Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03609-x OPEN

Performance and biases of Large Language Models in public opinion simulation

Yao Qu[1] & Jue Wang[1]✉

The rise of Large Language Models (LLMs) like ChatGPT marks a pivotal advancement in artificial intelligence, reshaping the landscape of data analysis and processing. By simulating public opinion, ChatGPT shows promise in facilitating public policy development. However, challenges persist regarding its worldwide applicability and bias across demographics and themes. Our research employs socio-demographic data from the World Values Survey to evaluate ChatGPT's performance in diverse contexts. Findings indicate significant performance disparities, especially when comparing countries. Models perform better in Western, English-speaking, and developed nations, notably the United States, in comparison to others. Disparities also manifest across demographic groups, showing biases related to gender, ethnicity, age, education, and social class. The study further uncovers thematic biases in political and environmental simulations. These results highlight the need to enhance LLMs' representativeness and address biases, ensuring their equitable and effective integration into public opinion research alongside conventional methodologies.

[1] School of Social Sciences, Nanyang Technological University, Singapore, Singapore. ✉email: wangjue@ntu.edu.sg

Introduction

Public opinion is crucial in shaping policy decisions, particularly in democratic societies, where it reflects the electorate's preferences, concerns, and priorities (Burstein, 2003). This feedback loop enables policymakers to remain attuned to their constituents' needs, fostering accountability and responsive governance (Hutchings, 2005). While traditional public opinion collection methods like surveys and interviews provide valuable insights, they are plagued by issues like low response rates, potential biases, and challenges in achieving representativeness. For instance, lengthy surveys in particular risk diminishing respondent engagement due to their extensive nature (Dillion et al., 2023). Fortunately, the recent advances in artificial intelligence (AI), especially large language models (LLMs) like ChatGPT, offer a novel approach to complementing traditional methods in public opinion collection as they are capable of swiftly responding to a multitude of questions (Lee et al., 2023). This efficiency, combined with the ability to process and analyze extensive text data, empowers LLMs to uncover insights into public sentiment often overlooked by conventional methods (Ray, 2023).

The role of generative LLMs in social science is increasingly recognized for its multifaceted applications. As noted by Korinek (2023), these models are instrumental in various tasks within psychological science, including editing academic papers and facilitating literature reviews. In the educational domain, Cowen and Tabarrok (2023) demonstrate how LLMs can simulate expert responses or create specific personas to deepen understanding of complex subjects like economics.

Recent research underscores the potential of LLMs in public opinion analysis. For instance, Argyle et al. (2023) demonstrated ChatGPT's ability to accurately reflect responses across various human subgroups, particularly in the context of presidential election behaviors. A notable correlation was observed between human responses and those generated by LLMs, referred to as 'silicon samples'. Similarly, Lee et al. (2023) found that LLMs can predict public opinions on global warming. However, Lee et al. (2023) emphasized the need for LLMs to incorporate a broader range of variables, including psychological factors, for a more precise simulation of opinions on complex issues like global warming. Additionally, Aher et al. (2023) and Horton (2023) explored LLMs' capacity to emulate specific personas, showing ChatGPT's proficiency in replicating human subject experiments with detailed demographic analysis. Complementary to this, studies by Brand et al. (2023) and Park et al. (2023) highlighted ChatGPT's skill in simulating consumer behavior and human actions in various scenarios. These studies collectively highlight the sophisticated simulation capabilities of LLMs like ChatGPT, marking their significant role and expanding influence in public opinion research.

While the use of LLMs like ChatGPT in social science shows promise, three significant challenges necessitate further investigation. C1) *The global applicability and reliability of LLMs*. The prevalent use of U.S. surveys in existing studies (Argyle et al., 2023; Lee et al., 2023) reflects the English-centric training data of ChatGPT. It leaves us with uncertainty regarding the model's effectiveness in navigating and accurately reflecting public opinion across diverse cultural, linguistic, and economic contexts. This gap in understanding poses a critical challenge in assessing the applicability and reliability of LLMs, such as ChatGPT, in public opinion analysis on a global scale. C2) *Demographic biases within LLMs*. Biases related to gender, race, education, age, and income, inherent in LLMs due to training on internet-based content, may not sufficiently represent diverse perspectives. For instance, Martin (2023) suggested a tendency in ChatGPT's responses to favor liberal and privileged viewpoints. Therefore, identifying and addressing specific areas of unfair representation, particularly in terms of socio-economic diversity, merits further research to ensure equitable AI development. C3) *Complexity and Choice Variability in LLM Simulations*. A notable research gap exists in assessing LLMs, like ChatGPT, for their capability to replicate complex decision-making across various topics. This gap encompasses a limited insight into the models' adaptability to distinct decision dynamics, such as environmental versus political issues, and the influence of increased choice complexity on simulation accuracy. Closing this gap is essential for evaluating the boundaries and efficacy of LLMs in diverse and complex societal contexts.

The study aims to tackle these challenges through a three-fold approach. Firstly, we explore the impact of cultural, linguistic, and economic development differences in AI simulation accuracy (for C1). This objective directly addresses the gap related to the predominance of English and U.S.-centric data in AI models. The study assesses how these biases influence public opinion representation in diverse contexts and their subsequent effects on policy decisions across different countries. Building upon this foundation, the second aim is to analyze the implications of demographic biases within AI simulations (for C2). This aim focuses on understanding how demographic biases in AI affect the inclusivity and representativeness of public policies, ensuring that diverse demographic perspectives are accurately reflected. Finally, we assess AI simulation accuracy in diverse issues and explore ideological and choice complexity biases in policy implications (for C3). This involves a focused examination of three aspects: the variation in AI simulation accuracy between topics like environmental and political issues, the influence of ideological biases on policy-related simulations, and the effect of choice complexity on simulation fidelity. These explorations are essential for guaranteeing that AI-driven policies are founded on a realistic, unbiased, and comprehensive grasp of complex societal issues.

The contributions of this paper are summarized as follows: The theoretical significance of this research lies in its potential to enrich public opinion theories by examining the parallels and discrepancies between human biases in opinion formation and AI biases in opinion simulation. This provides insights into AI's role and potential impact on public policy. On an empirical level, the study aims to empirically analyze biases related to culture, language, economy, demographics, and themes in AI-simulated public opinions. It seeks to highlight the complexities and challenges AI tools encounter in accurately representing diverse viewpoints.

Recognizing the challenges of adding value ethically to AI, especially in capturing the diversity and complexity of global public opinions, this study's outcomes inform the creation of more sophisticated AI applications in public policy. It underscores the need to develop policies informed by a balanced and inclusive representation of public opinions, essential for efficient governance in areas like environmental protection, economic development, and political processes.

Materials and methods

Tool: ChatGPT. Advancements in AI and natural language processing (NLP) have led to the development of LLMs, which are reshaping the landscape of content creation and text generation (Mathew, 2023). ChatGPT, a prominent example of these models developed by OpenAI, stands at the forefront of this transformation. Built on the Generative Pre-trained Transformer (GPT) architecture, ChatGPT excels at mirroring human-like language capabilities (Chan, 2023). It leverages vast datasets to generate contextually appropriate responses, showing the power

of LLMs in understanding and generating nuanced text (Ray, 2023). Inspired by the method of Argyle et al. (2023), we utilize ChatGPT to generate 'Silicon Sample Data' to assess the correspondence between simulated responses and real survey results across different research settings.

Survey data source. The World Values Survey (WVS), initiated in 1981, surveys socio-cultural, political, and moral values globally, covering nearly 100 countries and representing about 90% of the global population (Inglehart et al., 2014). The WVS's standard questionnaire ensures data consistency across diverse linguistic, economic, and cultural regions, making it valuable for comparative analyses like ours. This uniformity is crucial in our study to attribute any variation in AI-simulated responses to the AI's interpretation instead of differences in question phrasing. Besides, the WVS questionnaire covers a broad spectrum of topics, including economic, political, religious, and social values, making it useful for various research fields. It enables comparisons between responses on potentially biased topics like environmental issues and political questions, assessing AI simulation biases across different themes. Moreover, with interviews with nearly 400,000 respondents, the WVS is one of the largest studies of its kind (Inglehart et al., 2014). It provides detailed demographic data for each respondent, which is important for examining demographic representation biases in AI simulations and how well AI models mirror public opinion across diverse subgroups. In this study, we use data from WVS Wave Six (2010–2014). The time of the survey varied by country; it was conducted in Japan in 2010, the United States in 2011, Sweden in 2011, Singapore in 2012, South Africa in 2013, and Brazil in 2014.

