Evaluating ChatNetZero, an LLM-Chatbot to Demystify Climate Pledges

Angel Hsu[1,2], Mason Laney[1,2], Diego Manya[1,2], Ji Zhang[3] Linda Farczadi[3]
[1]The University of North Carolina at Chapel Hill, [2]Data-Driven EnviroLab, [3]Arboretica
angel.hsu@unc.edu, mlaney@cs.unc.edu, diego.manya@unc.edu
james@arboretica.com, linda@arboretica.com

Abstract

This paper introduces and evaluates ChatNetZero, a large-language model (LLM) chatbot developed through Retrieval-Augmented Generation (RAG), which uses generative AI to produce answers grounded in verified, climate-domain specific information. We describe ChatNetZero's design, particularly the innovation of anti-hallucination and reference modules designed to enhance the accuracy and credibility of generated responses. To evaluate ChatNetZero's performance against other LLMs, including GPT-4, Gemini, Coral, and ChatClimate, we conduct two types of validation: comparing LLMs' generated responses to original source documents to verify their factual accuracy, and employing an expert survey to evaluate the overall quality, accuracy and relevance of each response. We find that while ChatNetZero responses show higher factual accuracy when compared to original source data, experts surveyed prefer lengthier responses that provide more context. Our results highlight the importance of prioritizing information presentation in the design of domain-specific LLMs to ensure that scientific information is effectively communicated, especially as even expert audiences find it challenging to assess the credibility of AI-generated content.

1 Introduction

In the era of generative AI, the proliferation of climate change misinformation presents a significant challenge, impeding both scientific discourse and efforts to distinguish between credible and non-credible climate actions. Although scientific consensus has identified the imperative to achieve "net-zero" emissions by mid-century, widespread disagreement over its definition and implementation remains (Fankhauser et al., 2022). For example, according to the latest Pew Research Center Poll, two-thirds of Americans say that the US should use a mix of energy sources, including fossil fuels, which are fundamentally incompatible with a net-zero world (Tyson et al., 2022). A surge of over 11,000 private and subnational entities have committed to respective decarbonization pledges, albeit with varying degrees of credibility (Institute, 2024; UNFCCC, 2023). Across the world, citizens and regulators are increasingly resorting to litigation to combat false and disingenuous net-zero claims (Carrington, 2023).

With more users relying on artificial intelligence-driven large language models (LLMs) like Google's Gemini (Gemini Team, 2024) and OpenAI's ChatGPT (OpenAI, 2022) to obtain primary information, it is a foregone conclusion that the public will turn to these resources to gain a deeper understanding of what governments and businesses are doing on climate change and decarbonization. These tools, however, are not attune to the rapidly evolving landscape of climate policy, specifically the nuances of net-zero goals, where definitions and interpretations of credibility are evolving daily. Since non-state actors report climate actions in a variety of formats (e.g., press releases, PDF reports, spreadsheets, websites, etc.) (Hsu and Rauber, 2021), even the task of assembling a coherent dataset to analyze and compare entities' climate change strategies is challenging. Generative AI is prone to "hallucination," where models produce seemingly real responses that could fail to correspond to any actual input, posing potential risks of hazardous and legally-disputable claims (Alkaissi and McFarlane, 2023).

Here we introduce and evaluate ChatNetZero—an LLM-based chatbot developed through Retrieval-Augmented Generation (RAG)—which employs generative AI to produce answers to users' questions that are grounded in verified information (Lewis et al., 2021). It is designed to analyze unstructured net-zero related text documents and serve as a question-answering platform for climate policy-specific information.

ChatNetZero is able to accurately answer questions relating to broad net-zero domain knowledge, such as different terminology used to articulate net-zero commitments, as well as specific details on an entity's net-zero pledge and its content. To evaluate ChatNetZero's ability to provide accurate, high-quality responses, we assess its responses in two ways - comparing the generated responses to original expert texts; and engaging climate policy experts to evaluate its responses compared to other population chatbots, including GPT-4 (OpenAI, 2023), Gemini (Gemini Team, 2024), and Coral (Cohere, 2023), as well as the climate-domain specific ChatClimate (Vaghefi et al., 2023).

2 Background

2.1 The science of net zero

The concept of "net zero" refers to the equilibrium between human-caused greenhouse gas emissions and their removal, either through natural means such as carbon sinks (like oceans, land, and forests) or engineered methods like carbon capture and storage or direct air capture. Although rooted in climate science since the 2000s, its significance surged politically with the 2015 Paris Agreement. This historic Accord marked the first global commitment to limit the temperature rise to 1.5°C above pre-industrial levels, necessitating net-zero emissions mid-century (IPCC, 2018) and inspiring non-governmental actors to undertake their own net-zero initiatives (UNFCCC, 2023).

