

A critical appraisal of equity in conversational AI: Evidence from auditing GPT-3's dialogues with different publics on climate change and Black Lives Matter

Version: September 27, 2022

Kaiping Chen, Anqi Shao, Jirayu Burapachee, Yixuan Li
University of Wisconsin-Madison

*Kaiping Chen (corresponding)
Email: kchen67@wisc.edu

Abstract

Autoregressive language models, which use deep learning to produce human-like texts, have become increasingly widespread. Such models are powering popular virtual assistants in areas like smart health, finance, and autonomous driving, and facilitating the production of creative writing in domains from the entertainment industry to science communities. While the parameters of these large language models are improving, concerns persist that these models might not work equally for all subgroups in society. Despite growing discussions of AI fairness across disciplines, there is a lack of systemic metrics to assess what equity means in dialogue systems and how to engage different populations in the assessment loop. Grounded in theories of deliberative democracy and science and technology studies, this paper proposes an analytical framework for unpacking the meaning of equity in human-AI dialogues. Using this framework, we conducted an auditing study to examine how GPT-3 responded to different sub-populations on crucial science and social topics: climate change and the Black Lives Matter (BLM) movement. Our corpus consists of over 20,000 rounds of dialogues between GPT-3 and 3290 individuals who vary in gender, race and ethnicity, education level, English as a first language, and opinions toward the issues. We found a substantively worse user experience with GPT-3 among the opinion and the education minority subpopulations; however, these two groups achieved the largest knowledge gain, changing attitudes toward supporting BLM and climate change efforts after the chat. We traced these user experience divides to conversational differences and found that GPT-3 used more negative expressions when it responded to the education and opinion minority groups, compared to its responses to the majority groups. To what extent GPT-3 uses justification when responding to the minority groups is contingent on the issue. We discuss the implications of our findings for a deliberative conversational AI system that centralizes diversity, equity, and inclusion.

Introduction

Intelligent assistants have become an inseparable part of our daily lives in recent years, from chatbots used in financial services and smart healthcare to conversational AI systems such as Alexa, Siri, and Google Assistant (1-3). This wide and growing adoption of intelligent assistants, i.e., conversational AI systems, profoundly influence people's lives through affecting human agency in controlling and making decision with technologies (4). For instance, these smart systems not only address simple information-seeking questions from users (5) but also assist in high-stake decision-making scenarios such as surgery (6), collision prevention (7), and criminal justice (8), or serving as educational tools in health persuasion and pro-social behavior (9).

Along with this rapid application, scholars and practitioners have questioned to what extent these intelligent systems are built in consideration of diversity, equity, and inclusion (DEI) (10-13). Recent scholarship has shown that these emerging machine learning systems perform poorly when it comes to accurately identifying speeches from, and images of racial minorities (11). Not only in performance accuracy, but anecdotal evidence has also burgeoned with concerns that AI chatbots may cross the boundaries into sentient (14) and conspiracy responses (15).

While AI fairness has become a heated interdisciplinary discussion in recent years (16), the root of the problem, inequality in conversational systems, is not new. Inequality in communication has been a challenge throughout human history (17). Historically, this problem even goes back to Ancient Athens when philosophers such as Plato and Aristotle argued that democratic dialogues are key to achieving equity in society, but yet they excluded certain populations (such as women, foreigners, and slaves) from participating in these dialogues (18-20). The exclusion of voices from certain social groups in communication systems, whether in the arena of politics (21), social issues (22), or media (23) persists today. As scholars of democracy emphasize, inclusive dialogues should not only consist of participants from diverse backgrounds and life experiences but also should be a process where people can express, listen to, and respond to diverse viewpoints (24). Yet, inequality is "always in the room" because different languages carry different socioeconomic and cultural powers, and those whose language styles signify power often dominate the conversations (25).

Powered by large language models (LLMs), conversational AI is a form of communication system. Like in communications between humans, it faces the challenge of ensuring equity and inclusion in dialogues. What distinguishes this new form of communication technology (between humans and intelligent agents) from other traditional communication systems (e.g., interpersonal communication, mass media) are its powerful and profound implications for every sector of people's lives. Yet, how these new communication AI systems work (e.g., algorithms, training dataset) is often a black box to researchers due to industrial proprietorship (26, 27). As warned by science and technology studies (STS) scholars, emerging technologies can amplify existing social disparities (wealth, power) and discrimination such technologies are designed and deployed without input from different publics engaged in a democratic manner (28-30). However, there is no systematic body of work to examine how technological innovations such as conversational AI might reflect and reinforce these existing gaps, nor were there principles to articulate the rules of the relationship between technologies and humans (28). Understanding how emerging communication technologies like conversational AI empower or disempower different social groups can bring both theoretical significance in extending theories of (deliberative) democracy to new media technologies (31) and practical implication for how to design and deploy these technologies to benefit all social members equally (32).

As "machines powered by artificial intelligence increasingly mediate our social, cultural, economic and political interactions (p.477)", research is urgently needed to understand the behavior of AI systems to reap their benefits and minimize harms (33). Computer science scholars have started to audit different conversational AI systems. For instance, Koenecke and her colleagues (11) examined several on-the-market automated speech recognition (ASR)

systems (from Apple, IBM, Google, Amazon, and Microsoft) and found that the accuracy of recognizing African American speech is much lower than other race and ethnicity groups. Similar findings on race and gender bias in ASR systems were shown by other scholars (34, 35). While there is a rising call and institutional efforts to investigate equity in large language models (36, 37), several knowledge gaps have hindered our understanding of this high-stake issue.

First, there is little understanding of what equity means and should look like in conversational AI systems. Most empirics on LLM have investigated equity in terms of gender, race, and ethnicity. However, as extensive research in social science has shown, inequality and a “spiral of silence” in conversations also happen among other social groups such as those that are limited in their language skills and education (25), and those who hold minority viewpoints on an issue (38). Second, most empirical works have focused on speech recognition systems (11, 34, 35) rather than dialogue systems (i.e., chatbots). Dialogue systems are more complicated than speech recognition because it requires these large language models to not only accurately understand what the users say, but also provide relevant responses to engage in conversations with users in a way that involves more interaction and intelligence to address users’ needs. However, few studies have explored how these large language models will respond to different social groups when they engage in conversations that are much more complicated than speech recognition. Studying conversations beyond speech recognition is vital because of the prevalent use of Alexa/Siri systems in the wild. Further, anecdotal evidence suggests that there are several unintended consequences of LLM when they dialogue with humans on different issues, including using gender stereotypes and openly racist language, spewing conspiratorial misinformation, and amplifying the values and viewpoints of certain societal groups rather than benefiting or appreciating humanity as a whole (15, 39).

Drawing on theories of deliberative democracy and science and technology communication, this paper first proposes an analytical framework to assess equity in conversational AI. We then conducted algorithm auditing to collect a large-scale conversational dataset between one of the most advanced conversational AI systems — GPT-3 — and online crowd workers. OpenAI’s GPT-3 is an autoregressive language model family that is capable of human-like text completion tasks; tasks that can be tweaked to generate conversations. GPT-3 was able to achieve its best performance due to its large model capacity with 175 billion parameters and massive training data (40). Our algorithm auditing on GPT-3 provides empirical evidence to address three research questions; questions that aim to evaluate the extent of equity in a conversational AI system.

RQ1. *How do user experiences differ among different subpopulations in the American society when they have dialogues with GPT-3 about crucial science and social issues (i.e., climate change and BLM)?*

RQ2. *How does GPT-3 respond to different populations in our society on crucial science and social issues?*

RQ3. *How are conversational differences correlated with users’ experiences and learning outcomes with GPT-3 on crucial science and social issues?*

Theorizing DEI in Conversational AI: An Analytical Framework for Assessment

Theories from studies on deliberative democracy and science and technology studies (STS) provide valuable principles for assessing inclusivity and equity in a communication system. Although developed for evaluating human communications, they offer lessons on how a human-machine communication system that centers on DEI should be designed and function. In this

paper, we highlight several important principles to consider when assessing equity in conversational AI systems (Figure 1).

First and foremost is *engaging a diverse group of users* in the assessment loop. An extensive body of STS research has shown how certain populations have been excluded from science and technology innovations in various ways, including from the early stage of the R&D process to throughout the deployment and monitoring process of emerging technologies (29, 30). For instance, looking into the history of the development of Natural Language Processing (NLP) algorithms, the researcher found that the training dataset of early NLP comes from the corpus of the Wall Street Journal in the late 80s (41), which is not natural language used by ordinary people in their everyday life. While STS scholars highlight a public engagement approach toward responsible innovation, people of color, for example, are seldom engaged in these technology processes (29). As scholars have flagged, achieving responsible innovation requires us to rethink the meaning of “public” in terms of different “publics” (42, 43) and to engage marginalized populations as creators rather than merely users (44). STS researchers have consistently found that different publics hold varying perceptions of the benefits and risks of emerging technologies (45), which influence their trustworthiness and adoption of modern technologies (46). Diversity in publics not only calls for attention to studying how modern technologies affect populations with different genders and races and ethnicities, but also populations with varying levels of education, language skills, political ideologies, disability, and who hold divergent opinions on science and social issues (24, 42, 44). Therefore, when assessing equity in conversation AI systems, it is important to examine beyond gender, race, and ethnicity — the often-stressed attributes in AI fairness scholarship — to investigate how populations with limited education and language skills, and who hold minority views on an issue engage with these intelligent agents compared to the majority populations in these demographic and attitudinal attributes.