Simulation input parameters

Target variables. The first target variable, V81, assesses prioritization between the economy and the environment. It asks respondents to choose among statements: 1. Emphasis on protecting the environment, 2. Emphasis on economic growth, 3. No answer to environmental versus economic priorities. This variable is primarily used in the first two studies focusing on country comparisons and demographic biases. The survey questions for this and the below variables are available in Table S3 in the Supplementary Materials.

The second target variable is political election voting behavior, measured by question V228: "If there were a national election tomorrow, for which party on this list would you vote?" Respondents can choose from major political parties in their country, along with options like uncertainty or not voting. For example, in the United States, options include 1. Democrat, 2. Republican, 3. Other party, and 4. No answer/Don't know/I would not vote. This variable is introduced in the third study, where both environmental and political questions are utilized for thematic comparison.

Demographic Variables. Key demographic variables include ethnicity (V254), sex (V240), age (V242), education level (V248), and social class (V238). Ethnicity options are country-specific and reflect major ethnic groups for respective countries. Sex is coded as 1 for males and 2 for females. Age is a continuous variable. Education levels range from no formal education to a university degree. Social class is self-identified, with options like upper class, middle class, or lower class.

Covariates. For the environmental issue, we choose covariates that are frequently included in environmental surveys and have precedent in prior research (Lee et al., 2023), including:

- Membership in Environmental Organizations (V30): This assesses active (2), inactive (1), or non-membership (0) in various organizations, including environmental ones.
- Environmental Consciousness (V78): This measures respondents' identification with the statement "Looking after the environment is important to this person; to care for nature and save life resources." Responses range from 1 (very much like me) to 6 (not at all like me).
- Financial Support for Ecological Organizations (V82): This variable inquiries about donations to ecological organizations in the past two years, coded as 1 (yes) and 2 (no).
- Participation in Environmental Demonstrations (V83): This variable assesses involvement in environmental demonstrations in the past two years, with responses coded as 1 (yes) and 2 (no).
- Confidence in Environmental Organizations (V122): This measures confidence levels in environmental organizations, ranging from a great deal of confidence (1) to none at all (4).

As mentioned in the limitation, there are few covariates associated with the political question. We only identified one covariate, which is Political Ideology (V95): In political matters, people talk of 'the left' and 'the right.', if 1 means extremely left and 10 means extremely right, where would you place your views?

Simulation process

Model and API setting. Our study employs the GPT-3.5 Turbo model, due to GPT-3.5's superior efficiency in processing large data volumes and faster response capabilities, essential for our extensive simulation research. Moreover, despite the commonly held view that human morality is a challenging aspect for language models to grasp, Russell (2019) and Dillion et al. (2023) discovered a notable alignment between GPT-3.5's responses and human moral judgments. This congruence of GPT-3.5 can help enhance the accuracy and relevance of our simulations in replicating complex human ethical considerations. Note that we acknowledge that our findings are specific to the version of the language model used and do not necessarily reflect the capabilities or biases of all LLMs.

The impact of temperature settings on language model outputs varies depending on the task. As noted by Boelaert et al. (2024), in scenarios where responses are limited to predetermined options, such as in our experiments, temperature variations have minimal effect on outcomes. This contrasts with full-answer generation tasks, where temperature can influence next-token probabilities. Despite the limited impact in our case, we follow the recommendations of Guilherme and Vincenzi (2023) and Davis et al. (2024), who suggest that lower temperatures produce more consistent outputs. Consequently, we set the OpenAI API's temperature to 0.2 for our survey simulation.

Prompt design. We adopt an interview-styled format for generating AI responses that simulate human participants. The process begins with converting raw survey data, including demographic information and other covariates, into a format understandable by the AI model. We assign specific codes to each demographic attribute and then translate these codes into descriptive sentences. For example, 'V240-1' translates to "You are male." These sentences form a comprehensive demographic profile for each respondent, starting with "Please assume that you are…" Regarding the target question, our approach differs across studies. Initially, we focus exclusively on the environmental protection versus economic growth question for country comparison and demographic biases. For the third study, both the environmental question and the political election voting decision question are employed for thematic comparison.

We then integrate the demographic profile and the target question into a single prompt, guiding the AI to respond as a

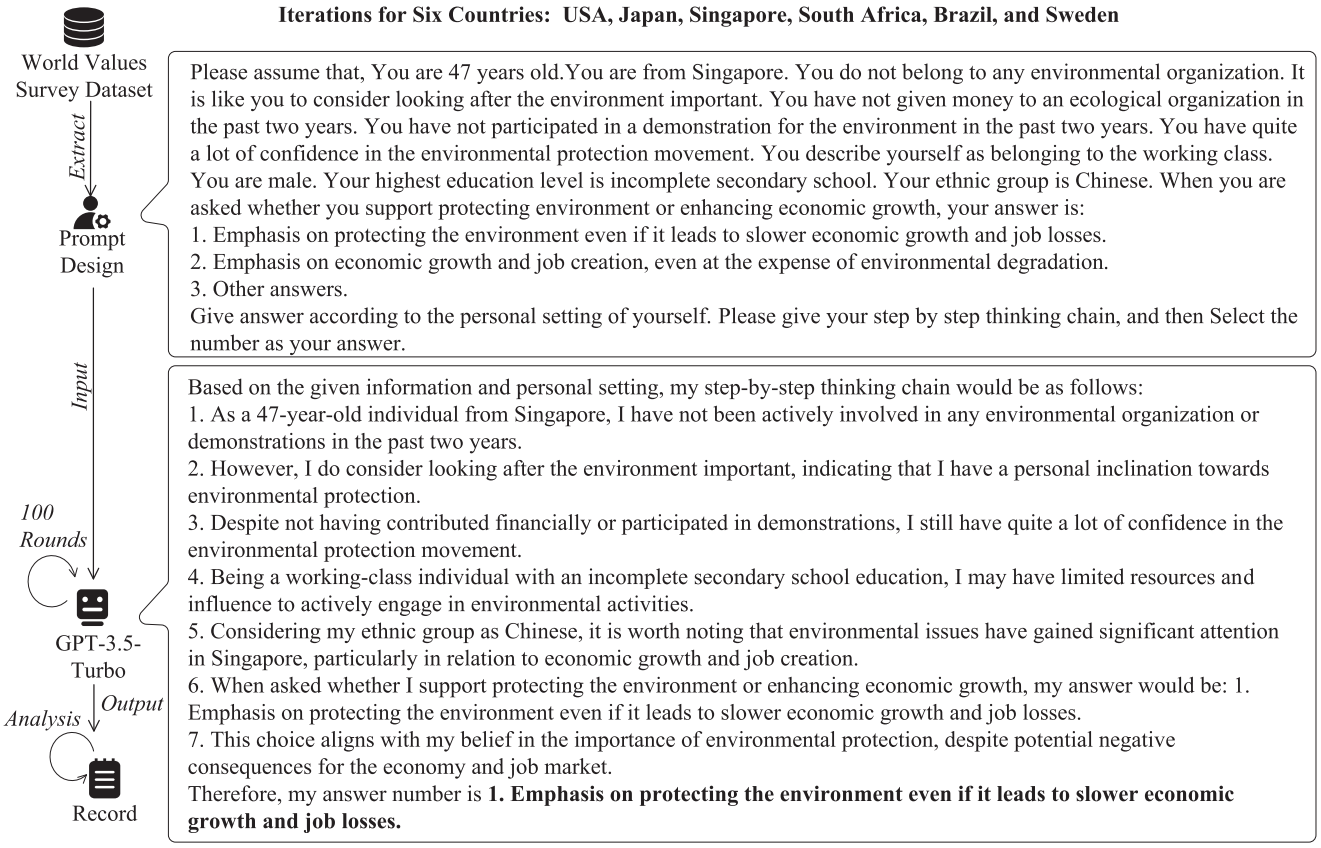

Fig. 1 ChatGPT simulation process flowchart. The top section outlines the procedure for generating ChatGPT-simulated responses using socio-demographic prompts derived from the World Values Survey Dataset for six countries. The bottom section displays the step-by-step rationale of ChatGPT to give an answer based on the demographic information.

person with specific demographic characteristics. For example, "Assuming you are a 30-year-old female with a university degree and middleclass status, when asked whether you support protecting the environment or enhancing economic growth, what is your choice: (1) emphasis on protecting the environment, (2) emphasis on economic growth, or (3) neither/other?"

