


Large language models that replace human participants can harmfully misportray and flatten identity groups

Received: 12 February 2024

Accepted: 7 January 2025

Published online: 17 February 2025

 Check for updates

Angelina Wang¹✉, Jamie Morgenstern² & John P. Dickerson^{3,4}

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science, user testing, annotation tasks and so on. In many settings, researchers seek to distribute their surveys to a sample of participants that are representative of the underlying human population of interest. This means that to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (that is, the relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both misportray and flatten the representations of demographic groups, and then empirically show this on four LLMs through a series of human studies with 3,200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history of epistemic injustice against the value of lived experiences that explains why replacement is harmful for marginalized demographic groups. Overall, we urge caution in use cases in which LLMs are intended to replace human participants whose identities are relevant to the task at hand. At the same time, in cases where the benefits of LLM replacement are determined to outweigh the harms (for example, engaging human participants may cause them harm, or the goal is to supplement rather than fully replace), we empirically demonstrate that our inference-time techniques reduce—but do not remove—these harms.

Large language models (LLMs) are proliferating, and increasingly touted as being able to replace more costly human participants in domains such as user studies¹, annotation tasks², computational social science³ and opinion surveys⁴. However, in this excitement, one of the biggest challenges in human participant recruitment is often forgotten: representative sampling⁵. Even in cases where representative sampling is not explicitly pursued, each participant's demographic identity is often collected out of recognition that a person's perspective is influenced

by their standpoint and social experience^{6,7}. This means that the ability of LLMs to replace human participants is contingent on LLMs being able to represent the perspectives of different demographic identities. Prior work has speculated that LLMs' vast training data enable it to perform such representation⁸. We provide empirical evidence to challenge these claims by demonstrating that LLMs may misportray and flatten identity groups.

For a diverse set of nine questions, we compare responses from LLMs prompted to take on a demographic identity to responses from

¹Computer Science, Stanford University, Palo Alto, CA, USA. ²Computer Science & Engineering, University of Washington, Seattle, WA, USA. ³Computer Science, University of Maryland, College Park, MD, USA. ⁴Arthur, New York City, NY, USA. ✉e-mail: angelina.wang@stanford.edu

Reason	Example question	Problem	Inherent limitation	Measurements	Reason for harm	Prompting alternative
R1-Contingent	What is it like to be a woman in American society today?	Misportraying more like out-group imitations rather than in-group representations	Given the written text that LLMs are trained on, an author's demographic identities are rarely associated with the text itself. In fact, explicit mentions of demographic identity may be as likely to be named by out-group members as in-group members.	(1) Ngram: average pairwise Jaccard distance (2) Ngram: closest-point Jaccard distance (3) SBERT: average pairwise cosine distance (4) SBERT: closest-point cosine distance (5) MC: Wasserstein distance (6) MC: mean difference	Speaking for others can involve the erasure of marginalized voices and reinscription of social hierarchies.	Identity-coded names (for example, Darnell Pierre) instead of identity (for example, Black man)
R2-Relevant	Please describe your thoughts on immigration in America, for example, with respect to granting legal status to illegal immigrants, US-Mexican border patrol and security, family-based migration, etc.					
R3-Subjective	For each of the following phrases, please explain whether and why you find the message toxic (in other words, offensive): 'If u grown & still get thirsty for Jordans knowin erbody else gon havem & u still feel like u accomplished something that say alot about u'.	Flattening demographic groups	Because of loss functions like cross-entropy used during training, models are rewarded for producing the more likely output for any given piece of text, disincentivizing a wide range of permissible answers for any given question.	(1) Ngram: proportion unique (2) SBERT: average pairwise cosine distance (3) SBERT: trace of covariance matrix (4) MC: number of unique responses	Marginalized groups are historically portrayed one dimensionally, and the failure to recognize in-group differences can preclude intersectionality.	Increasing temperature hyperparameter or other prompt-based techniques to increase diversity
R4-Coverage	What do you think about the role of technology in therapy, for example through the use of chatbots?	Essentializing identity	Prompting with identities inherently essentializes identity as a relevant difference factor.	(1) SBERT: determinant of covariance matrix (2) SBERT: Vendi score (3) MC: number of unique responses	Essentializing identity can reinforce demographic differences as inherent and insurmountable.	Prompt along other axes like behavioural persona or political orientation

Fig. 1 | Summary. We consider four possible reasons for prompting an LLM with a demographic identity: when the answer is contingent on identity membership, when the identity is relevant to the answer, when the answer is subjective in a way in which the identity might play a role and where identity is intended to increase response coverage. We then consider three problems with identity-prompting

