

Climate Change from Large Language Models

Anonymous ACL submission

Abstract

Climate change poses grave challenges, demanding widespread understanding and low-carbon lifestyle awareness. Large language models (LLMs) offer a powerful tool to address this crisis, yet comprehensive evaluations of their climate-crisis knowledge are lacking. This paper proposes an automated evaluation framework to assess climate-crisis knowledge within LLMs. We adopt a hybrid approach for data acquisition, combining data synthesis and manual collection, to compile a diverse set of questions encompassing various aspects of climate change. Utilizing prompt engineering based on the compiled questions, we evaluate the model’s knowledge by analyzing its generated answers. Furthermore, we introduce a comprehensive set of metrics to assess climate-crisis knowledge, encompassing indicators from 10 distinct perspectives. These metrics provide a multifaceted evaluation, enabling a nuanced understanding of the LLMs’ climate crisis comprehension. The experimental results demonstrate the efficacy of our proposed method. In our evaluation utilizing diverse high-performing LLMs, we discovered that while LLMs possess considerable climate-related knowledge, there are shortcomings in terms of timeliness, indicating a need for continuous updating and refinement of their climate-related content.

1 Introduction

The climate crisis, exacerbated by fossil fuel burning, deforestation, and industrial processes, poses a grave global threat. Its impacts range from rising sea levels to intensified weather events and biodiversity loss. Addressing this crisis is urgent, prompting widespread efforts to reduce greenhouse gas emissions and adopt more sustainable practices (Schimanski et al., 2024). In this context, large language models (LLMs) like GPT-4 (OpenAI, 2023) can play a vital role in raising awareness and educating the public about the climate emergency.

LLMs have the potential to reach a global audience and provide accurate, up-to-date information on the causes and consequences of the climate crisis. They can also engage in discussions (Stede and Patz, 2021) with users, answering questions and addressing concerns related to climate change.

Existing LLMs have access to a significant amount of information related to the climate crisis, but this knowledge is often underutilized due to the models’ lack of interpretability. Furthermore, the quality of climate crisis-related responses generated by LLMs has not been thoroughly evaluated, which limits their potential to provide valuable insights to researchers, policymakers, and other stakeholders involved in addressing climate issues. Existing methodologies for evaluating LLMs in general domains are inadequate for climate-crisis knowledge. This paper aims to analyze the challenges and opportunities associated with leveraging LLMs for climate crisis knowledge and propose a methodology to extract and assess the quality of this knowledge in an explainable way. Our approach involves eliciting climate crisis knowledge from LLMs through designed prompts and evaluating the quality of this knowledge using comprehensive metrics.

Extracting climate crisis knowledge from LLMs is a non-trivial task due to limited interpretability. Our approach aims to improve understanding and evaluation of this knowledge, enabling a more human-interpretable assessment of their capabilities. We symbolize the parameter knowledge in the text through elaborately designed prompts. To assess the knowledge accurately, we require a substantial number of relevant questions and answers. We developed a pipeline to generate and curate such questions by combining outputs from LLMs with public datasets. We then utilize LLMs to provide answers to these questions.

The second challenge is evaluating knowledge related to the climate crisis. Prior studies have

primarily relied on perplexity to assess generated content, but this approach falls short in accurately capturing knowledge from a human cognitive perspective. Certain research efforts have resorted to human evaluation, an approach that can be both costly and time-consuming. Other studies have attempted to utilize classifiers to grade answers, yet these methods prove inadequate for accurately evaluating knowledge pertinent to the climate crisis. To address this issue, we propose a method to automatically evaluate the knowledge of LLMs related to the climate crisis by evaluating the quality of questions and answers. We first propose 5 metrics for evaluating questions (importance, clarity, relevance, difficulty, and innovation) and another 5 metrics for evaluating answers (relevance, depth, readability, innovation, and timeliness). We leverage high-performing LLMs to score questions and answers, then average the scores for comprehensive assessment. This integrated approach enhances evaluation accuracy and reliability.

The contributions of this paper are as follows:

- (1) We propose a method to symbolize and assess the knowledge of climate crisis within LLMs.
- (2) We present an approach to collect questions and answers related to the climate crisis and use LLMs to automatically evaluate the LLMs' knowledge related to the climate crisis.
- (3) We introduce 5 question metrics and 5 answer metrics for objective scoring. Experimental findings validate the effectiveness of our method and highlight the limitations of LLMs in this context.

2 Related Work

2.1 Large Language Models for Climate Change

Global climate change is a significant challenge that necessitates a multidisciplinary approach. Artificial intelligence (AI) and natural language processing (NLP) technologies, such as ChatGPT, have potential applications (Kraus et al., 2023) in climate research, including model parameterization, data analysis, scenario generation, and evaluation. These techniques contribute to enhancing the accuracy of climate predictions and provide robust tools for researchers and policymakers. Machine learning (ML) workloads (Rolnick et al., 2022) are rapidly growing in importance, but their carbon footprint is a concern. Google has managed to keep ML training energy use below 15% of total energy

use over the past three years by implementing best practices. It is suggested that these practices be adopted throughout the ML field to significantly reduce the carbon footprint of model training. The application of LLM technology contributes to accurately analyzing the trends and impacts of climate change, providing strong support for sustainable development in the field of ESG (Environment, Social, and Governance) (Stammbach et al., 2022), and promoting the achievement of a green and low-carbon future.

