ARTICLE

Check for updates

ChatClimate: Grounding conversational AI in climate science

Saeid Ashraf Vaghefi [1,2,3,4 ✉], Dominik Stammbach [5], Veruska Muccione [2,6], Julia Bingler[7,8], Jingwei Ni[1,5], Mathias Kraus[9], Simon Allen [2,10], Chiara Colesanti-Senni[1], Tobias Wekhof [1,11], Tobias Schimanski[1], Glen Gostlow[1], Tingyu Yu[1], Qian Wang[1], Nicolas Webersinke[9], Christian Huggel [2] & Markus Leippold [1,12]

Large Language Models have made remarkable progress in question-answering tasks, but challenges like hallucination and outdated information persist. These issues are especially critical in domains like climate change, where timely access to reliable information is vital. One solution is granting these models access to external, scientifically accurate sources to enhance their knowledge and reliability. Here, we enhance GPT-4 by providing access to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6), the most comprehensive, up-to-date, and reliable source in this domain (refer to the 'Data Availability' section). We present our conversational AI prototype, available at www.chatclimate.ai, and demonstrate its ability to answer challenging questions in three different setups: (1) GPT-4, (2) ChatClimate, which relies exclusively on IPCC AR6 reports, and (3) Hybrid ChatClimate, which utilizes IPCC AR6 reports with in-house GPT-4 knowledge. The evaluation of answers by experts show that the hybrid ChatClimate AI assistant provide more accurate responses, highlighting the effectiveness of our solution.

[1] Department of Banking and Finance, University of Zürich, Zürich, Switzerland. [2] Department of Geography, University of Zürich, Zürich, Switzerland. [3] Eawag: Swiss Federal Institute of Aquatic Science, Dübendorf, Switzerland. [4] WMO: World Meteorological Organization, Geneva, Switzerland. [5] Center for Law and Economics, , ETH Zürich, Zürich, Switzerland. [6] Swiss Federal Research Institute WSL, Birmensdorf, Switzerland. [7] Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK. [8] Council on Economic Policies, Zürich, Switzerland. [9] Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany. [10] Institute for Environmental Sciences, University of Geneva, Geneva, Switzerland. [11] Center of Economic Research ETH Zürich, Zürich, Switzerland. [12] Swiss Finance Institute (SFI), Zürich, Switzerland. ✉email: saeid.vaghefi@geo.uzh.ch

Motivation. Large pre-trained Language Models (LLMs) have emerged as the de facto standard in Natural Language Processing (NLP) in recent years. LLMs have revolutionized text processing across various tasks, bringing substantial advancements in natural language understanding and generation[1–4]. Models such as LLaMA[5], T0[6], PaLM[7], GPT-3[1], or instruction fine-tuned models, such as ChatGPT[8], GPT-4[9] and HuggingGPT[10], have demonstrated exceptional capabilities in generating human-like text across various domains, including language translation, summarization, and question answering and have become a norm in many fields[11].

LLMs also excel at closed-book Question Answering (QA) tasks. Closed-book QA tasks require models to answer questions without any context[12]. LLMs like GPT-3/3.5 have achieved impressive results on multiple choice question answering (MCQA) tasks in the zero, one, and few-shot settings[13]. Recent works have used LLMs such as GPT-3[1] as an implicit knowledge base that contains the necessary knowledge for answering questions[14].

However, LLMs suffer from two major issues: hallucination[15] and outdated information after training has concluded[16]. These issues are particularly problematic in domains such as climate change, where it is critical to have accurate, reliable, and timely information on changes in climate systems, current impacts, and projected risks of climate change and solution space. Hence, providing accurate and up-to-date responses with authoritative references and citations is paramount. Such responses, if accurate, can help understand the scale and immediacy of climate change and facilitate the implementation of appropriate mitigation strategies.

Enhanced communication between government entities and the scientific community fosters more effective dialogue between national delegations and policymakers. A facilitated chat-based assisted feedback loop can be established by guaranteeing the accuracy of information sources and responses. This feedback loop promotes informed decision-making in relevant domains. For example, governments may ask a chatbot for feedback on specific statements in the report or request literature to support a claim. The importance of accurate and up-to-date information has been highlighted in previous studies as well[17–19].

By overcoming the challenges of outdated information and hallucination, LLMs can be used to extract relevant information from large amounts of text and assist in decision-making. Training LLMs is computationally expensive and has other negative downsides (see, e.g. [20,21]). To overcome the need for continuous training, one solution is to provide the LLMs with external sources of information (called long-term memory). This memory continuously updates the knowledge of an LLM and reduces the propagation of incorrect or outdated information. Several studies have explored the use of external data sources, which makes the output of LLMs more factual[22].

Contributions. In this paper, we introduce our prototype, ChatClimate (www.chatclimate.ai), a conversational AI designed to improve the veracity and timeliness of LLMs in the domain of climate change by utilizing the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (hereafter IPCC AR6)[23–26]. These reports offer the latest and most comprehensive evaluation of the climate system, climate change impacts, and solutions related to adaptation, mitigation, and climate-resilient development. Please refer to the 'Data Availability' section for a detailed list of the IPCC AR6 reports used in this study. We evaluate the LLMs' performance in delivering accurate answers and references within the climate change domain by posing 13 challenging questions to our conversational AI (hereafter chatbot)

across three scenarios: GPT-4, ChatClimate standalone, and a hybrid ChatClimate.

Findings. Our approach can potentially supply decision-makers and the public with trustworthy information on climate change, ultimately facilitating better-informed decision-making. Our approach underscores the value of integrating external data sources to enhance the performance of LLMs in specialized domains like climate change. By incorporating the latest climate information from the IPCC AR6 into LLMs, we aim to create models that provide more accurate and reliable answers to questions related to climate change. While our tool is effective in making complex climate reports more accessible to a broader audience, it is crucial to clarify that it does not aim to replace or engage in decision-making, either general or bespoke. The tool serves solely as a supplementary resource that helps distill and summarize key information, thereby supporting, but not substituting for, the complex and multifaceted process of informed decision-making on climate issues. Making these reports more accessible can contribute to the design of more effective policies. For example, easier understanding of worst-case scenarios can enable more targeted actions to prevent them. In summary, ChatClimate aims to make the information more accessible and the review process more efficient, without overstepping into the domain of policy or decision-making.

Results and Discussion

Chatbots and questions. We conducted three sets of experiments by asking hybrid ChatClimate, ChatClimate, and GPT-4 chatbots 13 questions (Table 1). Our team of IPCC AR6 authors then assessed the answers' accuracy. It is worth noting that our prototype's ability to provide sources for statements can facilitate the important process of trickle-back, which governments and other stakeholders often require in the context of IPCC reports.

Table 2 presents the returned answers from chatbots. The question "Is it still possible to limit warming to 1.5 °C?" targets mitigation, and the hybrid chatbot and ChatClimate explicitly return the greenhouse gas emission reduction amounts and time horizon while the GPT-4 answer is more general. All chatbots give a range of 2030 to 2052 as an answer to the question 'When will we reach 1.5 °C?'. Hybrid chatbot and GPT-4 add to the answers that reaching 1.5 °C depends on the emission pathways.

Evaluation of answers (accuracy score). Several studies focus on human-chatbot interaction effectiveness[27–30]. Evaluation involves factors such as relevance, clarity, tone, style, speed, consistency, personalization, error handling, and user satisfaction. This work, however, only examines the chatbot's performance on accuracy.

Expert Cross-Check of the Answers. Overall, the responses provided by the hybrid ChatClimate were more accurate than those of ChatClimate standalone and GPT-4. For the sake of brevity, we have provided a detailed analysis of Q1 and Q2 in Table 2 and only highlight the key issues for Q3–Q13. For instance, in Q1, we asked the bots whether it is still possible to limit global warming to 1.5 °C. Both the hybrid ChatClimate and ChatClimate systems referred to the amount of CO_2 that needs to be reduced over different time horizons to stay below 1.5 °C. However, GPT-4 provided a more general response. To verify the accuracy of responses generated by the ChatClimate bots, we cross-checked the references provided by both systems. We found that the ChatClimate bots consistently provided sources for their statements, as shown in Figs. 1 and 2, which is essential for verifying the veracity of the bot's answer. In Q2, we asked the bots about the time horizons when society may reach 1.5 °C. All three bots similarly referred to the time horizon of 2030 to 2052 based on the mitigation measures we take into

Table 1 The 13 designed questions for running the three sets of experiments.

