PLOS CLIMATE

RESEARCH ARTICLE

Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias

Sanguk Lee[1], **Tai-Quan Peng**[2]*, **Matthew H. Goldberg**[1], **Seth A. Rosenthal**[1], **John E. Kotcher**[3], **Edward W. Maibach**[3], **Anthony Leiserowitz**[1]

1 Yale Program on Climate Change Communication, Yale University, New Haven, Connecticut, United States of America, **2** Department of Communication, Michigan State University, East Lansing, Michigan, United States of America, **3** Center for Climate Change Communication, George Mason University, Fairfax, Virginia, United States of America

* winsonpeng@gmail.com

OPEN ACCESS

Citation: Lee S, Peng T-Q, Goldberg MH, Rosenthal SA, Kotcher JE, Maibach EW, et al. (2024) Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. PLOS Clim 3(8): e0000429. https://doi.org/10.1371/journal.pclm.0000429

Editor: Lily Hsueh, Arizona State University, UNITED STATES OF AMERICA

Received: February 22, 2024

Accepted: June 28, 2024

Published: August 7, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pclm.0000429

Data Availability Statement: All data used in this study are publicly available on OSF (https://osf.io/w36gn/).

Abstract

Large language models (LLMs) can be used to estimate human attitudes and behavior, including measures of public opinion, a concept referred to as algorithmic fidelity. This study assesses the algorithmic fidelity and bias of LLMs in estimating public opinion about global warming. LLMs were conditioned on demographics and/or psychological covariates to simulate survey responses. Findings indicate that LLMs can effectively reproduce presidential voting behaviors but not global warming opinions unless the issue relevant covariates are included. When conditioned on both demographic and covariates, GPT-4 demonstrates improved accuracy, ranging from 53% to 91%, in predicting beliefs and attitudes about global warming. Additionally, we find an algorithmic bias that underestimates the global warming opinions of Black Americans. While highlighting the potential of LLMs to aid social science research, these results underscore the importance of conditioning, model selection, survey question format, and bias assessment when employing LLMs for survey simulation.

Introduction

Public opinion on global warming influences policy-making decisions [1] and public behavior [2]. Scholars and policymakers typically use representative surveys to measure and understand public opinion [3]. While surveys are essential tools for understanding public opinion about global warming, their resource-intensive nature often limits the depth and inclusivity of analysis, potentially resulting in biases toward majority views and neglecting minority perspectives. Moreover, the scarcity of resources makes it challenging to encompass all practically significant variables in a single survey, limiting our grasp of the connections among diverse factors influencing public opinions on global warming.

Large Language Models (LLMs) like GPT have the potential to complement traditional survey methods by simulating survey responses with fewer resources and augmenting data from underrepresented sub-populations. Moreover, unlike conventional predictive methods, such

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

as regression models that depend solely on numeric data, LLMs excel in integrating semantic information (e.g., semantic information in a survey question). This unique ability is particularly advantageous in generating responses for untested global warming survey questions, thereby potentially complementing public opinion datasets [4]. However, these potentials of LLMs hinge upon their ability to accurately estimate public opinion on global warming. As an initial assessment, this study investigates the extent to which LLMs can accurately emulate and reflect multiple dimensions of public opinion about global warming.

LLMs have demonstrated significant potential for contributing to social science research. A recent development lies in their capacity to accurately replicate the perceptions, viewpoints, and behavior of the general population or specific subgroups, termed *algorithmic fidelity* [5]. Algorithmic fidelity refers to the extent to which LLMs' intricate web of connections among ideas, attitudes, and sociocultural contexts accurately reflects those found in various human subgroups [5]. By training on an extensive corpus of human-generated data that includes human perceptions and behaviors, LLMs may possess the capability to simulate diverse facets of public opinion.

Recent studies have yielded promising results. For instance, Argyle and colleagues found strikingly high correlations in voting behaviors during presidential elections between human samples and *silicon samples* derived from LLMs [5]. Silicon samples refer to samples synthesized by LLMs conditioned to thousands of sociodemographic backstories sourced from real human participants in surveys [5]. Similarly, Hwang and colleagues found that LLMs were able to accurately reflect public opinions on diverse political issues, including gun control, gender perspectives, economic inequality, trust in science, and so forth [6]. However, the majority of these studies have focused on the political domain, particularly presidential elections and support for political issues.

It remains uncertain whether LLMs can accurately represent public beliefs and attitudes about other important social issues, such as global warming. Global warming perceptions differ from political opinions in that, while both are subjective, climate change itself is grounded in scientific fact. Given this distinction, LLMs might perform differently when predicting public perspectives on global warming. LLMs are engineered to prioritize factual correctness through extensive training and alignment processes [7]. This focus on correctness could limit LLMs' ability to reflect the varied and sometimes erroneous human perspectives on global warming.

Algorithmic fidelity in Large Language Models (LLMs) is influenced by input conditioning and model choice. Accurately reflecting public opinion on global warming requires LLMs to be conditioned with detailed demographics and specific covariates to gain a more accurate perspective of individuals. Research shows that LLMs' predictive accuracy improves when they incorporate past opinion data, as well as demographics and ideology [6]. Additionally, different LLM versions, like GPT-3.5 and GPT-4, vary in their algorithmic fidelity [7–9]. Building on these studies, we assess algorithmic fidelity under different conditions and models. Specifically, we compare LLMs conditioned solely on demographics with LLMs conditioned on both demographics and issue-related covariates. For the sake of simplicity, we categorize political ideology and party affiliation as demographic variables. Furthermore, we examine the algorithmic fidelity of distinct LLM versions: GPT-3.5 and GPT-4.

Public perceptions of global warming are complex. They include, for example, beliefs about global warming, understanding of its causes, worry about the issue, policy support, behavior, and more. To gain deeper insights into how well LLMs represent diverse psychological aspects of global warming, we assess the evaluation metrics (e.g., accuracy, F1) and distribution of LLM predictions against nationally representative survey responses. A strong performance shown by evaluation metrics indicates robust algorithmic fidelity, leading to closely matching distributions between silicon and survey samples. The study draws on nationally representative climate change survey data collected in 2017 and 2021 as benchmarks for evaluating the algorithmic fidelity of LLMs.