LLMs, like GPT-3, are widely used in various fields, including entertainment, health, and finance (Zhu, 2023a). However, their performance can be uneven when interacting with different social groups (Vaghefi et al., 2023). Leippold (2023) suggest an analytical framework to evaluate fairness in human-AI conversations. By analyzing over 20,000 conversations about climate change and the Black Lives Matter movement, they find that GPT-3 performs well when engaging with educational and minority groups regarding viewpoints. These groups not only received accurate and unbiased information but also changed their attitudes and expressed support for related actions after the dialogue. LLMs have achieved remarkable results in AI, but they still use imprecise language in areas where accuracy is critical, such as climate change. Webersinke et al. (2021) overcome its limitations and improve reliability by treating LLM as a proxy for accessing multiple sources such as Climate-Watch and general Google searches for the latest accurate climate data.

Climate change poses a significant threat to human health, and effective, evidence-based policies are needed to mitigate or eliminate these risks. This necessitates the translation of scientific knowledge into policy. To address this challenge, Cheng et al. (2022) propose the development of domain-specific language models for climate and health to capture available knowledge and solve various tasks, such as identifying similarities between climate and health concepts, fact-checking, extracting relationships, and generating policy text. Biswas (2023) conducted a study on the application of ChatGPT in climate data analysis, scenario generation, and model evaluation. The research provided valuable tools for both researchers and policymakers. Paterson et al. (2022) interviewed GPT-3 on the topic of climate change. Their study highlights the capabilities of LLMs but also notes that they can sometimes generate incorrect or nonsensical responses,

a phenomenon known as hallucinations. The researchers will focus on strategies to prevent such hallucinations, making the models more reliable through techniques like reinforcement learning (Ti-wari et al., 2021), and exploring the potential applications of GPT-3 in finance (Caldecott et al., 2022; Zhu, 2022a) and other relevant domains.

2.2 Large Language Models for Human Evaluation

Large Language Models achieve controllability through human feedback mechanisms and fine-tuning the model to match human preferences. However, this approach has limitations, including complexity and instability. To address these challenges, Rafailov et al. (2023) proposed an algorithm called Direct Preference Optimization (DPO). DPO accurately optimizes the constrained reward maximization problem in a single stage by establishing a mapping between the reward function and the optimal policy. The application of LLMs in the medical field has sparked widespread discussion. However, they face challenges such as the potential spread of misinformation and the risk of data manipulation. Reddy (2023) evaluates the regulatory mechanisms that should be in place when applying LLMs to healthcare, as well as methods for assessing their performance and practical value. These efforts aim to ensure public trust in these models. Wang et al. (2023) highlight that large language models, including GPT-4, exhibit biases in assessing the quality of responses generated by different models. By altering the sequence of responses within a context, it is possible to manipulate the evaluation outcomes to favor one model over others. To address this issue, they developed a calibration framework that incorporates three straightforward and effective strategies: multi-evidence calibration, balanced position calibration, and human cycle calibration. These methods help to reduce evaluation bias and align the results more closely with human judgment.

KoLA (Yu et al., 2023) is a meticulously crafted knowledge-centric evaluation benchmark designed to assess the capabilities of LLMs. The benchmark features a four-tiered classification system for knowledge-related abilities (Zhu et al., 2023), which emulates human cognition. Additionally, it incorporates data from Wikipedia and other sources that are regularly updated. KoLA employs an evaluation methodology that utilizes both standard scores and self-comparison indicators. The authors

evaluated 21 open-source and commercial LLMs and conducted a thorough analysis of their findings. Chiang and Lee (2023) investigated whether large language models could serve as a substitute for human evaluation. The study compared the use of LLMs and human evaluators in assessing text quality for two natural language processing tasks. The findings indicate that the evaluation outcomes generated by LLMs align with those provided by human experts. The researchers discovered that the results from LLM evaluations remained consistent across different formats of task instructions and were deemed stable and reliable. The paper further discusses the limitations and ethical implications of using LLMs for assessment purposes.

3 Approach

We formalize the climate crisis knowledge evaluation task. Given a set of climate crisis questions $X^{(q)} = \{x_1^{(q)}, x_2^{(q)}, \dots\}$ and answers $X^{(a)} = \{x_1^{(a)}, x_2^{(a)}, \dots\}$, we use LLMs as evaluators to generate responses based on predefined metrics, which reflect the knowledge contained within the model. Unlike previous work, the innovation of this paper is that we propose an automatic LLM inference framework that evaluates the climate-crisis knowledge of LLMs from 10 different perspectives. The overview framework is shown in Figure 1. The timeline includes data acquisition, prompt engineering, question evaluation, response generation, and response evaluation. These modules can be processed in parallel. In this section, we first introduce the acquisition of Climate Crisis Questions and Answers, followed by an introduction to the Climate Crisis Knowledge Evaluation.

3.1 Acquisition of Climate-Crisis Q&A Dataset

Our proposed method for acquiring a large number of questions about the climate crisis involves a two-step process: question generation and question selection. Initially, we used the Llama2-70B (Touvron et al., 2023) model to generate 100,000 questions. This model has advanced language understanding and generation capabilities, enabling the creation of a diverse range of questions that cover various aspects of the climate crisis. After generating the questions, we perform a thorough classification and labeling process to facilitate efficient analysis and processing of the questions.