Questions	Difficulty (1–5)
	1:very-low, 2:low, 3:medium, 4:high, 5:very-high
Q1= "Is it still possible to limit warming to 1.5°C?"	3
Q2= "When will we reach 1.5°C?"	3
Q3= "What does overshoot mean?"	1
Q4= "Can we avoid overshooting 1.5°C?"	3
Q5= "Have emissions reductions fallen for some countries?"	3
Q6= "What are the issues with financing adaptation?"	4
Q7= "Where is the majority of climate finance going?"	4
Q8= "What are the options for scaling up adaptation and mitigation in developing countries?"	4
Q9= "Which regions will be disproportionally affected by climate change?"	5
Q10= "What is climate justice?"	2
Q11= "What is maladaptation?"	4
Q12= "Is there evidence of maladaptation?"	5
Q13= "Will glaciers in Scotland melt?"	5

Table 2 Comparison of generated answers to questions 1 and 2 from hybrid ChatClimate, ChatClimate, and GPT-4.

QAs comparison

hybrid ChatClimate	ChatClimate	GPT-4
Q1) Is it still possible to limit warming to 1.5°C?		
Yes, it is still possible to limit global warming to 1.5 °C, but it would require immediate, deep, rapid, and sustained reductions in greenhouse gas emissions, including CO_2 and methane (Page:31, IPCC AR6 WGIII Chapter03). Scenarios limiting warming to 1.5 °C with no or limited overshoot show that net energy system CO_2 emissions must fall by 35–51% by 2030 and 87–97% by 2050 (Page:43, IPCC AR6 WGIII Technical Summary). Net electricity sector CO_2 emissions would need to reach zero globally between 2045 and 2055 (Page:3, IPCC AR6 WGIII Chapter06). However, the likelihood of limiting warming to 1.5 °C with no or limited overshoot has decreased compared to the Special Report on Global Warming of 1.5 °C due to the rise in global greenhouse gas emissions since its publication (Page:4, IPCC AR6 WGIII Chapter03). Only a small number of published pathways achieve this target with-out overshooting over the 21st century (Page:31, IPCC AR6 WGIII Technical Summary). In summary, while still possible, limiting warming to 1.5 °C would require immediate and drastic action to reduce	It is still possible to limit warming to 1.5 °C with no or limited overshoot, but the likelihood has decreased since the Special Report on Global Warming of 1.5 °C (SR1.5) due to the rise in global GHG emissions since it was published. This has led to higher near- term emissions (2030) and higher cumulative CO_2 emissions until the time of net zero (medium confidence). Limiting warming to 1.5 °C requires deep, rapid, and sustained reductions in greenhouse gas emissions, including methane, alongside rapid reductions of CO_2 emissions to net zero IPCC AR6 WGIII Chapter03, Page:4, IPCC AR6 WGIII Technical Summary, Page:31.	Limiting global warming to 1.5 °C is becoming increasingly challenging due to the continued rise in greenhouse gas emissions and the slow pace of implementing effective climate policies. However, it is not entirely impossible. Achieving this goal would require rapid and far-reaching transitions in energy, land, urban infrastructure, and industrial systems. This includes drastically reducing greenhouse gas emissions, increasing renewable energy sources, improving energy efficiency, and adopting sustainable land use practices. The Intergovernmental Panel on Climate Change (IPCC) has outlined various pathways to achieve this target, but it requires immediate and ambitious action from governments, businesses, and individuals worldwide.
Accuracy: 5	**accuracy: 5**	**accuracy: 3**
Q2) When will we reach 1.5 °C?		
We are likely to reach 1.5°C of global warming be- tween 2030 and 2052 if the current rate of warming continues (high confidence) (IPCC AR6, Reference: IPCC AR6 WGI Chapter01, Page: 44). However, the exact timing may vary depending on the emissions re- duction pathways and efforts to achieve net-zero CO_2 emissions (In-house knowledge).	We are likely to reach 1.5 °C of global warming between 2030 and 2052 if it continues to increase at the current rate (high confidence). Reference: IPCC AR6 WGI Chapter01, Page: 44.	As an AI, I cannot predict the exact year when global temperatures will reach 1.5 °C above pre-industrial levels. However, according to the Intergovernmental Panel on Climate Change (IPCC), we could reach this threshold between 2030 and 2052 if global warming continues at its current rate. The actual timeline depends on various factors, including greenhouse gas emissions, climate policies, and global efforts to mitigate climate change.
accuracy: 5	**accuracy: 4**	**accuracy: 2**

account. The consistency of the answers shows that this time horizon has been mentioned in the training data of GPT-4 as well (e.g., IPCC AR6 WGI, which was released in August 2021 or Special Report of IPCC on Global Warming of 1.5 ºC, 2018).

Impact of prompt engineering on the answers. Prompting is a method for guiding the LLMs toward desired outputs[6,31]. To achieve the best performance of LLMs in NLP tasks, proper design of prompts is essential.

IPCC AR6, WGIII Chapter03, p31, →

Limiting warming to 1.5°C (>50%) or to 2°C (>67%) requires deep, rapid, and sustained reductions of other greenhouse gases including methane alongside rapid reductions of CO₂ emissions to net zero. This ensures that the warming contributions from non-CO₂ forcing agents as well as from CO₂ emissions are both limited at low levels. The AR6 WGI estimated remaining carbon budgets until the time of reaching net zero CO₂ emissions for a range of warming limits, taking into account historical CO₂ emissions and projections of the warming from non-CO₂ forcing agents (Box 3.4 in Section 3.3, AR6 WGI Section 5.5).

Executive Summary

Warming cannot be limited to well below 2°C without rapid and deep reductions in energy system carbon dioxide (CO₂) and greenhouse gas (GHG) emissions. In scenarios limiting warming to 1.5°C (>50%) with no or limited overshoot (2°C (>67%) with action starting in 2020), net energy system CO₂ emissions (interquartile range) fall by 87–97% (60–79%) in 2050. In 2030, in scenarios limiting warming to 1.5°C (>50%) with no or limited overshoot, net CO₂ and GHG emissions fall by 35–51% and 38–52% respectively. In scenarios limiting warming to 1.5°C (>50%) with no or limited overshoot (2°C (>67%)), net electricity sector CO₂ emissions reach zero globally between 2045 and 2055 (2050 and 2080). (*high confidence*) {6.7}

IPCC AR6, WGIII Chapter06, p3 ←

Many pathways in the literature show how to limit global warming compared to pre-industrial times to 2°C (>67%) with no overshoot or to limit warming to 1.5°C (>50%) with no or limited overshoot. The likelihood of limiting warming to 1.5°C with no or limited overshoot has dropped in AR6 compared to the *Special Report on Global Warming of 1.5°C* (SR1.5) because global GHG emissions have risen since the time SR1.5 was published, leading to higher near-term emissions (2030) and higher cumulative CO₂ emissions until the time of net zero (*medium confidence*). Only a small number of published pathways limit global warming to 1.5°C without overshoot over the course of the 21st century. {3.3, Annex III.II.3}

IPCC AR6 WGIII Technical Summary, p43, →

IPCC AR6, WGIII Chapter03, p4 ←

TS.5.1   Energy

A broad-based approach to deploying energy-sector mitigation options can reduce emissions over the next ten years and set the stage for still deeper reductions beyond 2030 (*high confidence*). There are substantial, cost-effective opportunities to reduce emissions rapidly, including in electricity generation, but near-term reductions will not be sufficient to limit warming to 2°C (>67%) or limit warming to 1.5°C (>50%) with no or limited overshoot. {6.4, 6.6, 6.7}