Methods

Survey sampling

Two nationally representative survey datasets were collected in October 2017 (N = 1,304) and September 2021 (N = 1,006). In these surveys, participants were asked to answer multiple questions related to global warming. These surveys were conducted under an exemption granted by the Institutional Review Board (IRB) of Yale University (IRB Protocol ID: 2000031972). Subsets of de-identified datasets were accessed for the purpose of this research on September 18, 2023. For each survey, the researchers obtained a distinct sample from the Ipsos KnowledgePanel, comprising U.S. adults aged 18 and over. This panel, which mirrors the U.S. population, was assembled using probability sampling methods. Panel members were recruited using various techniques, such as random digit dialing and address-based sampling, covering nearly all residential phone numbers and addresses in the U.S. Participants completed the survey forms online. Those without internet access were provided with computers and internet connectivity. Upon joining the Ipsos panel, members are informed that participation in every survey is voluntary, and that all data are collected and shared with clients in an anonymous format. Additionally, at the start of the survey, participants were notified that certain questions will pertain to their political opinions, offering them the option to withdraw if they so choose.

Silicon sample data collection

To generate silicon sample datasets, we used two versions of GPT (GPT-3.5 vs. GPT-4) and two sets of conditional inputs (demographics only vs. demographics and issue-related covariates). Specifically, silicon samples were generated using GPT-3.5-turbo-16k and GPT-4 through the OpenAI API, setting the temperature at 0.70 based on a prior study [5]. For models conditioned solely on demographics, we fed demographic information, such as race/ethnicity, gender, age, political ideology, political party affiliation, education, and residential state, into the models via prompts. Meanwhile, for the models conditioned on both demographics and covariates, additional covariates such as issue involvement in global warming, interpersonal discussions about the topic, and awareness of the scientific consensus, were included along with demographics. These covariates were selected because they appear commonly in both waves of the survey and served as important covariates in previous studies [10–13].

We utilized an interview format adapted from Argyle et al. [5] for our prompts (prompt examples are available in Tables A and B in S1 Text). At the system level, GPTs were instructed to act as an interviewee, guided by the directive: “You are an interviewee. Based on your previous answers, respond to the last question.” Subsequently, the simulated interview began. To establish a clear timeline of survey, the first prompt was phrased as: “Interviewer: What is the current year and month of this interview? Me: October 2017.” For the 2021 survey, "September 2021" was typed in. After setting the timeline, "Me" responses leading up to the final question were provided using actual survey data. For instance, regarding race/ethnicity, the prompt was framed as: “Interviewer: I am going to read you a list of five race categories. What race do you consider yourself to be? ’White, Non-Hispanic’, ’Black, Non-Hispanic’, ’2+ Races, Non-Hispanic’, ’Hispanic’, or ’Other, Non-Hispanic.’ Me: {race from survey response}.”

The final question is the target, for which GPTs supply an answer. For instance, regarding global warming beliefs with a binary answer option, it was phrased as, “Interviewer: What do you think: Do you think that global warming is happening? Would you say ’Yes’, or ’No’?” For other target questions about global warming, we provided comprehensive answer options that matched the survey. For global warming belief with multiple response options, the target question was phrased as, “Interviewer: What do you think: Do you think that global warming is happening? Would you say ’Yes’, ’No’, ’Don’t know’, or ’Refused’ to answer?” For the causation of global warming, the target question was phrased as, “Interviewer: Assuming global warming is happening, do you think it is ’Caused mostly by natural changes in the environment’, ’Caused mostly by human activities’, ’Caused by both human activities and natural changes’, ’Neither because global warming isn’t happening’, ’Other (Please specify)’, ’Don’t know’, or ’Refused’ to answer?” For global warming worry, the target question was, phrased as “Interviewer: How worried are you about global warming? Would you say you are ’Not at all worried’, ’Not very worried’, ’Somewhat worried’, ’Very worried’, or ’Refused’ to answer?”

Occasionally, GPTs generated answers that did not precisely match the listed options. We manually corrected these hallucinations. For instance, instead of a straightforward ‘Yes,’ GPT might produce, ‘Yes, I believe global warming is happening.’ Such responses were recoded to align with the intended options. Any deviations were easily identifiable and adjusted to fit within the given answer choices.

Survey measurements

Target variables. *Global warming belief*. To measure belief in global warming, we provided a brief definition of global warming: “Global warming refers to the idea that the world’s average temperature has been increasing over the past 150 years, may be increasing more in the future,” and then asked “Do you believe that global warming is happening?” with three response options: “No,” “Don’t know,” and “Yes.”

Global warming cause. We used a recoded version of the survey question. Originally, the survey asked: “Assuming global warming is happening, do you think it is. . .” with five answer options: “Caused mostly by human activities,” “Caused mostly by natural changes in the environment,” “None of the above because global warming isn’t happening,” “Other (Please specify),” “Refused.” This measure was then recoded to incorporate open-ended responses, expanding the original five answer choices to seven categories: “Caused mostly by human activities,” “Caused mostly by natural changes in the environment,” “Caused by human activities and natural changes,” “Neither because global warming isn’t happening,” “Don’t know,” “Other (Please specify),” and “Refused.” This recoded version was used in the LLM prompt.

Global warming worry. This was measured with a question asking “How worried are you about global warming?” with four response options: “Not at all worried,” “Not very worried,” “Somewhat worried,” and “Very worried.”

Demographics. Demographic details such as race, ethnicity, gender, age, education, and residential state were provided by Ipsos, based on answers provided when enrolling panel members. Race and ethnicity used five categories: “White, Non-Hispanic,” “Black, Non-Hispanic,” “Other, Non-Hispanic,” “Hispanic,” and “2+ Races, Non-Hispanic.” Gender included two categories: “Male,” and “Female.” Age was segmented into four groups: “18–29,” “30–44,” “45–59,” and “66+.” Education was segmented into four categories: “Less than high school,” “High school,” “Some college,” “Bachelor’s degree or higher.” Residential state includes 50 states and the District of Columbia of the U.S.