Major questions, however, continue to surround the credible and scientific implementation of net-zero pledges, particularly regarding whether entities intend to completely eliminate emissions or plan to offset them by purchasing questionable credits from reductions elsewhere. Assessing the legitimacy of these commitments is challenging due to the prevalence of greenwashing, where numerous companies and government bodies engage in superficial efforts that mislead the public. Additionally, the public often lacks the necessary tools to discern credible or genuinely high-integrity climate pledges, as such evaluations typically require expert knowledge.

2.2 Previous applications of NLP for climate change

The potential of generative AI and LLMs to significantly improve access to climate-related information is rapidly gaining recognition, evidenced by the increasing number of initiatives to develop climate-domain specific LLMs and chatbots in recent years. ClimateBERT was one of the first specialized transformer-based language models that was pre-trained on over 2 million climate-related texts, including news sites, research articles and company climate reports (Webersinke et al., 2022). The authors found that ClimateBERT outperformed a base LLM without domain-adaptive training (DistilRoBERTa) in text classification tasks identifying whether a text contained climate-related material. ClimateBERT-NetZero builds on ClimateBERT by detecting net zero or reduction targets in texts, leveraging the Net Zero Tracker data as a labeled dataset to pretrain the ClimateBERT classifier, which results in superior predictive performance compared to larger models (Schimanski et al., 2023). ChatClimate (Vaghefi et al., 2023) is a chatbot that instructs GPT-4 to only provide answers based on the IPCC's climate science reports (IPCC, 2023). Others (i.e., ClimSight, see Koldunov and Jung (2024)) are experimenting with ways of combining physical climate data and LLMs to make data and information from climate models, including large-scale global precipitation and weather data, more accessible to users.

2.3 Limitations of climate-related LLMs

The development of domain-specific LLMs and general LLM applications highlights a growing demand for resources to enhance understanding of climate science and the actions taken by governments and businesses to address climate change and decarbonization. Beyond the well-documented hallucination problem, climate-related LLM applications are susceptible to replicating or exacerbating greenwashing, especially when trained on self-reported climate action data, which is often at risk of 'net-zero greenwashing' due to misalignment between climate pledges and corporate actions (InfluenceMap, 2023). The climate domain is also particularly prone to misinformation and political polarization in social media and other outlets (de Freitas Netto et al., 2020; Thapa Magar et al., 2024), a particular challenge for even climate-related LLMs to distinguish (Leippold et al., 2024).

3 Methods

3.1 Data sources

We worked with experts from the Net Zero Tracker to identify the most relevant and credible documents with which to supply ChatNetZero. Since we do not want to contaminate the data retrieval process with possible greenwashing or falsehoods from the entities themselves (e.g., a company's own corporate responsibility report or a government's own climate action strategy), we initially only use four sources of information to ground ChatNetZero's beta pilot:

- **The United Nations High-Level Expert Group (HLEG) report** on *Integrity Matters: Net-Zero Emissions Commitments of Non-State Entities* (HLEG, 2022): The HLEG report provides ten recommendations for companies, financial institutions, and subnational governments to establish credible net-zero pledges that are aligned with scientific scenarios and recommendations.
- **Net Zero Tracker database** and **Net Zero Stocktake reports**: The Net Zero Tracker (zerotracker.net) is the most comprehensive platform evaluating more than 4,000 entities' net-zero and decarbonization efforts. These entities include all national governments, all regions in the G20, all cities with a population greater than 500,000, and the Forbes Global 2000 companies. The dataset evaluates whether an entity has declared a net-zero or similar decarbonization pledge in addition to more than a dozen indicators assessing their integrity. We also include the Net Zero Stocktake reports, which are annual reports assessing the status and trends of net zero targets in the database (Net Zero Tracker, 2022, 2023).
- NewClimate Institute's **Corporate Climate Responsibility Monitor Reports** (New Climate Institute, 2022, 2023): These reports authored by the NewClimate Institute, a German-based climate policy think tank, evaluate the credibility of net-zero targets and policies set by 25 multinational companies, including Maersk, IKEA, Apple, Google, and H&M, among others.

Table 1 describes a summary of the final data used to train ChatNetZero. While these documents and data sources are not the singular authorities regarding net-zero and decarbonization policy, they represent a set of consistent and coherent benchmarks to ground ChatNetZero. Other documents, including The Oxford Principles for Net Zero Aligned Carbon Offsetting (Allen et al., 2020) or British Standard Institute's Net Zero Target-Setting Standards (Institute, 2023) may represent diverging viewpoints (i.e., regarding the use of offsets when an entity cannot meet its own emission reduction targets solely through its internal efforts) and were not used for ChatNetZero's pilot, but would not necessarily be excluded from future model design and development.

Description	Number
Number of spreadsheet chunks	21,154
Number of report chunks	5,355
Number of tokens in spreadsheet data	1,781,790
Number of tokens in report data	342,908

Table 1: Summary of Data Chunks and Tokens

3.2 ChatNetZero Design

To tackle the limitations of generic LLMs (i.e., hallucination), we developed a Retrieval-Augmented Generation design combined with other customized algorithms, including query processing, analytical text transformation, and chunk ranking algorithms. ChatNetZero also provides references with each answer that includes active hallucination checks that provide specific document and page references to users (see below sections 3.2.2 and 3.2.3). ChatNetZero's workflow is illustrated in Figure 1, and we describe each algorithmic module in greater detail below.