To improve the authenticity of our simulations, we used prompts in the native languages of non-English speaking countries—Sweden, Brazil, and Japan—drawing directly from the questionnaires in local languages available in the WVS database and enabled the ChatGPT to respond in the language of the query. This method preserves the original context and meaning, enhancing the accuracy of our cross-linguistic analysis of ChatGPT performance. For other countries in our study, where English is the primary language and the questionnaires were administered in English, we continued to use English prompts. Moreover, for each sample, we conducted 100 simulations considering the variability inherent in the model's responses.

To validate our simulation, we instruct the AI to provide a reasoning chain before its final answer, ensuring responses mimic human-like thought processes. Additionally, we direct the AI to forego politically correct answers, favoring responses based on an assumed personal setting. The AI-simulated response, typically a chosen numerical option, is then extracted and recorded. Figure 1 shows the process from raw survey data conversion to AI-generated responses, as well as the reasoning chain of ChatGPT before giving an answer.

Comparative design. The literature on bias in AI systems reveals varied detection methods. Delobelle et al. (2021) questioned the generality of using fixed templates and specific seeds, while Caliskan et al. (2017) emphasized the role of training data in introducing biases into AI. Akyürek et al. (2022) noted bias metrics' inconsistency, potentially leading to contradictory findings. Liu et al. (2022) discussed the operational difficulties in developing bias classifiers and the often-limited access to a model's word embeddings that are essential for thorough bias assessment.

In the context of AI systems, particularly language models like ChatGPT, algorithmic fidelity would imply the model's ability to reflect the diversity of human opinions, cultural nuances, and socio-cultural dynamics in its responses or outputs (Argyle et al., 2023; Lee et al., 2023). For instance, if a language model is used to simulate public opinion, high algorithmic fidelity would mean that the opinions generated by the model closely align with the actual distribution of opinions across different populations. The concept is crucial in evaluating the effectiveness and reliability of AI systems in applications where reflecting human-like understanding and behaviors is important.

In line with the theoretical framework of algorithmic fidelity, we posit that an unbiased AI should accurately reflect the wide range of opinions represented in the WVS, showing the diversity and proportionality inherent in a global, multicultural sample. Consequently, our operational definition of bias is centered around the extent of deviation in the AI's depiction of public opinion from the empirically observed distribution of responses within the WVS. To assess this, we utilize agreement not as a direct bias metric, but as a tool to assess the degree of alignment between ChatGPT's responses and the actual WVS outcomes.

Thus, the detection of biases stems from a comparative analysis that scrutinizes agreement scores across various countries, demographic segments, and thematic areas. Through

Table 1 Country comparison by culture, economy, and language.

Country	Cultural Background	Economic Status	Dominant Language(s)
USA	Western	Developed	English
Japan	East Asian	Developed	Japanese
Singapore	Southeast Asian	Developed	English
South Africa	African	Developing	English, Afrikaans
Brazil	Latin American	Developing	Portuguese
Sweden	Nordic	Developed	Swedish

This table categorizes six countries by their cultural background, economic status, and dominant languages to contextualize the dataset used for ChatGPT's simulation analysis.

examination of the variations in agreement scores among these groups, we identify which simulations most accurately mirror the surveyed populations and which may display signs of bias. Higher agreement levels in certain groups, as opposed to others, suggest a lower propensity for bias in the model's representations of those particular groups' opinions.

Cultural, linguistic, and economic bias evaluation. Cultural, linguistic, and economic biases in AI models like ChatGPT, primarily stem from their internet-based training data, which is heavily skewed towards specific cultures, languages, and economic perspectives (Ray, 2023). The strategic selection of Japan, Singapore, the U.S., South Africa, Sweden, and Brazil for this study, as detailed in Table 1, aims to encompass a broad spectrum of cultural, economic, and linguistic contexts. This facilitates a thorough analysis of ChatGPT's performance and biases across varied global settings.

Demographic bias assessment. The study investigates the presence of gender, racial, age, educational, and income biases in AI models such as ChatGPT, likely originating from biases in the training data (Ray, 2023). We assess these biases through simulated interactions with ChatGPT among varied demographic groups within the United States, specifically analysing responses to environmental issues.

Complexity and choice variability. We continue to address the potential ideological bias in AI models like ChatGPT (Ray, 2023). This entails examining three key aspects: the difference in AI simulation accuracy across topics such as environmental and political issues, the presence of ideological biases for different topics, and how choice complexity affects simulation fidelity.

Data analysis. To measure the correspondence between the simulated responses and the real survey results, our analysis primarily employs Cohen's Kappa, a robust measure adjusting for chance agreement, thus providing a more accurate assessment of ChatGPT's responses compared to actual survey results. A Kappa value of 1 indicates perfect agreement, while a value of 0 indicates no agreement beyond what is expected by chance. Negative values indicate less agreement than expected by chance.

In support of Cohen's Kappa, we also utilize Cramer's V, which measures the strength of association between two nominal variables independent of table size, offering values from 0 (no association) to 1 (perfect association). This method complements Kappa by assessing the overall correspondence between variables.

Finally, we assess the Proportion Agreement, a fundamental measure that determines the percentage of instances where two evaluators provide identical classifications. While this method yields a straightforward calculation of agreement, it lacks the capacity to account for coincidental concurrence. Consequently, a high rate of agreement does not invariably mean a substantive association, as it might merely reflect chance alignment. This limitation renders the Proportion Agreement a supplementary tool rather than a focal point of our analysis, particularly in comparison to Cohen's Kappa and Cramer's V.

Together, these statistical methods provide a thorough analytical framework. Our focus, however, is on Cohen's Kappa for its robust adjustment for chance, a vital factor in analyzing AI response patterns. We conducted 100 simulations per respondent to calculate agreement and used the mean of these calculations as the overall agreement level for each prompt. This method reduced the variability in the model's responses, yielding a more reliable consensus estimate.

Results

This research has provided insights into the capabilities and limitations of LLMs like ChatGPT for simulating public opinion across various cultural, economic, linguistic, demographic, and thematic contexts. Our findings highlight that while LLMs show promise in replicating public opinions, particularly in contexts like the United States where the model's training data is more robust, there are notable limitations in its global applicability and reliability. Moreover, our analysis within the United States uncovered unfair representation of specific demographic groups. This disparity suggests that current LLMs, including ChatGPT, may inherently possess biases influenced by the demographic representation in their training data. The underrepresentation or misrepresentation of certain groups, especially marginalized communities, raises concerns about the equitable use of LLMs in public opinion research. Last, the study reveals that ChatGPT favors liberal choices more in political than environmental simulations, that its simulation accuracy is higher for political behaviors than complex environmental decisions, and that increased choice complexity reduces the model's simulation accuracy. These findings highlight the importance of addressing inherent biases and the incorporation of more diversified training materials in AI models for reliable application across various topics and countries.

Comparative study across countries. Figure 2 presents the distribution of Cohen's kappa values across each country, derived from 100 simulation iterations. The mean value of these results is calculated and reported. Figure 3 illustrates the differences in ChatGPT's ability to simulate survey responses across countries based on Cohen's Kappa score. A higher score shows a higher level of agreement in simulation. The results on the other two measures – Cramer's V and Proportion Agreement – are available in Table S1 in the Supplementary Materials.