Political ideology. This was measured with a question asking “In general, do you think of yourself as. . .” with six response options: “Very liberal,” “Somewhat liberal,”, “Moderate, middle of the road,” “Somewhat conservative,” “Very conservative.”

Political party. We employed a two-step method to gauge political party. First, participants were asked to identify themselves as “Republican,” “Democrat,” “Independent,” “Other,” or

"No party/Not interested in politics." Those who chose "Independent" or "Other" were then asked a second question: whether they were more aligned with the "Republican party," "Democratic party," or "Neither." If participants initially identified as Republican or Democrat, or if they leaned towards one of these parties in the secondary question, they were categorized accordingly. Those who answered "Independent" in the first question or "Neither" in the second question were categorized into "Independent/Other." Participants who responded with "No party/Not interested in politics" were categorized into "No party/Not interested."

Global warming covariates. *Issue involvement in global warming*. This was measured with a question asking "How important is the issue of global warming to you personally?" with six response options: "Not at all important," "Not too important," "Somewhat important," "Very important," "Extremely important," and "Refused."

Interpersonal discussion about global warming. This was measured with a question asking "How often do you discuss global warming with your family and friends?" with five response options: "Never", "Rarely", "Occasionally", "Often," and "Refused."

Awareness of scientific consensus. This was measured with a question asking "Which comes closest to your own view?" with five response options: "Most scientists think global warming is not happening", "There is a lot of disagreement among scientists about whether or not global warming is happening", "Most scientists think global warming is happening," "Don't know enough to say," and "Refused."

Pilot test. As a pilot test, we assessed LLMs' algorithmic fidelity in predicting presidential election voting behaviors to replicate a prior study [5]. We conducted this replication using our own dataset to explore the extent to which algorithmic fidelity is generalizable across different datasets. Our results largely confirm that LLMs, when conditioned on individual demographics, can replicate voting behaviors effectively. The detailed information about the pilot test is presented in S1 Text. Fig A and Table C in S1 Text provides the results of the pilot test.

Results

Algorithmic fidelity of global warming belief: From binary choice to polynomial choice

We investigated whether LLMs demonstrate a high level of algorithmic fidelity for beliefs in global warming. As an initial examination, we limited our sample to respondents who answered either "Yes" or "No" to the question of whether global warming is happening and constrained GPT models to these binary responses. The average accuracy of GPTs across the models, conditions, and years was 85% (*SD* = 3.41), which suggests that GPTs predict the publics' belief that global warming is happening with high accuracy.

Accuracy, while an intuitive measure of correct predictions, can be misleading in datasets with skewed distributions, such as our data on belief that global warming is happening. For example, if a majority of survey participants respond with "Yes" and GPTs predict all cases as "Yes," they can still appear highly accurate. To address this issue, we use an additional evaluation metric called a F1 score. A F1 score accounts for both precision (i.e., the proportion of correct positive predictions among all predictions labeled as positive) and recall (i.e., the proportion of actual positive cases that are correctly identified by the model), providing a balanced evaluation of data that is unevenly distributed. We assessed the models using a F1 score and a Macro-Average F1 score (MAF1) which averages F1 scores across answer options.

When conditioned only on demographics, GPT models showed compromised prediction accuracy, with high F1 scores for "Yes" predictions (F1 range: .91-.92) but low or non-existent F1 scores for "No" predictions (F1 range: NA-.08). Interestingly, these models seem to assume a universal belief in global warming, an assumption that does not accurately reflect the diversity of real-world viewpoints in the U.S. To enhance the algorithmic fidelity of LLMs, we introduced additional covariates relevant to global warming, such as issue involvement, interpersonal discussion about global warming, and awareness of the scientific consensus about global warming. When GPT-4 was conditioned on both demographics and these additional covariates, its MAF1 improved from .49 to .82 in the 2017 survey and from unavailable to .85 in 2021. Similarly, under the same conditionings, GPT-3.5's MAF1 increased from unavailable to .53 in 2017 and to .65 in 2021.

We then evaluated the effects of adding a third response option, "Don't know," instead of limiting GPTs to a simple "Yes/No" decision, mirroring the approach commonly used in public opinion surveys. The introduction of an additional response option decreased the accuracy of both GPTs. The average accuracy across models, conditions, and years dropped to 75% (*SD* = 3.70), lower than in binary scenarios. GPTs conditioned only on demographics failed to generate "No" or "Don't know" responses, resulting in nonexistent F1 scores for these categories. GPTs conditioned on both demographics and covariates showed improved performance. Nevertheless, GPT-4 had more difficulty with "Don't Know" (2017 F1: .16; 2021 F1: .20) than "No" (2017 F1: .58; 2021 F1: .60). GPT-3.5 performed poorly with both "No" (2017 F1: .24; 2021 F1: .21) and "Don't Know" (2017 F1: .32; 2021 F1: .34).

Fig 1 displays the response distributions from survey participants and silicon samples regarding their belief that global warming is happening, with two (upper panel) and three (lower panel) response choices. Similar to the binary version, models that were conditioned solely on demographics with three answer choices overestimated the proportion of individuals who believe global warming is happening, compared to the actual survey results. When GPTs were conditioned on both demographics and global warming covariates, response distributions aligned more with the survey data. The evaluation metrics for binary and multiple choice are available in Tables D and E in S1 Text, respectively.

Algorithmic fidelity of global warming cause

We then evaluated GPT models' responses regarding the causes of global warming against survey participant answers. The survey provided the options “Human,” “Nature,” “Both,” “Global warming isn't happening,” among others (labels simplified here, full details in the Methods section). These options were replicated in our prompts. The average accuracy across GPT models, conditions, and years was 51% (*SD* = 7.42). MAF1 scores were not available, as models failed to produce F1 scores for some answer options, preventing MAF1 calculation.

Fig 2 compares response distributions from the survey and silicon samples regarding global warming causation. Notably, GPT-4, conditioned only on demographics, greatly overestimates the number of individuals attributing global warming to human activities compared to GPT-3.5 under the same conditions. This was unexpected, considering GPT-4's general superiority in cognitive tasks. However, adding covariates to demographic conditioning aligns the response distribution more closely with the survey data. The evaluation metrics are available in Table F in S1 Text.