Warming cannot be limited to 2°C or 1.5°C without rapid and deep reductions in energy system CO₂ and GHG emissions (*high confidence*). In scenarios limiting warming to 1.5°C (>50%) with no or limited overshoot (*likely* below 2°C), net energy system CO₂ emissions fall by 87–97% (interquartile range 60–79%) in 2050. In 2030, in scenarios limiting warming to 1.5°C with no or limited overshoot, net CO₂ and GHG emissions fall by 35–51% and 38–52% respectively. In scenarios limiting warming to 1.5°C with no or limited overshoot (*likely* below 2°C), net electricity sector CO₂ emissions reach zero globally between 2045 and 2055 (2050 and 2080) (*high confidence*). {6.7}
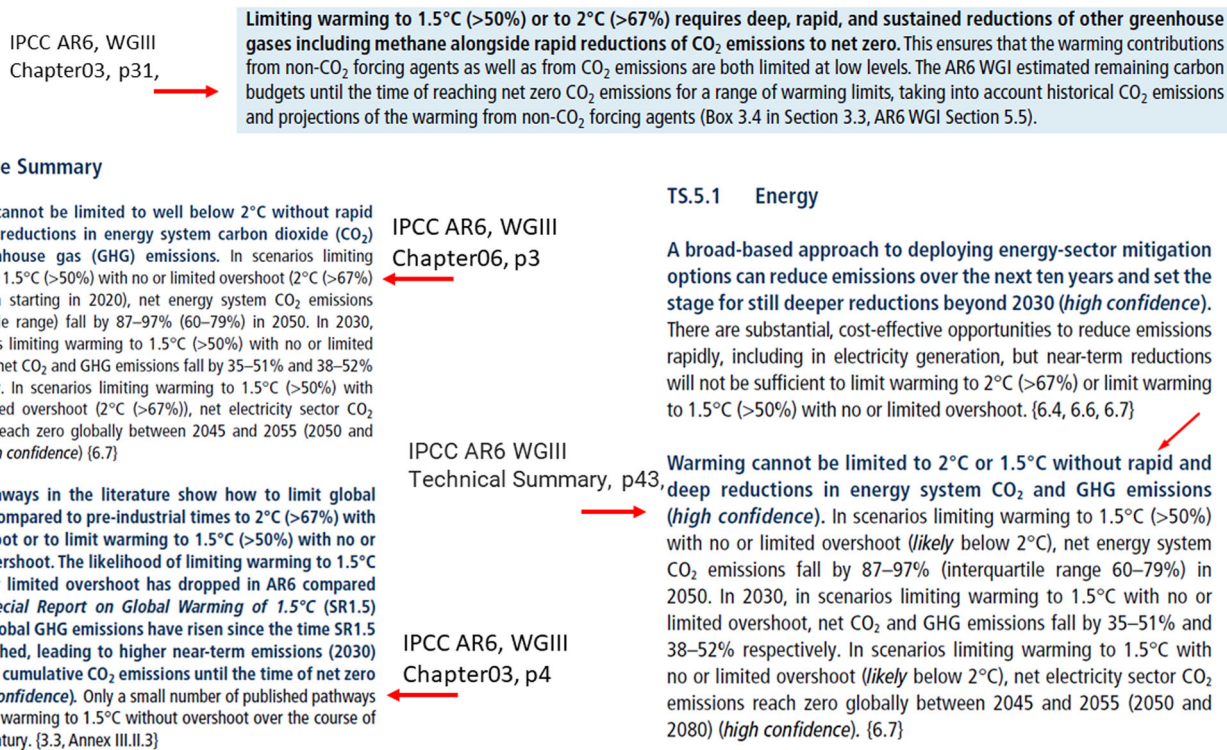
Fig. 1 Cross-checking of the references for Question 1. The red arrows show a part of the text that has been referenced. All the remaining text shows IPCC AR6 contents.
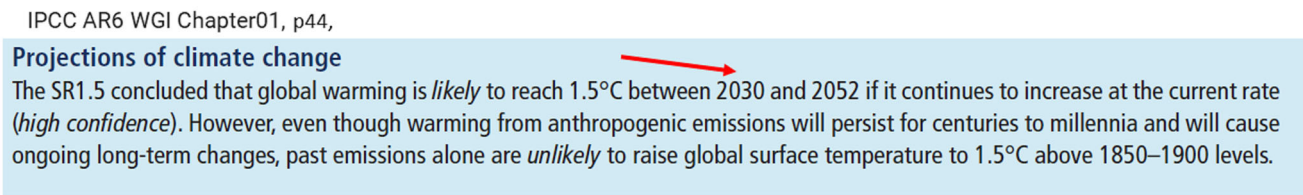
IPCC AR6 WGI Chapter01, p44,

Projections of climate change
The SR1.5 concluded that global warming is *likely* to reach 1.5°C between 2030 and 2052 if it continues to increase at the current rate (*high confidence*). However, even though warming from anthropogenic emissions will persist for centuries to millennia and will cause ongoing long-term changes, past emissions alone are *unlikely* to raise global surface temperature to 1.5°C above 1850–1900 levels.

Fig. 2 Cross-checking of the references for Question 2. The red arrows show a part of the text that has been referenced. All the remaining text shows IPCC AR6 contents.

This can be accomplished either through manual engineering[32] or automatic generation[33]. The main goal of prompt engineering is to optimize LLMs' performance across different NLP tasks[33]. To illustrate the impact of prompt engineering, we present two crafted prompts (Boxes 1 and 2) along with their corresponding retrieved answers to Question 2. These examples serve to highlight how variations in prompt design can noticeably influence the information retrieved (Table 3).

Also, we demonstrate how, by adjusting the retrieval hyperparameters, we can extract extra knowledge from long-term memory. We retrieved 10 and 15 nearest vectors to the user's query vector. With this, GPT-4 can obtain more information and provide more accurate answers when accessing the outside world. In Table 4, we present the results, demonstrating that in the K-15 scenario, where we retrieved 15 nearest vectors to the user's query vector, GPT-4 was able to access and provide more comprehensive information about the regions most affected by climate change. This underscores the importance of prompt engineering.

Expert cross-check of the QA scenarios Q3-13. We provide the detailed results of Q3-Q13 in the Supplementary Information (Table S1–S11). It is important to note that hallucination issues are more effectively addressed by the ChatClimate standalone version, which draws answers exclusively from IPCC AR6. For instance, we ask in Q13: "Will glaciers melt in Scotland?". Both GPT-4 and the Hybrid ChatClimate hallucinate. The Hybrid ChatClimate provides the following answer: Yes, it is very likely that glaciers in Scotland will melt. Glaciers are projected to continue to lose mass under all emissions scenarios (very high confidence) and this includes glaciers in various regions, such as the Arctic and Antarctic (high confidence) (IPCC AR6 WGI Chapter 08, Page: 68). Although the specific mention of Scotland is not provided in the referenced materials, the overall trend of glacier melting applies to glaciers worldwide, including those in Scotland (In-house knowledge). This response inaccurately implies the presence of glaciers in Scotland, whereas it is well-documented that Scotland does not currently have any glaciers. The country's glaciers melted many thousands of years ago, with the last ice vanishing around 14,000 years ago at the end of the last Ice Age[34]. However, the Standalone version accurately responds: There is not enough information provided to specifically answer the question about glaciers in Scotland. The primary reason for this behavior lies in the configuration of the standalone LLM to exclude the use of in-house knowledge and rely solely on the IPCC AR6 reports. Furthermore, we believe automated fact-checking methods (e.g. [35,36]) can further help address hallucinations in a post-processing step where we first perform an

---

Box 1 ▌ Input prompt-1 for hybrid ChatClimate

```
"system", "content": You are a Q&A bot, an intelligent system that answers user questions based on the information
provided by the user above the question and your in-house knowledge. There are five pieces of extra information above
the user question. The user question is in the final line. When you use the user
information, always indicate the Page and Reference in your answer. The Page and Reference are located below each
piece of information. Additionally, let us know which part of your answer is from the IPCC information and which part
is based on your in-house knowledge by writing either (IPCC AR6) or (In-house knowledge). If the information cannot be
found in the information provided by the user or your in-house knowledge, please say 'There is not enough
information'.
"role": "user", "content": External knowledge + Question
```