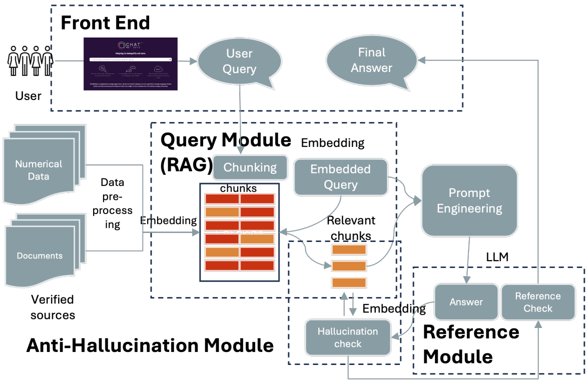

Figure 1: ChatNetZero's Query, Reference, and Anti-Hallucination module workflow.

3.2.1 RAG Module

Our RAG module entails a multi-step process to chunk, embed, query process, and customize responses to a user-inputted question.

Chunking All source documents, including Excel and PDF data, is converted into plain text, which is then segmented into chunks ranging from 50 to 1,000 words. To maintain the source data structure, each chunk encapsulates entire paragraphs. Each chunk is then embedded into a large, high-dimensional numerical vector, which represents the meaning of the text (Mikolov et al., 2013). For traceability, every chunk is tagged with its originating document's name and page number, facilitating later checks against potential hallucination.

Embedding We embed both the chunked source documents and the user queries using OpenAI's `text-embedding-ada-002` model. This embedding space is used to perform semantic search between user queries and chunks from the source documents, allowing us to find the most relevant chunks to inform ChatNetZero's final output response (see section 3.2.1). fate chose a small chunk size because when the response chunks were significantly longer than the user query, we found that the semantic search performed poorly, resulting in the selection of chunks that were not relevant to the user's question.

Query processing We designed ChatNetZero to handle two types of user queries: actor-specific queries, where an individual entity or multiple entities such as a government or company are named; and generic queries, where a user asks a question that doesn't identify a specific entity. For actor-specific queries, we developed an algorithm to recognize if the user's query mentions a specific actor included in the Net Zero Tracker data and to then prioritize that data for answering. The algorithm handles irregular spelling, abbreviation, and translation of actor names, and has enhanced capability to cover all actors mentioned in a long query which effectively combats the "laziness" of LLMs when answering long questions (Guo et al., 2023). For actor-specific queries, at most 25 embedded chunks are retrieved from the backend, with between 1 and 5 chunks per actor. If the number of chunks exceeds 25, we reduce it to 25 while ensuring that each entity retains at least one chunk. We limit the number of chunks per entity to a maximum of 5. This process involves enforcing at least one chunk per entity specifically from the NZT Excel data, not from embedded reports. Additionally, our rule-based algorithm builds on the top-k chunk algorithm. Initially, we employ a Named Entity Recognition (NER) algorithm to identify any actors mentioned in the query, then we select chunks related to these actors from the NZT Excel data, and finally, we choose the top-k chunks from the remaining embedded documents. For generic queries, we select the top 10 most relevant chunks from the report documents, as determined by semantic similarity to the query.

Prompt Engineering We then take the retrieved chunks during the query processing step and combine them with the user prompt and then send it to OpenAI's GPT-4 Turbo model. We use a temperature of 0.0 in order to ensure that the model produces reliable and consistent output. The model is instructed to follow a set of guidelines (see Figure 2) designed to facilitate clear and truthful answers.

Architecture ChatNetZero's backend utilizes the LangChain architecture (Harrison, 2022) to allow for future interchange of LLMs without affecting the algorithmic process.

1. Your response must be precise, thorough, and solely based on the textual information provided.
2. Do not use embellishing language. Keep your answer as similar as possible to the original data.
3. If an entity is mentioned in the query, be sure to include mention of it in the answer.
4. Only use the pieces of information that you need to formulate a detailed answer.
5. If you are unsure, simply acknowledge the lack of knowledge, rather than fabricating an answer.
6. Keep your ANSWER within 100 words.

Figure 2: Guidelines included in the prompt given to GPT-4 Turbo

3.2.2 Anti-Hallucination Module

We developed an anti-hallucination module that first processes the raw output of the GPT-4 Turbo LLM after the Prompt Engineering step described above by dividing it into sentences and embedding them using the same process as described above (Embedding). Each vectorized output sentence is then compared against selected chunks from the RAG Module to verify its origin; sentences that cannot be traced to an original chunk are then excluded, given untraceable sentences' high potential for hallucination. To evaluate the performance of the anti-hallucination module, we conducted several assessments with the Net Zero Tracker team, which is comprised of climate science and policy experts and 300 volunteers who have helped to collect and validate data on the Tracker's 4,000+ entities.

