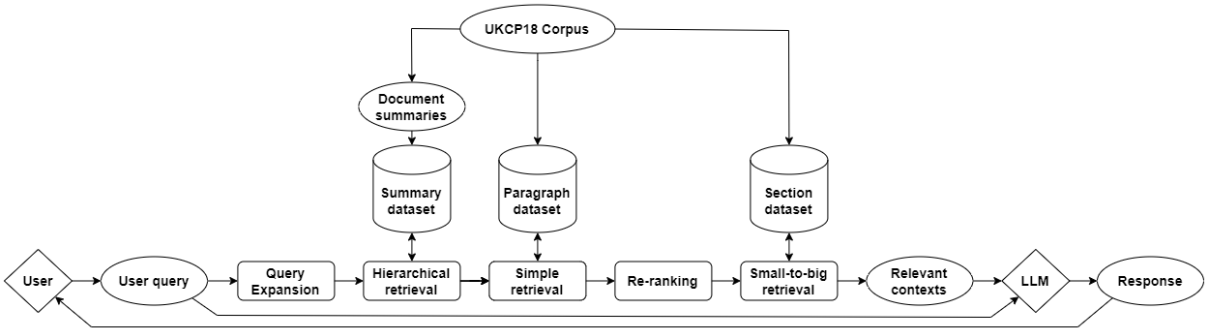


Figure 2: Overview of the RAG framework for model S2BH-CHR-MQG3.

chunk re-ranking, and query expansion. A total of 14 RAG pipelines were evaluated across the four components (see Table 1) and a locally optimal solution for each component was identified.

Since the four components work together in a functional RAG pipeline, component evaluation was performed sequentially in four experimental phases that each identified the best option for one component. This is based on the assumption that the components are independent of each other. The best component option found in each phase was adopted for subsequent phases of testing. This heuristic approach greatly reduces the number of test combinations (as high as 128 considering all possible combinations). Evaluation during this process used automated metrics that are described below; outcomes are presented in the Results section. The final RAG solution chosen for the UKCP Chatbot is visualized schematically in Figure 2 and incorporates the best components selected by this process, with an additional layer of human evaluation/testing (see Section 3.5).

3.3.1 Document segmentation (chunking)

This first phase considers the best low-level chunking strategy and how many chunks are needed in the prompt. It uses a simple retrieval approach based on cosine similarity between vector embeddings of the query and context chunks. We evaluated RAG pipelines that use the top-5 or top-10 most-similar chunks (following ClimateGPT (Thulke et al., 2024) and ChatClimate (Vaghefi et al., 2023)) chosen by two different chunking strategies (fixed-length F, or paragraph-based P). Fixed length and paragraph has the length of 1024 tokens. Later we consider larger-sized section chunks and summary chunks to improve final retrieval, but we do not use them in the initial retrieval stage.

3.3.2 Enhanced Information Retrieval

We introduce two retrieval strategies, *small-to-big* and *hierarchical*, and a third strategy that combines them, to enhance the simple retrieval of top-k relevant paragraph chunks above. These three retrieval algorithms combine multiple chunking strategies (including section and summary chunks) to enhance the final outputs of information retrieval. These enhanced retrieval methods provide more information to answer a user question by extracting longer document sections based on the smaller chunks found by simple retrieval (small-

to-big) and by pre-selection of relevant documents prior to simple retrieval (hierarchical). These methods localize the relevant documents and sections to reduce inclusion of irrelevant information.

Small-to-big contextual expansion. Here small paragraph chunks are enhanced with bigger section chunks to increase the amount of relevant content found during information retrieval. First, 10 paragraph chunks are identified (using the P10 model, which provides the most relevant information) and then the document sections containing those paragraphs are also retrieved. The top-5 sections (fitting within the 16k context-length limit of GPT-3.5) most similar to the user query are then used to create the final prompt for question-answering. Expanding from paragraphs to sections increases the relevant/specific information extracted from the corpus and thereby enables better answers to be generated.

Hierarchical filtering. Here a pre-filter is applied to consider only the top-k most relevant documents for initial retrieval of paragraph chunks, creating a two-stage (or hierarchical) retrieval process. We set $k = 20$ to include a large number of documents and allow a more diverse set of chunks to be retrieved. Relevant documents are identified by first creating a summary of each document and then using cosine similarity between the embeddings of the user query and each document summary. The P10 model for simple retrieval is then applied to all paragraph chunks from the top-k relevant documents. This approach can prevent the spurious inclusion of paragraphs from irrelevant documents.

The two approaches above are then combined, leading to pipelines using hierarchical filtering (with 15 or 20 documents retained) followed by small-to-big retrieval. This helps extract the relevant sections from the most relevant documents.

3.3.3 Chunk ranking

Ranking (cf. re-ranking) is prioritization amongst the matching chunks selected by a retrieval method; it applies a rule or strategy to re-order the selected chunks and decide which ones will be included in the prompt. Here we tested four re-ranking strategies which we applied to the combined, hierarchical & small-to-big model, which was chosen as the candidate model for this phase based on the automated evaluation results.