Box 2 ▌ Input prompt-2 for ChatClimate

```
"system", "content": As a climate scientist and specialized Q&A bot with expertise

in climate change, climate science, environmental science, physics, and energy science, your primary objective
is:

  1. Provide accurate and comprehensive answers to the user's question using the
     inputs provided by the user above the question and your in-house knowledge.

  2. When synthesizing the information above the question (and also do your best to eliminate redundant parts),
     ensure you cite the 'Reference', 'Page', and 'URL' where the information is sourced.

  3. If your answer draws from your in-house knowledge, signify this by appending (In
     -house knowledge) rather than citing a specific reference.

  4. In cases where sufficient information is lacking to address a question, reply with 'There is not enough info to
     answer the question.'

  5. It's imperative to maintain accuracy and refrain from creating information. If any aspect is unclear, do not
     create answers about that aspect.

Please proceed with the information and question provided above the question.
You will be first presented with all the extra information from a user. At the
end, you are presented with the original user question.
More specifically, with this structure: five piece of information is provided above
the user question. With this structure: some texts, \nReference:, \nPage, ORG,
\nURL:, \n\n--------------\n\n and finally user question.
ALWAYS return URL
"role": "user", "content": External knowledge + Question
```

---

automated fact-check of the chatbot response before returning it to the user.

Personalized-GPTs or GPT-n, risk management. Domain-specific chatbots and conversational AI tools provide easy access to accurate information. However, potential risks from external data sources, such as inaccuracies or biases, should be acknowledged. In this study, we develop and implement domain-specific chatbots for the climate change domain. We compare three chatbot scenarios and found that the hybrid ChatClimate provids more accurate answers to 13 sample questions. We evaluate the answers

internally, benefitting from the expert knowledge of co-authors. Since training LLMs is resource-intensive[9], integrating them with the outside world by providing long-term memory and prompt engineering could yield better results with fewer resources. However, creating long-term memory requires caution. We used the IPCC AR6 as a comprehensive and reliable source to build external memory for LLMs, highlighting the importance of such databases for chatbot accuracy. Although there is an ongoing debate about pausing LLM training for some months until proper regulations are established, we believe that regulating LLM training, fine-tuning, and incorporating it into applications is necessary. Specifically,

Table 3 Comparison of LLMs response to different prompts.

Prompts

Prompt 1	Prompt 2
You are a Q&A bot, an intelligent system that answers user questions based on the information provided by the user above the question and your in-house knowledge. There are five pieces of extra information above the user question. The user question is in the final line. When you use the user information, always indicate the Page and Reference in your answer. The Page and Reference are located below each piece of information. Additionally, let us know which part of your answer is from the IPCC information and which part is based on your in-house knowledge by writing either (IPCC AR6) or (In-house knowledge). If the information cannot be found in the information provided by the user or your in-house knowledge, please say that 'There is not enough information'.	As a climate scientist and specialized Q&A bot with expertise in climate change, climate science, environmental science, physics, and energy science, your primary objective is: 1) Provide accurate and comprehensive answers to the user's question using the inputs provided by the user above the question and your in-house knowledge. 2) When synthesizing the information above the question (and also do your best to eliminate redundant parts), ensure you cite the 'Reference', 'Page', and 'URL' where the information is sourced. 3) If your answer draws from your in-house knowledge, signify this by appending (In-house knowledge) rather than citing a specific reference. 4) In cases where sufficient information is lacking to address a question, reply with 'There is not enough info to answer the question.' 5) It's imperative to maintain accuracy and refrain from creating information. If any aspect is unclear, do not create answers about that aspect. Please proceed with the information and question provided above the question. You will be first presented with all the extra information from a user. At the end, you are presented with the original user question. More specifically, with this structure: five pieces of information are provided above the user question. With this structure: some texts, Reference:, Page, ORG, URL:, and finally, the user question. Always return URL.
Results **ChatClimate Hybrid Prompt-1** The best estimate for reaching 1.5 °C of global warming lies in the first half of the 2030s in most of the considered scenarios and modeled pathways. However, it's more likely than not to reach 1.5 °C between 2021 and 2040 even under the very low GHG emission scenarios (SSP1-1.9) (Page 33, CLIMATE CHANGE 2023 Synthesis Report Summary for Policymakers Page 18, SYNTHESIS REPORT OF THE IPCC SIXTH ASSESSMENT REPORT (AR6) Longer Report) (IPCC AR6). This is based on current projections and could change depending on various factors, including the rate of greenhouse gas emissions and the effectiveness of mitigation strategies (In-house knowledge).	**ChatClimate Hybrid Prompt-2** 1. According to the IPCC's Sixth Assessment Report, the best estimate for reaching 1.5 °C of global warming lies in the first half of the 2030s in most of the considered scenarios and modeled pathways. However, it's important to note that global warming is more likely than not to reach 1.5 °C between 2021 and 2040 even under the very low GHG emission scenarios (SSP1-1.9), and likely or very likely to exceed 1.5 °C under higher emissions scenarios. These estimates are based on current trends and could change depending on the effectiveness of mitigation and adaptation actions implemented in the near term. References: CLIMATE CHANGE 2023 Synthesis Report Summary for Policymakers, Page: 33, ORG: IPCC-SYR, URL: (https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCCAR6SYRSPM.pdf) SYNTHESIS REPORT OF THE IPCC SIXTH ASSESSMENT REPORT (AR6) Longer Report, Page: 18, ORG: IPCC-SYR, URL: (https://report.ipcc.ch/ar6syr/pdf/IPCCAR6SYRLongerReport.pdf) CLIMATE CHANGE 2023 Synthesis Report Summary for Policymakers, Page: 56, ORG: IPCC-SYR, URL: (https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCCAR6SYRSPM.pdf)
accuracy: 3.5	**accuracy: 5**

external database integration and prompt engineering should be considered in regulations for chatbots. Furthermore, training LLM models on huge amounts of data has a potentially very high carbon footprint and we have little knowledge about the carbon footprint embedded in LLMs such as GPT-4[37]. However, inference and the use of already trained LLM models is less energy intensive.

Database setup: access to different databases. With the results of ChatClimate, we show how retrieval-augmented LLMs can be updated with more recent information. However, the design of the retrieval system plays a pivotal role in the effectiveness of question-answering systems, particularly when specialized knowledge is required. To illustrate the impact of this design aspect, we scrutinized various database configurations. Generally, we constrain the retrieved information to the top-K results, which are selected based on the highest similarity metrics between the query vector and vectors sourced from climate databases (i.e., IPCC WGI, WGII, WGIII reports and the Synthesis Report 2023). While this approach ensures that sufficient information is retrieved to answer the question, it can be further tailored. For example, if there's a

need to include specific reports or additional data layers in the query results, our system offers unique flexibility. Instead of utilizing a single, centralized database, we can separate it into multiple specialized databases. This design allows for the option to direct queries individually to each database, enabling more precise and context-specific responses. We demonstrate the efficacy of this methodology by designing three separate databases: the first focuses on IPCC reports, the second exclusively includes the IPCC Synthesis Report, and the third incorporates recent publications from the World Meteorological Organization (WMO) (Table 5). This is only to demonstrate how updated science on top of the IPCC AR6 cycle could enhance the retrieval of information, and we do not claim that we have added all the new reports. There are many other sources that we did not include in our study (see 'Further development' in the 'Limitations and Future Works' section), and we only relied on the IPCC AR6 reports.

Limitations and future works
Hallucination prevention. Model hallucination is still an eminent unresolved problem in NLP. Although we have tried to force LLMs

Table 4 Cross-checking of the hybrid ChatClimate references for Question 9 with k-5, k-10, and k-15 nearest for extra knowledge retrieval.