The second assessment dimension examines *equity through user experiences and learning*. In particular, we ask whether there is a gap between different users regarding how they feel about the chat experiences with the conversational AI system, and to what extent their knowledge about an issue grows after the chat. User experience is a key criterion adopted by human-computer interaction (HCI) designers and researchers; scholars emphasize the importance of evaluating users’ satisfaction, intention to use the system again, and intention to recommend the system to others (47, 48). Thus, a key component of equity is to understand whether people’s experiences of engaging with these systems vary based on their demographic and attitudinal attributes. Besides user experiences, learning is another key criterion; computer-mediated communication scholars use it to study how much knowledge gains users achieve after they engage with a chatbot system about a topic area (49). For example, researchers find that chatbots designed with human-like persuasive and mental strategies are more likely to persuade users to increase their knowledge of healthy and pro-social behaviors (9, 50, 51). An equitable conversational AI system should bring comparable user experiences to different populations and nurture social learning.

In addition to examining user experiences and learning focus by surveying people’s sense of equity after the chat, it is equally critical to understand equity during the dialogue process. Thus, the third assessment dimension investigates *equity in the deliberation styles* by asking whether there is a gap in the styles and sentiments when a conversational AI system responds to different users. A democratic communication system, termed communicative democracy (52), includes a diverse set of language styles: the use of greeting languages, the use of justification (i.e., deliberative languages (53)) when expressing an opinion (justification includes citing facts as well as personal stories (54)), and the use of rhetoric (e.g., appeals to emotion). These elements have guided scholars to analyze how people talk to each other. Applying to human-AI communication, a conversational AI system that embodies equity should respond to different social groups comparably with greetings, the use of justification in expressing opinions, and emotions. In a word, disparities such as the frequency of negative emotions or opinion expression without

justifications need to be minimized when a conversational AI system responds to populations who differ in their education level, race and ethnicity, opinion toward an issue, etc.

Algorithm Auditing Designs

By operationalizing our theoretical framework, this paper is, to our knowledge, *the first study to audit how GPT-3 has conversations with different subpopulations in U.S. on crucial science and social issues*. Recent scholarship on algorithm auditing and HCI informed several key design decisions in our data collection (detailed in Methods and Materials). First, like many HCI studies that stress user experiences as one key criterion for assessing a chatbot system (47, 48), we designed a series of user experience questions in our post-survey. Second, to measure participants' knowledge gains after engaging with a chatbot system, we draw from the design of asking knowledge questions in pre- and post-surveys to measure how participants' attitudes toward an issue change (e.g., knowledge gains) (24). Moreover, across many auditing studies, the researchers emphasize the importance of an organic data collection design that allows participants to directly interact with the algorithms (55, 56). We built a user interface for crowd workers to directly converse with GPT-3 model. The interface integrates GPT-3 API, which gives real-time responses to our participants. Finally, we ensured diversity in our participant recruitment by using screening questions and pilot testing to understand the demographic distribution of the online crowdsourcing platform.

Our algorithm auditing study focuses on two crucial topics, topics that our participants had conversations about with the GPT-3 model: climate change and Black Lives Matter (BLM). The first topic represents a classic controversial science issue, where research has shown how public perceptions toward it vary across populations, especially when there is a constant minority group that holds doubtful and denial attitudes toward climate change is real and human-induced and are difficult to persuade to think otherwise (57). Studying the topic of climate change thus provides an excellent opportunity to examine not only how GPT-3 responds to this group compared to the opinion majority, but also whether there might be social learning and attitudinal changes after the chat. The second topic represents a heated social issue, which has raised continuous attention from the media and public in the recent decade (58).

Results

A Substantive User Experiences Gap in Engaging with GPT-3 for the Opinion and the Education Minority Groups

For RQ1, which investigates how users' experiences differ after having dialogues with GPT-3 about crucial science and social issues, there is a substantive divide in user experiences in the opinion and the education minority groups. Moreover, we do not observe a significant user experience divide between the race and ethnicity minority vs. the majority, nor between male vs. female participants.

Table 1 and Table 2 present the results of quantile regressions of our dependent variables (i.e., user experiences with GPT-3) over our major demographic and attitudinal predictors. These quantile regressions allow us to examine for each quantile of our dependent variables, how our major variables of interest (i.e., populations that consist of the majority vs. minority for each demographic and opinion attribute) are associated with their user experiences with GPT-3.

After dialoguing with GPT-3 about the climate change issue, opinion minorities, compared to opinion majorities, consistently reported more negative experiences including lower ratings, lower satisfaction, worse learning experiences, and less intention to continue the chat or recommend this chatbot to others. These negative relationships are stronger at lower quantiles (e.g., for users' learning experiences from the bot, $\beta_{q0.25} = -1.02$; $\beta_{q0.50} = -0.75$; $\beta_{q0.75} = -0.48$), suggesting

that the user experience gap between the majority vs. the minority opinion groups becomes more appreciable for users with worse experiences with the chatbot.

Education minorities, compared to the education majority group, also reported more negative experiences with GPT-3. They reported lower ratings, worse learning experiences, and less intention to continue the chat. The pattern of the user experience gap between the education majority vs. minority group also became larger for users with worse experiences (i.e., lower quantiles of DV).

When it comes to user experiences after dialogues on BLM, we found that participants' opinions toward BLM are significantly associated with their user experiences with GPT-3. Opinion minorities (i.e., those who hold less agreement that Black lives matter) reported much lower user experiences with GPT-3 compared to the opinion majorities. The effect size of the user experience gap for opinion minorities in BLM is even larger compared to that for the opinion minorities in the climate change discussions.

Knowledge Gains are Significant for the Opinion and the Education Minority Groups

Although the opinion minority group and the education minority group reported much lower user experiences with GPT-3 on both issues, their attitudes toward climate change and BLM significantly changed toward a positive direction after the chat (RQ2). Table 3 presents the OLS regression where the dependent variable is knowledge gain, measured by the difference between participants' post-chat and pre-chat attitudes toward an issue.

For the education minorities asked to discuss climate change, their knowledge gain about climate change is 0.07 points higher compared to the education majority groups (on a 1-5 scale). The results suggest that GPT-3 has a much larger educational value for the education minority populations. For the BLM topic, there is no significant difference between the education minority vs. majority group in terms of their knowledge gains on the topic.

On both issues, the opinion minority groups also changed their views by going toward the supportive side of both topics after the chat, and their knowledge gains are 0.2 points and 0.12 points higher than the opinion majority groups for both issues. Considering that we measured their attitude toward each issue on a 1-5 scale, a 0.2 points difference between the majority vs. minority group is substantive as it accounts for nearly 4% more knowledge gains for the minority groups.

To elucidate why the opinion and the education minority groups showed more supportive attitudes toward both issues post-chat (i.e., much larger knowledge gains), we conducted issue stance analyses on 1) all the responses GPT-3 gave to our participants during the climate change / BLM discussions as well as 2) all the responses our participants raised to GPT-3. These stance analyses allow us to unpack how GPT-3's and participants' stances on climate change / BLM changed over the course of the dialogues. Figure 2 shows the change in issue stance for climate change dialogues (left panel) and the BLM dialogues (right panel), respectively for GPT-3 (blue lines) and our participants (orange lines). The y-axis represents the average probability of a supportive stance toward an issue among all the sentences in a round of conversation. For instance, in terms of the climate change issue, we found that at the beginning of the conversations (round 2), about 20% of our participant's discussions expressed clear support for the issue, and near the end of the conversations, about 25% of participants' discussions expressed supportive attitudes. For GPT-3, the number of responses that expressed direct support for this issue declined a little bit over the course. For both issues, there is a trend toward convergence in the supportive stance between GPT-3 and our participants. The results from the stance analyses further explain

the knowledge gains we observed among our participants in their post-chat attitudes toward the issues.

The Black Box of How GPT-3 Responds to Different Sub-Populations: A Close Look at its Deliberation Styles and Sentiment

Different from the majority of literature — most of which studies gender and race inequality in intelligent systems — we found that the conversational differences in how GPT-3 responds to different populations lie more in the *issue opinion* and the *education level* of our participants. Taking the climate change issue as an example, the top panel in Figure 3 presents the prevalence of different topics/themes from analyzing GPT-3's responses to participants. The bottom panel presents the topics that are indications of deliberative styles such as how GPT-3 used justification when it responded to different participants, who vary in their opinions toward the issue, education level, language skillset, and race and ethnicity. The red dots represent the average effects of a user's attributes on whether a topic is more or less likely to be discussed.

We found that GPT-3 was more likely to cite scientific research as well as give more external links in its responses to the opinion minority group than the opinion majority group. For the education minorities, GPT-3 also used more logical reasoning and citation of scientific research to back the fact that (human-caused) climate change is happening, compared to the responses it gave to the education majority group. Below is an example of a typical GPT-3 response that referred participants to scientific evidence:

GPT-3: "I think it is a real phenomenon. Scientists have studied it closely and they have discovered the facts of climate change and its impacts on the Earth."

However, when it comes to the BLM discussions, we found that GPT-3 was more likely to use preference-based responses without providing justifications (i.e., topic 5, topic 7 in Figure 4) when they responded to the education and the opinion minorities. Below are two examples of typical GPT-3 responses that were given to our participants that is preference-based.

GPT-3: "I do not think it would be a good idea to talk about this. As much as I do like to help you, this is a matter we truly disagree on."

GPT-3: "The absolute truth is that I don't know the answer to this question. Although in some dark corner of the world there is probably a man who does know the answer."