3.2.3 Reference Module

This module enables automated validation of ChatNetZero's outputs and ensures traceability to one of the original data sources (see above). If a sentence successfully passes the anti-hallucination algorithm, the module appends a citation to the corresponding report to the generated response, including the page number and sentence position of the matched content. The module's output, presented to the user via a web application, includes references for each sentence. These references link directly to the original pages of the source material so users can manually check and validate ChatNetZero's generated response.

3.2.4 Enhanced Analytical Capabilities

To address LLMs' inherent limitations in mathematical tasks, we developed an algorithmic process that enhances the model's utility in interpreting and responding to queries requiring analytical analysis (for example, "How many companies in Germany have pledged a net-zero target?"). The algorithm restructures the Net Zero Tracker dataset, which tracks over 30 net-zero variables for over 4,000 actors, from an Excel tabulated format into optimized natural language sentences. This transformation enables the numerical data to be retrieved using the same process as text chunks, enabling the LLM to utilize numerical net-zero data and significantly enhancing the range of questions that ChatNetZero can deliver to its users.

3.3 Validation

3.3.1 Factual Evaluation

To assess the factual accuracy of ChatNetZero, we prompted four other large language models, including ChatClimate, GPT-4 Turbo, Gemini 1.0 Ultra, and Coral with Web Search with eight questions (Figure 3) that relate to factual statements regarding details of specific climate actors' net zero or climate pledges. We used reputable sources—such as official policy documents and corporate reports—as ground truth reference material. The evaluation strictly assessed factual accuracy by determining if responses (found in Appendix A) exactly matched the reference material. We analyzed two aspects of the LLM responses. First, whether the LLM provided a direct and correct answer to the question provided:

- If the question asked about conditions for the use of offsets for B company, we evaluated whether the LLM provided a direct answer to that question (i.e, B Company has/doesn't have conditions in the use of offsets), regardless of other contextual or additional statements included in the answer.
- If the reference material indicated that a company's climate target was to reduce 30% emissions by 2050, we expect a correct answer to include both figures (i.e., the 30% and 2050 target year) when describing the climate targets of the company.

If the LLM provided an exact match to the data provided in the source material, we assigned a score of 1; if not, we scored the response 0.

Second, we evaluated each factual sentence in an LLM's answer individually either as 'Correct', 'Incorrect', or 'Unverifiable', regardless of whether they addressed the main question or if they were simply contextual statements. We report this score as the ratio of correct factual statements to the total number of factual statements.

1. How does Walmart's climate goals compare with Amazon's and other large retail stores?

2. How many nations in the world have a net zero target enshrined in law?

3. How many companies rule out the use of offsets / credits for their net zero targets?

4. Does 3M or Pfizer have any conditions on the use of offsets?

5. How do the United States, China, Wal-Mart, Apple and California compare in terms of their decarbonization efforts

6. How does Foxconn's climate goals compare with Fast Retailing's? Limit response to 100 words and use your most recent information, including databases and searching online.

7. How does VakifBank and Saudi Aramco compare in terms of their climate policy's end target status? Limit response to 100 words and use your most recent information, including databases and searching online.

8. How does Reliance Industries and Emaar Properties compare in terms of their climate interim targets? Limit response to 100 words and use your most recent information, including databases and searching online.

Figure 3: Domain-specific questions posed to each LLM for evaluating factual accuracy of responses.

3.3.2 Expert evaluation

Beyond assessing factual accuracy, we posed 12 questions (see Figure 4) to each large language model. We then anonymized and randomized their

1. Can a company pledge net zero by 2050 and still plan to utilize fossil fuels?

2. Can a company rely on offsets and still claim credible net zero?

3. Does 3M or Pfizer have any conditions on the use of offsets?

4. What are Scope 3 emissions and what categories of Scope 3 emissions should a company/subnational government include in a net-zero target?

5. If a city or subnational government doesn't have control of out of boundary emissions (e.g., electric utilities), how can it credibly set a net-zero target?

6. What is an example of a company (or country) that has produced a 'good plan' to achieve their target?

7. What constitutes a credible net-zero target?

8. What are examples of greenwashing in corporate net-zero targets?

9. What does it mean for a company's net-zero target to be 1.5C aligned?

10. Is Apple's net-zero target credible?

11. What does it mean for an entity to contribute a 'fair-share' of emissions reductions?

12. Is Wal-Mart greenwashing its climate commitments?

Figure 4: Domain-specific questions posed to each LLM for expert assessment of response quality, accuracy, and relevance.

responses (found in Appendix A) in a Qualtrics survey, which we distributed to 10 climate scientists and policy experts. While ChatNetZero was designed to include references for each response, and some LLMs (including Gemini and Coral) provide references as well, we removed these from the responses for the Qualtrics survey so that experts would evaluate the quality of the responses themselves. These experts were asked to evaluate each response across three dimensions: overall quality, factual accuracy, and relevance. Respondents were asked to evaluate each response on a scale of 1 to 5, with 5 being the highest and 1 being the lowest. They were also given the opportunity to provide qualitative comments.

