Large Language Models and Climate Information

# Evaluating the Quality of ChatGPT's Climate-related Responses

The use of ChatGPT and language models carries the risk of spreading false information. Although ChatGPT produces mostly correct information about climate related topics, the answers reflect societal misunderstandings about climate change.

By Jens Bergener, Maike Gossen, Marja Lena Hoffmann, Felix Bießmann, Marek Veneny and Ruben Korenke

## 1 Introduction

ChatGPT, as a generative large language model (LLM) with an intuitive chatbot interface, has changed the way online users acquire information. Unlike traditional search engines, where you search using keywords, it allows you to ask questions, leading to novel patterns of search behavior. It then generates personalized, human-like responses retrieved from its pre-trained model. The responses are not simply a list of links or snippets of human-written texts, as you might be accustomed to with search engines. Instead, they are generated as natural language text. LLMs work by prediction. When prompted by the user, they generate their response by predicting likely words, selecting one, and then repeating the process.

Since its release in November 2022, ChatGPT has captivated users around the world. So much so that just a few months later, in January 2023, it passed the 100 million user milestone, making it the fastest growing information platform in online history (Hu 2023). Its impressive capabilities make ChatGPT a potential source of information on a wide range of topics, including climate change.

The problem, however, is that ChatGPT is essentially a black box – no one, not even the developers themselves, can really tell you how it gets to a particular response. That may be fine for creative tasks, such as asking the program to write a haiku. But not when you ask it to deal with topics where factual and accurate information is crucial, such as the consequences of climate change. Despite decades of research and evidence suggesting that climate change poses a direct threat to human wellbeing and the health of the planet, accurate scientific information coexists with misinformation in public discourse and the media (Fischer et al. 2019). Therefore, it is important to examine the quality of ChatGPT's responses in areas where truthfulness and accuracy are required. Previous research has shown that ChatGPT correctly explained the concepts of current environmental research topics such as microplastics, life cycle assessment, and circular economy (Zhu et al. 2023). However, research points to potential pitfalls, such as distorted information, social bias, lack of deeper expertise, and lack of accountability in environmental decision-making (ibid.).

In addition, ChatGPT has been shown to suffer from hallucinations and to make factual claims that cannot be verified by any source. Hallucinations in this context refer to errors in the generated text that are semantically or syntactically plausible, but are actually incorrect or nonsensical. Moreover, previous work shows that ChatGPT tends to make meaningless guesses rather than reject unanswerable questions (Shen et al. 2023). Added to these concerns is the very real possibility that ChatGPT's ability to formulate detailed responses may lead everyday users, who lack the experience to detect factual errors in the model's responses, to blindly trust the responses it generates. When people are asked to assess the accuracy of LLM results across a wide range of topics, they tend to weight ChatGPT's advice more heavily if they are unfamiliar with the topic, have used ChatGPT in the past, or have previously received accurate advice from the model (Zhang 2023).

Previous research has approached evaluating the efficacy of ChatGPT responses by comparing them to available datasets on traditional neuro-linguistic programming application tasks such as multitasking, or by testing its performance on specific types of questions (e.g., Shen et al. 2023, Bang et al. 2023). While these assessments provide valuable insights into ChatGPT's capabilities, they do not reflect questions that address climate change issues.

Our goal is to advance the above research on ChatGPT's response competence in the context of climate change. In particular, we are interested in answering the following two research questions: Does ChatGPT provide accurate and relevant responses to questions about climate change? Are the responses provided by ChatGPT consistent over time?

## 2 Methodology

To answer our research questions, we developed an evaluation framework consisting of two main steps: Firstly, (1) crowdsourcing questions about climate change, and (2) evaluating ChatGPT responses for accuracy, relevance, and consistency.

DOI 10.14512/OEW380346

We crowdsourced a total of 95 questions within the team of the *Green Consumption Assistant* project and categorized them into themes. Reviewers signed up for a ChatGPT account and used ChatGPT's default "Free Research Preview" model. The default settings use a fine-tuned version of the GPT 3.5 LLM. The LLM was trained on a large corpus of text data, including books, articles, and websites, and then fine-tuned with human feedback using reinforcement learning (OpenAI 2022). We prompted the model on February 9 and 10, 2023, and asked the questions without additional hints (zero-shot prompts). For each prompt, we started a new chat with no previous chat history.