Algorithmic fidelity of global warming worry: From categorical answers to ordinal assessment

In the final phase, we asked the GPT models about their estimated level of worry about global warming. In the survey, this question was structured as an ordinal variable with four distinct categories: "Very Worried," "Somewhat Worried," "Not Very Worried," and "Not At All Worried." We incorporated this ordinal scale in the prompts provided to the GPT models. The average accuracy across models, conditions, and years was 48% (*SD* = 13.02). GPT-4, conditioned only on demographics, matched survey data poorly (2017 MAF1 = .22, 2021 MAF1 =

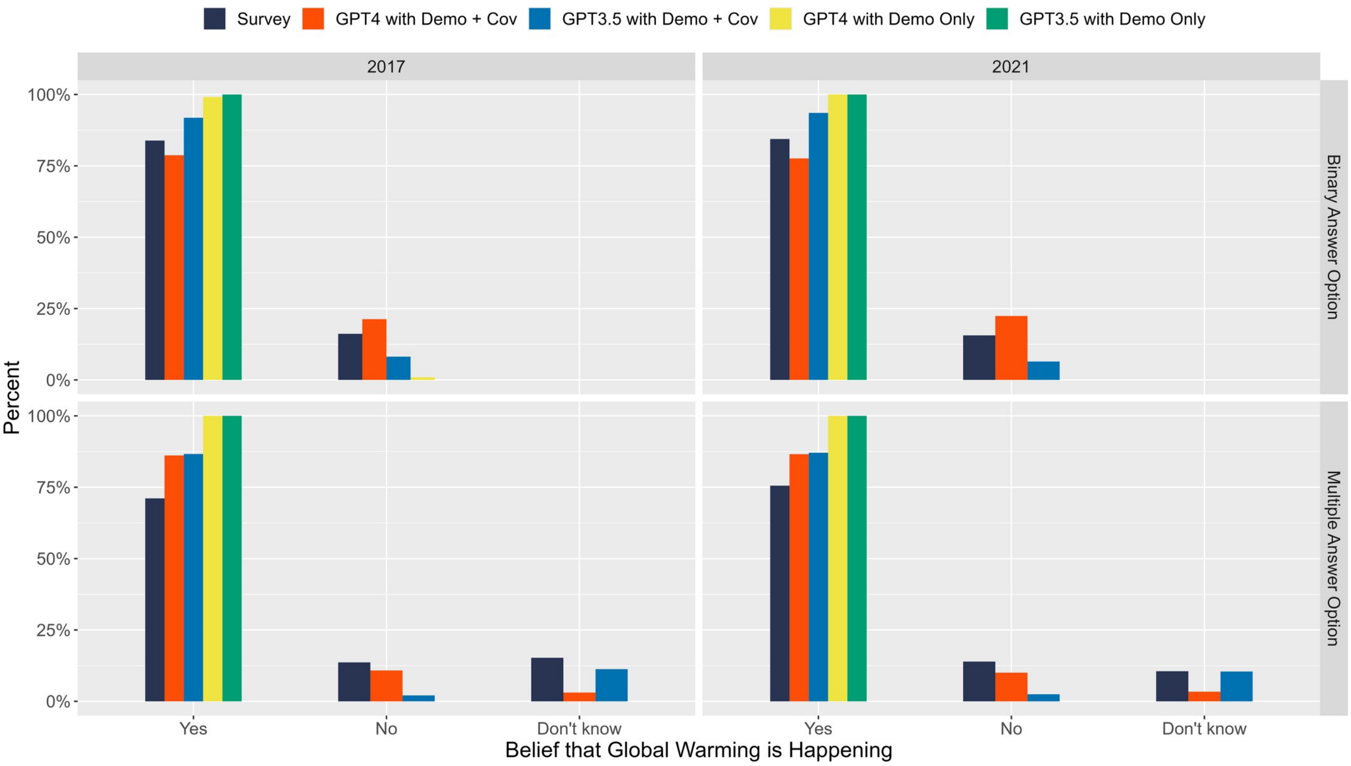

Fig 1. Belief that global warming is happening: Distributional comparison of survey and silicon samples. *Note*: "Demo Only" represents GPTs are conditioned solely on demographics and "Demo + Cov" represents GPTs are conditioned on demographics and covariates.