4 Results

4.1 Factual Evaluation

Table 2 provides a summary of the scores for our assessment of the factual accuracy of five LLM outputs, including ChatNetZero. Overall, our evaluation shows that ChatNetZero has higher factual scores for both its answers to the question itself and for the rest of additional information that provides more context or complements the main answer to the prompted question. For example, when asked, "How does Wal-mart's climate goals compare with Amazon's and other large retail stores?" (Figure 3, Question 1), ChatNetZero provided more factually accurate answers than the other LLMs (see Appendix A for the factual scores of individual responses). An exception was GPT-4, which displayed a similar level of accuracy. Both provided factually correct responses to the main question and had the majority of their factual statements verified as correct. However, when asked a similar question about two non-English entities such as Foxconn and Fast Retailing (Figure 3: Question 6), the factual accuracy scores of ChatNetZero were higher than all other LLMs, many of which were unable to provide complete answers, likely due to limitations in their training data.

4.2 Expert Evaluation

Across all 12 questions, experts evaluated Gemini Ultra—followed closely by GPT-4—as producing the highest quality responses overall (3.91±0.91), with the greatest relevance (4.0±0.96) and factual accuracy (3.9±0.91) (see Table 3). ChatNetZero yielded the lowest overall quality (2.64±0.87), relevance (2.92±0.94), and factual accuracy (2.94±1.07) of the LLMs evaluated. As Figure 5 illustrates, however, performance varied by question, and there were several questions where ChatNetZero was evaluated to have on average a comparable or better overall response compared to Gemini, such as Question 3: "Does 3M or Pfizer have any conditions on the use of offsets." We provide responses from ChatNetZero and the other LLMs evaluated in Appendix A.

We believe that the relatively low performance of ChatNetZero was in part due to its shorter average response length (110.5±8.91 words) compared to the other LLMs (Table 4). GPT had the largest average number of words per response (434.92±70.75 words), followed by Gemini (361.25±91.32 words). We found a generally positive correlation between an LLM's response length and the expert evaluated quality of the response (Figure 6), with the experts' evaluation of the factual accuracy of a response most closely related to the word length of the response (R^2=0.43).

5 Discussion

The design of ChatNetZero and our comparison of its outputs to one climate domain-specific LLM (ChatClimate) and other popular LLMs (Gemini,