Following our initial selection, we conducted an

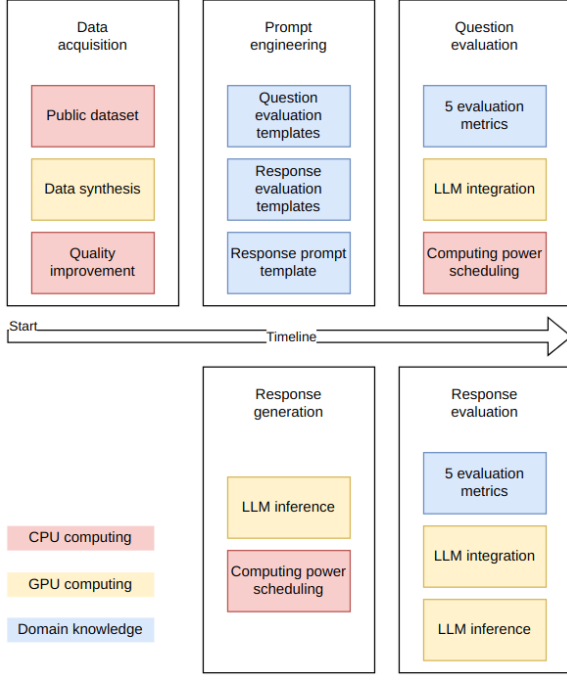


Figure 1: The schematic diagram of the proposed climate crisis knowledge evaluation framework

additional review to eliminate questions that were irrelevant or duplicates. We established a set of rules to guide this process, which was fully automated with no human intervention. To ensure the quality of the questions, we improved their quality through the following steps: (1) Removal of overlapping questions: Through semantic analysis, we identified redundant questions and employed an embedding-based question retrieval method to retain only unique questions, effectively eliminating duplicates based on a defined threshold, as shown in equation (1). (2) Climate crisis relevance assessment: We conducted a relevance analysis of each question to ensure its practical significance and value in the context of the climate crisis, leveraging keyword occurrence.

$$\alpha_{i,j} = \frac{h_i^{(q)T} h_j^{(q)}}{\|h_i^{(q)}\| \cdot \|h_j^{(q)}\|} \quad i, j \in m, i \neq j \quad (1)$$

where $h_i^{(q)} \in \mathbb{R}^d$ is determined using equation (2).

$$h_i^{(q)} = \mathcal{F}_{encoder}(x_i^{(q)} | \Theta) \quad (2)$$

where $\mathcal{F}_{encoder}(\cdot)$ is a language model for generating embeddings (Xiao et al., 2023). Θ represents the parameters of the model. $x_i^{(q)}$ is the sequence of text in the question.

After processing these questions, we obtained a valuable collection of 19,241 high-quality questions related to the climate crisis, about 5% of this data came from an external dataset. Since LLMs are pre-trained with the next token prediction task, as shown in equation (3), we subsequently leveraged Llama2-70B to generate corresponding answers for each question (Zhu et al., 2021). Our two-stage methodology effectively facilitated the accumulation of a substantial number of question-answer pairs.

$$p(x) = \prod_{i=1}^n p(w_i | w_{i-1}, \dots, w_1, \Theta) \quad (3)$$

where x is the input text and w_i represents the i -th token. Θ is the model parameter. In the following, we introduce a novel methodology for assessing knowledge about the climate crisis. Our approach aims to establish an objective and precise criterion for evaluating questions and answers related to this critical topic, leveraging the capabilities of multiple LLMs.

3.2 Evaluation of Climate-Crisis Knowledge

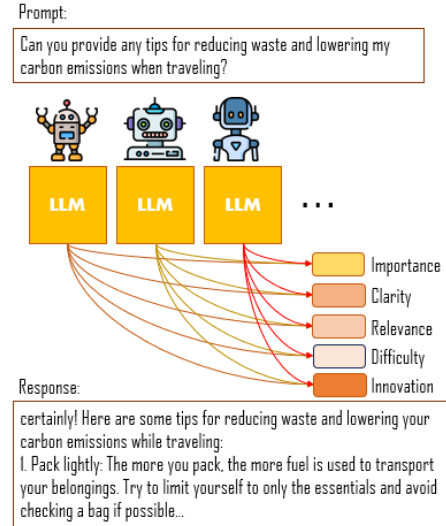


Figure 2: An illustration of utilizing multiple LLMs to automatically evaluate a question-answer pair in the context of climate change

We use multiple LLMs to generate scores for the questions, as shown in Figure 2. To allow the model to evaluate the responses from various aspects, we developed several prompt templates (Zhu, 2023b) for questions and answers, including different types of questions, so that the model can be evaluated from multiple perspectives. For instance,

the prompt might be: "Please assess the importance of the above question: How valuable is this question to the user? Can it help users express their needs and confusion?" or "Please rate the clarity of the above questions: Is the question clear and easy to understand?" In this way, the model can rate each question and answer based on its learned knowledge.

To evaluate the quality of the questions, we evaluate them from the following aspects: (1) Importance of the problem: How valuable is this problem to the user? Can it help users express their needs and confusion? (2) Clarity of the question: Is the question clear and understandable at a glance? (3) Relevance of the question: Is the question closely related to the topic? (4) Question difficulty: Is the question too difficult or too easy for users to understand or too simple to interest users? (5) Innovation of the question: Is the question novel and can it inspire users to think?

To evaluate the quality of answers, we evaluate the following aspects: (1) Relevance of the answer: Does the answer accurately answer the user's question and can it solve the user's needs? (2) Depth of answer: Does the answer provide enough detail so that users can fully understand and apply the information? (3) Answer readability: Is the answer written in plain language and clearly formatted for users to read and understand? (4) Innovation of the answer: Does the answer provide unique insights or solutions that will help users achieve better results on similar problems? (5) Timeliness of the answer: Is the content of the answer up-to-date and able to adapt to changing circumstances and needs?