QAs comparison

Hybrid K-5	Hybrid K-10	Hybrid K-15
Q6) Where is the majority of climate finance going?		
The majority of climate finance is going to a few large economies, with Brazil, India, China, and South Africa accounting for around one-quarter to more than a third depending on the year, a share similar to that represented by developed countries (IPCC AR6, WGIII Chapter 15, Page 18). Additionally, global climate finance is heavily focused on mitigation, with more than 90% on average between 2017–2020 being directed towards mitigation efforts (IPCC AR6, WGIII Chapter 15, Page 3).	Regions that will be disproportionally affected by climate change include East, Central, and West Africa, South Asia, Micronesia and Melanesia, and Central America (IPCC AR6, Page 81). These regions are highly vulnerable to climate hazards such as droughts, floods, and sea-level rise. In Asia, climate change hotspots with strong climate signals and high concentrations of vulnerable people are present in semi-arid, glacial-fed river basins and mega deltas (IPCC AR6, Page 10). South Asia and Africa are also identified as hotspots of multi-sectoral climate change risk, largely due to changes in water-related indicators affecting crop production (IPCC AR6, Page 81). Disproportionate impacts are exacerbated by factors such as poverty, hunger, conflict, and environmental degradation, which make countries and regions more vulnerable to climate change (IPCC AR6, Page 81). Vulnerable populations within these regions, including the poor, children, women, and marginalized Indigenous or ethnic minority people, often have limited resources and structures to support them during climate-related disasters (IPCC AR6, Page 27). Regions such as Africa and Central and South America are more likely to experience disproportionately higher impacts due to the increasing frequency of extreme temperatures, which can affect public health, agriculture, ecosystems, biodiversity, and infrastructure (IPCC AR6, Page 68). In addition, these regions are projected to face increased risks of flood, hunger, and water scarcity (IPCC AR6, Pages 75, 81, 110). In summary, regions that will be disproportionally affected by climate change include parts of Africa, South Asia, Central America, and specific climate-change hotspots in Asia. These regions face higher risks due to a combination of climate hazards, vulnerable populations, and limited resources for adaptation and support (In-house knowledge).	Regions that will be disproportionally affected by climate change include East, Central, and West Africa, South Asia, Micronesia and Melanesia, and Central America (IPCC AR6, WGII Chapter08, Page:81). These regions are highly vulnerable to climate hazards such as droughts, floods, and sea-level rise. In addition, Africa and Central and South America are more likely to experience disproportionately higher impacts due to their vulnerability, adaptation gaps, and development needs (IPCC AR6, WGII Chapter18, Page:68). South Asia and Africa are also identified as hotspots of multi-sectoral climate change risk, largely due to changes in water-related indicators affecting crop production (IPCC AR6, WGII Chapter16, Page:81). Climate change is expected to increase the magnitude and frequency of droughts and hurricanes in these regions, potentially forcing millions of people to leave their homes (IPCC AR6, WGII Chapter12, Page:80). Furthermore, vulnerable groups within these regions, such as the poor, children, women, and marginalized Indigenous or ethnic minority people, are differentially impacted by climate change (IPCC AR6, WGII Chapter08, Page:27). These vulnerable regions and countries often have limited resources and structures to support these groups when climate hazards occur, placing an additional burden on them (In-house knowledge). In summary, regions such as East, Central, and West Africa, South Asia, Micronesia and Melanesia, and Central America will be disproportionally affected by climate change due to their high vulnerability, exposure to climate hazards, and limited resources to support vulnerable populations (IPCC AR6, WGII Chapter08, Page:81).
accuracy: 3	**accuracy: 3**	**accuracy: 4**

not to hallucinate by using external databases, up-to-date references, and prompt engineering, it still requires human supervision. For example, cross-checking the references ensures that the model is not hallucinating. Issues around mitigation of hallucination have been more elaborated in literature[15,38]. In future work, we will analyze how likely it is for ChatClimate to hallucinate, and we intend to automate the supervision process to reduce human effort.

Sufficiency and completeness of ChatClimate's semantic search. The accuracy of the answers to user questions, as well as the sufficiency and completeness of these answers and the retrieved texts from external sources, depend on many factors. These factors include the top-k hyperparameter, the prompt, and data sources. ChatClimate answers questions based on the top-k relevant chunks retrieved. Therefore, there is a low chance that the semantic search neglects some critical text chunks for a question. In this study, we demonstrated the importance of having decentralized databases compared to a single centralized database where all data is stored

and retrieved. However, this is still an important open research direction for our future work. In future work, we plan to focus on enhancing the quality of retrieved information, specifically by examining the difference between sufficiently retrieved information to answer a question and completely retrieved information for a more comprehensive answer. Another aspect that we will consider in future work is the impact of chunk size on retrievals. This will be the subject of research focusing on paragraph-level splitting rather than sentence-level splitting for retrievals.

Multi-modal search. The current version of ChatClimate does not support querying from tables, and interpretation of figures is also not supported. This is an ongoing research topic in the field of NLP, where search extends beyond textual data to include images, tabular data, and data interpretation. In future work, we will develop a multi-modal LLM where people can upload images and also ask questions based on existing tables and figures in the report. We welcome contributions in this regard.

Table 5 Comparison of various databases in response to a question.

QAs comparison

Standalone IPCC SYR Standalone	ChatClimate (IPCC AR6 reports)	Standalone ChatWMO
Q2 extra analysis, When will we reach 1.5°C?		
The best estimate for reaching 1.5 °C of global warm- ing lies in the first half of the 2030s in most of the considered scenarios and modeled pathways. How- ever, it's more likely than not that global warming will reach 1.5 °C between 2021 and 2040, even under the very low GHG emission scenario (SSP1-1.9). It's also likely or very likely to exceed 1.5°C under higher emissions scenarios (Page 33, CLIMATE CHANGE 2023 Synthesis Report Summary for Policymakers Page 18, Synthesis Report Of The IPCC Sixth Assess- ment Report (AR6) Longer Report Page 56, Climate Change 2023 Synthesis Report Summary for Policy-makers).	According to the SR1.5 report by the Intergovernmental Panel on Climate Change (IPCC), global warming is likely to reach 1.5 °C between 2030 and 2052 if it continues to increase at the current rate. This conclusion is made with high confidence (Reference: IPCC AR6 WGI Chapter01, Page: 44).	Based on the information provided, it is estimated that the annual mean global near-surface temperature could exceed 1.5 °C above 1850–1900 levels for at least one year between 2023 and 2027, with a 66% chance of this happening. However, it is less likely (32%) that the five-year mean will exceed this threshold during the same period (WMO Global Annual to Decadal Climate Update 2023–2027, Page:6). The 1.5 °C level specified in the Paris Agreement refers to long-term warming over many years, and temporary exceedances are expected to occur with increasing frequency as global temperatures approach the long-term threshold. The Intergovernmental Panel on Climate Change (IPCC) has stated that to limit global warming to 1.5 °C, global anthropogenic CO_2 emissions need to reach net zero by around 2050 (Integrated Weather and Climate Services in Support of Net Zero Energy Transition (WMO-No. 1312), Page:18). However, there is not enough information to provide a specific year when the 1.5 °C threshold will be consistently exceeded.
Accuracy: 4	**Accuracy: 3**	**Accuracy: 5**

Chain of Thoughts (COTs). In this study, we did not fully explore the potential of chain of thoughts (COTs) by testing different prompts. However, we expect that implementing COTs will improve the accuracy of our system's outputs, which we plan in our future works.

Evaluation of LLMs responses. We acknowledge that the evaluation of responses was not the primary focus of this work, and we relied solely on expert knowledge to assess the model's performance. Additionally, further work is needed to provide a comprehensive description of the evaluation procedure, including aspects such as inter-annotator agreement and a more transparent explanation of query generation.

Fact-checking. Providing access for LLMs to various trustworthy resources can enhance the model's ability to perform fact-checking and provide well-grounded information to users. In ongoing research, we are exploring the potential of automated fact-checking methods (e.g.[35,36]). To this extent, we are building an authoritative and accurate knowledge base that can be used to fact-check domain-specific claims[39] or LLM-produced responses. In this knowledge base, we will also leverage statements from the IPCC AR6 reports to validate or refute claims related to climate change and other environmental issues.