Besides examining the differences in the deliberation styles of how GPT-3 responded to different populations, we investigated the sentiments and rhetoric of its responses. In the climate change discussions (Figure 5), GPT-3 used more words in its responses, as well as more analytical expressions to the education minorities, compared to its' responses to the education majority group. However, it also used fewer positive sentiments in responses to the education minorities. For BLM discussions (Figure 4), GPT-3 also used more words when it responded to the education minority group, compared to its responses to the education majority group. With the opinion minority group, GPT-3 used a more negative sentiment and fewer analytical expressions in its responses, echoing our findings about the deliberation style GPT-3 used where it tended to express its views with less justification when discussing the BLM issue with the opinion minority group. In short, we found that GPT-3 used fewer positive sentiments while having conversations with the opinion and the education minority group on both issues.

Correlation between GPT-3's Conversational Styles and User Experiences

For our final research question (RQ3), which examines how GPT-3's conversational styles might be associated with user experiences, we found that in climate change conversations, user experiences are positively associated with the number of words, positive emotions, analytics, and clout (confident) expressions used in GPT-3's responses (Table 4). These positive associations are stronger (i.e., with larger coefficients) for users with worse user experiences (i.e., lower

quantiles of our DVs), suggesting that these conversational features can bring more positive impact for those users who expressed worse experiences. We also noticed that GPT-3's use of negative words is negatively associated with participants' learning experience, intention to continue the chat, and recommend the chatbot. Similar patterns in the association between GPT-3's conversational styles and user experiences are also found in the BLM issue (Table 5). For instance, the numbers of words used in GPT-3's responses are positively associated with participants' ratings (e.g., this bot is human-like). Positive emotions used in the responses are associated with a higher rating of the bot and a higher intention to continue chatting with the bot.

Discussion

Despite rising awareness of measuring and improving fairness in AI models across disciplines in recent years (10, 12, 13, 59), there is a lack of systemic ways to assess what equity means in conversational AI — what a democratic conversation may look like when we extend communications between humans to intelligent agents. This paper offers a starting point to draw from theories in deliberative democracy, STS, and science communication to propose an analytical framework for evaluating equity in conversational AI. This framework first highlights that people in the assessment loop should be considered beyond the often-studied demographic traits such as gender, race, and ethnicity by expanding to consider people's education level, language skills, as well as their attitudinal traits of an issue. This framework unpacks equity, not only in terms of user experiences and learning across different populations with the conversational AI system but also in response styles (e.g., deliberation, sentiments) given by the conversational AI system to different populations. As one of the first studies to audit GPT-3's dialogues with different populations on crucial social and scientific issues, our empirical findings inform HCI designs in several ways regarding how to centralize DEI (60).

First, our findings about the opposing forces between user experiences and knowledge gains for certain minority populations reveal a potential dilemma facing conversational AI designs; while the opinion and the education minority groups reported much worse user experiences with GPT-3 in their post-survey compared to the opinion and the education majority groups, these two groups also achieved the largest knowledge gains, changing attitudes toward supporting human-induced climate change and BLM after the chat. This dilemma between uncomfortable user experience and positive education values of human-AI conversations speaks to classic communication theories of persuasion effects. In particular, researchers have found that certain communication strategies (e.g., using the loss frames, and fear appeals) make participants uncomfortable, yet they can be effective in persuading pro-social behavior and attitudes (61, 62). Successful persuasion has one or three effects on public attitudes: reinforcing beliefs, revising beliefs, and increasing knowledge. Across our findings, we found all these three impacts. For the opinion majority groups, their attitudes are reinforced toward being more supportive of climate change and BLM after the chat. In our opinion minority groups, for the climate change and BLM issues, their attitudes shifted towards being more supportive. This could be due to their experiencing cognitive dissonance — a phenomenon when people experience discomfort when they encounter opinions that are different from theirs (63). Cognitive dissonance sometimes can fuel people to update their beliefs (64). For our education minority groups, they became more informed after the chat and more supportive of human-induced climate change and BLM movements.

This trade-off between dissatisfactory user experience and positive knowledge gain motivates AI engineers and regulators to think more about the mission of a conversational AI system in terms of how to find a balance between the two. Towards that end, our open-ended post-survey, which asked what users had hoped to hear from GPT-3 on the topic they discussed, provides some potential user-centered answers. One key suggestion our participants offered for enhancing their user experiences with GPT-3 is to avoid repetitive answers to make the response less boring and richer in vocabulary. Other participants shared that they were disappointed when GPT-3 responded that it is not human, or it cannot understand what the human asked in the chat. Exploring how to vary the language styles of LLM to achieve this balance is of high importance for

conversational AI designers and engineers, as our participants expressed that they would lose trust in the AI system and would not use the system again in the future.

Second, examining AI equity beyond race, ethnicity, and gender will inform the design of conversational AI systems for more complicated tasks beyond simple Q&A exchanges. With more and more chatbots entering the area of “social-oriented” tasks (51) and “persuasive communication” (e.g., health apps) (9), understanding how an LLM would respond to populations who hold various opinions (i.e., value systems about social issues) and how these opinion groups perceive the benefits and limitations of conversational AI systems are critical for enhancing diversity in LLM training datasets. For instance, we found that while GPT-3 used more justification such as citing external links and scientific research when responding to the education and the opinion minorities on climate change, it also used more negative sentiment words in the responses, which might explain worse user experiences among these sub-populations. Differently, for BLM, we observed that GPT-3 used less justification when it responded to the education and the opinion minorities. These nuances in how these intelligent agent systems responded to different publics with varying deliberative styles and sentiments, and how their response styles also vary on the issue offer vital implications for studying equity in dialogue systems --- any single conversational aspect such as sentiment is not adequate to understand potential biases in dialogue systems; instead, we need a theory-driven approach, e.g., theories of deliberative democracy, to guide multi-dimensional metrics to investigate different aspects of dialogues that can contradict each other.

The results of GPT-3’s less use of justification and more negative sentiment toward the opinion and the education minority groups on the BLM issue can bring unintended consequences for these groups. As extensive literature in human-to-human communication notes, during public deliberation about controversial social issues, participants who hold minority opinions can go into a spiral of silence when they perceive their opinions are the minority in society and are thus less likely to speak out during the discussion, thus their voices are under-heard (38). Participants with limited education skills are also found to be less listened to during public deliberation (65). In a word, human-AI conversations often mirror inequalities in human communications. Breaking off from the persistent challenge facing humanity requires more listening to these minority voices throughout the AI system design process.

Our paper offers the starting point for a critical inspection of what equity means in conversational AI and the status of equity and inclusion in LLM to move forward, future research can expand the scope of the topic issues for auditing dialogues. Moreover, although collecting dialogues from existing chatbots is a challenge due to industry proprietary, it is still valuable to examine other LLM and verify how our findings hold and vary across different conversational AI systems. Finally, in our auditing study, participants typed their answers, which did not allow us to examine other important demographic traits such as the role of accents and tones in dialogue systems. User experiences from minority groups can differ when they use written vs. oral communications. These are all exciting areas toward a more deliberative conversational AI system.

Materials and Methods

We briefly describe our algorithm auditing design for data collection, measurements for user experiences, and methods for analyzing human-AI dialogues, as well as our statistical models to address the three RQs. Further details are provided in *Supplemental Information (SI)*.

Data and Auditing Design

We adopted Generative Pre-trained Transformer 3 (GPT-3) from OpenAI as the language generator for our chatbot. OpenAI’s GPT-3 is an autoregressive language model family that is capable of human-like text completion tasks; tasks that can be tweaked to generate

conversations. GPT-3 was able to achieve its best performance due to its large parameter capacity of 175 billion parameters. Although it is not the current state-of-the-art model, the model's architecture is based on the transformer architecture which is being used extensively in language models in the past couple of years. We have seen many successors such as LaMDA and MUM, which are based on the transformer as well. Since it is hard to interpret how complex models understand our inputs, we believe that our analysis can detect some patterns that model research will need to take into consideration. In terms of implementation, OpenAI provides us the text completion API, which we were able to utilize in a chatbot manner just like their demo to collect the dialogue dataset our participants had with GPT-3. For the GPT-3 models, we used the most capable version available at the time of data collection (text-davinci-001), and we provided details in SI Appendix x about how we used content filter configuration and the code to replicate our web application. We performed a series of tests to validate the consistency of the responses generated by GPT-3 such as whether it can recognize and give similar output on synonyms and double negatives. (See SI Appendix A 1.3).

Our auditing design follows three stages: a pre-dialogue survey, dialogues, and a post-dialogue survey.

Stage 1. The pre-dialogue survey measured participants' demographics including race/ethnicity, gender, age, income, education, and ideology. To assess their efficacy in chatbot-related experiences, we drew from existing measurements of consumer experience with technology (48) and public responses to AI (66). Participants were then divided into two groups to discuss one of the following crucial issues: climate change or BLM. Before the conversation, we measured participants' attitudes toward the two issues. For participants in the climate change group, we asked them questions like whether they perceive climate change as a real phenomenon; one that is due to human activities and has negative impacts. For participants in the BLM group, we asked them questions such as whether they support the movement and perceive the movement as necessary. The measurements achieved internal consistency with Cronbach's alphas ranging from 0.86 to 0.88. These attitudinal questions were asked again in the post-dialogue survey and were used to calculate participants' knowledge gains on the two issues.

Stage 2. Participants were then directed to our UI webpage (see SI Appendix A1.2) to have dialogues with GPT-3 on their assigned topic. Each participant was required to have anywhere between six to 12 rounds of conversation with the chatbot. The whole dialogue was organic. We did not manipulate the chatbot to fit either of the topics, hence the dialogue topics were initiated by participants (i.e., participants need to ask questions about climate change or express ideas about BLM first).