We use the model to automatically score the metrics mentioned above. For the question, we use equation (4).

$$x_{i,j}^{(r)} = \mathcal{F}_{decoder}(< x_i^{(q)}; [pre]^{(q)}; m_j^{(q)}; [suf]^{(q)} >) \quad (4)$$

where the prefix and suffix of the template are denoted as $[pre]^{(q)}$ and $[suf]^{(q)}$, respectively. The j -th metric for question evaluation is represented as $m_j^{(q)}$. The LLM is denoted as $\mathcal{F}_{decoder}(\cdot)$. The concatenation operation is represented as $<;>$.

For the answer, we use equation (5). Then we can get the generation content as the candidate data. Finally, we extract the model scores from the data using information extraction methods.

$$x_{i,j}^{(r)} = \mathcal{F}_{decoder}(< x_i^{(q)}; x_i^{(a)}; [pre]^{(a)}; m_j^{(a)}; [suf]^{(a)} >) \quad (5)$$

where the prefix and suffix of the template are denoted as $[pre]^{(a)}$ and $[suf]^{(a)}$, respectively. The j -th metric for answer evaluation is represented as $m_j^{(a)}$.

Then we manually check the model scores through random sampling. We found that the model's evaluation of the quality of generated responses is highly consistent with that of humans. We also discovered some potential problems. In some cases, the model may misunderstand the intent of the instruction, resulting in an invalid response. Additionally, because the model is trained on a massive amount of cross-domain text data, it may not fully understand certain aspects of the climate crisis or questions of a metaphorical nature. To address these issues, we can further fine-tune the LLMs in the future to improve their ability to understand complex questions and answers.

We anticipate that this methodology will foster a deeper comprehension of climate crisis-related issues among individuals and offer a fair and unbiased evaluation criterion. In practical scenarios (Zhu, 2022b), when users submit questions or answers, LLMs will seamlessly process them and assign a corresponding score, without human intervention, using predefined prompt templates. This not only empowers the system to deliver insightful answers but also assesses the quality of the information, ultimately assisting users in grasping the gravity of climate crisis-related topics. By employing carefully crafted prompt templates, our approach guarantees an objective and precise evaluation of climate crisis-related questions and answers, thus contributing significantly to heightening public awareness and encouraging greater participation in the fight against climate change.

4 Experiments

4.1 Dataset

We curated a comprehensive climate-crisis QA dataset that encompasses a vast array of questions and answers about climate change. This dataset boasts a total of 19,241 samples, of which 95% of the questions were intelligently generated using the Llama2-70B model. The remaining 5% of questions were carefully sourced from pertinent information gathered from the internet, including the "Reddit Climate Change Dataset" that captures discussions about climate change on Reddit up to September 1, 2022. This dataset comprises 620,908 posts and 4,600,698 comments. To ensure the qual-

ity and relevance of our dataset, we employed a rigorous two-step processing method to eliminate any overlapping content and enhance its relevance. The answers within this dataset are also automatically generated by the Llama2-70B model.

4.2 Hyper-parameters

We employ several high-performing LLMs for evaluation, with the temperature parameter set to 0.5 for all models, and a maximum length of 2048. We set do_sample to false to ensure reproducibility of results. For Llama2-70b, we use top_k = 250, top_p = 1, and repetition_penalty = 1. For Baichuan2-13b, we set top_k = 5, top_p = 0.85, and repetition_penalty = 1.05. For the remaining models, we adhere to their default configurations. The experimental environment consists of an Intel(R) Xeon(R) Platinum 8163 CPU @ 2.50GHz with 256G of memory, and 8 RTX 3090 (24G) GPUs.

4.3 Evaluation

We assess the quality of questions and answers utilizing LLMs, which assign a rating on a scale ranging from 0 to 10. A higher score signifies superior quality in each respective aspect. Specifically, we employ 5 distinct metrics to evaluate questions and another set of 5 metrics to evaluate answers.

4.4 Question Quality Evaluation

The following models were adopted to assess the quality of questions.

The ChatGLM3-6B model (Zeng et al., 2022) is built upon the GLM architecture and employs an autoregressive blank infilling training approach. This training method equips the model to handle complex scenarios with ease, including tasks like tool invocation (function call), code execution (code interpreter), and Agent-related tasks.

Mistral-7B (Jiang et al., 2023) uses grouped-query attention and sliding-window attention, and it employs a byte-fallback BPE tokenizer. The model is designed to strike a balance between efficiency and performance by creating a lightweight architecture.

Zephyr-7B (Tunstall et al., 2023) is based on the Mistral-7B model and employs the direct preference optimization (DPO) training approach.

The Baichuan2-13B model (Yang et al., 2023) is trained on a high-quality corpus of 2.6 trillion tokens. The model employs ALiBi linear bias tech-

nology to enhance computational efficiency effectively.

The Yi-34B model (01.AI, 2023) is based on the Llama model architecture and has been trained on both Chinese and English data. It uses approximately 3T-sized tokens and supports long-context technology.

Llama2-70B (Touvron et al., 2023), as proposed by Meta, is an open-source model architecture that has been trained using reinforcement learning with human feedback (RLHF). This training methodology is designed to align the model’s behavior with human preferences, ensuring both its usefulness and safety.

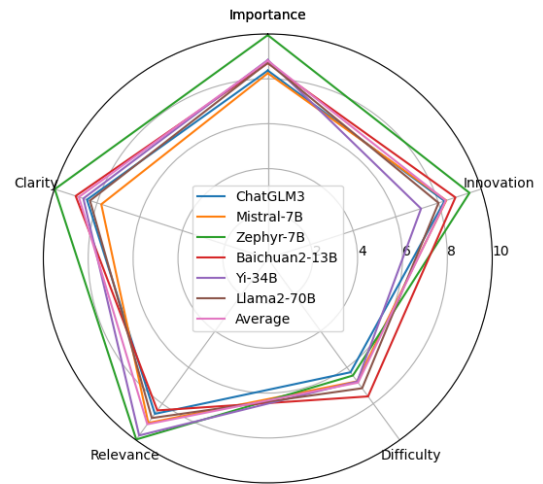


Figure 3: Visualization of question quality evaluation, with circles closer to the center indicating lower overall scores assigned by the model

As can be seen from the results in Table 1, the model evaluation results suggest that the quality of questions is highest concerning topic relevance but lowest concerning question difficulty. This means that integrating multiple LLMs can produce more credible results for climate-crisis knowledge.