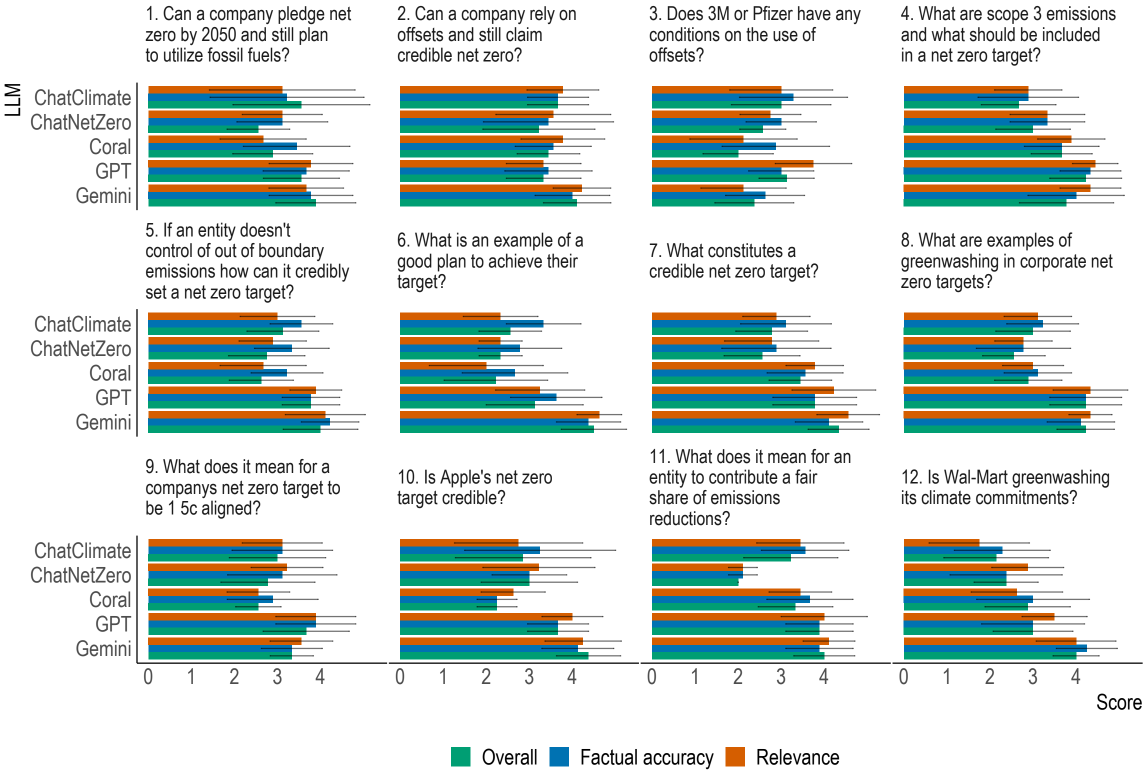

Figure 5: Average expert evaluation scores for overall quality, factual accuracy, and relevance of LLM responses to 12 climate policy and net-zero related questions. Bars show mean responses (scored from 1 to 5; with 5 being the highest) and lines show standard deviation from the means. Questions have been shortened for presentation. See Figure 4 for actual survey questions.

Model	Step 1	Step 2
ChatNetZero	0.75	0.79±0.15
ChatClimate	0.25	0.25±0.46
GPT	0.375	0.54±0.34
Gemini	0.375	0.35±0.44
Coral	0.375	0.65±0.34

Table 2: Summary results of factual evaluation of LLM responses to questions posed in Figure 3. Step 1 was determined by the following scoring: 1=Correct Answer; 0=Wrong or No Answer. Step 2 was determined as a ratio of correct factual statements to the total number of factual statements in the response.

Model	Relevance	Factual	Overall
ChatNetZero	2.92±0.94	2.94±1.07	2.65±0.87
ChatClimate	2.94±1.12	3.22±1.15	2.98±1.09
GPT	3.88±0.88	3.70±0.93	3.63±0.90
Gemini	4.00±0.96	3.90±0.91	3.91±0.91
Coral	2.94±1.08	3.17±1.04	2.87±0.93

Table 3: Mean expert ratings of LLM responses across 12 climate policy and net-zero questions (Figure 4).

Model	mean length	stdev
ChatNetZero	110.50	8.91
ChatClimate	167.00	80.68
GPT	434.92	70.75
Gemini	361.25	91.32
Coral	258.67	66.40

Table 4: Average word length of responses generated by LLMs evaluated in this study.

GPT-4, and Coral) underscores several findings about the use of LLMs in navigating the complex landscape of climate policy, particularly in relation to rapidly shifting and emerging concepts like “net zero.”

Length vs. perception of accuracy

While ChatNetZero was designed to deliver concise and accurate responses—confirmed by our factual evaluation comparing responses to source documents—our expert evaluation showed a preference for longer, more detailed answers that of-

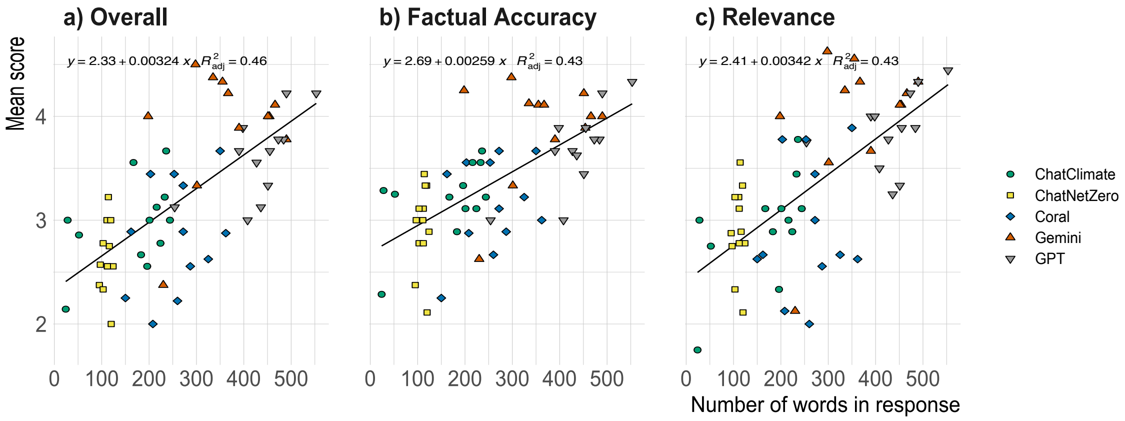

Figure 6: Comparison of word count of LLM-generated responses to climate policy and net-zero concepts versus expert evaluations of responses' a) overall quality; b) factual accuracy; and c) relevance. Experts were asked to evaluate each LLM response on a scale of 1-5, with 5 being the highest.

fer broader context, even if the added information isn't always accurate or verifiable (Tables 2, 3, 4). This "verbosity bias" (Saito et al., 2023) indicates that humans tend to prefer longer, more detailed answers, believing they are more accurate than concise ones. Similarly, (Chiesurin et al., 2023) found that users favor fluent, grammatical responses and sophisticated linguistic dialogue, even when these responses lack trustworthy information.