For each question, we attempted to assess the intent and level of specificity of ChatGPT's response and fact-check the response against reliable sources of information on climate change, such as the IPCC. We also assessed how well it responded to the question in terms of relevance (the response addresses all parts of the question and provides an adequate amount of information) and accuracy (the information in the output was factually correct).

We scored the responses for both relevance and accuracy on a scale of 1–10, with 1 being not accurate or relevant and 10 being very accurate or relevant. We then calculated the mean as well as the median to provide an overall score for each response. Furthermore, responses were randomly verified by a second reviewer.

To asses whether ChatGPT maintains consistency in its responses to the same prompt, we compared whether the output was consistent with the input at different times (first round of prompts on February 9 and 10, 2023, and second round of prompts on May 17, 2023).

It is worth noting that the original GPT-3.5 model we used in the first round was retired on May 10. The second round was conducted with the updated default model of Chat GPT. ChatGPT's capabilities may have changed over time as the model has been trained with feedback (OpenAI allows users to provide feedback using the thumbs up/thumbs down button if they feel a response is inappropriate). Since the language model is deterministic, we expect minor inconsistencies in the wording of responses, but not in the overall relevance and accuracy of the content.

## 3 Results

### 3.1 Evaluation of quality of ChatGPT's responses

The overall quality of the responses is quite satisfactory, with an average score of 8.25 on a scale of 1 to 10, considering the mean. This is made up of an average score of 8.01 for relevance and 8.49 for accuracy.

A look at the histograms (Figure 1) shows that the distribution of scores for relevance, accuracy and overall quality of ChatGPT's responses are dominated by high scores. More than 50 % of the responses had an accuracy of 10 and a relevance of 8. However, we also note that 6.25 % of the responses scored with an accuracy of 3 or less and 10 % of the responses had a relevance of 3 or less. This shows that the majority of responses are accurate and relevant, but there are also a substantial number of incorrect and irrelevant responses.

The evaluation of ChatGPT's responses revealed several strengths and weaknesses of the LLM. We randomly chose several records (responses with a rating of 5 or lower or with a rating of 8 or higher) and inspected the questions and ChatGPT's responses in more detail.

We observed that ChatGPT provides balanced and nuanced arguments, and concludes many responses with a comment that encourages critical consideration to avoid biased responses. For example, ChatGPT's response to the question *"How is life in the sea affected by climate change and how can the effect be reduced?"* mentions not only reducing greenhouse gas emissions, but also reducing non-climatic effects of human activities such as overfishing and pollution. Another example is the last part of the response to the question *"How will climate change impact me and my family in the coming decades?"*, which considers the equity aspect by mentioning that the effects of climate change will not be evenly distributed and that some populations, such as low-income communities and those in the Global South, are likely to be disproportionately affected.

Although the overall quality of ChatGPT's responses to our climate-related questions was high, we explored possible reasons for the cases in which our assessment resulted in lower scores for accuracy and relevance. We found that the most common error that resulted in a lower score for the accuracy crite-

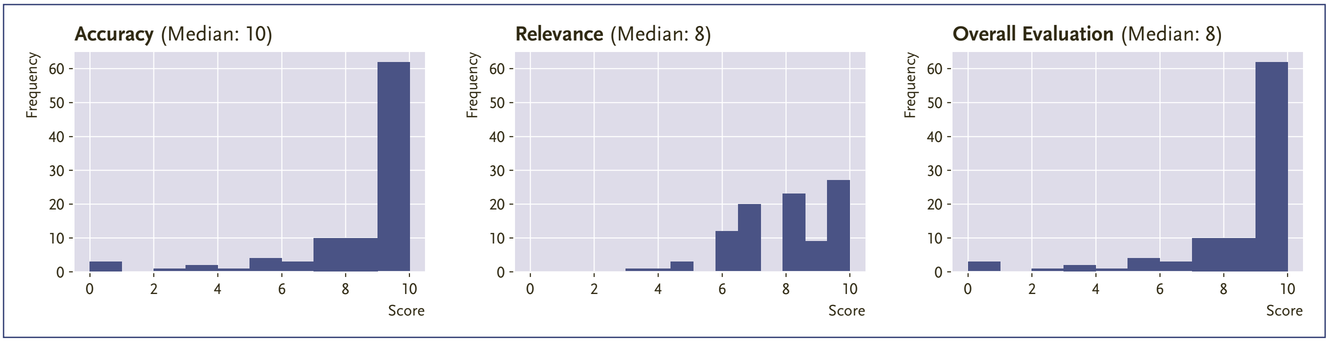

Figure 1: Histograms of frequency and score for relevance, accuracy, and overall quality of ChatGPT's responses