Lost-in-the-middle. Language models can struggle to parse information in a long prompt, most often missing relevant information placed in the middle of a long input sequence (Liu et al., 2024). This re-ranking strategy places the most relevant chunks at the beginning and the end of the prompt, moving the least relevant chunks to the middle, following (deepset, 2025a). Unlike many other rankers, it does not use the query and simply re-orders the list of retrieved chunks.

Cohere is a platform that provides relevance-based re-ranking language models (Shi and Reimers, 2024) trained on query-passage pairs in documents. Here we used the Cohere "rerank-english-v3.0" model[2] , which was fine-tuned to retrieve the most relevant passage for a given query.

Diversity ranking ranks a list of chunks based on the relevance to the query and the diversity of the information in each chunk. The greedy algorithm initially chooses the most similar chunk to the query and then iteratively adds chunks that are, on average, least similar to previously added chunks, until all chunks are ranked (deepset, 2025b). Following the implementation in deepset (2025b), we use a sentence BERT model (Reimers and Gurevych, 2019), here "all-MiniLM-L6-v2"[3], to embed the query and the chunks for ranking.

Reciprocal ranking (Rackauckas, 2024) is used with multiple query generation (see below). For each query, inverse rank scores are calculated for all retrieved chunks:

$$\texttt{reciprocal\_score} = \frac{1}{\texttt{rank} + k} \quad (1)$$

where `rank` is the similarity-based rank of the chunk and k is a smoothing factor. The final ranking is calculated using the mean value of all reciprocal scores for each chunk.

3.3.4 Query Expansion

Retrieval responses are highly dependent on the exact phrasing of the query, so this phase seeks to diversify phrasing to give a more consistent retrieval of information (Rackauckas, 2024). An LLM (GPT-3.5) is utilised to generate multiple versions of the original query, keeping the meaning but varying how it is written. Each version is then used for information retrieval and the combined responses are used collectively to generate an answer to the original query. Here we tested

[2] https://huggingface.co/Cohere/rerank-english-v3.0

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

RAG pipelines using 3 or 5 versions of the original query. All chunks retrieved were collated and ranked together using a re-ranker (Cohere) and the original query to determine the top-k (here k=10) chunks used in the prompt. Here we use the S2BH-CHR model as the candidate based on the evaluation of the re-ranking stage of the pipelines. The prompt for generating multiple queries is presented in Appendix B (prompt-1).

3.3.5 Prompt construction

The complete prompt for the proposed RAG framework comprises a detailed *system instruction* and a *user prompt*. The system instruction includes the context for the chatbot (the UKCP archive), the task (question-answering), detailed constraints to ensure that answers are generated only from provided chunks from the UKCP corpus, and the steps to create an answer. The user prompt includes a structured format of the chunks, the query, and an answer mark ("ANSWER:") to prompt the system to generate an answer. The full prompt is presented in Appendix B (prompt-2).

3.4 Evaluation

A combination of automated and human evaluation was used to assess the quality of the UKCP Chatbot.[4] Automated evaluation metrics were used to compare 14 RAG pipeline variants and ChatGPT (GPT-3.5, chosen as a strong baseline example of a general-purpose LLM). Human experts then evaluated four RAG pipelines identified by automated evaluation, to determine the best pipeline overall and characterise user perceptions of the system.

3.4.1 Evaluation data

Two datasets are used for the automated evaluation of outputs from the UKCP Chatbot: (1) A dataset of 250 **synthetic QCA triplets**; and (2) An anonymised dataset of 50 **authentic QA pairs** from the UKCP helpdesk. Details about the dataset creation are available in the Appendix C.

3.4.2 Automated Evaluation Metrics

Automated evaluation here aims to assign metrics to RAG pipeline responses to assess three important characteristics that good answers must contain for our use case: (i) answer relevance; (ii) answer faithfulness; and (iii) answer coverage. Answer relevance measures how well the response aligns with the intent of the user query. Answer faithfulness measures the extent to which the response is based on the source information (or conversely, how much it uses other unsupported content). For ***Relevance*** and ***Faithfulness*** scores, we use metrics provided by RAGAS (Es et al., 2024). As accurate answers to scientific questions (as here in the climate domain) often require a high level of specificity and detail, here we propose a new ***Coverage*** metric, which calculates the proportion of all the named entities, keywords, and numerical values from the context chunks that are given in a generated answer. Details of the above metrics are described in Appendix.D. For the ChatGPT coverage score we compare the answers to the groundtruth context.

We also compare RAG-system scores to those from a baseline LLM (ChatGPT/GPT-3.5). Since relevance and faithfulness are defined using context chunks from a RAG system, the only metric that can be directly compared to a non-RAG LLM is coverage. We compute two metrics: (i) ***ChatGPT mean coverage score*** and (ii) ***Proportion of answers with coverage*** $>$ ***ChatGPT***: the percentage of answers by each RAG pipeline that have a higher coverage score than ChatGPT.