The United States displays a moderate Cohen's Kappa score of 0.239, indicating a reasonably good simulation of survey responses. On the other hand, Japan and South Africa's low Cohen's Kappa values of 0.024 and 0.006, respectively, highlight significant limitations in the model's accuracy within these contexts. The inconsistency suggests that the simulation's current assumptions—such as the uniform influence of cultural, economic, and social factors across different countries—may be flawed, indicating these elements are not sufficiently integrated or weighted in the model.

To better understand the correlation, we have transformed key aspects—culture, economy, and language—into binary variables using a 0 and 1 scheme, as detailed in Table 2. Cultural background is coded as "Western" (1) or "Not Western" (0), economic status as "Developed" (1) or "Developing" (0), and dominant language as "English" (1) or "Not English" (0). Each of

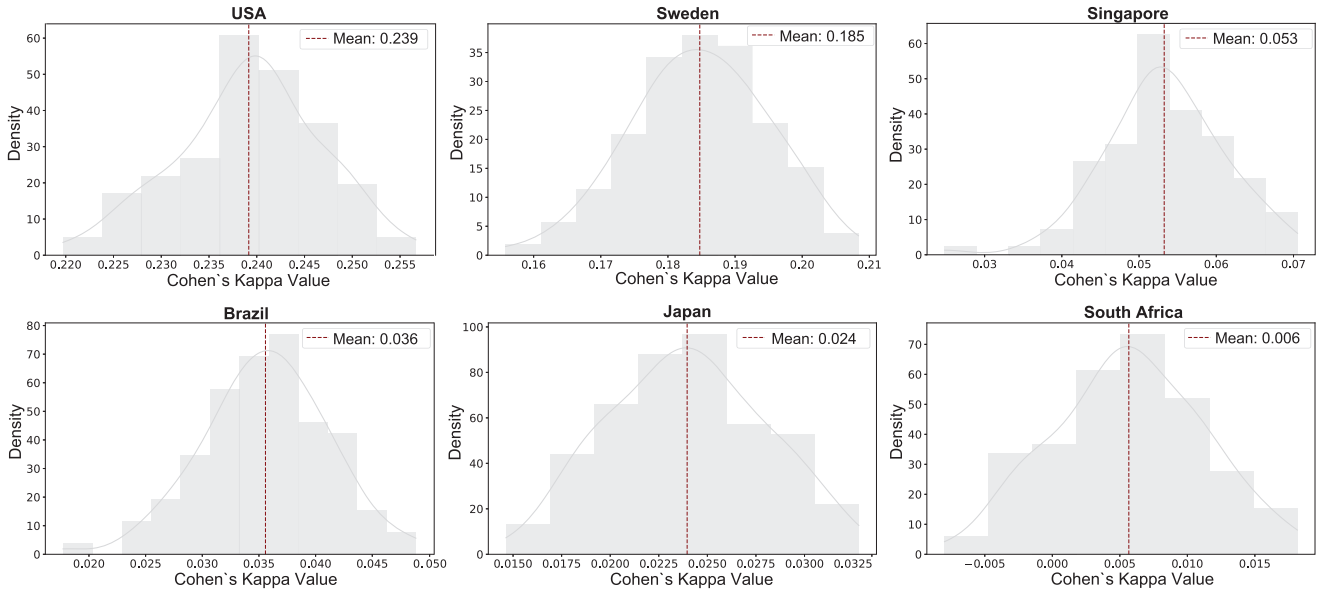

Fig. 2 Distribution of Cohen's Kappa for 100 simulations with mean highlighted. This figure illustrates the variability and central tendency in Cohen's Kappa statistics through 100 simulations for six different countries: USA, Sweden, Singapore, Brazil, Japan, and South Africa. The density plots demonstrate the distribution of the kappa values, while the dashed vertical lines indicate the mean kappa value for each country, providing a reference for the central location of the data within each simulation set.

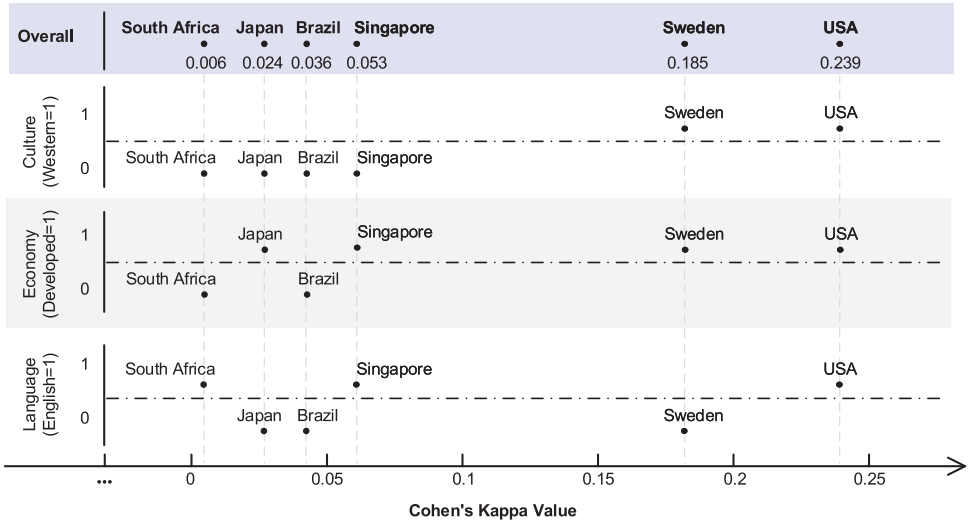

Fig. 3 ChatGPT simulation accuracy by country. The horizontal axis quantifies Cohen's Kappa values, and the vertical axis segregates countries into different categories based on culture, economic development, and primary language.

Table 2 Country comparison: cultural, economic, and linguistic aspects.

Country	Culture (Western = 1)	Economy (Developed = 1)	Language (English = 1)	Combined Code
USA	1	1	1	(1, 1, 1)
Japan	0	1	0	(0, 1, 0)
Singapore	0	1	1	(0, 1, 1)
South Africa	0	0	1	(0, 0, 1)
Brazil	0	0	0	(0, 0, 0)
Sweden	1	1	0	(1, 1, 0)

Each country is coded with binary indicators for being Western (1 for Western), Developed (1 for Developed economies), and English-speaking (1 for English as a primary language).

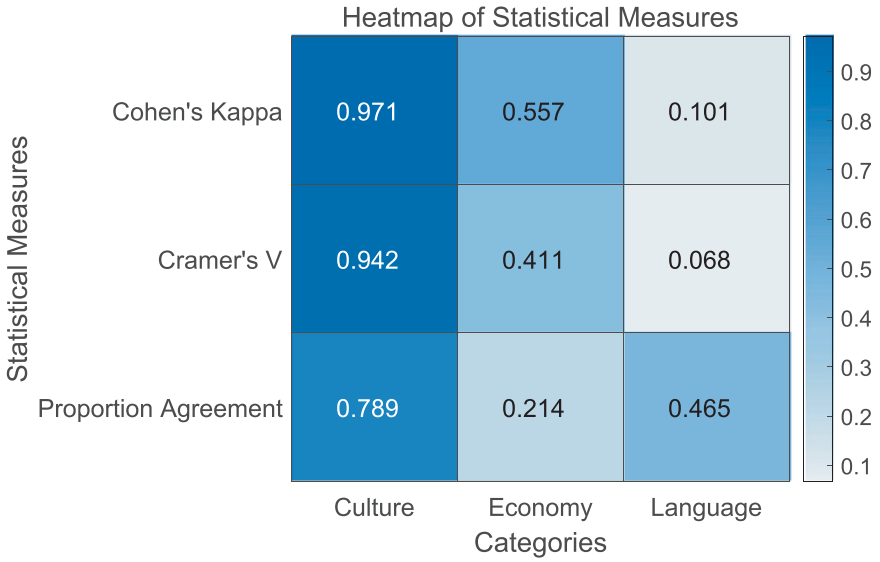

Fig. 4 Simulation correlation heatmap with cultural, economic, and linguistic factors. In the heatmap, dark blue indicates a strong positive correlation, whereas lighter blue suggests a weaker positive correlation.

these six countries uniquely represents a code formed by combining the three categories.