Stage 3. The post-dialogue survey assessed participants' evaluations of their experiences with the chatbot (i.e., *user experience*). Five sets of questions were provided for participants to evaluate their user experiences (1) ratings of the chatbot, (2) satisfaction with the dialogue, (3) learning experience with the chatbot, and their intention to (4) continue the chat or (5) recommend the chatbot to others. These evaluation questions were measured on a 5-point Likert scale that asks the participants to indicate to what extent they agree with the statements (47, 48). The ratings of GPT-3 were measured by letting participants evaluate statements including whether the chatbot is natural, friendly, humanlike, and smart. User satisfaction was measured by asking whether they think the conversational experience is satisfying, understandable, fun, enjoyable, and entertaining. Perceived learning experiences were measured by asking whether they think the chatbot helped them learn the topic quickly and improved their confidence in the topic discussed. Intention to continue the chat was measured by asking whether they would like to continue the dialogue, use the chatbot in their daily life, and use the chatbot frequently. Intention to recommend the chatbot to others was measured by asking their likelihood to recommend this chatbot to their friends or family for the topic they discussed, and to other people who need to talk about this topic. Cronbach's alphas suggest high internal consistency in these measurements, ranging from 0.88 to 0.94. We further had our participants respond to open-

ended questions, writing down what they expected to hear from GPT-3, but GPT-3 failed to provide.

We conducted several rounds of pilot tests in November 2021 to refine our survey instruments and our participant-GPT-3 chat UI design. Then we entered the real launch, during which we recruited 4,240 anonymous participants from the crowdsourcing platform Amazon Turk (MTurk) from December 2021 to February 2022. To ensure data quality, those who failed the attention check questions or completely stayed off-topic during dialogues (e.g., typing random words when asked to talk about climate change) were marked as invalid workers and thus excluded from the analyses. In the end, we had 3,290 valid participants, resulting in 26,211 rounds of dialogues with GPT-3. Table S1 in SI presents the demographic distribution of participants in our sample as well as their pre-chat attitudes toward climate change and BLM.

We recoded participants' race and ethnicity, primary language, education level, and prior attitudes on climate change and BLM into majority vs. minority groups (i.e., a binary variable). 21.49% of participants self-identified as non-white in our recorded sample and were recoded as the minority group for the race & ethnicity variable. The rest 80% of participants who self-identified as white were recoded as the majority group. 8.81% of our participants' first language is not English. There is also a significant education attainment divide in our collected sample. Approximately 82.22% of our sample have at least acquired a bachelor's degree (i.e., the majority group). We recoded participants with lower than bachelor's degrees as education minorities. Considering the high average agreement towards climate change in our sample, we recoded participants' opinions toward climate change into minorities vs. majorities, using the 1st quarter of the average attitude score as the threshold, where a higher score means more support for climate change facts. Similar dividing criteria were also applied to participants in the BLM topic to divide them into the opinion minority vs. majority group.

Analysis Method

We conducted quantile regressions to analyze the relationship between users' demographics, attitudes toward an issue, and their user experiences after the chat (RQ1) as well as the association between GPT-3's conversational features and the resulting user experiences (RQ3). Our data distribution exhibits a highly bi-modal style that most user experience items gathered around either the lower quantile (i.e., extremely negative) or the higher quantile (i.e., extremely positive). In this case, ordinary least squares (OLS) models are not an optimal modeling option since these dependent variables are not normally distributed. Therefore, we turned to quantile regressions to generate more robust and informative models.

To analyze dialogues in this study (RQ2), we conducted topic modeling and linguistic analyses. Specifically, we used structural topic modeling (STM) (67), which produces the most prevalent topics and associated keywords from a set of documents based on the latent Dirichlet allocation (LDA) approach, while also paying attention to prior differences in users' demographics and attitudes. We generated ten topics each for chatbot responses in the climate change and BLM dialogues respectively, with participants' demographic divides and the order of rounds in the dialogue as covariates.

In linguistic analyses, we specified five prespecified dictionaries for keyword analyses from the Linguistic Inquiry and Word Count (LIWC) software (68). These include word counts of positive emotion and negative emotion, analytic words showing logical and formal thinking, clout words showing leadership and bold confidence, and authentic words showing honesty and genuineness. We also included the word count of the response as a simple indicator of the response length from the chatbot.

To elaborate on our findings about why our participants, on average, changed their attitudes toward more believing in climate change and BLM, we performed stance analyses on GPT-3's responses as well as participants' responses during their dialogues. We manually coded 1,390

prompts on the climate change topic, and 1,742 prompts from the BLM topic into dichotomous variables: *supporting climate change/BLM or not*. The prompts were randomly and evenly selected from participants' inputs and chatbot responses. We leveraged Recurrent Neural Networks (RNN) with a gated recurrent unit (GRU) to train two classifiers to identify climate change and BLM stances respectively (69). The models showed good performance in stance detection, with an F1 score of 0.74 for climate change prompts and 0.78 for BLM prompts. We then used this stance model to calculate the probability of whether each prompt was supportive of the topic and predict the binary stance in a one-hot manner. Further details of the codebook and training processes are in SI Appendix A 3.4.

Acknowledgments

Support for this research was provided by American Family Insurance through a research partnership with the University of Wisconsin–Madison’s American Family Insurance Data Science Institute. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not reflect the views of American Family Insurance.

References

1. Q. Q. Chen, H. J. Park, How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems* (2021).
2. S. Tian *et al.*, Smart healthcare: making medical care more intelligent. *Global Health Journal* **3**, 62-65 (2019).
3. T.-H. Wen *et al.* (2020) Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*.
4. R. Fanni, V. E. Steinkogler, G. Zampedri, J. Pierson, Enhancing human agency through redress in Artificial Intelligence Systems. *AI & SOCIETY*, 1-11 (2022).
5. E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, R. Socher, A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems* **33**, 20179-20191 (2020).
6. N. Mirchi *et al.*, The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PloS one* **15**, e0229596 (2020).
7. S. J. Cachumba, P. A. Briceño, V. H. Andaluz, G. Erazo (2019) Autonomous driver assistant for collision prevention. in *Proceedings of the 2019 11th International Conference on Education Technology and Computers*, pp 327-332.
8. R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art. *Social Method Res* **50**, 3-44 (2021).
9. J. Zhang, Y. J. Oh, P. Lange, Z. Yu, Y. Fukuoka, Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *J. Med. Internet Res.* **22**, e22845 (2020).
10. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq (2017) Algorithmic decision making and the cost of fairness. in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp 797-806.
11. A. Koenecke *et al.*, Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* **117**, 7684-7689 (2020).
12. M. D. McCradden, S. Joshi, M. Mazwi, J. A. Anderson, Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* **2**, e221-e223 (2020).
13. S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
14. The Sentinel (2022) Google wants you to chat with its Artificial Intelligence chatbot at your own risk.
15. S. Johnson (2022) AI is mastering language. Should we trust what it says?
16. R. Reich, M. Sahami, J. M. Weinstein, H. Cohen (2020) Teaching computer ethics: A deeply multidisciplinary approach. in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pp 296-302.
17. K. A. Raaflaub, *Equalities and inequalities in Athenian democracy* (Princeton University Press Princeton, NJ, 1996).
18. M. Schofield, *Plato: political philosophy* (Oxford University Press, 2006).
19. M. H. Hansen, *The Athenian democracy in the age of Demosthenes: structure, principles, and ideology* (University of Oklahoma Press, 1999).
20. W. Von Leyden, *Aristotle on equality and justice: His political argument* (Springer, 1985).
21. J. J. Mansbridge, *Beyond adversary democracy* (University of Chicago Press, 1983).
22. A. Gutmann, *Liberal equality* (CUP Archive, 1980).
23. K. Chen, J. Jeon, Y. Zhou, A critical appraisal of diversity in digital knowledge production: Segregated inclusion on YouTube. *New Media & Society*, 14614448211034846 (2021).
24. J. Fishkin, *When the people speak: Deliberative democracy and public consultation* (Oup Oxford, 2009).
25. A. Lupia, A. Norton, Inequality is always in the room: language & power in deliberative democracy. *Daedalus* **146**, 64-76 (2017).