3.5 Human Evaluation

Four selected RAG pipelines, chosen as the top-performing pipeline from each of the four phases of automated evaluation, were tested by subject matter experts. An initial screening was conducted by climate experts in the author group to choose the best two of these pipelines for further testing. Interactive evaluation was then performed by a panel of experts (n=10) recruited from UK Met Office staff. Details are given in Appendix E.

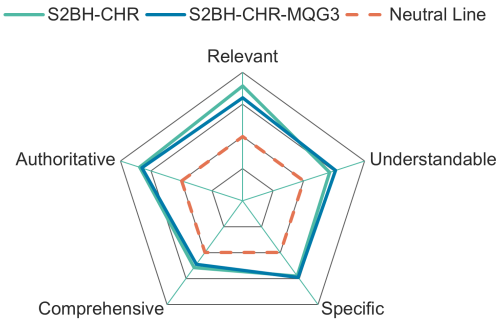
Two-stage survey design. Panelists received a survey in two stages, with access to the live chatbot given in the second stage.

In the first stage, two preliminary questions assessed participant background: (Q1a) Duration of professional experience in climate science; and (Q1b) Self-assessed familiarity with UKCP18 data. Next, participants evaluated the quality of answers provided by the chatbot. Four question-answer pairs were chosen from the 50 authentic QA pairs dataset, that could be easily reviewed without extensive additional knowledge. Answers given by the two selected pipelines were provided

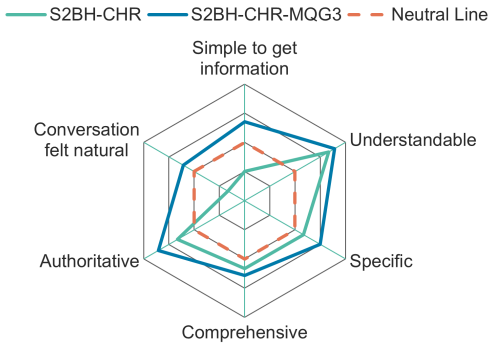
[4] Evaluation data, user testing survey, and implementation of the RAG pipeline are available at https://github.com/arjun8009/UKCP-Repo-pub.

Table 2: Automated evaluation metrics calculated for all RAG pipelines

Model Variation	Phase	Faithfulness Mean Score (Sources)	Relevance Mean Score	Coverage Mean Score	Answers with Coverage>ChatGPT	ChatGPT Mean Coverage
F10	1	0.90	0.93	0.34	79.00%	0.26
P10	1	0.89	0.94	0.34	78.00%	0.26
F5	1	0.89	0.93	0.32	72.00%	0.26
P5	1	0.86	0.93	0.32	77.00%	0.26
H20	2	0.81	0.93	0.35	78.00%	0.26
S2BH15	2	0.91	0.94	0.36	85.00%	0.26
S2BH20	2	0.92	0.94	0.36	83.00%	0.26
S2B	2	0.92	0.94	0.34	81.00%	0.26
S2BH-CHR	3	0.93	0.96	0.38	87.00%	0.26
S2BH-DVR	3	0.91	0.94	0.34	84.00%	0.26
S2BH-LST	3	0.90	0.93	0.36	82.00%	0.26
S2BH-REC	3	0.92	0.95	0.35	86.00%	0.26
S2BH-CHR-MQG3	4	0.91	0.90	0.32	81.00%	0.26
S2BH-CHR-MQG5	4	0.92	0.90	0.36	77.00%	0.26



(a) Average human ratings of answer quality (n=10 participants). Scaled from strong disagree (inner) to strong agree (outer).



(b) Average human ratings of interaction quality (n=10 participants). Scaled from strong disagree (inner) to strong agree (outer).

Figure 3: Human evaluation of answer quality and interaction quality for RAG pipelines SB2H-CHR and SB2H-CHR-MQG3.

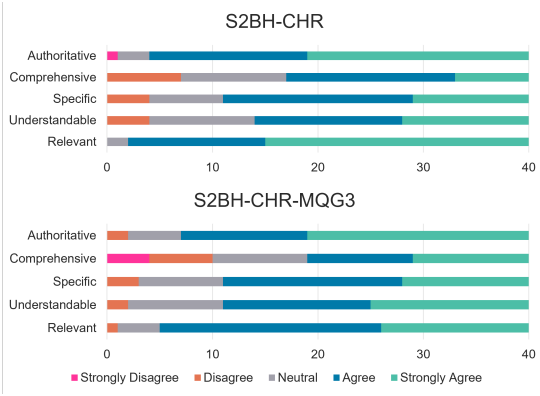
alongside the original human answer. Participants then used a standard Likert scale (1 - strong disagree; 2 - disagree; 3 - neutral; 4 - agree; 5 - strong agree) to assess RAG-pipeline answers on five quality metrics: (Q2a) Relevant; (Q2b) Understandable; (Q2c) Specific; (Q2d) Comprehensive; and (Q2e) Authoritative. Below each Likert scale, a free text box asked participants to explain their ratings and provide additional qualitative feedback.