We use Pearson correlation coefficients to identify linear relationships between various factors and ChatGPT's simulated results. These coefficients, which vary between −1 and 1, elucidate both the strength and type of these relationships. Coefficients near 1 or −1 denote strong positive or negative correlations, respectively, whereas a coefficient around 0 indicates a lack of significant correlation.

Figure 4 demonstrates the correlations between different simulation result metrics and the binary categories of cultural background, dominant language, and economic status. In the heatmap, dark blue indicates a strong positive correlation, whereas lighter blue suggests a weaker positive correlation. The heatmap analysis underscores the substantial influence of culture on ChatGPT's simulation accuracy, with a high Cohen's Kappa correlation of 0.971, indicating a strong predictive relationship. Complementary to this, the correlation with Cramer's V and Proportion Agreement is also notable, recorded at 0.942 and 0.789, respectively, reinforcing culture's pivotal role. In contrast, economic factors reveal a moderate correlation through a Cohen's Kappa value of 0.557, suggesting its influence is considerable but not as pronounced. Furthermore, language demonstrates its impact with a Cohen's Kappa correlation of 0.101, confirming its relevance, albeit to a lesser extent than cultural and economic factors. These correlations highlight the significance of integrating diverse socio-cultural and economic considerations for enhancing the fidelity of ChatGPT simulations in reflecting public opinion.

Demographic representation in the United States. Since the previous result shows that ChatGPT's effectiveness in simulating survey responses is most prominent in the United States, we further explore the demographic subpopulation representation within this country using the environmental issue survey question. Here, we only highlight Cohen's Kappa results using Fig. 5 since our analysis across Cohen's Kappa, Cramer's V, and Proportion Agreement demonstrated a consistent pattern. The corresponding results from the other two measures are available in Table S2 in the Supplementary Materials.

Figure 5 reveals distinct patterns in the alignment between ChatGPT's simulated and actual responses across U.S. demographics regarding the priority between economy and environment. Males show slightly higher agreement and association than females. Among ethnic groups, white and other ethnicities exhibit more robust correspondence. Older age groups demonstrate notably stronger alignment, indicating age-related variability. In terms of social class, the upper and middle classes align more closely with the simulations. Additionally, the group with university education displays a higher fidelity in response alignment, suggesting a correlation between higher education and response predictability.

These trends are in line with Dillion et al. (2023), who observed that GPT models tend to mirror the viewpoints of individuals with higher incomes and education. Furthermore, Ray's (2023) review corroborates our findings regarding gender and ethnic biases. However, our study diverges when it comes to age representation; we find that older age groups align more with ChatGPT's Turbo-3.5. In contrast, Santurkar et al. (2023) noted that the 65+ demographic is poorly represented by current language models. This discrepancy might be attributed to the specific models used, as each may possess unique biases (Dillion et al., 2023), influencing the representation of different age groups.

Moreover, while the previous studies (Dillion et al., 2023; Ray, 2023) primarily examined political issues, our research extends into environmental issues. It not only reaffirms the existence of these demographic biases but also suggests their pervasiveness across different spheres. This implies a more widespread issue of unfair representation of different demographic subpopulations in AI models, warranting careful consideration and action.

Comparative analysis of topic-related results

Accuracy of political vs. environmental issue simulations. We compare ChatGPT's simulation accuracy on two different issues within the United States: environmental protection versus economic development, and political voting decisions. Research by Lee et al. (2023) successfully forecasts political outcomes using demographic data solely, suggesting a simpler decision-making process compared to environmental issues, which appear less predictable from demographics alone. We assess this by comparing political decisions and environmental decisions, both with and without additional covariates. Our comparative analysis between political and environmental decisions involved distinct sets of covariates: five for environmental decisions, including

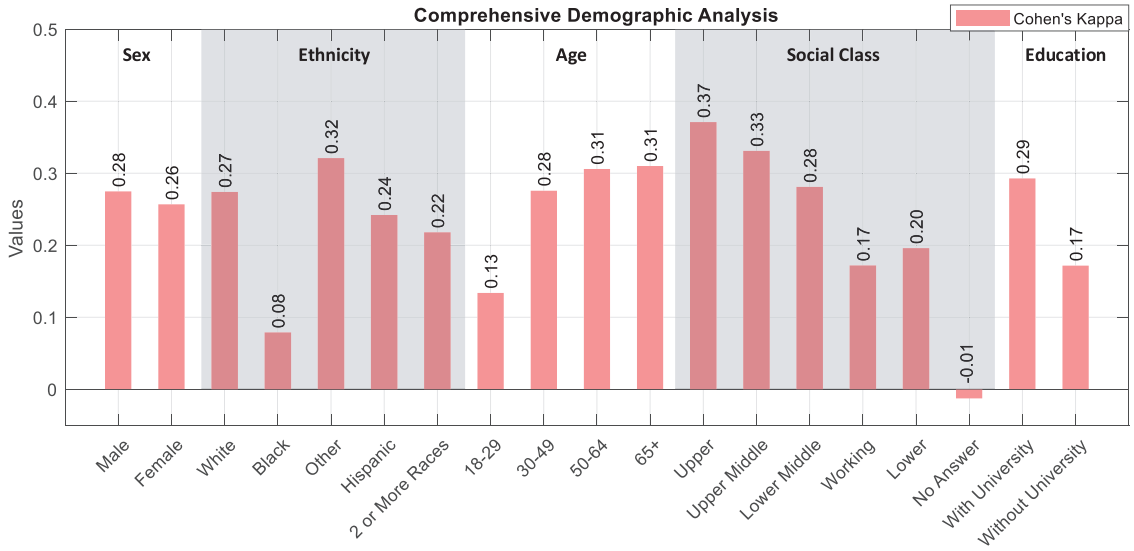

Fig. 5 Comprehensive demographic analysis. The bar chart quantifies the demographic representation of ChatGPT's simulation accuracy using Cohen's Kappa. Each bar reflects the level of agreement between ChatGPT's responses and human judgment across different demographic sectors: sex, ethnicity, age, social class, and education.

Table 3 Environmental vs political issue simulations.

Simulation	Covariates	Cohen's Kappa	Cramer's V	Proportion Agreement
Environmental Issue	With Multiple Covariates	0.270	0.274	66.83%
	Without Covariates	0.000	.	38.07%
Political Issue	With One Covariate	0.306	0.324	65.92%
	Without Covariate	0.145	0.263	54.71%

This table compares the accuracy of ChatGPT's simulations for environmental and political issues. The results indicate marginally higher accuracy for political issues.

Membership in Environmental Organizations, Environmental Consciousness, Financial Support for Ecological Organizations, Participation in Environmental Demonstrations, and Confidence in Environmental Organizations; and one for political decisions, Political Ideology. To enable a direct comparison amidst the diverse response options of the original survey for V228 and V81, our study focused exclusively on respondents who voted for either "Democrats" or "Republicans" in the political voting question (Lee et al., 2023). This binary categorization was also applied to the ChatGPT simulations. Similarly, for environmental issues, we confined our analysis to participants who voted with a preference for either economic or environmental priorities, ensuring a uniform framework of binary options to isolate the model's performance from response complexity.

Table 3 indicates that political simulations demonstrate higher accuracy compared to environmental simulations, both with covariates and without covariates. This inherent predictability of political behavior is supported by Lee et al. (2023), who found demographics alone to be a strong predictor. In the comparison involving covariates, political simulations are modeled using just a single covariate, whereas environmental simulations incorporate multiple covariates but still fail to achieve comparable accuracy. This discrepancy suggests additional complexities in modeling environmental decision-making. Therefore, our study reinforces the notion that simulating environmental decision-making is inherently more challenging than predicting political behavior.

Ideological bias in ChatGPT simulations across different topics. We investigated ChatGPT's potential bias towards liberal ideologies in simulations related to environmental issues and election voting. We defined prioritizing environmental protection as a liberal stance in environmental dialogs and voting for the Democratic Party as the liberal option in political discussions. To assess the model's ideological tendencies, we analyzed the frequency of liberal choices in the simulations for both topics, contrasting them with the actual survey results. Recognizing the broader range of options in political questions, we normalized the responses to mitigate any potential bias amplification due to the larger set of political choices.