As shown in Figure 3, the curve closer to the outside indicates a higher overall score for the model. Zephyr-7B gives the highest overall score of question quality, while Mistral-7B and Yi-34B give the lower overall score of question quality. This means that different models have different standards for knowledge about the climate crisis, and we tend to choose models with more stringent standards.

Table 1: Question quality evaluation scores (0-10) across 5 dimensions

Models	Importance	Clarity	Relevance	Difficulty	Innovation
ChatGLM3-6B	8.37	8.47	8.56	6.27	8.28
Mistral-7B	8.25	7.80	9.06	6.77	8.36
Zephyr-7B	9.95	9.98	9.97	6.45	9.45
Baichuan2-13B	8.70	9.00	8.37	7.60	8.79
Yi-34B	8.84	8.64	9.75	6.78	7.18
Llama2-70B	8.72	8.33	8.79	7.15	8.00
Average	8.81	8.85	9.13	6.84	8.34

Table 2: Answer quality evaluation scores (0-10) across 5 dimensions

Models	Relevance	Depth	Readability	Innovation	Timeliness
ChatGLM3-6B	9.92	8.42	8.82	8.73	8.39
Mistral-7B	8.67	8.98	9.00	8.55	8.22
Zephyr-7B	9.98	9.89	9.95	9.75	9.95
Baichuan2-13B	8.73	8.41	8.48	8.23	6.65
Yi-34B	9.65	9.00	9.22	7.54	8.17
Llama2-70B	9.11	9.10	9.35	8.97	9.12
Average	9.34	8.97	9.14	8.63	8.42

4.5 Answer Quality Evaluation

As can be seen from the results in Table 2, the model evaluation results suggest that the quality of the answers is the highest in terms of relevance, but is low in terms of question timeliness. This means that LLMs can understand climate crisis knowledge and are accustomed to generating relevant responses, but contain insufficient timely information.

As shown in Figure 4, the curve closer to the outside indicates a higher overall score for the model. Among the models evaluated, Zephyr-7B gives the highest overall score for answer quality, while Baichuan2-13B gives a lower overall score for answer quality. We can find that different models have different sensitivity to the timeliness of answers.

4.6 Computing Efficiency Analysis

We compare LLMs and conduct experiments using 4-bit quantization to ensure optimal efficiency. We evaluate model performance using 10 different prompts and set a maximum sequence length of 2048.

As shown in Table 3, ChatGLM3-6B has the fastest inference speed, while Llama2-70B has the lowest speed. Due to the high GPU memory needs,

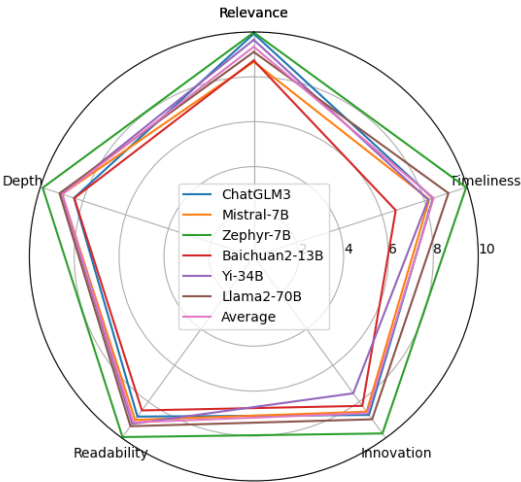


Figure 4: Visualization of answer quality evaluation, with circles positioned closer to the center indicating lower overall scores assigned by the model

we averagely split different layers of Llama2-70B to 8 GPUs, so it has extra time consumption to communicate among PCIe GPUs. Mistral-7B and Zephyr-7B perform poorly compared to similarly sized models. For the GPU memory consumption, we found that Llama2-70B has the best GPU memory utilization efficiency (0.51GB/1B). ChatGLM3-6B has the worst memory utilization efficiency (0.80GB/1B).

Table 3: Efficiency of Inference for Various LLMs

Models	Time (S)	Memory (GB)
ChatGLM3-6B	26.66	4.80
Mistral-7B	179.09	5.37
Zephyr-7B	325.14	5.33
Baichuan2-13B	44.12	11.79
Yi-34B	70.39	19.48
Llama2-70B	709.63	36.20

4.7 Case Study

We use the question "Can you provide any tips for reducing waste and lowering my carbon emissions when traveling?" as an example to assess the responses of the evaluation models.

As shown in Table 4 located in the Appendix, each of the models adopted is capable of providing high-quality responses. Among them, Yi-34B stands out for offering the most comprehensive suggestions, while Llama2-70B is known for providing a response that is both concise and effective.

5 Conclusion

This paper introduces an automated framework for evaluating the climate-crisis knowledge of LLMs. Our proposed approach assesses climate-crisis knowledge based on the quality of symbolized questions and their corresponding answers. The evaluation process is crafted to be both robust and comprehensive, encompassing a two-stage question acquisition strategy and an answer generation procedure. Furthermore, we have devised an automated evaluation methodology along with a comprehensive set of metrics, including 5 for question evaluation and 5 for answer evaluation. Experimental findings indicate that our approach holds significant value in assessing LLMs' knowledge pertaining to climate change.