The second stage assessed the usability and "conversationality" of the selected pipelines. Based on their evaluations in the first stage, participants were asked to interact with their preferred RAG pipeline via an online chatbot interface (see Appendix F). Users were tasked with a realistic scenario involving the use of UKCP18 data (see Appendix E for details) and asked to retrieve relevant information from the chatbot. They then evaluated their experience using Likert scales for six usability metrics: (Q3a) Simple to get information; (Q3b) Understandable; (Q3c) Specific; (Q3d) Comprehensive; (Q3e) Authoritative; (Q3f) Conversation felt natural. Free text boxes allowed further detail to be provided.

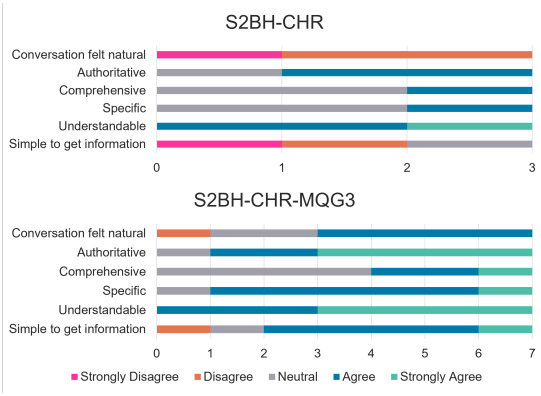
4 Results

4.1 Automated evaluation

Table 2 shows the automated evaluation metrics for all 14 RAG pipelines that were tested. All the proposed RAG models perform better than ChatGPT (GPT-3.5) in terms of the percentage of answers having a higher coverage Score, showing substantial improvements in mean scores (72 to 87 % of the answers generated by the RAG pipelines had a higher number of relevant keywords, entities and numbers). This finding is a clear validation of the RAG approach for this use case, showing



(a) Response breakdown for answer quality (n=40; 10 participants rating answers to 4 questions).



(b) Response breakdown for interaction quality (n=10 participants).

Figure 4: Human evaluation of answer quality and interaction quality for RAG pipelines SB2H-CHR and SB2H-CHR-MQG3: response breakdown

that the general-purpose LLM is unable to perform as well as RAG systems with additional domain-specific information.

From automated evaluation metrics (Table 2) we also conclude that, while small, the differences between pipelines do allow marginally better candidates to be identified. Since our four automated evaluation phases tested qualitatively different pipelines, sequentially introducing more complexity to the RAG framework, we chose the best pipeline from each phase for additional human evaluation. Evaluation phase 1 focused on the chunking strategy. We found that a higher number of chunks yields a higher coverage score and paragraph-based chunking produces better faithfulness and relevance scores. Therefore, pipeline P10 is adopted as the best candidate from phase 1. Evaluation phase 2 looked at the retrieval component. Here a combination of small-to-big and hierarchical methods gave the best outputs, so pipeline S2BH20 is chosen as the best candidate from phase 2. Phase 3 of automated evaluation considered chunk ranking approaches, with results showing that the coherence-based re-ranking strategy has the best performance. Hence pipeline S2BH-CHR is taken forward from phase 3. In Phase 4, we examined query diversification as a method for improving retrieval, finding that it boosts the faithfulness score significantly. Since both variant pipelines performed similarly, we chose S2BH-CHR-MQ3 due to its lighter computational load (few synthetic queries per answer). In this phase we observe a lower relevance score. Multiple query generation involves generating different versions of the same query and hence the generated answer contains information from various chunks that would not have been in the top 10 chunks if the original query was used. Hence the generated questions by the relevance metric can be slightly different from the original questions as the information can contain additional details. Therefore we observe a decrease in relevance. Overall we chose four pipelines for human evaluation: P10, S2BH20, S2BH-CHR, and S2BH-CHR-MQ3.

4.2 Human Evaluation

Initial screening by climate experts in the author team showed that the more complex RAG pipelines identified by automated evaluation phases 3 and 4 (S2BH-CHR and S2BH-CHR-MQG3) outperformed the simpler pipelines from phases 1 and 2 (P10 and S2BH20). Therefore S2BH-CHR and S2BH-CHR-MQG3 were further evaluated by the panel of subject matter experts.

The first stage of human evaluation by our panel of subject matter experts considered the quality of answers provided by the two RAG pipelines for four authentic questions received by the UKCP helpdesk. Figure 3a and results breakdown in Figure 4a show that for both pipelines, participants agreed that answers were authoritative, comprehensive, specific, understandable, and relevant. The weakest aspect across both pipelines was the comprehensiveness of answers. While pipeline S2BH-CHR appears to marginally outperform pipeline S2BH-CHR-MQG3 on answer quality, 7/10 users said that overall they preferred the responses generated by S2BH-CHR-MQG3.