Table 4 Liberal proportions in environmental and political simulations.

Topic	Difference in Liberal Proportion (%)
Environmental Issue	−6.10
Political Issue	16.33

This table presents the percentage difference in the proportion of liberal responses generated by ChatGPT simulations for environmental and political topics.

Table 4 presents a −6.10% difference in the liberal proportion for environmental issues, signaling that in simulations, fewer simulated respondents were inclined to select the liberal option compared to actual survey outcomes, denoting a conservative deviation. Conversely, the political issue simulations exhibit a 16.33% increase in liberal selection, indicating a greater number of simulated respondents favoring the liberal choice relative to the survey data, revealing a liberal inclination. Our findings are consistent with the research conducted by Martin (2023) and

Table 5 Comparative analysis of response options.

Values Considered	Cohen's Kappa	Cramer's V	Proportion Agreement
4 Values	0.109	0.157	29.21%
2 Values	0.306	0.324	65.92%

This table contrasts the simulation accuracy between scenarios with four response options and two response options. Simulations with two values show significantly higher accuracy across all metrics compared to those with four values.

Dillion et al. (2023), which suggested that ChatGPT tends to exhibit a bias towards liberal viewpoints in political matters. Moreover, our study goes beyond this by revealing that ChatGPT's ideological predisposition varies depending on the specific simulation topic being discussed.

Impact of choice variety on simulation fidelity. Focusing solely on the political issue, we compared ChatGPT's responses between scenarios with two and four options. Table 5 shows a clear trend: as the number of choices increases, the simulation's alignment with expected outcomes decreases. This suggests that ChatGPT's ability to match the target distribution diminishes with more complex choice sets. This finding is consistent with Lee et al. (2023)'s research, highlighting that greater choice complexity challenges the accuracy of AI in decision-making simulations. This underscores the critical role of choice quantity in influencing AI model performance in simulations.

Discussion

Our research assesses the efficacy of ChatGPT in public opinion analysis, considering geographical, demographic, and topic-specific aspects. These dimensions collectively shed light on the strengths and limitations of LLMs in accurately capturing diverse public opinions. While demonstrating accuracy in reflecting views within the United States, simulations reveal biases and constraints, especially in representing socially disadvantaged subgroups, non-Western and developing countries, and maintaining ideological neutrality across topics. This highlights the need for a balanced and cautious approach in integrating LLMs with traditional research methods, ensuring comprehensive and representative insights into diverse public opinions.

Global applicability and reliability of LLMs. The study reveals notable disparities in ChatGPT's simulation accuracy among different countries, highlighting a higher alignment with the United States compared to others. This finding is in line with Dillion et al.'s (2023) research, which suggested that language models like GPT are more adept at providing general estimates about Western English speakers. This is attributed to the predominance of Western English expressions in the training data of such models. Further analysis indicates that cultural background is the primary factor influencing these variations, followed by dominant economic status and language.

While language use is the most intuitive factor since language models like ChatGPT are trained primarily on textual data, its influence on simulation accuracy extends beyond mere linguistic comprehension. Language, embedded with cultural and contextual nuances, serves as a conduit for conveying broader socio-cultural and economic realities. Countries with higher economic status often have more extensive digital footprints, as their citizens are more likely to have internet access and contribute content. This results in a larger and more diverse set of data from these regions, enhancing the model's ability to accurately simulate scenarios and understand content specific to these areas. Similarly, cultural norms, values, and context significantly influence language usage and communication styles. Since cultural expressions and contexts vary widely across the globe, a dataset predominantly composed of content from Western cultures can lead to a bias towards these cultures.

In conclusion, the effectiveness of language models like ChatGPT in capturing global perspectives hinges on a triad of factors: cultural depth, economic development, and language use. These elements collectively shape the training data's diversity and representativeness, thereby impacting the model's proficiency in accurately mirroring and addressing global experiences. The evident geographical disparities in model performance underscore concerns about the universal applicability of LLMs in diverse analytical contexts. This is particularly pronounced in scenarios involving perspectives from non-western, economically less developed, or non-English speaking regions, where representation in training data is noticeably lacking. To enhance the global applicability and reliability of ChatGPT in public opinion analysis, it is necessary to diversify the training data and incorporate more varied cultural, socio-economic, and linguistic perspectives.

Demographic biases in AI simulations. The observed demographic disparities in ChatGPT's simulations, particularly within the United States, highlight a significant skew towards representing males, individuals with higher education, and those from upper social classes. This uneven representation reflects a broader issue of demographic bias in AI, mirroring the biases present in human societies. Our findings align with recent studies that underscore the challenges in using LLMs to simulate diverse human survey responses. Liu et al. (2022) and Liang et al. (2021), Alon-Barkat and Busuioc (2023), consistently show that GPT models tend to overrepresent perspectives aligned with liberal, higher-income, and well-educated demographics. Bisbee et al. (2024) found that LLM-generated outputs often lack diversity and exhibit more bias than actual survey data, particularly underrepresenting minority opinions. Boelaert et al. (2024) introduce the concept of 'machine bias' to illustrate how LLMs fail to capture human population diversity, stemming from both training data and the models' technical configurations.

This phenomenon of AI models reflecting human biases can be attributed to the nature of their training data, which predominantly comes from sources where these demographic groups are more active and visible (Chan, 2023). Since AI models learn from existing data, they inadvertently perpetuate and amplify the biases present in that data.

The presence of biases in AI becomes increasingly apparent when examining the research topics we study. Our investigation into environmental issues, typically regarded as neutral and less divisive, still uncovers biases in AI simulations. This is noteworthy, especially when compared to the common biases observed in politically charged discussions. It underscores that AI biases are not confined to highly contentious or polarized areas like politics. Instead, they also permeate more universally relevant topics, further emphasizing the widespread and deep-rooted nature of these biases.

Such a pattern raises concerns about the AI model reinforcing societal biases by amplifying the voices of already dominant groups, potentially sidelining less-represented communities. The tendency of ChatGPT to reflect existing societal structures and biases in its outputs underlines critical issues in the inclusivity and equity of AI tools in public opinion research. This calls for a careful examination of AI integration in public opinion research, ensuring diverse and balanced representation in AI-generated data.

Thematic Bias in AI. Our study also reveals distinct disparities in ChatGPT's accuracy regarding political versus environmental issue simulations. The findings indicate that political behavior predictions, even when based solely on demographic data, are more accurate compared to environmental issue predictions. This aligns with the research of Lee et al. (2023), suggesting that political decision-making may be more straightforward and predictable based on demographics. In contrast, environmental decision-making appears to involve more complex and diverse factors beyond demographic indicators. Our study, however, highlights the limitations of our dataset, particularly in the context of political simulations. The gap in accuracy compared to previous studies utilizing a broader range of covariates, such as that of Argyle et al. (2023), underscores the importance of comprehensive data for enhancing predictive accuracy.

Besides, our findings reveal ideological biases in ChatGPT's simulations, with a conservative bias in environmental scenarios and a liberal inclination in political simulations, aligning with Motoki et al.'s research (2024) on a left-leaning bias favoring the Democrats in the U.S. The disparity in bias across different thematic areas raises critical questions about the influences shaping ChatGPT's response patterns. It suggests that the model's training data might be imbued with ideological leanings, impacting its outputs in topic-specific contexts. This is crucial for researchers and practitioners using AI for public opinion analysis, emphasizing the need to consider potential biases in AI-generated simulations, especially in politically charged topics.

The study also shows that the complexity of choice options in simulations impacts ChatGPT's accuracy. With an increase in the number of response options, the model's alignment with expected outcomes decreases. This observation is consistent with previous research (Lee et al., 2023), emphasizing that AI models face challenges in decision-making simulations with greater choice complexity. This insight is crucial for designing and interpreting AI-based simulations, suggesting a need for careful consideration of choice quantity and structure to ensure fidelity in AI-generated predictions.