*“It is more important than ever to ensure that climate change information is accurate.”*

rion was caused by hallucinatory facts. For instance, ChatGPT’s response to the question *“Which percentage of recyclable waste is really recycled by Germany?”* is correct in broad strokes but not correct in the details. According to the Federal Environment Agency of Germany, the recycling rate for municipal waste rose from 56 % in 2002 to 67.4 % in 2020 (Umweltbundesamt 2022) while ChatGPT claims that it was 63 % in 2020. The existence of hallucinatory facts has also been recognized by other studies (e.g. Jang/Lukasiewicz 2023).

In some cases, ChatGPT confuses the meaning of very similar words and therefore presents incorrect results based on wrong reasoning (e.g. Arctic Sea vs. Antarctic Sea) or presents fake references that do not exist in real . For example, ChatGPT quotes a source named “International Association for Sustainable Transportation (IASTS)”, which in reality does not exist at all. The problem that ChatGPT generates false or fabricated information such as made-up references or fabricated DOI or URL links is a widespread concern (e.g. Zhu et al. 2023). In some cases we also found errors in referencing, i.e. ChatGPT referencing  scientific sources and literature, but drawing wrong conclusions from them.

Lower scores for relevance were caused by ChatGPT’s responses that did not include the most important and relevant information. For instance, ChatGPT’s response to the question *“How can I better understand my climate impact?”* suggests specific ways to reduce emissions even though the question was about better understanding of individual climate impacts. In addition, responses received lower scores if they did not address all parts of the question, contained irrelevant or too much information, or contained contradictory or confusing information. For example, ChatGPT’s response to the question *“How can I live sustainably on a budget?”* does not mention behaviors that have a large impact and are easy on the budget, such as questioning the need for consumption and incorporating sufficiency.

### 3.2 Evaluation of output consistency of ChatGPT’s responses

In addition to assessing the relevance and accuracy of ChatGPT responses, we also assessed consistency over time. The consistency of responses varied by question. While for some questions the response is almost identical in both cases, for others it is very different. In these cases, the variation sometimes went beyond the wording and included new or different content, which in some cases resulted in a change in the main message and tone. In addition, the comparison revealed that in some cases the second round responses took a more thoughtful tone and pointed out the limitations or contradictions of the responses.

Another interesting observation was that while we did not systematically analyze the results of the second round, a first impression from the responses was that those who changed tended to change for the better and gave more detailed, accurate, or appropriate responses. For example, on the question *“What area is needed for wind and solar farms to replace fossil fuels, assuming the economy continues to electrify?”*, the first run of ChatGPT produced a rather vague response that did not provide specific estimates, but instead listed a number of factors that influence the exact amount of land needed for wind and solar farms. In the second run, ChatGPT responded with a list of relevant factors as well, but additionally added concrete estimates that could also be found in external literature.

## 4 Discussion

Our evaluation showed that the rating of ChatGPT’s responses was remarkably good. Many climate-related questions received correct, balanced responses that scored well in terms of relevance and accuracy. However, there were a few questions that resulted in responses that were of a lower quality. These cases reveal major drawbacks of ChatGPT as a source of information and require special attention and consideration.

We can draw several important lessons from cases where the scoring of responses resulted in low accuracy.

First, ChatGPT’s inaccurate responses often had a plausible-sounding tone. Many linguistic subtleties can cause a piece of information to be incorrect, and ChatGPT seems unaware of these nuances. In part, this may be related to the well-studied characteristic of deep neural networks that as their capacity increases, these algorithms are often overconfident about their incorrect predictions (Guo et al. 2017). For this reason, text generators such as ChatGPT are often called “stochastic parrots” (e.g. Bender et al. 2021). Because they are trained to give responses that feel right to people, the confident answering style can fool individuals into thinking the output is correct. According to OpenAI, it is a challenge to solve this problem because there is currently no source of truth during the training of the model. When the model is trained to be more careful, it rejects questions that it can answer correctly. Meanwhile, supervised training misleads the model because the ideal response depends on what the model knows, not on what the human demonstrator knows. The consequence of this behavior is that detecting incorrect facts in the ChatGPT output requires expert knowledge about the domain of the response. However, studies show that lack of knowledge in a particular area is often the reason why users ask ChatGPT for help in the first place (Zhang 2023).