The second stage of human evaluation by the panel asked users to interact with their preferred

chatbot to complete a typical task related to the UKCP data archive. Results are shown in Figure 3b and Figure 4b. Participants using pipeline S2BH-CHR (n=3; 30%) reported negative (worse than neutral) outcomes for two performance criteria. It was hard to get the information they needed and the conversation felt unnatural. It should be noted that the number of users testing this pipeline was small and outcomes may be unreliable. Participants using the more popular pipeline S2BH-CHR-MQG3 (n=7; 70%) all reported positive outcomes on all criteria, but conversationality and simplicity of getting information were again the weakest aspects. Overall, across both pipelines tested performance was generally positively rated, with S2BH-CHR-MQG3 receiving stronger ratings on interaction quality.

The free text boxes in the user surveys gave some useful qualitative feedback. Users reported that the perceived weakness around "conversationality" arose from the repetition of phrases, which made the chatbot feel artificial. Broader questions were seen to be more successfully answered than specific questions; one user commented that "The chatbot did not have access to the underlying data, just a headline message. This made answers vague and less authoritative." While not all users were able to find the exact information they needed, they were impressed by the chatbot's ability to suggest relevant topics that fell slightly outside the initial scope of the questions they had posed. Furthermore, there were some areas of information where users reported not receiving information that they expected and knew to exist. These user comments provide areas for future improvement of the RAG-based chatbot.

5 Conclusion

In this work, we develop an LLM-based RAG framework with systematic evaluation to create a tool (the UKCP Chatbot) to increase access and understanding of complex climate information. A heuristic phased design approach was utilized to identify the optimal design for the RAG system, with evaluation of multiple recently reported strategies for chunking, retrieval, re-ranking, and query expansion. This process was complemented by two-stage automated and human evaluation. The best pipeline was identified as S2BH-CHR-MQG3 (see Table 2 and Figure 3). The resulting chatbot provides accurate and trustworthy information from the UKCP archive.

Limitations

Two main limitations of our RAG-based system were identified. First, a *lack of conversational ability* was observed during human evaluation. Due to the amount of retrieved information and relatively large size of generated answers, earlier portions of the conversation history were pruned to reduce context length, making the chatbot "forget" past questions and answers; this made it less conversational. Second, *answer completeness* is another possible weakness. Results from the automated coverage metric and human evaluation both indicate that answers provided by the chatbot, while normally correct, are in some cases incomplete. In some instances, the retrieved information may be comprehensive, but the LLM might fail to incorporate it all into the summary response. In other cases, retrieval may omit portions of relevant information. These identified limitations call for further research on RAG systems to improve conversational ability and answer completeness, without compromising the trustworthiness/accuracy of outputs.

In future work, we will explore the use of multimodal RAG frameworks, since the UKCP18 archive is originally a multimodal database that includes reports, images, maps, and raw climate data. We also aim to refine our testing methodology with new metrics to account for factual accuracy. Also, human evaluation in this study focused on a small number of subject matter experts; in future, we aim to extend the evaluation to a more diverse set of user groups and gain more comprehensive insights into the performance of the chatbot.

Ethics Statement

We do not identify any ethical issues for this exploratory study. The UKCP18 archive is available to the public (https://www.metoffice.gov.uk/research/approach/collaboration/ukcp/data/index), published under the Open Government Licence (https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). The UKCP Chatbot is a prototype and not publicly available. It is currently undergoing internal evaluation at the Met Office. Before any public release, it will be thoroughly assessed by a wider

stakeholder group and subject to further ethical and governance review.

References

ADEPT. 2019. Preparing for a changing climate: Good practice guidance for local government. https://www.adeptnet.org.uk/system/files/documents/Good%20Practice%20Guide%20ADEPT%202019f.pdf.

Anglian Water. 2020. Anglian water's climate change adaptation report 2020. https://www.anglianwater.co.uk/SysSiteAssets/household/in-the-community/climate-change-adaptation-report-2020.pdf. Accessed: 2024-12-05.

CSIRO and Bureau of Meteorology. 2015. Climate change in australia information for australia's natural resource management regions: Technical report.

deepset. 2025a. LostInTheMiddleRanker. https://docs.haystack.deepset.ai/docs/lostinthemiddleranker. Accessed: 2025-02-26.

deepset. 2025b. SentenceTransformersDiversityRanker . https://docs.haystack.deepset.ai/docs/sentencetransformersdiversityranker. Accessed: 2025-02-26.

Department for Environment, Food and Rural Affairs. 2024. Accounting for the effects of climate change: Supplementary green book guidance. https://assets.publishing.service.gov.uk/media/6645e47e993111924d9d3655/Accounting_for_the_effects_of_climate_change.pdf. Accessed: 2024-12-05.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Environment Agency. 2024. Flood risk assessments: climate change allowances. https://www.gov.uk/guidance/flood-risk-assessments-climate-change-allowances. Accessed: 2024-12-04.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

A.M. Fischer, K.M. Strassmann, M. Croci-Maspoli, A.M. Hama, R. Knutti, S. Kotlarski, C. Schär, C. Schnadt Poberaj, N. Ban, M. Bavay, U. Beyerle, D.N. Bresch, S. Brönnimann, P. Burlando, A. Casanueva, S. Fatichi, I. Feigenwinter, E.M. Fischer, M. Hirschi, M.A. Liniger, C. Marty, I. Medhaug, N. Peleg, M. Pickl, C.C. Raible, J. Rajczak, O. Rössler, S.C. Scherrer, C. Schwierz, S.I. Seneviratne, M. Skelton, S.L. Sørland, C. Spirig, F. Tschurr, J. Zeder, and E.M. Zubler. 2022. Climate scenarios for switzerland CH2018 – approach and implications. *Climate Services*, 26:100288.

Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. Unlearning climate misinformation in large language models. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 178–192, Bangkok, Thailand. Association for Computational Linguistics.

Global Framework for Climate Services. 2025. What are climate services? https://gfcs.wmo.int/what-are-climate-services. Accessed: 2025-02-25.

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

Intergovernmental Panel On Climate Change (IPCC). 2023. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1 edition. Cambridge University Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Jason A. Lowe, Dan Bernie, Philip Bett, Lucy Bricheno1, Simon Brown, Daley Calvert, Robin

Clark, Karen Eagle, Tamsin Edwards2, Giorgia Fosser, Fai Fung, Laila Gohar, Peter Good, Jonathan Gregory, Glen Harris, Tom Howard, Neil Kaye, Elizabeth Kendon, Justin Krijnen, Paul Maisey, Ruth McDonald, Rachel McInnes, Carol McSweeney, John F.B. Mitchell, James Murphy, Matthew Palmer, Chris Roberts, Jon Rostron, David Sexton, Hazel Thornton, Jon Tinker, Simon Tucker, Kuniko Yamazaki, and Stephen Belcher. 2018. UKCP18 science overview report.

Met Office. 2025. Ukcp archive. https://www.metoffice.gov.uk/research/approach/collaboration/ukcp. Accessed: 2025-02-25.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1).

Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756, Singapore. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Sylvie Shi and Nils Reimers. 2024. Introducing rerank 3: A new foundation model for efficient enterprise search & retrieval. https://cohere.com/blog/rerank-3. Accessed: 2024-12-05.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

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

Nicolas Webersinke. 2022. ClimateBERT: A pretrained language model for climate-related text. In *AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

A UKCP corpus pre-processing: extraction, segmentation, cleaning

Data extraction. Many UKCP documents are in PDF format with complex layouts, figures, tables, and multi-column text. Automatic text extraction often produced outputs that were fragmented or out of order. Inconsistent formatting styles made it difficult to develop a single automated extraction process that maintained the integrity of content and structure or correctly extracted captions for images and tables. Careful manual checking and intervention were used to correct formatting issues, remove irrelevant data (e.g., page numbers), and ensure content integrity.

Document segmentation (chunking). Four datasets were created by segmenting each document using different chunking approaches: *fixed-length*, *paragraph*, *section*, and *summary* methods. Each "chunk" is derived from one of the original UKCP documents, in a size/format that an LLM can effectively process. Each chunk also includes metadata specifying the originating UKCP document name, page, and section from which it was sourced. The details of each chunking method are given below.

Fixed-Length: Chunks have a fixed length of 1,000 characters. This is efficient but does not account for semantic/structural boundaries within document content. Chunks are often cut off mid-paragraph, leading to incomplete representations of topics/ideas.

Paragraph: Chunks represent each paragraph within the corpus. This preserves the natural semantic boundaries within documents and can potentially give more meaningful retrieval results. Each chunk varies in length.

Section: Chunks represent each section within the corpus (defined as content given under a single heading). This preserves continuity between adjacent paragraphs and might improve retrieval quality by delivering larger chunks of related content. Each chunk varies in length.

Summary: Each chunk is a LLM-generated summary of a UKCP document created using a two-step approach. Firstly, each section of the document is summarised by an LLM (here GPT-3.5) to extract key points. Secondly, all section summaries are combined (by the same LLM) into a single cohesive summary for the entire document. Each chunk varies in length.

Data cleaning. Several processes were used to ensure the quality and consistency of the extracted chunks: (i) Removal of irrelevant or extraneous elements such as page numbers, footnotes, and headers; (ii) Correction of text extraction errors, such as erroneous characters; (iii) Correction of image/table captions and their linking to corresponding visual/numerical content. Final datasets were manually checked to rectify any remaining inconsistencies.

Data representation. The embedding model used to represent the query and chunk was "text-embedding-ada-002", if not otherwise specified in the main text of the paper.

B Prompts

Below are the prompts for generating multiple queries for query expansion (prompt-1) and for generating answers (prompt-2).