https://doi.org/10.1371/journal.pclm.0000429.g001

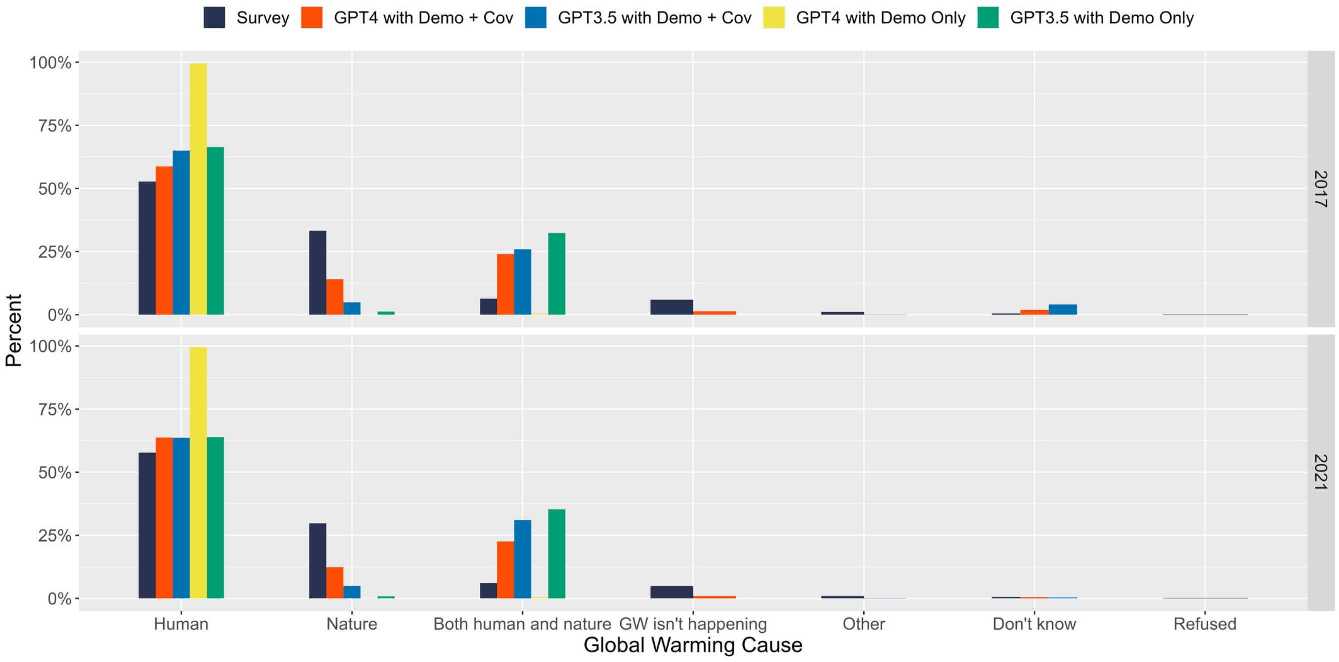

Fig 2. Global warming cause: Distributional comparison of survey and silicon samples. *Note*: "Demo Only" represents GPTs are conditioned solely on demographics and "Demo + Cov" represents GPTs are conditioned on demographics and covariates.

https://doi.org/10.1371/journal.pclm.0000429.g002

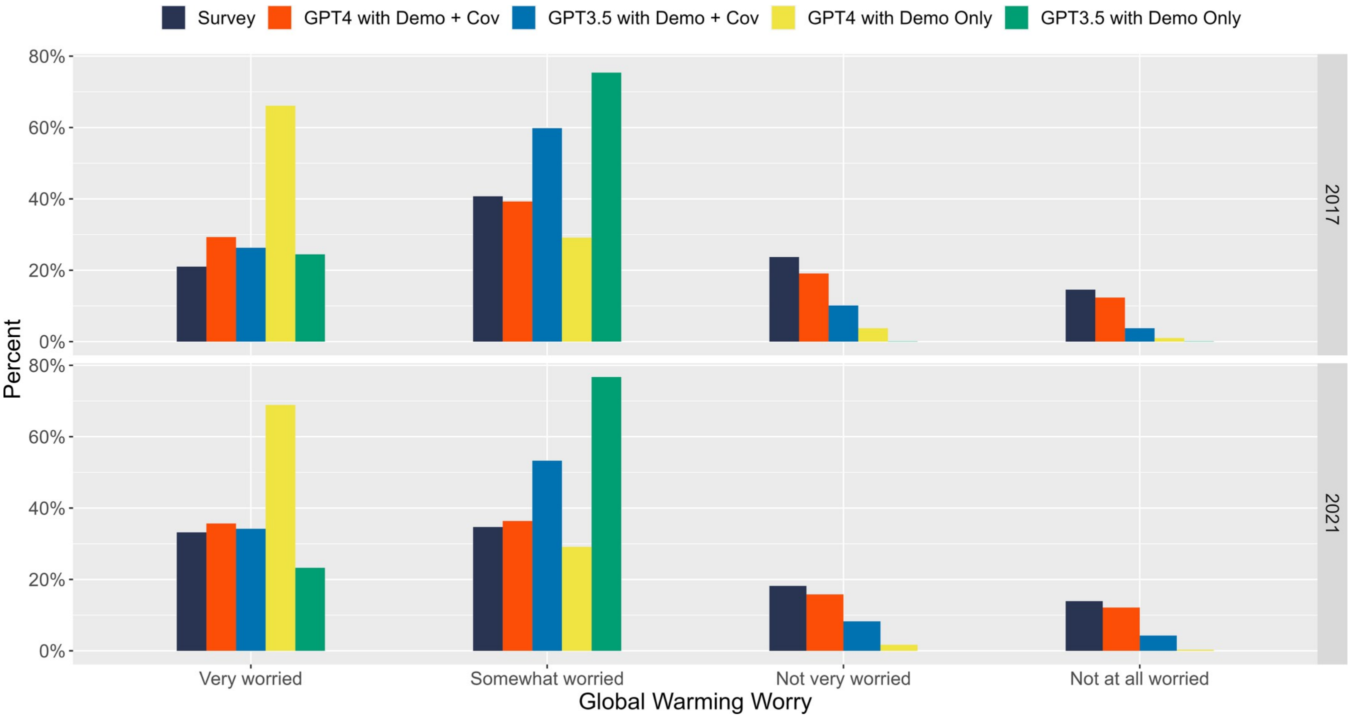

Fig 3. Global warming worry: Distributional comparison of survey and silicon samples. *Note*: “Demo Only” represents GPTs are conditioned solely on demographics and “Demo + Cov” represents GPTs are conditioned on demographics and covariates.

https://doi.org/10.1371/journal.pclm.0000429.g003

.22). GPT-3.5 with the same conditioning failed to produce any responses for "Not very worried" and "Not at all worried," rendering MAF1 unavailable for both years. Adding additional covariates to demographics improved fidelity, with GPT-4 outperforming GPT-3.5 (2017 MAF1 = .65 and .47; 2021 MAF1 = .54 and .50, respectively).

Fig 3 illustrates how the interplay between conditions and model version impacts algorithmic fidelity. Both GPT-4 and GPT-3.5, when conditioned only on demographics, overestimated the number of individuals worried about global warming. Consistent with earlier findings, GPT-4’s estimations were more extreme compared to GPT-3.5, particularly in over-representing those "very worried" about global warming. However, GPT-4 conditioned on both demographics and covariates displayed a response distribution more closely aligned with survey data than the similarly conditioned GPT-3.5. The evaluation metrics are available in Table G in S1 Text. Fig B and Table H in S1 Text summarizes the assessment of overall distributions between survey and silicon samples across models and survey items based on Kullback-Leibler Divergence (KLD).