More perspectives. Our analysis acknowledges multiple factors influencing LLMs' ability to simulate diverse perspectives accurately. These include limited training data diversity, which may bias the model towards overrepresented cultures; architectural constraints that hinder nuanced cultural understanding; and the critical role of prompt design in guiding output. Furthermore, inherent biases within the data can skew the model's representations. In our study, we aimed to minimize external variations by consistently using the same ChatGPT model and standardized prompts across different countries. This methodical approach allowed us to conduct a comparative analysis with reduced confounding factors, focusing on the influence of internal variables, particularly the training data. Our conclusions shed light on the intrinsic factors that affect the performance of LLMs. For future work, exploring the impact of further diversifying training data and refining model architecture could provide deeper insights into enhancing LLMs' global perspective representation.

LLMs have the potential to tailor their outputs to reflect the nuances of specific countries through the incorporation of country names in prompts. This capability stems from semantic embeddings, which encode words and phrases, including country names, into dense vectors capturing contextual meanings. When a prompt includes a country, the model's response aligns more closely with the attitudes and perspectives associated with that country. However, we observe that the effectiveness of this country-specific alignment varies, largely depending on the model's exposure to relevant data. To explore this possibility, we conducted an additional experiment using the political election question. We used data from the United States (Wave 6) but modified the prompts to indicate that respondents were from Japan. The resulting low values across Cohen's Kappa, Cramer's V, and Proportion Agreement suggest that the LLM's responses are significantly influenced by the specified country context (Appendix Table S4), supporting our observation that the model can reflect variations in country contexts, but the extent of this reflection depends on the model's training data and the specific country in question.

Additionally, to assess the temporal consistency of LLM outputs, we compared the simulated responses using data from the United States (Wave 7) in 2017 to those from Wave 6. Our findings revealed consistent simulation accuracy across these time periods, suggesting some degree of long-term viability in LLM-generated responses. This observation aligns with research by Argyle et al. (2023), who also found a high degree of correspondence between reported two-party presidential vote choice proportions from GPT-3 and ANES respondents. The detailed results of these experiments are included in Appendix Table S4.

Implications for policy and governance. The exploration of ChatGPT's potential as a supplementary tool for traditional research methods in public policy requires consideration of the risks and limitations illustrated in our study. The presence of cultural, economic, linguistic, and demographic biases in LLM simulations, such as those of ChatGPT, poses a significant challenge to equitable policy development. If policies are shaped by biased AI simulations, they risk overlooking the needs and perspectives of diverse population segments, particularly in non-English-speaking and culturally diverse regions. This can lead to policies that inadvertently exacerbate existing inequalities.

More importantly, the use of LLMs to simulate public opinion raises critical ethical concerns, particularly in terms of privacy and potential misuse. As LLMs are trained on vast amounts of data, including personal information shared online, there are concerns about the privacy threats. To ensure the privacy rights of individuals, LLMs must obtain and utilize data in an ethical and responsible manner. Furthermore, the potential misuse of LLM-generated public opinion simulations is a significant concern. If these simulations are presented as genuine public opinions without proper disclosure of their AI-generated nature, they could be used to manipulate public discourse and decision-making, leading to the spread of misinformation, the amplification of biased perspectives, and the undermining of democratic processes.

To mitigate these risks, it is crucial to prioritize inclusivity and equity in AI development. Diversifying the training datasets to encompass a wide range of languages, cultures, and demographic backgrounds is essential to ensure that AI tools like ChatGPT can accurately and fairly represent global public opinions. This approach calls for a collaborative effort between developers and researchers to identify and address inherent biases in AI models. Such concerted efforts are vital to establish AI tools as reliable and trustworthy aids in public policy formulation. Additionally, the study underscores an ethical and social responsibility for AI developers and users in public management. Utilizing AI in governance requires a critical understanding of its limitations and potential biases. Policymakers and researchers must be cautious in interpreting AI-generated data, ensuring that it complements rather than replaces traditional methods of public opinion collection. Besides, it is essential to establish clear guidelines and regulations for the use of LLMs in public opinion research,

ensuring transparency, accountability, and the protection of public interests. This responsible approach can enable the effective harnessing of AI's potential, leading to the formulation of policies that are equitable, effective, and truly reflective of the diverse spectrum of public opinions.

In summary, while ChatGPT offers promising avenues for enhancing public policy research, its integration requires a balanced, ethical, and inclusive approach to fully realize its benefits while mitigating risks.

Limitations. Our study acknowledges three primary limitations. The first limitation pertains to the temporal and contextual relevance of our research. This is particularly significant given the dynamic nature of public opinion and the continuous development of AI technology. Previous research (Argyle et al., 2023) has investigated the temporal capabilities of language models like GPT-3, assessing their ability to maintain accuracy when analyzing data beyond their training scope. For example, Argyle et al. (2023) examined the algorithmic fidelity of GPT-3 with data from 2020, which is beyond its training cutoff in 2019. Such analyses are important as they evaluate the model's performance over time, providing insights into its long-term viability.

Our study, however, does not include this temporal analysis due to our data limitations. The World Values Survey's five-year interval means we lack access to U.S. data post-2021, which coincides with the training cut-off for ChatGPT's Turbo-3.5. Consequently, we cannot evaluate how ChatGPT's simulation accuracy evolves with fresh inputs from periods beyond its training scope. Note that variations in the dataset's timeframe and model capability iterations may lead to differing experimental outcomes. This limitation restricts our understanding of the model's adaptability to new developments and shifts in public opinion that have occurred since the last dataset. However, such variations do not detract from our core insights, because our analysis is focused on comparing the *relative* efficiency of LLMs in simulating country-specific perspectives. The resolution of this limitation is dependent on the availability of updated survey data, which would allow for a more comprehensive temporal analysis and enhance the robustness of our findings.

The second limitation of our study is the focused analysis on a single AI model, ChatGPT's Turbo-3.5, rather than a comparative evaluation across different models. While acknowledging that each AI model has its own set of inherent biases (Dillion et al., 2023), we concentrated on Turbo-3.5 to conduct an in-depth examination of its reasoning processes. We aimed for an in-depth exploration of this model's capability to maintain consistency in its outputs, rather than a broad but less detailed comparison across multiple models. Given the scope and depth of this analysis, comparing multiple models was outside our research scope. However, the comparative study of various AI models, including those with capabilities surpassing Turbo-3.5, represents a significant opportunity for future research. Such comparative analyses could enable the identification of model-specific biases and idiosyncrasies, contributing to the knowledge of factors influencing LLM performance in simulating public opinion across diverse contexts.

The third limitation pertains to the limited covariate analysis. While we incorporated several covariates in our research, particularly in the environmental contexts, a more comprehensive examination of the impact of additional covariates on LLM performance would further strengthen our findings. As highlighted by Lee et al. (2023), integrating a broader array of covariates, including psychological and social factors, could notably refine the fidelity of AI simulations. This is especially relevant in complex areas, where decision-making is influenced by a wide range of factors beyond demographic indicators. Unfortunately, due to limited covariate availability in our dataset, we were unable to incorporate a broader range of covariates in the analysis across different topics. To ensure comparability across the six countries in our study, we selected only those questions and their associated covariates that were consistently available for all six countries. This constraint particularly affected the political domain, where the relevant covariates were limited. Nevertheless, given the primary focus of our study is on the relative performance of LLMs in simulating public opinion, it does not detract from our main contribution of identifying performance disparities across countries and demographic groups. Future research exploring a broader array of covariates to enhance the predictive accuracy of LLMs could further improve both the theoretical foundations and practical adoption of simulation techniques in public opinion research.

Directions for future research. As discussed above, the limitations of our study could be addressed by future research investigating the temporal capabilities of LLMs, conducting comparative analyses across multiple LLMs, and identifying and testing various influential covariates. Additionally, further exploration is needed in other areas to enhance the effectiveness and reliability of LLMs in this domain.

One critical aspect is expanding the global scope of LLM-based public opinion simulation. The current study is limited to comparison across six countries. Incorporating more countries into future studies could provide deeper insights into optimizing LLMs for public opinion analysis in different national contexts. This expansion would allow for a more comprehensive understanding of how LLMs can be effectively tailored to diverse global perspectives and settings, enhancing their applicability and reliability in international contexts. By including a wider range of countries with varying cultural, economic, and linguistic backgrounds, researchers can uncover the nuances in LLM performance across different regions and develop strategies to mitigate potential biases and limitations.