Second, consistent with previous research suggesting that ChatGPT often fails to generate logically correct predictions (Jang/Lukasiewicz 2023) and that GPT-4 often generates explanations that contradict its own outputs for different inputs in similar contexts (Bubeck et al. 2023), we also observed inconsistencies in responses over time in some cases. Had we based our evaluation on the second-round responses, the results would likely have been better. Clearly, the large language model, as measured by the relevance and quality of the responses, has improved over time.

Third, our work supports the notion of biases ingrained within large-scale LLMs. In this context, a bias is understood as the presence of systematic misrepresentations that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns, and results from the training data (Ferrara 2023). Some of ChatGPT's incorrect responses to our questions reflected larger societal misconceptions about effective action on climate change, such as overvaluing single actions with small impacts at the expense of more consequential contributions. Responses also sometimes seemed overly optimistic about technological solutions as the central way to mitigate climate change.

Fourth, the fact that the specifics of the training data and model development are not publicly available poses great risks to the public use of LLMs. The collective efforts of the open scientific community will hopefully help to further explore the limitations of models such as ChatGPT that have been highlighted in this study. But it is not only the evaluation by the scientific community that contributes to a more responsible approach to the potential of LLMs. Recently, there have been a number of impressive advances in open versions of conversational agents similar to ChatGPT (Chiang et al. 2023, Geng et al. 2023). Importantly, these open solutions can achieve performance close to that of ChatGPT, but at a much lower computational cost. One factor mentioned in previous research that has contributed to these improvements is the focus on the quality of the training data. Similar to recent advances in image generation with public Deep Learning models (Rombach et al. 2022), for which the availability and quality of training data was fundamental, there is reason to believe that investment in data curation (Gebru et al. 2021) and documentation of model development (Mitchell et al. 2019) will help improve the trustworthiness of LLM applications.

## 5 Limitations and Future Work

A limitation of our study approach is that human expert evaluation is difficult to scale across different subject areas and text volumes. Furthermore, human expert evaluation is not resistant to potential errors. In future evaluations, this could be minimized by including a larger number of experts in the evaluation process. Regarding the evaluation of the consistency of the model, we only compared two runs that were performed a few months apart. For a more comprehensive assessment of the consistency of ChatGPT, further studies could conduct the run more frequently with shorter and similar time intervals. In addition, for a more accurate assessment of consistency, it would be important to establish more adjectival criteria against which multiple responses could be compared over time.

Another limitation arises from the composition of the 95 questions we asked ChatGPT. While these cover a wide range of climate-related topics, we did not create them or their wording based on actual user data. This could result in potential misrepresentation of the climate-related topics that users would actually ask ChatGPT, and possibly irrelevant responses from the model due to specifics in the wording of the questions. For future studies, it would be interesting to consider user behavior before compiling the list of questions and assess how ChatGPT users interact with ChatGPT and how they would interpret and use the model's responses to climate-related questions.

## 6 Conclusion

The findings of our study show that with the proliferation of AI-powered texts, source verification is more important than ever to ensure that climate change information is accurate. Although the majority of responses appeared plausible, we discovered some responses with biased information and hallucinated facts. Detecting these incorrect responses requires detailed expertise in the area of interest. A major challenge in this context is that while modern generative LLMs have become much better at producing authentic-sounding texts, the evaluation of these texts remains difficult to automate. Modern conversational agents such as ChatGPT have been trained using sophisticated reinforcement learning methods that take advantage of human feedback. However, most of these humans have not been trained to judge the factual accuracy of individual responses, but rather how correct the response sounds to them.

Although LLMs should be adequately deployed in each use case and weighed against the energy consumption and emissions involved in training the models, their use has the potential to revolutionize the way information about climate change is communicated. Their ability to process and analyze large amounts of data and provide easy-to-understand responses to everyday questions could make them a valuable source of climate change information. We conclude that, at best, LLMs can help people understand and communicate information about climate change because they operate at an amazing scale. At worst, LLMs repeat past climate change communication issues because they echo misconceptions about climate change, spread false information, and fuel misinformation.

### References

Bang, Y./Cahyawijaya, S./Lee, N./Dai, W./Su, D./Wilie, B./Lovenia, H./Ji, Z./Yu, T./Chung, W./Do, Q. V./Xu, Y./Fung, P. (2023): A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. DOI: 10.48550/arxiv.2302.04023

Bender, E. M./Gebru, T./McMillan-Major, A./Shmitchell, S. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 610–623. DOI: 10.1145/3442188.3445922