The primary contribution of this paper is the proposal of an automated framework to evaluate

climate-crisis knowledge in LLMs, without reliance on human intervention. Looking forward, we aim to leverage this technique in the development of an online climate crisis knowledge system that utilizes our methodologies to provide users with real-time, expert-level Q&A services. Our research introduces novel concepts and methodologies that address challenges in the field of climate crisis, thereby enriching the research and applications of AI in this critical domain.

References

- 01.AI. 2023. Yi. <https://github.com/01-ai/Yi>.
- Som S Biswas. 2023. Potential use of chat gpt in global warming. *Annals of biomedical engineering*, 51(6):1126–1127.
- Ben Caldecott, Matthew McCarten, Christophe Christaen, and Conor Hickey. 2022. Spatial finance: practical and theoretical contributions to financial analysis. *Journal of Sustainable Finance & Investment*, pages 1–17.
- Kaiping Cheng et al. 2022. How gpt-3 responds to different publics on climate change and black lives matter: A critical appraisal of equity in conversational ai.
- David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of ACL*, pages 15607–15631. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Mathias Kraus, Julia Anna Bingler, Markus Leippold, Tobias Schimanski, Chiara Colesanti Senni, Dominik Stammach, Saeid Ashraf Vaghefi, and Nicolas Webersinke. 2023. Enhancing large language models with climate resources. *arXiv preprint arXiv:2304.00116*.
- Markus Leippold. 2023. Thus spoke gpt-3: Interviewing a large-language model on climate finance. *Finance Research Letters*, 53:103617.
- OpenAI. 2023. *Gpt-4 technical report*. *Preprint*, arXiv:2303.08774.
- David Patterson, Joseph Gonzalez, Urs H  lzle, Quoc Le, Chen Liang, Llu  s-Miquel Mungu  a, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28.

628	Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano	Shengyi Huang, Leandro von Werra, Cl��mentine	685
629	Ermon, Christopher D Manning, and Chelsea Finn.	Fourrier, Nathan Habib, Nathan Sarrazin, Omar San-	686
630	2023. Direct preference optimization: Your language	seviero, Alexander M. Rush, and Thomas Wolf. 2023.	687
631	model is secretly a reward model. <i>arXiv preprint</i>	Zephyr: Direct distillation of lm alignment . <i>Preprint</i> ,	688
632	<i>arXiv:2305.18290</i> .	<i>arXiv:2310.16944</i> .	689
633	Sandeep Reddy. 2023. Evaluating large language mod-	Saeid Ashraf Vaghefi, Dominik Stambach, Veruska	690
634	els for use in healthcare: A framework for transla-	Muccione, Julia Bingler, Jingwei Ni, Mathias	691
635	tional value assessment. <i>Informatics in Medicine</i>	Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias	692
636	<i>Unlocked</i> , page 101304.	Wekhof, Tobias Schimanski, et al. 2023. Chatcli-	693
637	David Rolnick, Priya L Donti, Lynn H Kaack, Kelly	mate: Grounding conversational ai in climate science.	694
638	Kochanski, Alexandre Lacoste, Kris Sankaran, An-	<i>Communications Earth & Environment</i> , 4(1):480.	695
639	drew Slavin Ross, Nikola Milojevic-Dupont, Natasha	Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai	696
640	Jaques, Anna Waldman-Brown, et al. 2022. Tack-	Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui.	697
641	ling climate change with machine learning. <i>ACM</i>	2023. Large language models are not fair evaluators.	698
642	<i>Computing Surveys (CSUR)</i> , 55(2):1–96.	<i>arXiv preprint arXiv:2305.17926</i> .	699
643	Tobias Schimanski, Andrin Reding, Nico Reding, Julia	Nicolas Webersinke, Mathias Kraus, Julia Anna Bin-	700
644	Bingler, Mathias Kraus, and Markus Leippold. 2024.	gler, and Markus Leippold. 2021. Climatebert: A	701
645	Bridging the gap in esg measurement: Using nlp to	pretrained language model for climate-related text.	702
646	quantify environmental, social, and governance com-	<i>arXiv preprint arXiv:2110.12010</i> .	703
647	munication. <i>Finance Research Letters</i> , 61:104979.	Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas	704
648	Dominik Stambach, Nicolas Webersinke, Julia Bin-	Muennighoff. 2023. C-pack: Packaged resources	705
649	gler, Mathias Kraus, and Markus Leippold. 2022.	to advance general chinese embedding . <i>Preprint</i> ,	706
650	Environmental claim detection. <i>Available at SSRN</i>	<i>arXiv:2309.07597</i> .	707
651	4207369.	Aiyuan Yang, Bin Xiao, Bingning Wang, Borong	708
652	Manfred Stede and Ronny Patz. 2021. The climate	Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang,	709
653	change debate and natural language processing. In	Dong Yan, Fan Yang, et al. 2023. Baichuan 2:	710
654	<i>Proceedings of the 1st Workshop on NLP for Positive</i>	Open large-scale language models. <i>arXiv preprint</i>	711
655	<i>Impact</i> , pages 8–18.	<i>arXiv:2309.10305</i> .	712
656	Prayag Tiwari, Hongyin Zhu, and Hari Mohan Pandey.	Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao,	713
657	2021. Dapath: Distance-aware knowledge graph rea-	Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiao-	714
658	soning based on deep reinforcement learning. <i>Neural</i>	han Zhang, Hanming Li, et al. 2023. Kola: Carefully	715
659	<i>Networks</i> , 135:1–12.	benchmarking world knowledge of large language	716
660	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	models. <i>arXiv preprint arXiv:2306.09296</i> .	717
661	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang,	718
662	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti	Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,	719
663	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton	Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b:	720
664	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	An open bilingual pre-trained model. <i>arXiv preprint</i>	721
665	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	<i>arXiv:2210.02414</i> .	722
666	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	Hongyin Zhu. 2022a. Financial data analysis applica-	723
667	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	tion via multi-strategy text processing. <i>arXiv preprint</i>	724
668	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	<i>arXiv:2204.11394</i> .	725
669	Isabel Kloumann, Artem Korenev, Punit Singh Koura,	Hongyin Zhu. 2022b. Metaaid: A flexible frame-	726
670	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-	work for developing metaverse applications via ai	727
671	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	technology and human editing. <i>arXiv preprint</i>	728
672	tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	<i>arXiv:2204.01614</i> .	729
673	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	Hongyin Zhu. 2023a. Fqp 2.0: Industry trend anal-	730
674	stein, Rashmi Rungta, Kalyan Saladi, Alan Schelten,	ysis via hierarchical financial data. <i>arXiv preprint</i>	731
675	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	<i>arXiv:2303.02707</i> .	732
676	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	Hongyin Zhu. 2023b. Metaaid 2.5: A secure frame-	733
677	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,	work for developing metaverse applications via large	734
678	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	language models. <i>arXiv preprint arXiv:2312.14480</i> .	735
679	Melanie Kambadur, Sharan Narang, Aurelien Ro-	Hongyin Zhu, Hao Peng, Zhiheng Lyu, Lei Hou, Juanzi	736
680	driguez, Robert Stojnic, Sergey Edunov, and Thomas	Li, and Jinghui Xiao. 2023. Pre-training language	737
681	Scialom. 2023. Llama 2: Open foundation and fine-	model incorporating domain-specific heterogeneous	738
682	tuned chat models . <i>Preprint</i> , <i>arXiv:2307.09288</i> .		
683	Lewis Tunstall, Edward Beeching, Nathan Lambert,		
684	Nazneen Rajani, Kashif Rasul, Younes Belkada,		