26. I. Freiling, N. M. Krause, D. A. Scheufele, K. Chen, The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication* **71**, 686-714 (2021).
27. E. Ruane, A. Birhane, A. Ventresque (2019) Conversational AI: Social and Ethical Considerations. in *AICS*, pp 104-115.
28. S. Jasanoff, *The ethics of invention: Technology and the human future* (WW Norton & Company, 2016).
29. R. Owen, J. R. Bessant, M. Heintz, *Responsible innovation: managing the responsible emergence of science and innovation in society* (John Wiley & Sons, 2013).
30. D. H. Guston, Building the capacity for public engagement with science in the United States. *Public Understanding of Science* **23**, 53-59 (2014).
31. L. Bernholz, H. Landemore, R. Reich, *Digital technology and democratic theory* (University of Chicago Press, 2021).
32. J. Rhee, *The robotic imaginary: The human and the price of dehumanized labor* (U of Minnesota Press, 2018).
33. I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477-486 (2019).
34. A. B. Wassink, C. Gansen, I. Bartholomew, Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication* **140**, 58-70 (2022).
35. J. P. Bajorek, Voice recognition still has significant race and gender biases. *Harvard Business Review* **10** (2019).
36. R. Bommasani *et al.*, On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
37. C. Potts (2022) Percy Liang on the center for research on foundation models' first and next 30 years.
38. E. Noelle-Neumann, The spiral of silence a theory of public opinion. *Journal of communication* **24**, 43-51 (1974).
39. L. Lucy, D. Bamman (2021) Gender and representation bias in GPT-3 generated stories. in *Proceedings of the Third Workshop on Narrative Understanding*, pp 48-55.
40. L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681-694 (2020).
41. Y. A. Loukissas, *All data are local: Thinking critically in a data-driven society* (MIT Press, 2019).
42. J. Hudson, D. Hudson, P. Morini, H. Clarke, M. C. Stewart, Not one, but many "publics": public engagement with global development in France, Germany, Great Britain, and the United States. *Development in practice* **30**, 795-808 (2020).
43. N. Fraser, "Rethinking the public sphere: A contribution to the critique of actually existing democracy" in *Public Space Reader*. (Routledge, 2021), pp. 34-41.
44. P. Rangnekar (2021) Decoding the "Encoding" of Ableism in Technology and Artificial Intelligence. (Science Connect).
45. H. S. Boudet, Public perceptions of and responses to new energy technologies. *nature energy* **4**, 446-455 (2019).
46. L. Frewer, Risk perception, social trust, and public participation in strategic decision making: Implications for emerging technologies. *Ambio*, 569-574 (1999).
47. Q. V. Liao *et al.* (2018) All work and no play? Conversations with a Question-and-Answer Chatbot in the Wild. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp 1-13.
48. A. Venkatesh *et al.*, On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625* **4**, 60-68 (2018).
49. F. Colace *et al.*, Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research* **7**, 528-533 (2018).
50. W. Shi *et al.* (2020) Effects of persuasive dialogues: testing bot identities and inquiry strategies. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp 1-13.
51. L. Qiu *et al.*, Towards socially intelligent agents with mental state transition and human utility. *arXiv preprint arXiv:2103.07011* (2021).
52. I. M. Young, *Inclusion and democracy* (Oxford University press on demand, 2002).

53. J. Steiner, *The foundations of deliberative democracy: Empirical research and normative implications* (Cambridge University Press, 2012).
54. K. Chen, How deliberative designs empower citizens' voices: A case study on Ghana's deliberative poll on agriculture and the environment. *Public Understanding of Science* **30**, 179-195 (2021).
55. D. Metaxa *et al.*, Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human-Computer Interaction* **14**, 272-344 (2021).
56. A. Papoutsaki *et al.* (2015) Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. in *Third AAAI Conference on Human Computation and Crowdsourcing*.
57. S. J. O'Neill, M. Boykoff, Climate denier, skeptic, or contrarian? *Proceedings of the National Academy of Sciences* **107**, E151-E151 (2010).
58. R. R. Mourão, D. K. Brown, Black Lives Matter coverage: How protest news frames and attitudinal change affect social media engagement. *Digit. Journal.* **10**, 626-646 (2022).
59. N. Science, T. C. S. C. o. A. Intelligence, *The national artificial intelligence research and development strategic plan: 2019 update* (National Science and Technology Council (US), Select Committee on Artificial ..., 2019).
60. Microsoft Research (2019) Fairness and interpretability in AI: Putting people first.
61. G. J. Y. Peters, R. A. Ruiters, G. Kok, Threatening communication: A qualitative study of fear appeal effectiveness beliefs among intervention developers, policymakers, politicians, scientists, and advertising professionals. *Int. J. Psychol.* **49**, 71-79 (2014).
62. C. Yan, J. P. Dillard, F. Shen, Emotion, motivation, and the persuasive effects of message framing. *Journal of Communication* **62**, 682-700 (2012).
63. L. Festinger, Cognitive dissonance. *Scientific American* **207**, 93-106 (1962).
64. E. Harmon-Jones, J. W. Brehm, J. Greenberg, L. Simon, D. E. Nelson, Evidence that the production of aversive consequences is not necessary to create cognitive dissonance. *Journal of personality and social psychology* **70**, 5 (1996).
65. S. W. Rosenberg, *Deliberation, participation and democracy* (Springer, 2007).
66. S. Cave, K. Coughlan, K. Dihal (2019) " Scary Robots" Examining Public Responses to AI. in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp 331-337.
67. M. E. Roberts, B. M. Stewart, D. Tingley, Stm: An R package for structural topic models. *Journal of Statistical Software* **91**, 1-40 (2019).
68. R. L. Boyd, A. Ashokkumar, S. Seraj, J. W. Pennebaker, The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022).
69. K. Cho *et al.*, Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

Figures and Tables

Table 1. User Experience Gaps after Human-GPT-3 Chat on Climate Change

User experience	Quantile	Intercept	Gender (Male)	Race/ethnicity minority	Education minority	Language minority	Opinion minority	Liberal	Conservative	R ²
Rating of the chatbot										
	q _{0.25}	3.75*** (0.49)	-0.04 (0.07)	-0.11 (0.10)	-0.57*** (0.11)	0.17 (0.15)	-0.37*** (0.10)	-0.04 (0.12)	0.46*** (0.11)	0.10
	q _{0.50}	3.51*** (0.31)	-0.02 (0.05)	0.01 (0.07)	-0.33*** (0.07)	0.01 (0.10)	-0.32*** (0.07)	0.01 (0.08)	0.27*** (0.07)	0.07
	q _{0.75}	3.76*** (0.37)	-0.05 (0.06)	-0.05 (0.08)	-0.25*** (0.08)	0.03 (0.11)	-0.24*** (0.08)	-0.12 (0.09)	0.04 (0.09)	0.04
	OLS	4.01*** (0.32)	-0.05 (0.05)	-0.05 (0.07)	-0.33*** (0.07)	0.11 (0.10)	-0.32*** (0.07)	-0.01 (0.08)	0.26*** (0.08)	0.09
Satisfaction with the dialogue										
	q _{0.25}	2.46*** (0.30)	0.01 (0.04)	0.08 (0.06)	-0.19** (0.07)	0.11 (0.09)	-0.46*** (0.06)	0.05 (0.07)	0.27*** (0.07)	0.08
	q _{0.50}	2.47*** (0.23)	0.01 (0.03)	-0.01 (0.05)	-0.06 (0.05)	0.12 (0.07)	-0.45*** (0.05)	-0.04 (0.06)	0.06 (0.05)	0.09
	q _{0.75}	2.34*** (0.28)	-0.03 (0.04)	0.02 (0.06)	0.06 (0.06)	0.10 (0.09)	-0.31*** (0.06)	-0.05 (0.07)	0.02 (0.07)	0.09
	OLS	3.00*** (0.29)	-0.01 (0.04)	-0.02 (0.06)	-0.05 (0.07)	0.17* (0.09)	-0.42*** (0.06)	-0.02 (0.07)	0.18** (0.07)	0.07
Learning experience with the chatbot										
	q _{0.25}	3.24*** (0.58)	0.13 (0.09)	0.07 (0.12)	-0.90*** (0.13)	-0.05 (0.17)	-0.79*** (0.12)	0.12 (0.14)	0.52*** (0.13)	0.16
	q _{0.50}	3.13*** (0.27)	0.02 (0.04)	0.06 (0.06)	-0.44*** (0.06)	-0.08 (0.08)	-0.71*** (0.06)	0.06 (0.07)	0.28*** (0.06)	0.11
	q _{0.75}	2.65*** (0.29)	0.01 (0.04)	0.12* (0.06)	-0.20** (0.07)	-0.07 (0.09)	-0.43 (0.06)	-0.02 (0.07)	0.12 (0.07)	0.04
	OLS	4.07*** (0.36)	0.02 (0.05)	0.10 (0.07)	-0.37*** (0.08)	-0.01 (0.11)	-0.66*** (0.08)	0.04 (0.09)	0.32*** (0.08)	0.16
Intention to continue the interaction										
	q _{0.25}	3.72*** (0.51)	0.01 (0.08)	0.03 (0.1)	-0.87*** (0.11)	-0.01 (0.15)	-0.75*** (0.11)	-0.19 (0.12)	0.34** (0.12)	0.19
	q _{0.50}	2.94*** (0.30)	0.05 (0.04)	0.04 (0.06)	-0.74*** (0.07)	0.09 (0.09)	-0.59*** (0.06)	0.05 (0.07)	0.21** (0.07)	0.12
	q _{0.75}	3.12*** (0.25)	-0.04 (0.04)	0.10* (0.05)	-0.42*** (0.06)	0.13 (0.08)	-0.49*** (0.05)	-0.02 (0.06)	0.08 (0.06)	0.09
	OLS	3.74*** (0.36)	0.01 (0.05)	0.09 (0.07)	-0.50*** (0.08)	0.13 (0.11)	-0.61*** (0.08)	-0.05 (0.09)	0.30*** (0.08)	0.19
Intention to recommend the chatbot to others										
	q _{0.25}	3.21*** (0.52)	0.07 (0.08)	-0.09 (0.11)	-0.83*** (0.12)	0.31* (0.16)	-0.73*** (0.11)	-0.12 (0.12)	0.40** (0.12)	0.17
	q _{0.50}	2.30*** (0.31)	0.01 (0.05)	-0.01 (0.06)	-0.61*** (0.07)	0.07 (0.10)	-0.59*** (0.07)	-0.01 (0.07)	0.21** (0.07)	0.10
	q _{0.75}	2.69*** (0.27)	-0.02 (0.04)	0.01 (0.06)	-0.37*** (0.06)	0.17* (0.08)	-0.47*** (0.06)	0.01 (0.07)	0.09 (0.07)	0.07
	OLS	3.31*** (0.38)	0.01 (0.06)	0.03 (0.08)	-0.46*** (0.08)	0.14 (0.11)	-0.59*** (0.08)	-0.05 (0.09)	0.32*** (0.09)	0.16

Note: The bottom row in each section of the table shows OLS regression as a reference. The number of observations: 1693. The table presents part of the full regression models. In the full regression models, we also controlled for other demographic variables including participants' age, income level, efficacy with using chatbots, as well as the language styles of each participant such as the word count of their average input, their use of positive and negative emotion words, use of analytical words, clout, and authentic expressions in conversations.