Algorithmic bias assessment across sub-populations

Here we examine how GPT models represent beliefs that global warming is happening across different sub-populations. The focus was on high-fidelity models in which GPT-4 conditioned on both demographics and covariates for binary beliefs about global warming.

Table 1 details the accuracy and MAF1 results. Interpretations were based on MAF1, where a score below 0.70 is typically considered inadequate. The overarching findings indicate that GPT-4 accurately predicted belief that global warming is happening among diverse sub-populations. However, certain sub-populations were less accurate. GPT-4 was less precise in predicting Non-Hispanic Blacks’ belief that global warming is happening in both 2017 (MAF1 =

Table 1. Accuracy and MAF1 of GPTs for global warming belief across sub-populations.

	GW Belief		GW Belief	
	2017 GPT-4 Demo + Cov		2021 GPT-4 Demo + Cov	
Variables	**Acc**	**MAF1**	**Acc**	**MAF1**
Race/ethnicity				
2+ Races, Non-Hispanic	.94	.86	1.00	1.00
2017 *n* = 32				
2021 *n* = 18				
Black, Non-Hispanic	.90	**.62**	.92	**.60**
2017 *n* = 104				
2021 *n* = 73				
Hispanic	.90	.77	.92	.82
2017 *n* = 132				
2021 *n* = 101				
Other, Non-Hispanic	.96	.91	.89	**.64**
2017 *n* = 51				
2021 *n* = 38				
White, Non-Hispanic	.88	.82	.90	.86
2017 *n* = 786				
2021 *n* = 669				
Gender				
Female	.89	.80	.92	.85
2017 *n* = 568				
2021 *n* = 454				
Male	.89	.84	.90	.85
2017 EL *n* = 431				
2021 EL *n* = 401				
2017 *n* = 537				
2021 *n* = 445				
Age				
18–29	.91	.85	.89	.72
2017 *n* = 145				
2021 *n* = 113				
30–44	.88	.78	.94	.87
2017 *n* = 243				
2021 *n* = 171				
45–59	.88	.78	.89	.84
2017 *n* = 299				
2021 *n* = 230				
60+	.89	.85	.91	.87
2017 *n* = 418				
2021 *n* = 385				
Political ideology				
Moderate, middle of the road	.90	.74	.93	.82
2017 *n* = 424				
2021 *n* = 380				
Somewhat conservative	.81	.80	.81	.80
2017 *n* = 206				
2021 *n* = 165				

(*Continued*)

Table 1. (Continued)

	GW Belief		GW Belief	
	2017 GPT-4 Demo + Cov		2021 GPT-4 Demo + Cov	
Variables	Acc	MAF1	Acc	MAF1
Somewhat liberal	.96	**.68**	.98	.83
2017 *n* = 262				
2021 *n* = 168				
Very conservative	.76	.72	.83	.82
2017 *n* = 99				
2021 *n* = 84				
Very liberal	.97	.78	.98	-
2017 *n* = 104				
2021 *n* = 88				
Political party				
Democrats	.96	**.62**	.97	.74
2017 *n* = 535				
2021 *n* = 423				
Independent/Other	.89	.84	.91	.84
2017 *n* = 108				
2021 *n* = 96				
No party/Not interested	.81	.73	.84	.78
2017 *n* = 105				
2021 *n* = 50				
Republicans	.81	.81	.83	.83
2017 *n* = 354				
2021 *n* = 321				
Education				
Bachelor’s degree or higher	.92	.84	.93	.86
2017 *n* = 440				
2021 *n* = 358				
High school	.86	.81	.88	.83
2017 *n* = 272				
2021 *n* = 226				
Less than high school	.88	.74	.98	.96
2017 *n* = 64				
2021 *n* = 42				
Some college	.88	.82	.89	.84
2017 *n* = 329				
2021 *n* = 273				

https://doi.org/10.1371/journal.pclm.0000429.t001

.62) and 2021 (MAF1 = .60). Further analysis reveals that GPT-4 underestimated Non-Hispanic Blacks believing that global warming is happening. GPT-4 also underrepresented Non-Hispanic Others’ belief in global warming in 2021 (MAF1 = .64), although this requires further investigation due to the limited sample size and heterogeneity of the results within the subgroup. Table I in S1 Text includes the assessment of algorithmic bias for presidential election voting behaviors. Individual F1 scores for each answer option for presidential election and global warming belief (binary choice) across sub-populations are available in Table J in S1 Text. Moreover, Accuracy and MAF1 scores for global warming belief (multiple choice), cause, and worry across sub-populations are available in Table K in S1 Text.