739 knowledge into a unified representation. *Expert Sys-*
740 *tems with Applications*, 215:119369.

741 Hongyin Zhu, Prayag Tiwari, Ahmed Ghoneim, and
742 M Shamim Hossain. 2021. A collaborative ai-
743 enabled pretrained language model for aiot domain
744 question answering. *IEEE Transactions on Industrial*
745 *Informatics*, 18(5):3387–3396.

746 **A Appendix**

Table 4: Comparison of responses from different models

Models	Response
ChatGLM3-6B	<p>当然可以，以下是一些减少浪费和降低碳排放的旅行建议：</p> <ol style="list-style-type: none"> 1. 选择环保的交通工具：选择公共交通工具、自行车或步行，而不是自驾或飞机。火车和巴士通常比飞机更环保，尤其是长途旅行。 2. 飞行时选择直飞航班：直飞航班通常比转机航班更环保，因为它们减少了燃料消耗和碳排放。 3. 优化旅行计划：提前规划行程，避免不必要的绕路和重复旅行，减少碳足迹。 4. 住宿选择：选择环保型酒店或民宿，这些地方通常采取措施减少能源消耗和浪费。 5. 自带餐食：避免经常在外就餐，而是自带食物和饮料，这样可以减少一次性餐具的使用。 6. 减少一次性用品的使用：尽量避免使用一次性物品，如塑料袋、瓶子和餐具。携带可重复使用的物品，如水壶、餐具和购物袋。 7. 回收和节约资源：在住宿和旅行时，尽量回收利用废弃物，并节约用水和用电。 8. 支持环保旅行：选择参加环保旅行团或活动，这些活动通常会考虑对环境的影响。 9. 购买碳抵消：如果必须乘坐飞机等高排放交通工具，可以考虑购买碳抵消，以补偿碳排放。 10. 分享行程：如果可能，与其他人分享行程，这样可以减少整体碳足迹。 <p>记住，每次旅行都会对环境产生影响，但通过采取上述措施，您可以减少对环境的影响，并帮助保护我们的地球。</p>
Mistral-7B	<p>Sure, here are some tips for reducing waste and lowering your carbon emissions when traveling:</p> <ol style="list-style-type: none"> 1. Pack light: Try to pack only what you need for your trip. This will reduce the amount of weight you need to carry, which will in turn reduce your carbon emissions. 2. Use reusable containers: Instead of using disposable plastic bags or containers, bring your own reusable containers for food and other items. This will reduce waste and save money. 3. Bring your own water bottle: Instead of buying bottled water, bring your own reusable water bottle. This will reduce waste and save money. 4. Use public transportation: Instead of renting a car or taking taxis, use public transportation to get around. This will reduce your carbon emissions and save money. 5. Stay in eco-friendly accommodations: Look for accommodations that are eco-friendly, such as hotels that use renewable energy sources or have recycling programs. 6. Eat locally: Instead of eating at chain restaurants, try to eat locally. This will reduce your carbon emissions and support the local economy. 7. Use reusable bags: Instead of using plastic bags, bring your own reusable bags for shopping and other activities. This will reduce waste and save money. 8. Use energy-efficient electronics: Bring energy-efficient electronics, such as a portable charger and a laptop with a long battery life. This will reduce your energy consumption and save money. <p>By following these tips, you can reduce your waste and lower your carbon emissions when traveling.</p>
Zephyr-7B	<p>Certainly! Here are some tips for reducing waste and lowering your carbon emissions when traveling:</p> <ol style="list-style-type: none"> 1. Pack light: The more you pack, the more fuel is needed to transport your luggage. Pack only the essentials and avoid overpacking. 2. Choose eco-friendly accommodations: Look for accommodations that prioritize sustainability, such as those with green certifications, solar panels, or recycling programs. 3. Use reusable water bottles and bags: Bring a reusable water bottle and bag to avoid using single-use plastics. Many destinations have water refill stations, so you can fill up your bottle for free. 4. Eat local and seasonal: Eating locally sourced and seasonal foods reduces the carbon footprint of transportation and supports the local economy. 5. Use public transportation: Use public transportation, such as buses, trains, and subways, instead of renting a car. This reduces your carbon emissions and saves you money. 6. Offset your carbon emissions: Consider offsetting your carbon emissions by investing in carbon-reducing projects, such as reforestation or renewable energy. 7. Support sustainable tourism: Choose activities that prioritize sustainability, such as nature walks, bike tours, and cultural experiences. 8. Reduce water usage: Turn off the tap when brushing your teeth, take shorter showers, and reuse towels to reduce water usage. 9. Support local businesses: Support local businesses, such as small shops and restaurants, to reduce the carbon footprint of transportation and support the local economy. 10. Educate yourself: Learn about the local culture and environment to better understand the impact of your actions and make informed decisions. <p>Remember, every small action can make a big difference in reducing waste and lowering your carbon emissions when traveling.</p>