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 2. User Experience Gaps after Human-GPT-3 Chat on BLM

User experience	Quantile	Intercept	Gender (Male)	Race/ethnicity minority	Education minority	Language minority	Opinion minority	Liberal	Conservative	R ²
Rating of the chatbot										
q0.25	3.95***	(0.47)	0.02 (0.08)	-0.05 (0.10)	-0.29** (0.11)	0.19 (0.15)	-0.64*** (0.09)	0.07 (0.12)	0.48*** (0.11)	0.13
q0.50	4.33***	(0.32)	-0.05 (0.05)	0.08 (0.07)	-0.20** (0.08)	0.12 (0.10)	-0.6*** (0.07)	0.09 (0.08)	0.32*** (0.08)	0.09
q0.75	4.95***	(0.30)	-0.09 (0.05)	0.04 (0.07)	-0.10 (0.07)	0.07 (0.09)	-0.48*** (0.06)	0.04 (0.08)	0.23** (0.07)	0.05
OLS	4.72***	(0.31)	-0.04 (0.05)	0.05 (0.06)	-0.09 (0.07)	0.09 (0.09)	-0.30*** (0.06)	0.06 (0.07)	0.34*** (0.07)	0.16
Satisfaction with the dialogue										
q0.25	3.08***	(0.39)	0.03 (0.06)	0.02 (0.09)	-0.22* (0.09)	0.30* (0.12)	-0.74*** (0.08)	0.04 (0.100)	0.20* (0.09)	0.12
q0.50	2.17***	(0.20)	-0.03 (0.03)	0.10* (0.05)	-0.10* (0.05)	0.08 (0.06)	-0.43*** (0.04)	0.12* (0.05)	0.20*** (0.05)	0.09
q0.75	2.43***	(0.21)	-0.02 (0.03)	0.08 (0.05)	0.13* (0.05)	0.06 (0.07)	-0.35*** (0.04)	-0.01 (0.05)	0.09 (0.05)	0.10
OLS	3.28***	(0.28)	-0.04 (0.05)	0.14** (0.06)	0.01 (0.05)	0.15* (0.08)	-0.54*** (0.05)	-0.01 (0.07)	0.18*** (0.06)	0.14
Learning experience with the chatbot										
q0.25	4.15***	(0.61)	-0.03 (0.10)	-0.01 (0.13)	-0.45*** (0.15)	-0.07 (0.19)	-1.02*** (0.12)	0.04 (0.15)	0.41** (0.15)	0.21
q0.50	3.12***	(0.33)	0.02 (0.05)	0.01 (0.07)	-0.26*** (0.08)	0.10 (0.10)	-0.75*** (0.07)	0.08 (0.08)	0.24 (0.08)	0.12
q0.75	2.72***	(0.23)	0.01 (0.04)	0.11* (0.05)	-0.15** (0.06)	0.03 (0.07)	-0.48*** (0.05)	0.05 (0.06)	0.11 (0.06)	0.07
OLS	4.18***	(0.34)	-0.01 (0.06)	0.03 (0.07)	-0.16** (0.08)	0.09 (0.11)	-0.77*** (0.07)	0.02 (0.08)	0.31*** (0.08)	0.24
Intention to continue the interaction										
q0.25	4.65***	(0.55)	-0.08 (0.09)	0.10 (0.12)	-0.51*** (0.13)	0.26 (0.17)	-1.09*** (0.11)	-0.05 (0.14)	0.43 (0.13)	0.25
q0.50	3.43***	(0.30)	-0.05 (0.05)	0.10 (0.07)	-0.57*** (0.07)	0.18 (0.10)	-0.87*** (0.06)	0.1 (0.08)	0.33*** (0.07)	0.18
q0.75	2.70***	(0.25)	0.04 (0.04)	-0.02 (0.06)	-0.42*** (0.06)	0.34*** (0.08)	-0.66*** (0.05)	0.07 (0.07)	0.16 (0.06)	0.13
OLS	4.21***	(0.33)	-0.01 (0.05)	0.10 (0.07)	-0.34*** (0.08)	0.29*** (0.10)	-0.91*** (0.06)	0.08 (0.08)	0.43*** (0.08)	0.29
Intention to recommend the chatbot to others										
q0.25	4.58***	(0.57)	0.03 (0.09)	0.02 (0.13)	-0.65*** (0.14)	0.45 (0.18)	-1.20*** (0.11)	0.05 (0.14)	0.53*** (0.14)	0.27
q0.50	3.82***	(0.29)	-0.01 (0.05)	0.04 (0.07)	-0.75*** (0.07)	0.22* (0.09)	-0.97*** (0.06)	0.09 (0.07)	0.31*** (0.07)	0.16
q0.75	2.53***	(0.29)	0.01 (0.05)	0.03 (0.07)	-0.35*** (0.07)	0.21* (0.09)	-0.61*** (0.06)	0.06 (0.07)	0.16* (0.07)	0.08
OLS	4.52***	(0.35)	0.04 (0.06)	0.10 (0.07)	-0.40*** (0.08)	0.25** (0.11)	-0.94*** (0.07)	0.04 (0.08)	0.38*** (0.08)	0.27

Note: The bottom row in each section of the table shows OLS regression as a reference. The number of observations: 1594. The table presents part of the full regression models. In the full regression models, we also controlled for other demographic variables including participants' age, income level, efficacy with using chatbots, as well as the language styles of each participant such as the word count of their average input, the use of positive and negative emotion words, use of analytical words, clout, and authentic expressions in conversations.

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 3. Knowledge Gains and Demographic/Attitude Variables for Climate Change and BLM

Topic	Intercept	Gender (Male)	Race/ethnicity minority	Education minority	Language minority	Opinion minority	Liberal	Conservative	R ²
Climate change	-0.25* (0.1)	0.02 (0.02)	0.04 (0.02)	0.07** (0.03)	0.07* (0.03)	0.2*** (0.02)	0.01 (0.03)	-0.03 (0.03)	0.06
BLM	-0.06 (0.11)	0.02 (0.02)	-0.01 (0.03)	-0.02 (0.03)	0.05 (0.04)	0.12*** (0.03)	0.04 (0.03)	-0.03 (0.03)	0.03

Note: The table presents part of the full regression models. In the full regression models, we also controlled for other demographic variables such as participants' gender, as well as the language styles of each participant such as their use of positive and negative emotion words in conversations.

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 4. Correlation between GPT-3's Conversational Styles and User Experiences (Climate Change Discussions)

User item	experience	Quantile	Intercept	Word count (log)	Positive emotion (log)	Negative emotion (log)	Analytic	Clout	Authentic	(pseudo)R2
Ratings of chatbot										
		Q _{0.25}	2.79*** (0.42)	0.12* (0.06)	0.18*** (0.06)	-0.13 (0.09)	0.01*** (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.11
		Q _{0.50}	3.21*** (0.29)	0.06 (0.05)	0.16*** (0.04)	-0.07 (0.06)	0.01* (0.01)	0.01*** (0.01)	-0.01 (0.01)	0.07
		Q _{0.75}	3.27*** (0.32)	0.1* (0.05)	0.08* (0.04)	-0.04 (0.07)	0.01 (0.01)	0.01* (0.01)	-0.01 (0.01)	0.05
		OLS	3.52*** (0.32)	0.15** (0.05)	0.11** (0.04)	-0.12 (0.07)	0.01* (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.11
Satisfaction with chatbot										
		Q _{0.25}	2.43*** (0.33)	0.12* (0.05)	0.14** (0.05)	-0.11 (0.07)	0.01*** (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.10
		Q _{0.50}	2.13*** (0.24)	0.05 (0.04)	0.06 (0.03)	-0.05 (0.05)	0.01** (0.01)	0.01* (0.01)	-0.01** (0.01)	0.10
		Q _{0.75}	2.04*** (0.26)	0.08* (0.04)	0.1** (0.04)	-0.02 (0.06)	0.01* (0.01)	0.01 (0.01)	-0.01* (0.01)	0.10
		OLS	2.59*** (0.28)	0.15*** (0.04)	0.12** (0.04)	-0.06 (0.06)	0.01*** (0.01)	0.01* (0.01)	-0.01*** (0.01)	0.11
Learning experience from chatbot										
		Q _{0.25}	3.13*** (0.56)	0.25** (0.08)	0.18 (0.07)	-0.54*** (0.12)	0.02*** (0.01)	0.01 (0.01)	-0.01*** (0.01)	0.17
		Q _{0.50}	2.8*** (0.26)	0.1** (0.04)	0.08* (0.04)	-0.16** (0.06)	0.01*** (0.01)	0.01*** (0.01)	-0.01*** (0.01)	0.11
		Q _{0.75}	2.49*** (0.25)	0.11** (0.04)	0.04 (0.04)	-0.1 (0.06)	0.01** (0.01)	0.01 (0.01)	-0.01 (0.01)	0.05
		OLS	3.26*** (0.36)	0.2*** (0.06)	0.1* (0.05)	-0.27*** (0.08)	0.01*** (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.17
Intention to continue the chat										
		Q _{0.25}	3.46*** (0.65)	0.06 (0.1)	0.18* (0.08)	-0.28* (0.14)	0.01* (0.01)	0.01** (0.01)	-0.01** (0.01)	0.16
		Q _{0.50}	2.9*** (0.33)	0.02 (0.05)	0.09* (0.05)	-0.1 (0.07)	0.01 (0.01)	0.01* (0.01)	-0.01* (0.01)	0.10
		Q _{0.75}	2.9*** (0.24)	-0.01 (0.04)	0.07* (0.03)	-0.08 (0.05)	0.01 (0.01)	0.01 (0.01)	-0.01** (0.01)	0.09
		OLS	3.87*** (0.37)	0.02 (0.06)	0.12* (0.05)	-0.2** (0.08)	0.01** (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.16
Intention to recommend the chatbot to others										
		Q _{0.25}	3.14*** (0.6)	0.04 (0.09)	0.24** (0.08)	-0.43*** (0.13)	0.02*** (0.01)	0.01** (0.01)	-0.01*** (0.01)	0.14
		Q _{0.50}	2.57*** (0.4)	0.03 (0.06)	0.09 (0.05)	-0.15 (0.08)	0.01* (0.01)	0.01** (0.01)	-0.01** (0.01)	0.09
		Q _{0.75}	2.45*** (0.25)	0.09* (0.04)	0.07* (0.04)	-0.11* (0.06)	0.01* (0.01)	0.01* (0.01)	-0.01** (0.01)	0.08
		OLS	3.1*** (0.39)	0.09 (0.06)	0.12* (0.05)	-0.23** (0.08)	0.01*** (0.01)	0.01*** (0.01)	-0.01*** (0.01)	0.15