Further development. We continually improve ChatClimate and welcome community feedback on our website www.chatclimate.ai to enhance its question-answering capabilities. Our goal is to provide accurate and reliable climate change information, and we believe domain-specific chatbots like ChatClimate play a crucial role in achieving it. It is important to keep ChatClimate up-to-date by automating the inclusion of new information from the scientific literature. To ensure the continual relevance and accuracy of ChatClimate, we plan to carry out regular updates. Specifically, these updates will be conducted in alignment with the release of comprehensive global assessments such as those from the IPCC. In

particular, upon the release of any report from the Assessment Report 7th cycle, the relevant information will be integrated into our database to enhance ChatClimate's knowledge base.

Conclusion

In this study, we demonstrate how some limitations of current state-of-the-art LLMs (GPT-4) can be mitigated in a Question Answering use case. We demonstrate improvements by giving the LLMs access to data beyond their cut-off date of training. We also show how proper prompt engineering using domain expertise makes LLMs perform better. These conclusions are reached by comparing GPT-4 answers with our Hybrid and Standalone ChatClimate models. In summary, our study demonstrates that the hybrid ChatClimate outperformed both GPT-4 and ChatClimate standalone in terms of the accuracy of answers when provided access to the outside world (IPCC AR6). The higher performance can be attributed to the integration of up-to-date and domain-specific data, which addresses the issues of hallucination and out-dated information often encountered in LLMs. The results underline the importance of tailoring models to specific domains. The main findings of our work are summarized as follows:

1. The hallucination and outdated issues of LLMs could be refined by giving those models access to the knowledge beyond their training phase time and instructing LLMs on how to utilize that knowledge.

2. The core idea behind ChatClimate—providing long-term memory and external data to LLMs—remains valid, regardless of which GPT model is current. This is because there will always be reports (or other PDFs) published after the cutoff date for training LLMs, and ChatClimate can provide proper access to these documents, even without access to update the LLM itself. Similar arguments are made in[22,40].

3. With proper prompt engineering and knowledge retrieval, LLMs can provide sources of the answers properly.

4. Hyperparameter tuning during knowledge retrieval and semantic search plays an important role in prompt engineering. We tested this by K-5, K-10, and K-15 nearest pieces of knowledge to the question in the semantic search between the question and the database.

5. Regulating LLM training, fine-tuning, and incorporating it into applications are necessary. Specifically, external database integration and prompt engineering should be considered in regulations for chatbots. We emphasize the importance of regulation for checking the outcomes of domain-specific chatbots. In such domains, users may not have enough knowledge to verify answers or cross-check references, making biased data or engineered prompts potentially harmful to end users.

6. Our AI-powered tool makes climate information accessible to a broader community and may assist decision-makers and the public in understanding climate change-related issues. However, it is intended to complement, not replace, specialized local knowledge and custom solutions essential for effective decision-making.

7. Retraining LLMs is computationally intensive, thereby generating a high amount of CO_2 emissions. However, inference is comparatively less resource-demanding. In the retrieval-augmented framework we proposed, the necessity for frequent retraining of LLMs is substantially reduced. Consequently, the necessity to integrate new information through retraining is reduced. In evaluating the actual CO_2 emissions, we reference the GPT family of models. However, OpenAI has not disclosed any information regarding their training procedures[9]. Nonetheless, we advocate for the LLM community to adopt climate-aware workflows to address this concern.

8. Our findings not only emphasize the importance of leveraging climate domain information in QA tasks but also highlight the need for continued research and development in the field of AI-driven text processing.

Methods

ChatClimate pipeline. In this study, we develop a long-term memory database by transforming the IPCC AR6 reports into a searchable format. The reports are converted from PDFs to JSON format, and each record in the database is divided into smaller chunks that LLM can easily process. The choice of the batch size for embeddings is a hyperparameter. We insert data into the vector database in batches of 100 vectors, based on the guidelines provided by Pinecone VD. We utilize OpenAI's state-of-the-art text embedding model to vectorize each data chunk. Prior to injection into the database, we implement an efficient indexing mechanism to optimize retrieval times and facilitate effective information retrieval. Consequently, we can implement a semantic search that identifies the most relevant results based on the meaning and context of each query.

To elaborate, first, we created our database using IPCC AR6 reports (7 PDFs please see Supplementary Information for more details). Second, to enable the Large Language Models (LLMs) to access this long-term memory and to make these PDFs usable information for LLMs, we employed a PDF parser to digitize the pages of these reports and segment them into manageable text chunks. These chunks were then used to populate our external database, which feeds into the LLMs. Furthermore, we used an embedding model and a tokenizer to convert each chunk into a numeric vector, which was stored in our vector database.

When a user poses a question, it is first embedded and then indexed using semantic similarity to find the top-k nearest vectors corresponding to the inquiry. The dot product of two vectors is utilized to analyze the similarity between vector embeddings, which is obtained by multiplying their respective components and summing the results.

After identifying the nearest vectors to the query vector, we decode the numeric vectors to text and retrieve the corresponding text from the database. The textual information is then used to refine and improve LLM prompts. Augmented queries are posed to the GPT-4 model through instructed prompts, which enhance the user experience and increase the overall performance of our chatbot. Figure 3 shows the pipeline of ChatClimate.

Tools and external APIs. The first tool used in this study is a Python-based module that transforms IPCC AR6 reports from PDFs to JSON format (PDF parser) and preprocesses the data, utilizing the powerful pandas library to access and manipulate data stored in dataframes.

The second tool is the LangChain Python package (https://github.com/langchain-ai/langchain), which retrieves data from the JSON and chunks the extracted text into smaller sizes, ready for embedding. LangChain is a lightweight layer that transforms sequential LLM interactions into a natural conversation experience.

The third tool employed is OpenAI's embedding model 'text-embedding-ada-002,' which vectorizes all chunks of the IPCC AR6 JSON files. Vector embeddings have proven to be a valuable tool for a variety of machine learning applications, as they can efficiently represent objects as dense vectors containing semantic information.

The fourth tool involves storing the generated vectors in a database, allowing for efficient storage and retrieval of the vector embeddings.

The fifth and final tool used is the GPT-4 'chatcompletion' endpoint with instructed prompts, which provides answers to questions by leveraging the indexed vector embeddings.

Input prompts and ChatBots. The importance of prompt engineering for LLMs has been addressed in previous work[41–43]. We designed three prompts to compare the answers of our chatbots (i.e., ChatClimate, hybrid ChatClimate, and GPT-4). The prompt used in our study consists of a series of instructions that guide the completion of a chat with GPT-4 on how to answer a provided question. The prompt is structured to allow the chatbot to access external resources while using its in-house knowledge. Overall, the prompt is designed to guide how to answer the questions given the availability of external and/or in-house knowledge. We demonstrate the three prompts used in this study in Boxs 3, 4, and 5.

Hybrid ChatClimate. In this first scenario, the prompt starts with five pieces of external information retrieved from long-term memory, followed by a question that was asked by the user. The prompt instructs the chatbot to provide an answer based on the given information while using its own knowledge. Moreover, the chatbot is structured to prioritize IPCC AR6 for answers, referencing the names and pages of corresponding IPCC reports (Working Group I, II, III chapters, summary for policymakers, technical summary, and synthesis reports).

ChatClimate. In the second scenario, the prompt starts with five pieces of external information retrieved from long-term memory, followed by a question that was asked by the user. The prompt instructs the chatbot to provide answers only based on IPCC AR6.

GPT-4. In the last scenario, the prompt does not provide any extra information or instruction on how to provide answers and can be considered the baseline behavior.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Background

Large Language Models. LLMs have transformed NLP and AI research over the last few years[44]. They show surprising capabilities of generating creative text, solving basic math problems, answering reading comprehension questions, and much more. These models are based on the transformer architecture[45] and are trained on vast quantities of text data to identify patterns and connections within the data. Some notable examples of these models include GPT and BERT family models, which have been widely used for various NLP tasks[1–4,9]. The recent breakthroughs with models like T0[6], LLaMA[5], PaLM[7], GPT-3[1], and GPT-4[9] have further highlighted the potential of LLMs, with applications including chatbots[8,9] and virtual assistants[46]. However, LLMs suffer from hallucination[15], which refers to mistakes in the generated text that are simply made up or semantically incorrect. This can lead to vague or inaccurate responses to questions. Moreover, most of these models are trained on text prior to 2022

and thus have not been updated with new data or information since then[15].