Input prompt-1 for generating multiple queries

query system instruction =
Instructions:
1. You will be provided with a question from the user.
2. Your task is to generate multiple search queries related to this input question.
3. You must maintain the context of the original question and you must not exclude any key information from the question.
4. Phrase each query in a different way, but ensure that you do not deviate from the original meaning of the question.
5. Output <NUMBER> new queries in the form of a list. Do not deviate from this format.
Follow these steps before providing your final response:
Step 1: Take your time to thoroughly understand the provided question.
Step 2: Generate your new queries, ensuring that each new query is written in a distinctly different way to each other query.
Step 3: Reason step-by-step about whether the all of the key information from the original question can be found in each of the new queries. If there is key information found in the original question which cannot be found in any given new query, then you must replace this query by generating a new one. You must then follow these steps again.
Step 4: You may provide your final generated queries to the user. Do not output anything else.
query user prompt = QUESTION:<QUESTION>

Input prompt-2 for generating answers

system instruction = Instructions:
1. You are an expert on United Kingdom Climate Projections (UKCP). UKCP is a set of tools and data that demonstrates how the UK climate may change in the future.
2. UKCP18 is a set of climate model projections for the UK produced by the Met Office. It builds upon the previous set of projections (UKCP09) to provide the most up-to-date assessment of how the climate of the UK may change over the 21st century.
3. You will be provided with a question from the user, for which you will attempt to find the answer.
4. You will be provided with excerpts which are sourced exclusively from UKCP18 literature.
5. You MUST read all of the excerpts to understand the context for answering the question.
6. You will provide an EXPERT-LEVEL written response which comprehensively answers the question, using only information from the provided excerpts.
7. You should assume that you do not have access to any other sources of information.
8. UNDER NO CIRCUMSTANCES should you use information from any other source (such as the internet) to generate your responses.
9. The response you provide will be cross-checked with the excerpts provided to you. If there is information within the response which is not found in the excerpts, you will lose credibility.
10. You will be provided with the chat history of the conversation in your messages. You must follow the chat history to understand the context of the conversation.
11. If you cannot answer the question using information from the excerpts, you may ask once for more information from the user. If this additional information does not help you to find the answer from the excerpts, gently respond that you are unsure about the answer and recommend that they contact the Met Office's UKCP help desk.
12. You must not repeat or summarize the question which was asked to begin your response. Only respond with the answer, request for more information, or the statement that you cannot answer the question.
Follow these steps before providing your final response:
Step 1: Take your time to thoroughly understand the provided question.
Step 2: Take your time to thoroughly understand the provided excerpts, which are delimited by the following token: <SEP>.
Step 3: Generate an expert-level written response which comprehensively answers the question using only the excerpts provided. If you are unable to create a response that comprehensively answers the question using the provided excerpts, ask the user once for more information. If this additional information does not help you to find the answer from the excerpts, gently respond that you are unsure about the answer, recommend that they contact the Met Office's UKCP help desk and stop following these steps.
Step 4: Reason step-by-step about whether the all of the information in the response can be found in excerpts provided. If there is information found which cannot be found in the excerpts, then you must generate a new response and follow these steps again for the new response.
Step 5: You may provide your response to the user.
user prompt = EXCERPTS: <EXCERPTS>
QUESTION: <QUESTION>
ANSWER:

C Evaluation data creation

Synthetic QCA triplets. A dataset of question-context-answer (QCA) triplets was synthesized using the RAGAS package (Es et al., 2024), which takes contextual documents as input and uses an LLM to generate derived question-and-answer pairs. RAGAS can generate several types of questions. Three types of questions were created for this dataset. *Simple* questions are intended to be straightforward to

answer using the given context. *Reasoning* questions re-write a simple question such that reasoning is needed to answer it effectively. *Multi-context* questions re-phrase a simple question such that information from multiple context sections is needed to formulate an answer. RAGAS outputs are question-context-answer triplets. For this study, RAGAS was parameterized using GPT-4 as the LLM and the *section* chunks as content. A sample of 500 section chunks was randomly split into groups of 5, and then 10 question-answer pairs were generated for each section chunk, using a **1:2:2** ratio for simple, reasoning, and multi-context question types. The resulting dataset of 1,000 QCA triplets was sampled for 250 QCA triplets used for evaluation.

Examples :

Question : "Which UKCP18 model better represents Scotland's winter snow variability?"

Answer : The CPM better represents Scotland's winter snow variability, particularly in terms of lying snow and snowfall over the Scottish mountains.

Question : "What's PoT's role in estimating rare climate events?"

Answer : The PoT (peaks over threshold) method involves using all events exceeding a specified threshold in a given season, thus considering more of the data, and avoiding the risk of missing multiple extremes that may occur in close proximity. It also excludes any seasons which happen not to contain any extreme events.

Authentic QA pairs. A dataset of 50 question-answer (QA) pairs was derived from real questions received by the UKCP helpdesk and the answers provided by subject matter experts. The QA pairs were anonymized, cleaned and formatted, and manually selected to represent a diverse range of typical questions. The authentic QA pairs lacked contexts and were only used for human evaluation where the subject matter experts decided the correctness of the extracted contexts and the answers. (Examples in the github link under human evaluation survey form)

D Evaluation metrics: Relevance, Faithfulness, and Coverage

We follow the work RAGAS (Es et al., 2024) to use LLMs to measure the answer relevance and answer faithfulness. We further propose a metric to measure answer coverage. The detailed metric settings are described below.