Note: The table presents part of the full regression models. In the full regression models, we also controlled for participants' demographic variables including gender, age, income, efficacy with using chatbots, race and ethnicity, language skills, education level, and their opinions toward climate change.

*** p < 0.01, ** p < 0.05, * p < 0.1

Table 5. Correlation between GPT-3's Conversational Styles and User Experiences (BLM Discussions)

User experience item	Quantile	Intercept	Word count	Positive emotion	Negative emotion	Analytic	Clout	Authentic	(pseudo)R2
Ratings of chatbot									
	q _{0.25}	4.16*** (0.53)	0.23** (0.09)	0.16* (0.07)	-0.05 (0.12)	0.01 (0.01)	0.01 (0.01)	-0.01** (0.01)	0.13
	q _{0.50}	4.41*** (0.34)	0.11* (0.06)	0.09* (0.05)	-0.12 (0.08)	0.01 (0.01)	0.01 (0.01)	-0.01** (0.01)	0.08
	q _{0.75}	4.39*** (0.34)	0.06 (0.06)	0.08 (0.05)	-0.04 (0.08)	0.01 (0.01)	0.01 (0.01)	-0.01* (0.01)	0.05
	OLS	4.11*** (0.33)	0.19*** (0.05)	0.13** (0.05)	-0.1 (0.07)	0.01* (0.01)	0.01 (0.01)	-0.01** (0.01)	0.15
Satisfaction with chatbot									
	q _{0.25}	3.37*** (0.59)	0.1 (0.1)	0.08 (0.08)	-0.14 (0.13)	0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	0.11
	q _{0.50}	2.21*** (0.27)	0.06 (0.04)	0.05 (0.04)	-0.05 (0.06)	0.01* (0.01)	0.01 (0.01)	-0.01 (0.01)	0.09
	q _{0.75}	2.14*** (0.23)	0.05 (0.04)	0.06 (0.03)	0.03 (0.05)	0.01** (0.01)	0.01 (0.01)	-0.01 (0.01)	0.11
	OLS	2.88*** (0.29)	0.1* (0.05)	0.12** (0.04)	0.01 (0.07)	0.01*** (0.01)	0.01* (0.01)	-0.01 (0.01)	0.13
Learning experience from chatbot									
	q _{0.25}	3.46*** (0.63)	0.1 (0.1)	0.06 (0.09)	-0.1 (0.14)	0.02*** (0.01)	0.01 (0.01)	-0.01* (0.01)	0.20
	q _{0.50}	3.35*** (0.39)	0.07 (0.06)	0.03 (0.05)	-0.22 (0.09)	0.01*** (0.01)	0.01 (0.01)	-0.01 (0.01)	0.11
	q _{0.75}	2.71*** (0.2)	0.07* (0.04)	0.05 (0.03)	-0.03 (0.05)	0.01*** (0.01)	0.01 (0.01)	-0.01** (0.01)	0.07
	OLS	3.65*** (0.36)	0.13* (0.06)	0.08 (0.05)	-0.17 (0.08)	0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	0.22
Intention to continue the chat									
	q _{0.25}	4.2*** (0.77)	0.04 (0.12)	0.1 (0.1)	-0.25 (0.17)	0.01* (0.01)	0.01 (0.01)	-0.01 (0.01)	0.22
	q _{0.50}	3.43*** (0.33)	-0.06 (0.05)	0.09* (0.05)	-0.07 (0.07)	0.01** (0.01)	0.01 (0.01)	-0.01 (0.01)	0.16
	q _{0.75}	2.74*** (0.29)	-0.04 (0.05)	0.06 (0.04)	-0.06 (0.07)	0.01* (0.01)	-0.01 (0.01)	-0.01 (0.01)	0.12
	OLS	3.97*** (0.36)	-0.01 (0.06)	0.12* (0.05)	-0.11 (0.08)	0.01*** (0.01)	0.01 (0.01)	-0.01** (0.01)	0.24
Intention to recommend the chatbot to others									
	q _{0.25}	4.65*** (0.59)	0.09 (0.09)	0.13 (0.08)	-0.18 (0.13)	0.01*** (0.01)	-0.01 (0.01)	-0.01** (0.01)	0.24
	q _{0.50}	3.69*** (0.29)	0.01 (0.05)	0.11** (0.04)	-0.11 (0.07)	0.01*** (0.01)	-0.01 (0.01)	-0.01*** (0.01)	0.15
	q _{0.75}	2.46*** (0.25)	0.07 (0.04)	0.07* (0.04)	-0.06 (0.06)	0.01* (0.01)	-0.01 (0.01)	-0.01** (0.01)	0.09
	OLS	4.15*** (0.38)	0.08 (0.06)	0.13** (0.05)	-0.14 (0.08)	0.01*** (0.01)	0.01 (0.01)	-0.01** (0.01)	0.24

Note: The table presents part of the full regression models. In the full regression models, we also controlled for participants' demographic variables including gender, age, income, efficacy with using chatbots, race and ethnicity, language skills, education level, and their opinions toward BLM.

*** p < 0.01, ** p < 0.05, * p < 0.1

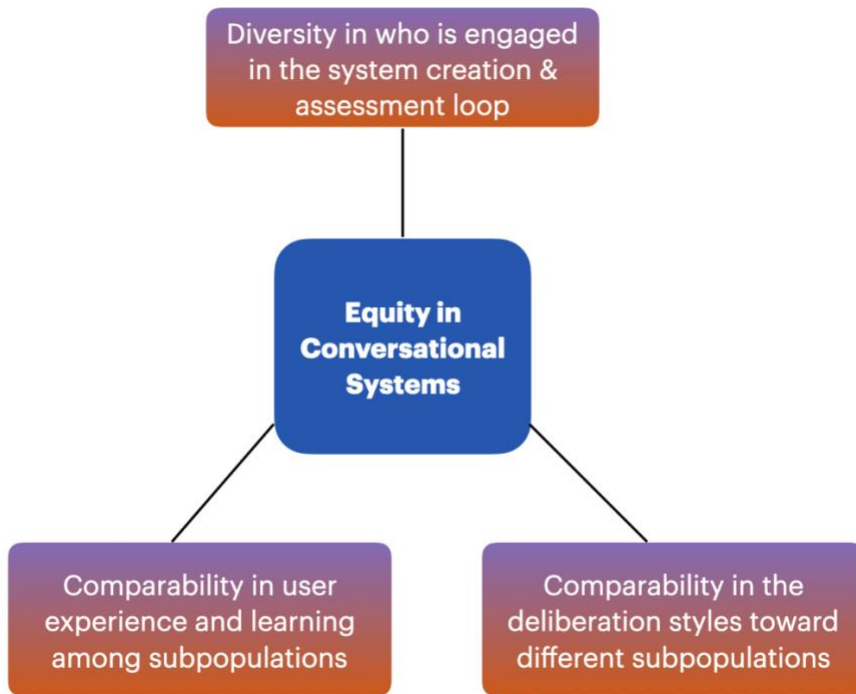
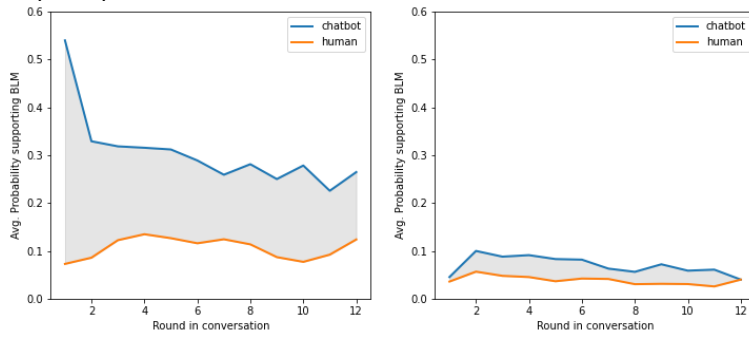
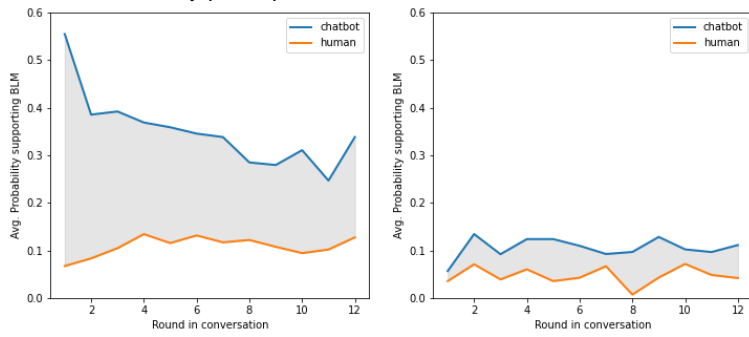