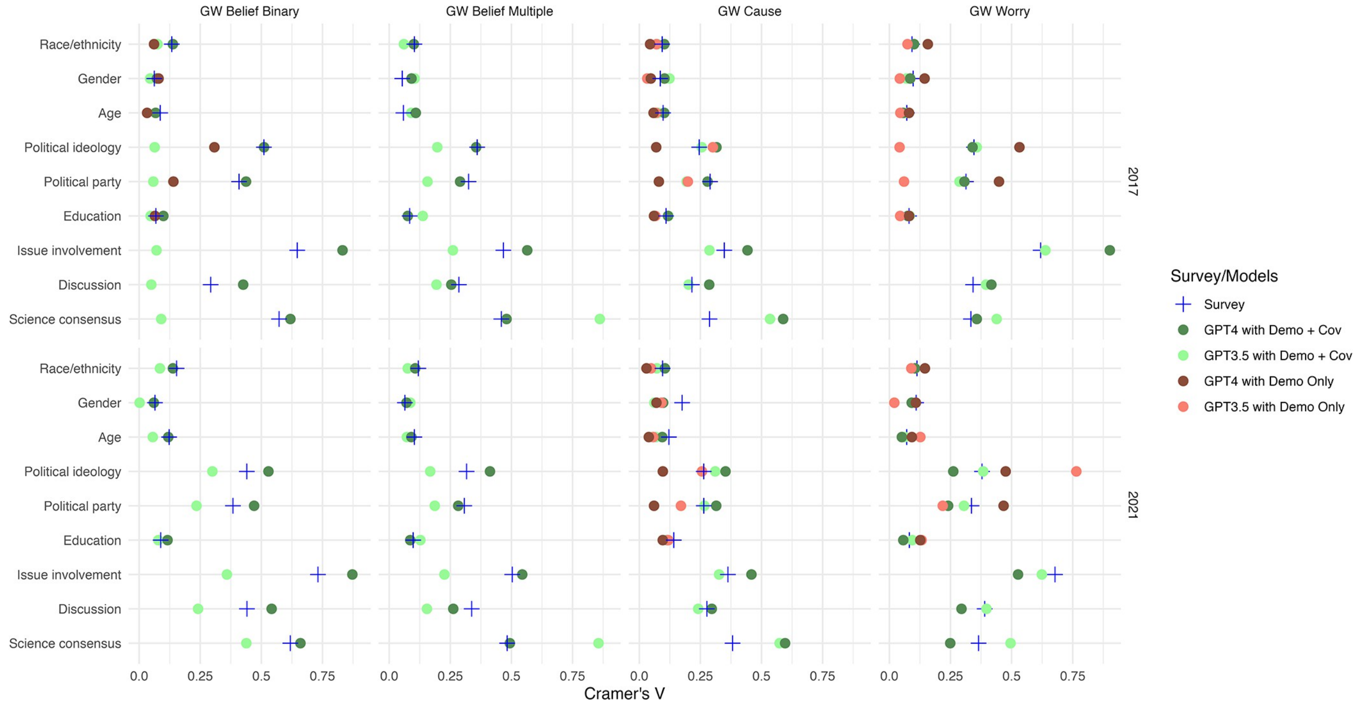

Fig 4. Cramer's V correlations in survey vs. GPTs. Note: Cramer's V could not be estimated for some models because there was no variation in the outcome (e.g., GPT-3.5 with Demo Only for GW Belief).

https://doi.org/10.1371/journal.pclm.0000429.g004

Correlational pattern correspondence assessment

GPT-generated responses "reflect underlying patterns of relationships between ideas, demographics, and behavior that would be observed in comparable human-produced data [5]." Accordingly, we investigated how GPT outputs correspond with correlations of demographics and covariates with the target variables in the survey data. We used Cramer's V to measure association strength. Fig 4 illustrates Cramer's V values between variables for both surveys and GPT models. GPT-4, conditioned on demographics and covariates, most closely matched survey data in association patterns. This is evidenced by the smallest difference in Cramer's V between the model and survey across demographic variables (To calculate the average difference in Cramer's V, we only consider Cramer's V values associated with demographic variables. Cramer's V values associated with global warming covariates are excluded from this calculation because they are not consistently available across all models. Including them could potentially lead to an unfair comparison between models.) and years (mean Cramer's V difference (hereafter, *diff*) = .04, *SD* = .05). It was followed by GPT-3.5 with the same conditioning (*diff* = .09, *SD* = .12), GPT-3.5 with only demographics (*diff* = .10, *SD* = .09), and GPT-4 with only demographics (*diff* = .11, *SD* = .11). Not surprisingly, global warming covariates tend to be more strongly correlated than demographics with the target variables. Models accounting for these covariates display better performance that models without them, indicating that integrating relevant background information is essential to achieve a high level of algorithmic fidelity.

Discussion

This research investigates the algorithmic fidelity and bias of Large Language Models (LLMs) by simulating public opinion about global warming and comparing the synthesized data with survey data. Built on prior research investigating algorithmic fidelity in the political realm [5], we extend our analysis to the topic of global warming. The overall findings reveal that LLMs exhibit promising capabilities in predicting public opinion about global warming. Nevertheless, our findings also identify several concerns about employing LLMs in global warming survey research.

Our study finds that incorporating issue-related covariates alongside demographics significantly enhances algorithmic fidelity in global warming research. LLMs conditioned with both are notably better at predicting individual beliefs about global warming than those using only demographic data. This improvement highlights LLMs' ability to incorporate psychological factors, which are often more indicative of global warming beliefs than demographics alone [11]. However, LLMs that rely solely on demographic information perform poorly even though demographic factors like age, education, and political views are known to be associated with global warming beliefs [11].

The version of the LLM also influences its algorithmic fidelity. GPT-4, conditioned on demographics alone, tends to overestimate the belief that global warming is human-caused and worry about global warming, more so than GPT-3.5. This overestimation is reduced when both demographics and issue-related covariates are used, with GPT-4 then offering more accurate predictions and closer alignment with survey results. This indicates that GPT-4's advanced training and alignment may both enhance and limit fidelity, especially for science-based subjects like global warming. GPT-4's design, which includes a thorough alignment process, aims for improved accuracy over GPT 3.5 [7]. With adequate data, such as covariates, GPT-4's sophisticated integration can potentially produce more accurate estimations of individual opinions than its predecessor. However, in the absence of comprehensive data, GPT-4 may overly rely on scientifically aligned opinions that could diverge from actual views. While this notion is conceivable, it lacks robust empirical support, and ongoing research is essential to understand the factors that influence the differing responses of models developed through distinct processes.

It is also crucial to acknowledge that the lack of transparency from AI developers complicates our ability to understand what influences LLM responses, particularly under uncertainty (e.g., when using only demographics). It raises questions about whether massive training data favor certain attitudes toward climate change, or if the post-training adjustments guided by human feedback lack representativeness. To enhance the reliability of LLMs for social science research, transparency in LLM development is imperative.

Our research indicates a potential algorithmic bias in LLMs regarding certain sub-populations, echoing prior findings [9]. LLMs, particularly those refined with human feedback, often reflect opinions more characteristic of liberal, higher-income, higher-educated, non-religious individuals, and those not adhering to religions such as Buddhism, Islam, or Hinduism [9]. In our study, LLMs perform poorly in predicting beliefs about global warming of non-Hispanic Black Americans. This discrepancy is not explained by sample size, as accuracy remains higher for other racial and multi-racial groups with smaller samples. LLMs often reflect the biases inherent in their training data, which can lead to biased outputs. These biases are particularly pronounced when considering groups that are underrepresented in the data sets. The implications of such biases are not confined to any single nation but are of global concern, as the perspectives of individuals from developing or economically disadvantaged countries may be underrepresented. The root of this problem lies in the limited availability and use of diverse training data, as well as potential biases in the human feedback that guides LLMs' learning processes. This situation emphasizes the need for a more inclusive approach in gathering data and curating feedback, ensuring a wider range of voices are heard and accurately reflected in the technology's outcomes. Therefore, a thorough investigation of LLMs' algorithmic biases, especially against marginalized groups and countries, is essential for ensuring fairness and accuracy across broader domains of social science research.