Baichuan2-13B	<p>Sure, here are some tips to reduce waste and lower your carbon emissions when traveling:</p> <ol style="list-style-type: none"> 1. Use public transportation or carpooling instead of renting a car. This will not only save you money but also reduce your carbon footprint. 2. Bring reusable water bottles, coffee mugs, and shopping bags. Avoid using single-use plastic items as much as possible. 3. Book accommodation that has eco-friendly practices like energy efficiency and water conservation. 4. Eat locally sourced food while traveling. This reduces the need for transportation and minimizes waste. 5. Choose accommodations with energy-efficient lighting and appliances. This will help in reducing energy consumption. 6. Minimize your use of air conditioning and heating. Wear layers of clothing to regulate your body temperature. 7. Unplug chargers and devices not in use. This will help in reducing energy consumption. 8. Consider staying at an eco-friendly hotel or resort that uses renewable energy sources. 9. Use energy-efficient devices like LED lights and energy-saving appliances. 10. Support local businesses and cultural activities. This will help in promoting sustainable tourism.
Yi-34B	<p>Yes, there are several tips to reduce waste and lower carbon emissions when traveling:</p> <ol style="list-style-type: none"> 1. Plan Your Trip Carefully: Choose modes of transportation that have lower carbon emissions, such as trains or buses over flying, especially for shorter distances. Combine multiple destinations into one trip to reduce the number of flights taken. 2. Use Public Transportation: Opt for local buses, trains, or subways instead of renting a car or using taxis. 3. Ride Sharing: Use carpooling or ridesharing services to reduce the number of vehicles on the road. 4. Eco-Friendly Accommodations: Choose hotels or hostels that have eco-friendly practices, such as energy-efficient lighting, water-saving devices, and recycling programs. 5. Bring Your Own Gear: Bring your own reusable water bottle, coffee mug, and cutlery to reduce the use of single-use items. 6. Pack Lightly: Travel with only what you need to reduce the weight and volume of your luggage, which can lower fuel consumption for planes and vehicles. 7. Offset Your Carbon Footprint: Consider investing in carbon offsets to compensate for the emissions you produce during your trip. 8. Eat Locally and Seasonally: Support local farmers and reduce the carbon footprint associated with food transportation by eating locally sourced and seasonal produce. 9. Shop Wisely: Avoid buying items that require a lot of packaging or that will end up in the trash. 10. Leave No Trace: If you're camping or hiking, follow the leave-no-trace principles, which include packing out all your trash and being mindful of the environment. 11. Travel Off-Peak: Avoid traveling during peak seasons when possible, as fewer travelers mean less environmental impact. 12. Be Mindful of Water Usage: Limit your water usage, especially in places where water is scarce. <p>By following these tips, you can help minimize your environmental impact while traveling and contribute to a more sustainable tourism industry. Remember that every little bit helps, and by making small changes, you can make a big difference.</p>
Llama2-70B	<p>certainly! Here are some tips for reducing waste and lowering your carbon emissions while traveling:</p> <ol style="list-style-type: none"> 1. Pack lightly: The more you pack, the more fuel is used to transport your belongings. Try to limit yourself to only the essentials and avoid checking a bag if possible. 2. Use public transportation: Instead of renting a car or taking taxis, use public transportation to get around. Not only is it more environmentally friendly, but it's also a great way to experience the local culture. 3. Bring a reusable water bottle: Instead of buying bottled water, bring a reusable water bottle and fill it up at a water fountain or filtered water station. 4. Avoid single-use plastics: Refuse single-use plastics like straws, bags, and utensils. Instead, opt for reusable alternatives like metal straws, cloth bags, and bamboo utensils. 5. Choose eco-friendly accommodations: Look for hotels, hostels, and vacation rentals that have implemented sustainability practices such as using solar power, reducing water waste, and recycling. 6. Reduce food waste: Plan your meals in advance and try to avoid ordering too much food. Also, consider choosing local, seasonal ingredients which have a lower carbon footprint. 7. Offset your carbon emissions: Consider offsetting your carbon emissions from your flight by investing in carbon offset projects that reduce greenhouse gas emissions elsewhere, such as reforestation or renewable energy projects. <p>I hope these tips help you reduce your impact on the environment while traveling!</p>