NLP and climate change. NLP techniques have been widely used in the analysis of text related to climate change. Applications range from financial climate disclosure analyses[17,47], detecting stance in media about global warming[48], detecting environmental claims[49] and climate claims fact-checking[50,51].

Question answering and chatbots. Question-answering (QA) systems and chatbots have become increasingly popular. They can provide users with relevant and accurate information in a conversational setting. The importance, limitations, and future perspectives of conversational AI have been addressed in the literature from the open domain[52,53] to domain-specific chatbots[54]. When presented with a question in human language, chatbots automatically provide a response. Although numerous information retrieval chatbots accomplish this task, transformer-based LLMs have become the de-facto standard in QA[1,4,8,9]. In the context of climate change, QA systems and chatbots can help bridge the gap between complex scientific information and public understanding
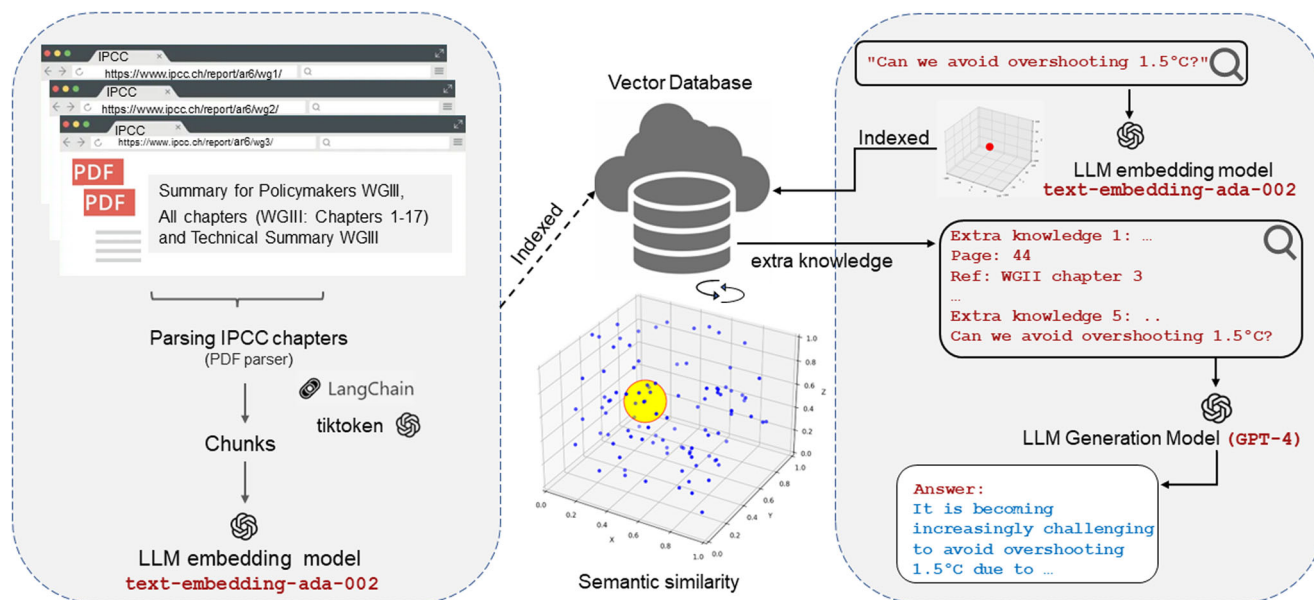

Fig. 3 ChatClimate Data Pipeline: from creating external memory, receiving questions to accurate answers from IPCC AR6. The black arrows show the sequence of tasks in the ChatClimate pipeline. Langchain is the Python library we used for splitting text into smaller chunks. Tiktoken is OpenAI's tokenizer. 'text-embedding-ada-002' is the embedding mode from OpenAI. GPT-4 is the large language model.

Box 3 ▌ Input prompt for hybrid ChatClimate

```
"system", "content": You are a Q&A bot, an intelligent system that answers user questions based on the information
provided by the user above the question and your in-house knowledge. There are five pieces of extra information above
the user question. The user question is in the final line. When you use the user
information, always indicate the Page and Reference in your answer. The Page and Reference are located below each
piece of information. Additionally, let us know which part of your answer is from the IPCC information and which
part
is based on your in-house knowledge by writing either (IPCC AR6) or (In-house knowledge). If the information cannot
be
found in the information provided by the user or your in-house knowledge, please say 'There is not enough
information'.
"role": "user", "content": External knowledge + Question
```

Box 4 ∥ Input prompt for ChatClimate

```
"system", "content": You are a Q&A bot, an intelligent system that answers user questions ONLY based on the
information provided by the user. There are five pieces of extra information above the user question. The user
question is in the final line. When you use the user information, always indicate the Page and Reference in your
answer. The Page and Reference are located below each piece of information. If the information
cannot be found in the information provided by the user, please say 'There is not enough information'
"role": "user", "content": External knowledge + Question
```

Box 5 ∥ Input prompt for GPT-4

```
"system", "content": "You are Q&A bot. A highly intelligent system that answers user questions"
"role": "user", "content": Question
```

---

by providing concise and accessible answers to climate-related questions. Such systems can also facilitate communication between experts, policymakers, and stakeholders, enabling more informed decision-making and promoting climate change mitigation and adaptation strategies[49,55]. As the field of NLP and its application to climate change[17,56] continues to advance, it is expected that QA systems and chatbots will play an increasingly important role in disseminating climate change information and fostering public engagement with climate science.

Long-term memory and agents for LLMs. One solution for enhancing the capabilities of LLMs in QA tasks is to fine-tune them on different datasets, which could be resource-wise expensive[52]. However, an alternative approach involves using agents that access the LLMs' long-term memory, retrieve information, and insert it into a prompt to guide the LLMs more effectively[57,58]. These agents can decide which actions to perform, such as utilizing various tools, observing their outputs, or providing responses to user queries[59]. This approach has been shown to improve the accuracy and efficiency of LLMs in a range of domains, including healthcare and finance[59]. Domain-specific chatbots use a similar concept, where an agent has access to a carefully curated in-house database (long-term memory) to answer domain-specific questions[60]. These chatbots can provide customized responses based on the available information in their database, allowing for more accurate and relevant answers to user queries.

Data availability
Our primary data sources in this study, which provide the LLMs with knowledge beyond their training phase, are the following reports from IPCC AR6. 1. Summary for Policymakers from each of the Working Groups (I, II, III): 3 pdfs 2. All chapters (WG I: Chapters 1-12, WG II: Chapters 1-18, Cross-Chapters 1-12, WG III: Chapters 1-17) and Technical Summary from each of the three working groups (https://www.ipcc.ch/report/ar6/wg1, https://www.ipcc.ch/report/ar6/wg2, https://www.ipcc.ch/report/ar6/wg3). 3. The IPCC Synthesis Report 2023: 1 pdf. In this study, we did not consider the following special reports from the AR6 cycle: 1. Special Report on Climate Change and Land (2019), 2. Special Report on the Ocean and Cryosphere in a Changing Climate (2019), 3. Special Report on Global Warming of 1.5 C (2018). This exclusion is firstly due to the publication dates of these reports falling within the cut-off date of the LLMs we employed, and secondly, the AR6 Synthesis Report: Climate Change 2023, which we utilized in our study, encapsulates the most important portions of these three special reports. In the subsection 'Database Setup: Access to Different Databases' in the Discussion, we also used the following reports of WMO to demonstrate how updated science on top of the IPCC AR6 cycle could enhance information retrieval. We do not claim that we have added all the new reports. There are many other sources that we did not include in our study, and we only relied on the IPCC AR6 reports for ChatClimate. 1. 2022 State of Climate Services: Energy (WMO No. 130). 2. WMO Global Annual to Decadal Climate Update 2023–2027. 3. State of the Climate in Asia 2022 (WMO No. 1321) 4. State of the Climate in South-West Pacific 2021 (WMO No. 1302) 5. State of the Climate in Africa 2021 (WMO No. 1300)