We reviewed experts' qualitative comments to gain further insight. Notably, experts who provided additional comments regarding the length of the evaluated LLMs' responses said they favored longer answers provided by Gemini and GPT-4. This preference likely contributed to the higher scores for more extensive responses compared to the more concise bullet-point answers from ChatNetZero, which were designed for brevity but were seen as disrupting the flow of information and reducing readability, despite the high accuracy of ChatNetZero's responses compared to the lengthier responses of other LLMs like Gemini and GPT-4.

Balancing factual accuracy with contextual relevance

Our finding of ChatNetZero's higher factual accuracy but lower expert evaluation compared to other models suggests that while factual correctness can be achieved through grounding a general-purpose LLM, the utility of responses in practical scenarios also heavily depends on the completeness and contextual alignment of the information provided.

Expert feedback suggests that a model's ability to integrate accurate data into contextually relevant responses is essential. Responses that simply list facts without a nuanced understanding of the topic may fail to meet users' needs for clear, actionable insights. This is especially critical in complex areas like climate policy, where decisions depend not only on data but also on its interpretation within diverse socio-economic and environmental contexts. However, the preference for lengthier responses from ChatNetZero over shorter statements might also be due to the specific user group in our study. Since our respondents were limited to experts in climate science and policy, including a more diverse or less specialized participant base could lead to different results.

Distinguishing factual accuracy in LLM-generated responses

Our study further demonstrates the challenge of utilizing LLMs to distinguish between accurate and irrelevant or even hallucinated content. This distinction is critical, as misinformation or misinterpretations in such a technical and impactful field can lead to poor decisions and public misconceptions. Although ChatNetZero generally provides factually correct responses, distinguishing these from less relevant or lengthier, contextualized answers remains difficult. This issue was reiterated by (Bulian et al., 2023), who reviewed climate responses from several LLMs and found that while the models scored high in information presentation, they were weak in the quality of the content provided.

We found that many responses from other LLMs contained auxiliary statements that were not fact-based. While these statements enhance the answers' readability, they could potentially lead to misinterpretation about the validity or adequacy of the responses. For instance, in our earlier example about Walmart and Amazon (Figure 3: Question 1), one LLM stated, "As of 2023, Walmart, Amazon, and other large retail stores have been increasingly vocal and active in their commitments to sustainability and addressing climate change." This introductory statement, without specific evidence or references to their enhanced vocal and active roles, could potentially lead to issues like greenwashing if not carefully scrutinized.

This result from our study highlights key lessons from the science communication literature, which emphasizes the importance of information presentation: scientific information should be comprehensible, aid understanding through layout and visualizations, and use appropriate sources and references (Bulian et al., 2023; Jamieson et al., 2017). Since we removed the reference features from all LLMs in our human evaluation study, users were not presented with this third criterion for presentational adequacy of scientific information—sources and citations—which might have influenced their evaluation of the overall quality or factual accuracy of the LLM responses. As a result, users were neither able to individually verify the accuracy of the responses nor use this feature to gauge response quality. Future validation efforts could involve asking users to assess LLM responses in conjunction with the provided references and to evaluate the sources themselves.

5.1 Future implications

Although climate-specific LLMs can enhance the understanding and application of climate strategies, additional research is needed to explore how the framing, length, and presentation of responses affect users' comprehension and perception. For regulators, they promise efficient, accessible information to facilitate the examination and confirmation of climate commitments, with the potential to foster greater scrutiny and trust through transparency. Businesses and other entities could also use these tools as an important benchmarking platform to understand which competitors have developed high-integrity, credible climate efforts. For advocates and the public, they promise the ability to hold entities accountable, provide access to reliable information, and engage more effectively in climate action discussions.

6 Limitations

Our study here is not without its limitations. The number of experts responding to our study (around 10 in total) was relatively modest and mainly represented the academic, scientific and policy communities. In the future, we could expand user evaluation to other demographics, including business, government, activist, or non-expert audiences. Second, by removing references from the LLM-generated responses, it's unclear how their inclusion by ChatNetZero and Gemini might have influenced experts' views or the perceived quality, accuracy, or relevance of the responses. Last, different parameter tunings, for example an LLM's temperature parameter, could result in a range of different responses (Dahl et al., 2024), which may affect a user's interpretation. Future evaluation could test the sensitivity of a user's evaluation to the temperature parameter on ChatNetZero's responses. Last, we acknowledge that ChatNetZero still assumes a base-level understanding of climate change concepts, including net zero. Future versions should consider whether additional user prompting or response framing should be presented to assess the user's baseline knowledge or prompt them to specify their level of understanding.