Our study offers actionable guidance for integrating LLMs into climate change research and other fields. We suggest researchers should condition LLMs with domain-specific variables related to global warming, such as public engagement and scientific agreement, to enhance their relevance and accuracy. Because of the multifaceted nature of global warming perceptions, the inclusion of a broader range of covariates might further sharpen the models' precision for particular investigative goals. Additionally, our findings indicate that more sophisticated models like GPT-4 show greater fidelity than GPT-3.5 when equipped with potential covariates. Furthermore, reducing the scope of answer choices, particularly vague ones such as "Don't know," can potentially improve estimation accuracy. Although restricting answer options reduces LLMs' applicability in social science research, it still holds significant value in streamlining exploratory data collection. For instance, in our study, we acquired data from 1,304 synthetic samples at a considerably lower cost (i.e., approximately $2.08 using GPT-3.5 and $20.86 using GPT-4 as of September 2023) and time reduction when compared with traditional survey techniques. This underscores LLMs' cost-efficiency and speed. However, we do not argue that LLMs can replace traditional survey methods and empirical research. Conventional surveys are and will be an essential tool for assessing public opinion, where cost-effectiveness should not overshadow scientific rigor. Hence, we propose employing LLMs as a complementary instrument for preliminary investigations, survey design, outcome forecasting, and hypothesis generation. Meanwhile, established methods should be maintained for empirical studies to ensure the rigor of research findings.

Our research has limitations that suggest avenues for future exploration. First, we focused on closed-ended questions, overlooking the depth offered by open-ended questions. Open-ended responses could provide richer qualitative data, potentially increasing LLM algorithmic fidelity. Future studies should investigate the impact of incorporating open-ended question responses as conditional inputs on LLMs. Second, our research focuses on a narrow aspect concerning the impact of prompt format (e.g., the number of answer options) on algorithmic fidelity, leaving unexplored other facets of prompt structure that may affect fidelity, such as the order of response options [14]. Other elements like a question sequence and the quantity and arrangement of target questions could also influence algorithmic fidelity. Understanding how these survey design elements affect both human and LLM responses needs more attention in future research.

In conclusion, this research provides valuable insights into the algorithmic fidelity and bias of LLMs in simulating public opinions regarding global warming. This study offers practical guidance on conditioning prompts and selecting models to maximize fidelity in social science applications while emphasizing the importance of validating LLMs, particularly for minority groups. A nuanced approach is required to harness the power of LLMs while addressing their limitations through proactive algorithm auditing and bias mitigation.

Supporting information

S1 Text.
(DOCX)

Author Contributions

Conceptualization: Sanguk Lee, Tai-Quan Peng.

Formal analysis: Sanguk Lee, Tai-Quan Peng.

Investigation: Sanguk Lee, Tai-Quan Peng.

Methodology: Sanguk Lee.

Resources: Tai-Quan Peng, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, Anthony Leiserowitz.

Supervision: Tai-Quan Peng.

Writing – original draft: Sanguk Lee.

Writing – review & editing: Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, Anthony Leiserowitz.

References

1. Bromley-Trujillo R, Poe J. The importance of salience: public opinion and state policy action on climate change. J Pub Pol. 2020; 40: 280–304. https://doi.org/10.1017/S0143814X18000375
2. Doherty KL, Webler TN. Social norms and efficacy beliefs drive the Alarmed segment's public-sphere climate actions. Nature Clim Change. 2016; 6: 879–884. https://doi.org/10.1038/nclimate3025
3. Berinsky AJ. Measuring Public Opinion with Surveys. Annu Rev Polit Sci. 2017; 20: 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724
4. Kim J, Lee B. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. arXiv; 2023. Available: http://arxiv.org/abs/2305.09620
5. Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. Out of One, Many: Using Language Models to Simulate Human Samples. Polit Anal. 2023; 31: 337–351. https://doi.org/10.1017/pan.2023.2
6. Hwang E, Majumder BP, Tandon N. Aligning Language Models to User Opinions. arXiv; 2023. Available: http://arxiv.org/abs/2305.14929
7. Aher G, Arriaga RI, Kalai AT. Using large language models to simulate multiple humans and replicate human subject studies. 2023.
8. OpenAI. GPT-4 Technical Report. arXiv; 2023. Available: http://arxiv.org/abs/2303.08774
9. Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T. Whose Opinions Do Language Models Reflect? arXiv; 2023. Available: http://arxiv.org/abs/2303.17548
10. Goldberg MH, Van Der Linden S, Maibach E, Leiserowitz A. Discussing global warming leads to greater acceptance of climate science. Proc Natl Acad Sci USA. 2019; 116: 14804–14805. https://doi.org/10.1073/pnas.1906589116 PMID: 31285333
11. Hornsey MJ, Harris EA, Bain PG, Fielding KS. Meta-analyses of the determinants and outcomes of belief in climate change. Nature Clim Change. 2016; 6: 622–626. https://doi.org/10.1038/nclimate2943
12. Reser JP, Bradley GL. The nature, significance, and influence of perceived personal experience of climate change. WIREs Climate Change. 2020; 11: e668. https://doi.org/10.1002/wcc.668
13. van der Linden S, Leiserowitz AA, Feinberg GD, Maibach EW. The scientific consensus on climate change as a gateway belief: Experimental evidence. PLOS ONE. 2015; 10: e0118489. https://doi.org/10.1371/journal.pone.0118489 PMID: 25714347
14. Pezeshkpour P, Hruschka E. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv; 2023. Available: http://arxiv.org/abs/2308.11483