Code availability
The back-end code to reproduce ChatClimate is available at: https://github.com/saeedashraf/chatipcc

References
1. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
2. Devlin, J. Chang, M.-W. Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. (2019).
3. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
4. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog.* (2019).
5. Touvron, H. et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. (2023).
6. Sanh, V. et al. Multitask-prompted training enables zero-shot task generalization. arXiv, 2110.08207. (2021).
7. Chowdhery, A. et al. Palm: Scaling language modeling with pathways. arXiv Preprint arXiv:2204.02311. (2022).
8. OpenAI. *InstructGPT: AI for Generating Instructions.* https://openai.com/research/instructgpt/. (2023b).
9. OpenAI. GPT-4 Technical Report. Technical report, OpenAI. (2023a).
10. Shen, Y. et al. HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace. arXiv:2303.17580. (2023).
11. Larosa, F. et al. Halting generative AI advancements may slow down progress in climate research. *Nat. Clim. Change* **13**, 497–499 (2023).
12. Li, J., Zhang, Z. & Zhao, H. Self-prompting large language models for open-domain QA. *arXiv* **2212**, 08635 (2022).
13. Robinson, J. Rytting, C. M. and Wingate, D. Leveraging large language models for multiple choice question answering. arXiv:2210.12353. (2023).
14. Shao, Z. Yu, Z. Wang, M. and Yu, J. Prompting large language models with answer heuristics for knowledge-based visual question answering. arXiv:2303.01903. (2023).
15. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
16. Jang, J. et al. Towards continual knowledge learning of language models. In *ICLR.* (2022).
17. Bingler, J. A. Kraus, M. Leippold, M. and Webersinke, N. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Res. Lett.*, 102776. (2022).

18. Kumar, A., Singh, S. & Sethi, N. Climate change and cities: challenges ahead. *Front. Sustain. Cities* **3**, 645613 (2021).
19. Sethi, N., Singh, S. & Kumar, A. The importance of accurate and up-to-date information in the context of climate change. *J. Clean. Prod.,* **277**, 123304 (2020).
20. Bender, E. M. Gebru, T. McMillan-Major, A. and Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097. (2021a).
21. Weidinger, L. et al. Ethical and social risks of harm from Language Models. arXiv:2112.04359. (2021).
22. Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. arXiv:2112.04426. (2022).
23. IPCC. 2021. Climate Change The Physical Science Basis. *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*
24. IPCC. 2022a. Climate Change 2022: Impacts, Adaptation, and Vulnerability. *Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*
25. IPCC. 2022b. Climate Change 2022: Mitigation of Climate Change. *Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*
26. IPCC. 2023. *Climate Change 2023*: Synthesis Report. Geneva, Switzerland: IPCC. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)].
27. Abdar, M., Tait, J. & Aleven, V. The impact of chatbot characteristics on user satisfaction and conversational performance. *J. Educ. Psychol.* **112**(4), 667–683 (2020).
28. Luger, E. and Sellen, A. Towards a framework for evaluation and design of conversational agents. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2885–2891. ACM. (2016).
29. Przegalinska, A. Ciechanowski, L. Stroz, A. Gloor, P. and Mazurek, G. In bot we trust: A new methodology of chatbot performance measures. Business Horizons, **62**, 785–797. Digital Transformation and Disruption. (2019).
30. Ramachandran, D. Eslami, M. and Sandvig, C. A Framework for Understanding and Evaluating Automated Systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 154–164. (2020).
31. Schick, T. and Schütze, H. Exploiting cloze questions for few shot text classification and natural language inference. arXiv:2001.07676. (2021a).
32. Hendy, A. et al. How good are GPT models at machine translation? A comprehensive evaluation. arXiv:2302.09210. (2023).
33. Zhou, Y. et al. Large Language Models Are Human-Level Prompt Engineers. arXiv:2211.01910. (2023).
34. Clark, C. D. et al. Growth and retreat of the last British Irish Ice Sheet, years ago: the BRITICE- CHRONO reconstruction. *Boreas* **51**(4), 699–758 (2022).
35. Guo, Z., Schlichtkrull, M. & Vlachos, A. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* **10**, 178–206 (2022).
36. Vlachos, A. and Riedel, S. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. Baltimore, MD, USA: Association for Computational Linguistics. (2014).
37. Bender, E. M. Gebru, T. McMillan-Major, A. and Shmitchell, S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097. (2021b).
38. Ni, J. et al. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. arXiv:2307.15770. (2023).
39. Stammbach, D. Webersinke, N. Bingler, J. A. Kraus, M. and Leippold, M. *Environmental Claim Detection*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada. (2023).
40. Shi, W. et al. RE- PLUG: Retrieval-Augmented Black-Box Language Models. arXiv:2301.12652. (2023).
41. Kojima, T., Gu, S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv* 2205.11916, (2023).
42. Reynolds, L. and McDonell, K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380959. (2021).
43. Schick, T. and Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. arXiv:2009.07118. (2021b).
44. Fan, L. et al. A Bibliometric Review of Large Language Models Research from 2017 to 2023. arXiv:2304.02020. (2023).
45. Vaswani, A. et al. *Attention is All you Need*. In Guyon, I. Luxburg, U. V. Bengio, S. Wallach, H. Fergus, R. Vishwanathan, S. and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. (2017).
46. Jo, A. The Promise and Peril of Generative AI. Nature 614. (2023).
47. Luccioni, A. Baylor, E. and Duchene, N. Analyzing Sustainability Reports Using Natural Language Processing. arXiv:2011.08073. (2020).
48. Luo, Y. Card, D. and Jurafsky, D. Detecting Stance in Media On Global Warming. In *Findings of the Association for Computational Linguistics*: EMNLP 2020, 3296–3315. Online: Association for Computational Linguistics. (2020).
49. Stammbach, D. Zhang, B. and Ash, E. The Choice of Textual Knowledge Base in Automated Claim Checking. J. Data Inf. Qual., 15. (2023).
50. Diggelmann, T. Boyd-Graber, J. Bulian, J. Ciaramita, M. and Leippold, M. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*. (2020).
51. Webersinke, N. Kraus, M. Bingler, J. A. and Leippold, M. ClimateBert: A pretrained language model for climate-related text. arXiv:2110.12010. (2022).
52. Adiwardana, D. et al. Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977. (2020).
53. OpenAI. ChatGPT: A large-scale generative language model for conversational AI. (2022).
54. Lin, B. Bouneffouf, D. Cecchi, G. and Varshney, K. R. Towards healthy AI: Large language models need therapists too. arXiv:2304.00416. (2023).
55. Callaghan, M. et al. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nat. Clim. Change* **11**(11), 966–972 (2021).
56. Kölbel, J. F. Leippold, M. Rillaerts, J. and Wang, Q. Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure. *Available at SSRN 3616324*. (2020)
57. Kraus, M. et al. Enhancing large language models with climate resources. arXiv:2304.00116. (2023).
58. Nair, V. Schumacher, E. Tso, G. and Kannan, A. DERA: Enhancing large language model completions with dialog-enabled resolving agents. arXiv:2303.17071. (2023).
59. Schick, T. et al. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761. (2023).
60. Gerhard-Young, G., Anantha, R., Chappidi, S. & Hoffmeister, B. Low-resource adaptation of open domain generative chatbots. *arXiv*, 2108.06329. (2022).

Author contributions
All the authors, S.V., D.S., V.M., J.B., J.N., M.K., S.A., C.C., T.W., T.S., G.G., T.Y., Q.W., N.W., C.H., M.L., designed the study, wrote the initial manuscript, conducted the analysis, reviewed, edited and improved the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-023-01084-x.

Correspondence and requests for materials should be addressed to Saeid Ashraf Vaghefi.

Peer review information *Communications Earth & Environment* thanks Francesca Larosa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Heike Langenberg. A peer review file is available

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.