Figure 1. Toward equitable conversational AI: A framework for assessment

All participants



Education minority participants



Opinion minority participants

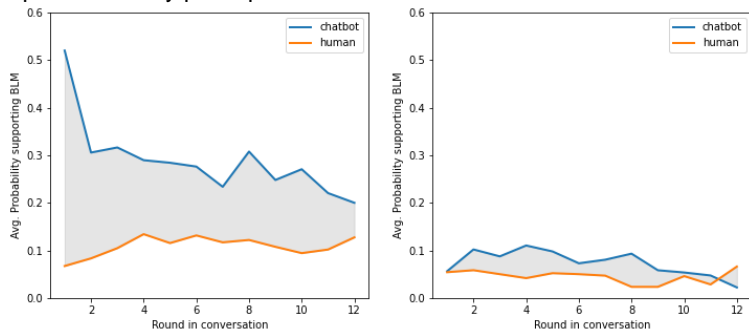


Figure 2. How stances evolved over conversation rounds for human participants and GPT-3

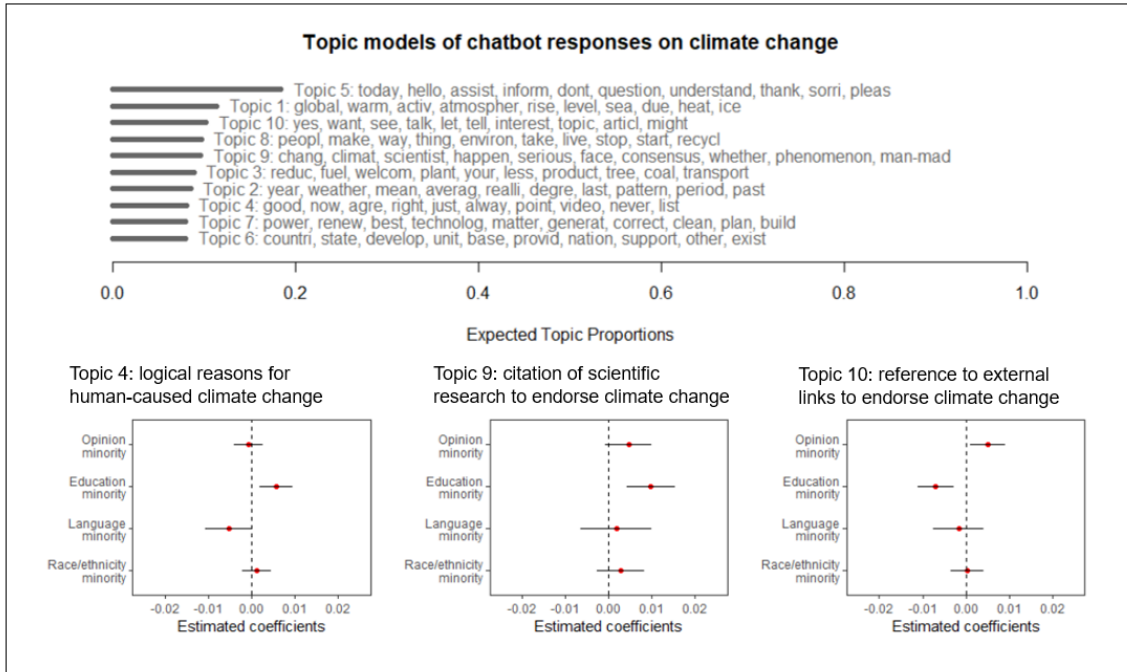


Figure 3. Effects of Participants' Demographics on GPT-3's Responses on the Climate Change Issue: STM analyses.

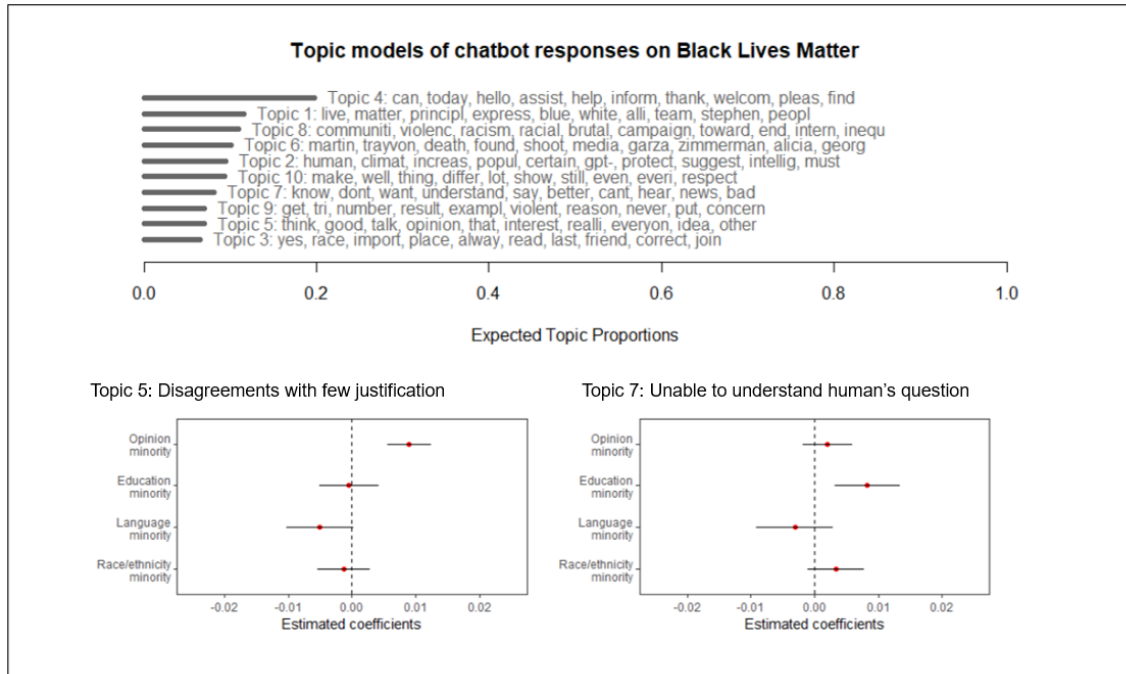


Figure 4. Effects of Participants' Demographics on GPT-3's Responses on the BLM Issue: STM analyses.

(a) Climate change dialogues

(b) Black Lives Matter dialogues

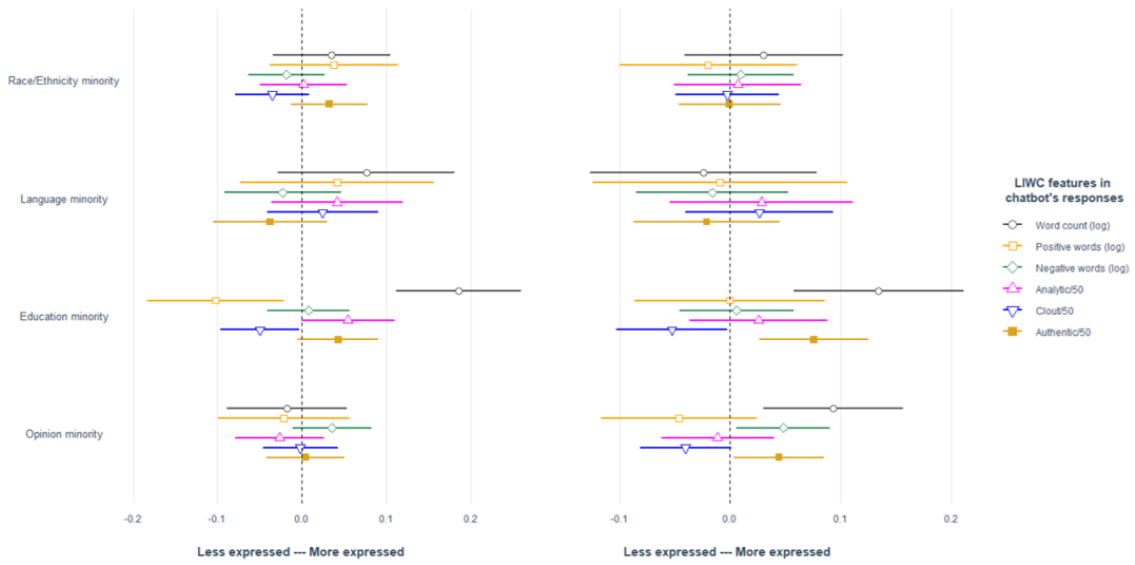


Figure 5. How GPT-3 Responds to Different Sub-populations: Analyses from LIWC

Note: We used $\log(x+1)$ to align the word count features (word count, positive word count, and negative word count) into a narrower distribution. While the Analytic, Clout, and Authentic scores are calculated by LIWC application into a normal distribution ranging from 0 to 100, we divided these scores by 50 to make the current visualization coherent in scales.