7 Conclusion

This study underscores the critical role and potential of specialized large language models like ChatNetZero in enhancing understanding and engagement in climate policy discourse. By demonstrating higher factual accuracy in handling complex climate-related content, ChatNetZero shows promise as a valuable tool for disseminating reliable information. However, challenges such as ensuring the presentation of information and contextual depth of responses remain. Future enhancements to ChatNetZero will consider options allowing users to customize response length and style. Adding more language could potentially compromise response quality, a challenge that will need further refinement. Addressing these challenges will be crucial for maximizing the utility of LLMs, ensuring that a customized LLM such as ChatNetZero provides transparency in its algorithmic processes to instill the trust and confidence required for any tool to impact decision making.

References

Hussam Alkaissi and Samy I. McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2). Publisher: Cureus.

Myles Allen, Kaya Axelsson, Ben Caldecott, Thomas Hale, Cameron Hepburn, Eli Mitchell-Larson, Yadvinder Malhi, Friederike Otto, and Nathalie Seddon. 2020. The Oxford Principles for Net Zero Aligned Carbon Offsetting 2020.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels Mede, Markus Leippold, and Nadine Strauss. 2023. Assessing Large Language Models on Climate Information. *arXiv preprint*. ArXiv:2310.02932 [cs].

Damian Carrington. 2023. Shell directors personally sued over 'flawed' climate strategy. *The Guardian*.

Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. *arXiv preprint*. ArXiv:2305.16519 [cs].

Cohere. 2023. Introducing coral, the knowledge assistant for enterprises.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93. ArXiv:2401.01301 [cs].

Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32(1):19.

Sam Fankhauser, Stephen M. Smith, Myles Allen, Kaya Axelsson, Thomas Hale, Cameron Hepburn, J. Michael Kendall, Radhika Khosla, Javier Lezaun, Eli Mitchell-Larson, Michael Obersteiner, Lavanya Rajamani, Rosalind Rickaby, Nathalie Seddon, and Thom Wetzer. 2022. The meaning of net zero and how to get it right. *Nature Climate Change*, 12(1):15–21. Publisher: Nature Publishing Group.

Google Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint*. ArXiv:2310.19736 [cs].

Chase Harrison. 2022. Langchain.

HLEG. 2022. Integrity Matters: Net Zero Commitments by Businesses, Financial Institutions, Cities and Regions. *United Nations' High-Level Expert Group on the Net Zero Emissions Commitments of Non-State Entities*.

Angel Hsu and Ross Rauber. 2021. Diverse climate actors show limited coordination in a large-scale text analysis of strategy documents. *Communications Earth & Environment*, 2(1):30. Publisher: Nature Publishing Group UK London.

InfluenceMap. 2023. "Net Zero Greenwash": The Gap Between Corporate Commitments and their Policy Engagement. Technical report.

British Standards Institute. 2023. BSI Net Zero Pathway.

NewClimate Institute. 2024. Corporate Climate Responsibility Monitor 2024. Technical report.

IPCC. 2018. Global Warming of 1.5°C.An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. Technical report, IPCC, Geneva.

IPCC. 2023. Summary for Policymakers. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, IPCC, Geneva.

Kathleen Hall Jamieson, Dan Kahan, and Dietram A. Scheufele. 2017. *The Oxford handbook of the science of science communication*. Oxford University Press.

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

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated Fact-Checking of Climate Change Claims with Large Language Models. *arXiv preprint*. ArXiv:2401.12566 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint*. ArXiv:1310.4546 [cs, stat].

Net Zero Tracker. 2022. Net Zero Stocktake 2022: Assessing the status and trends of net zero target setting across countries, sub-national governments and companies. Technical report, Net Zero Tracker.

Net Zero Tracker. 2023. Net Zero Stocktake 2023: Assessing the status and trends of net zero target setting across countries, sub-national governments and companies. Technical report, Net Zero Tracker.

New Climate Institute. 2022. Corporate Climate Responsibility Monitor 2022: Assessing the transparency and integrity of companies' emission reduction and net-zero targets. Technical report, New Climate Institute.

New Climate Institute. 2023. Corporate Climate Responsibility Monitor 2023: Assessing the transparency and integrity of companies' emission reduction and net-zero targets. Technical report, New Climate Institute.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity Bias in Preference Labeling by Large Language Models. *arXiv preprint*. ArXiv:2310.10076 [cs].

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets. *arXiv preprint*. ArXiv:2310.08096 [cs].

Neelam Thapa Magar, Binay Jung Thapa, and Yanan Li. 2024. Climate Change Misinformation in the United States: An Actor–Network Analysis. *Journalism and Media*, 5(2):595–613. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Alec Tyson, Cary Funk, and Brian Kennedy. 2022. Americans Largely Favor U.S. Taking Steps To Become Carbon Neutral by 2050. *Pew Research Center*.

UNFCCC. 2023. Race To Zero Campaign.

Saeid Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti Senni, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatIPCC: Grounding Conversational AI in Climate Science. *SSRN Electronic Journal*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text. *arXiv preprint*. ArXiv:2110.12010 [cs].

A Supplementary Material

Data from the factual and expert evaluations can be found on our Dataverse.