Relevance. This metric measures the relevance of the answer to the user query by an inverse method, using an LLM (GPT-4) to create alternate synthetic questions that could generate the answer and then measure their (cosine) similarity to the original user query. Mathematically, the metric is found as: $\texttt{relevance\_score}(g_i, q) = \frac{1}{N}\sum_{i=1}^{N} \texttt{cos}(E_{g_i}, E_q)$ where E_x is the embedding of a generated question g_i or the original query q, and $N = 3$ is the number of generated questions.

Faithfulness. This metric measures the extent that an answer uses only information that is contained in the chunks given as context. An evaluator LLM (GPT-3.5) is used to identify the sets of factual claims that are made in the provided answer and in the context chunks. Then the metric is defined as: $\texttt{faithfulness\_score} = \frac{|C_{\texttt{answer}}|}{|C_{\texttt{context}}|}$, where C_x is the set of claims present in either the answer or the context chunks.

Coverage. Accurate answers to scientific questions (as here in the climate domain) often require a high level of specificity and detail. This implies usage and adherence to numerical values, proper names, keywords and other entities. Here we propose a new *coverage* metric, which calculates the proportion of all the named entities, keywords, and numerical values from the context chunks that are given in a generated answer. Identification of entities, keywords, and numbers was performed using the trained model "en-core-web-sm"[5] in the SpaCy NLP package, and additional terms were identified using an LLM (GPT-4). All proper nouns and adverbs are considered as keywords. Coverage is then defined by: $\texttt{coverage\_score} = |\frac{|K_{\texttt{answer}} \bigcap K_{\texttt{context}}|}{|k_{\texttt{context}}|}$ where K_x is the set of all keywords, named entities, and numbers in either the context chunks or the answer.

[5] https://spacy.io/models/en#en_core_web_sm

E Human evaluation settings

Initial screening. The questions from the 50 authentic QA pairs dataset were posed to each RAG pipeline, then subject matter experts reviewed each response to evaluate its quality and validity. The two best-performing pipelines were determined based on subjective evaluations of answer correctness, framing, style and specificity. Responses from these selected pipelines were further screened to curate a pool of answers that could be easily reviewed without extensive contextual information. Four questions with answer pairs were selected.

Interactive evaluation. A panel of subject matter experts evaluated the two remaining RAG pipelines in a two-stage process conducted by survey and online access to the chatbot.

Panel recruitment. The panel (n=10) was recruited from Met Office staff to ensure a good baseline understanding of climate science. Within this group, there was a range of experience, with 4/10 panelists having over 10 years of experience using UKCP18 data.

User task for human evaluation of UKCP Chatbot Users were given access to the chatbot via an online user interface and asked to complete the following task within 30 minutes: *'Task - The Ministry of Defence (MoD) needs to construct 30 large buildings by 2030 in various locations around the UK coastline. The MoD would like to ensure that the buildings are suitably prepared to stay cool in the future. Use the chatbot to find relevant information and try to achieve this task.*'

F UKCP ChatBot interface

A screenshot of the user interface of the UKCP Chatbot is presented in Figure 5 below.

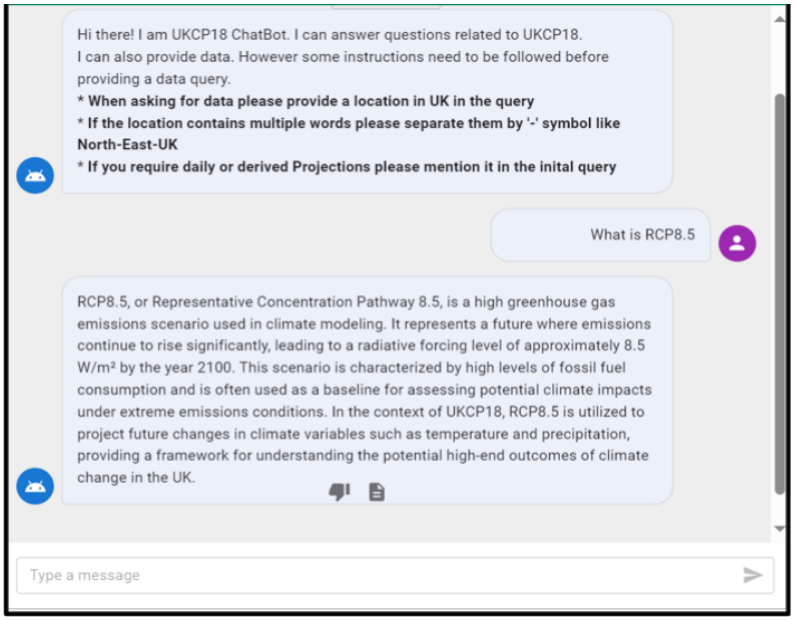


Figure 5: User Interface of the UKCP Chatbot