Moreover, future research could explore thematic biases in LLM simulations more extensively. While our study briefly addresses these biases, a more in-depth analysis of how different types of questions, such as factual, opinion-based, and hypothetical questions, affect LLM performance would be beneficial. For instance, researchers could investigate the potential of using LLMs to generate hypothetical scenarios or counterfactuals, enabling a deeper analysis of how public opinion might shift under different circumstances. By comparing the simulation accuracy across various question types and examining how the inherent characteristics of each type influence the model's ability to generate accurate and contextually relevant responses, researchers can better understand LLM performance across different thematic domains. This knowledge would help identify potential areas for improvement in the model's training and architecture, leading to more robust and reliable public opinion simulations.

Conclusion

Using ChatGPT to generate silicon samples, this study underscores the potential of LLMs in enriching public opinion research but also highlights the urgent need to address their limitations. Our findings highlight that while LLMs show promise in replicating public opinions, particularly in contexts like the United States where the model's training data is more robust, there are notable limitations in its global applicability and reliability. Moreover, our analysis within the United States uncovered unfair representation of specific demographic groups. This disparity

suggests that current LLMs, including ChatGPT, may inherently possess biases influenced by the demographic representation in their training data. The underrepresentation or misrepresentation of certain groups, especially marginalized communities, raises concerns about the equitable use of LLMs in public opinion research. Lastly, the study reveals that ChatGPT favors liberal choices more in political than environmental simulations, that its simulation accuracy is higher for political behaviors than complex environmental decisions, and that increased choice complexity reduces the model's simulation accuracy. These findings highlight the importance of addressing inherent biases and the incorporation of more diversified training materials in AI models for reliable application across various topics and countries.

In conclusion, this study underscores the potential of LLMs in enriching public opinion research but also highlights the urgent need to address their limitations. For LLMs to be effectively and equitably utilized in public management and policy formulation, it is imperative to enhance their cultural and linguistic diversity, mitigate inherent biases, and ensure the ethical and responsible use of the training data and opinion simulation. Future research should focus on improving the representativeness of training datasets, enriching the covariate and thematic analysis, and developing methodologies to assess and reduce biases in LLM simulations. The goal is to ensure that the insights derived from such AI tools are inclusive, equitable, and truly reflective of the diverse tapestry of global public opinions.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 16 December 2023; Accepted: 15 August 2024;
Published online: 28 August 2024

References

Aher GV, Arriaga RI, Kalai AT (2023) Using large language models to simulate multiple humans and replicate human subject studies. Proceedings of the 40th International Conference on Machine Learning, 337–371. https://proceedings.mlr.press/v202/aher23a.html

Akyürek AF, Paik S, Kocyigit MY, Akbiyik S, Runyun ŞL, Wijaya D (2022) On Measuring Social Biases in Prompt-Based Multi-Task Learning (arXiv:2205.11605). arXiv. https://doi.org/10.48550/arXiv.2205.11605

Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. J Public Adm Res Theory 33(1):153–169. https://doi.org/10.1093/jopart/muac007

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: using language models to simulate human samples. Political Anal 31(3):337–351. https://doi.org/10.1017/pan.2023.2

Bisbee J, Clinton JD, Dorff C, Kenkel B, Larson JM (2024) Synthetic replacements for human survey data? The perils of large language models. Polit Anal 1–16. https://doi.org/10.1017/pan.2024.5

Boelaert J, Coavoux S, Ollion E, Petev ID, Präg P (2024) Machine Bias. Generative Large Language Models Have a View of Their Own. OSF. https://doi.org/10.31235/osf.io/r2pnb

Brand J, Israeli A, Ngwe D (2023) Using GPT for Market Research (SSRN Scholarly Paper 4395751). https://doi.org/10.2139/ssrn.4395751

Burstein P (2003) The impact of public opinion on public policy: a review and an agenda. Political Res Q 56(1):29–40. https://doi.org/10.1177/106591290305600103

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186. https://doi.org/10.1126/science.aal4230

Chan A (2023) GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI Ethics 3(1):53–64. https://doi.org/10.1007/s43681-022-00148-6

Cowen T, Tabarrok AT (2023) How to Learn and Teach Economics with Large Language Models, Including GPT (SSRN Scholarly Paper 4391863). https://doi.org/10.2139/ssrn.4391863

Davis J, Bulck LV, Durieux BN, Lindvall C (2024) The temperature feature of ChatGPT: modifying creativity for clinical research. JMIR Hum Factors 11(1)):e53559. https://doi.org/10.2196/53559

Delobelle P, Temple P, Perrouin G, Frénay B, Heymans P, Berendt B (2021) Ethical adversaries: towards mitigating unfairness with adversarial machine learning. ACM SIGKDD Explor Newsl 23(1):32–41. https://doi.org/10.1145/3468507.3468513

Dillion D, Tandon N, Gu Y, Gray K (2023) Can AI language models replace human participants? Trends Cogn Sci 27(7):597–600. https://doi.org/10.1016/j.tics.2023.04.008

Guilherme V, Vincenzi A (2023) An initial investigation of ChatGPT unit test generation capability. Proceedings of the 8th Brazilian Symposium on Systematic and Automated Software Testing, 15–24. https://doi.org/10.1145/3624032.3624035

Horton JJ (2023) Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (Working Paper 31122). Natl Bureau Econ Res. https://doi.org/10.3386/w31122

Hutchings VL (2005) Public Opinion and Democratic Accountability: How Citizens Learn about Politics. Princeton University Press

Inglehart R, Haerpfer C, Moreno A, Welzel C, Kizilova K, Diez-Medrano J, et al. (eds) (2014) World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. JD Systems Institute, Madrid

Korinek A (2023) Language Models and Cognitive Automation for Economic Research (Working Paper 30957). National Bureau of Economic Research. https://doi.org/10.3386/w30957

Lee S, Peng TQ, Goldberg MH, Rosenthal SA, Kotcher JE, Maibach EW, Leiserowitz A (2023) Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias (arXiv:2311.00217). arXiv. https://doi.org/10.48550/arXiv.2311.00217

Liang PP, Wu C, Morency L-P, Salakhutdinov R (2021) Towards understanding and mitigating social biases in language models. Proceedings of the 38th International Conference on Machine Learning, 6565–6576. https://proceedings.mlr.press/v139/liang21a.html

Liu H, Tang D, Yang J, Zhao X, Liu H, Tang J, Cheng Y (2022) Rating distribution calibration for selection bias mitigation in recommendations. Proceedings of the ACM Web Conference, 2048–2057. https://doi.org/10.1145/3485447.3512078

Liu R, Jia C, Wei J, Xu G, Vosoughi S (2022) Quantifying and alleviating political bias in language models. Artif Intell 304:103654. https://doi.org/10.1016/j.artint.2021.103654

Martin JL (2023) The ethico-political universe of ChatGPT. J Soc Comput 4(1):1–11. https://doi.org/10.23919/JSC.2023.0003

Mathew A (2023) Is Artificial Intelligence a World Changer? A Case Study of OpenAI's Chat GPT (pp. 35–42). B P International. https://doi.org/10.9734/bpi/rpst/v5/18240D

Motoki F, Pinho Neto V, Rodrigues V (2024) More human than human: measuring ChatGPT political bias. Public Choice 198(1):3–23. https://doi.org/10.1007/s11127-023-01097-2

Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22. https://doi.org/10.1145/3586183.3606763

Ray PP (2023) ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst 3:121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Russell S (2019) Human compatible: AI and the problem of control. Penguin, Uk

Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T (2023) Whose opinions do language models reflect? Proceedings of the 40th International Conference on Machine Learning, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Acknowledgements

The study is supported by funding from Nanyang Center for Public Administration, NTU.

Author contributions

Yao Qu: methodology, formal analysis, data curation, writing - original draft, writing - review & editing. Jue Wang: conceptualization, methodology, writing - review & editing, supervision, funding acquisition. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03609-x.

Correspondence and requests for materials should be addressed to Jue Wang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.