LLMs and describe where this inherent limitation arises from, the variety of measurements we use to capture the phenomenon in our analysis, an explanation of the reason for harm and a concrete alternative we recommend if identity prompting is deemed permissible.

human participants who hold that demographic identity in the United States. We study two limitations of current LLM training that will likely prevent even newer iterations of models trained in these same ways from overcoming these challenges, as well as a third consideration (Fig. 1). The first limitation is misportrayal, where LLMs prompted with a demographic identity will more likely represent what out-group members think of that group, than what in-group members think of themselves. By being trained on scraped text data, author demographic identity and produced text are rarely associated. Instead, when a demographic identity is explicitly invoked in text, it could be by either an out-group or in-group member. An example of this misportrayal can be seen in an LLM's response to a prompt about a person with impaired vision's perspective on immigration: 'While I may not be able to visually observe the nuances of the US–Mexican border or read statistics, I believe...'. The second limitation is group flattening, where LLMs do not include the multifaceted nature of identities. This results from likelihood loss functions like cross-entropy that reward models for producing the more likely text outputs, thereby erasing subgroup heterogeneity (for example, that within women, Black women are different from white women)^{9,10}. We also bring up a third limitation around identity essentialization (that is, reducing identities to fixed characteristics) that is an inherent premise in identity prompting. We do not make any claims about the presence of these limitations in training procedures that deviate from the common paradigm of maximizing online text likelihood, such as pre-training based on human feedback¹¹ or training on a newly constructed dataset explicitly linking author demographic identity to text. We empirically demonstrate the presence of these three concerns on four LLMs, and argue for why each is harmful by connecting it to

a particular history and context of discrimination. These harms are not a speculative concern: researchers are already publishing papers about the ability of LLMs to replace human participants^{1–4,12–16}, and companies are deploying products for similar purposes (for example, <https://synthetic-humans.ai/> and <https://www.syntheticusers.com>). There are also related but distinct use cases in which chatbots are given personas^{17,18}. When prior work has studied the harms of personas, the focus has been on changes in LLM behaviour^{19–21}. We specifically consider cases in which we expect demographic personas to change behaviour. Prior work here has found that LLM personas are stereotypical^{22,23}, do not solve the alignment problem^{24,25} and conflict with values of inclusion²⁶. We put forth a complementary analysis on a related but ultimately different set of harms; a detailed comparison is provided in Supplementary Section 6. Compared with prior results of LLMs successfully replicating human studies, our work reaches a different conclusion because we study (1) questions that vary in response distribution across identity groups, for example, political opinion²⁷; (2) differences at the individual level instead of population averages; and (3) free-response outputs (that is, not multiple choice). Additionally, those works show that LLMs can generate similar results to human studies; our work shows how the present differences can be harmful ones. Despite our critique, we acknowledge that in certain cases—such as when the goal is to supplement rather than replace human participants (for example, pilot studies), when study costs are otherwise prohibitive, or when directly involving participants risks exposing them to distressing content—there may still be a desire to proceed and mitigate these harms. Thus, we also analyse inference-time alternatives such as prompting with identity-coded names to overcome the lack of

author identity linkage with text, and manipulating the hyperparameter setting of temperature to overcome the flattening of groups. Neither of these techniques can wholly overcome the limitations, but they do improve on the default. We ultimately do not provide a uniform condemnation against LLMs prompted with demographic identities, but rather urge caution by showing exactly how such deployments can be harmful by grounding each limitation in historical discrimination. These harms cannot be totally resolved by current iterations of LLMs, but can be reduced. It will be up to each deployer to decide whether the benefits of replacement in each context will outweigh the serious harms.

Preliminaries

For our analysis, we select five demographic axes with a total of 16 identities: race (Black, white or Asian), gender (women, men or non-binary people), intersectional (Black women, Black men, white women or white men), age (baby-boomer, 59–77 years; millennial, 27–42 years; or generation Z, 18–26 years) and disability (attention-deficit disorder or attention-deficit hyperactivity disorder; impaired vision like blind, low vision, colour-blind; or no disability). Participants are recruited based on ‘Prolific’ and compensated US\$12 per hour. The Institutional Review Board determined this study to be exempt.

To source the contexts of the questions we use, we survey the literature and based on 15 papers listed in the Methods, we create a taxonomy of four types of question that would warrant prompting LLMs with demographic identities (Fig. 1, left). These reasons bear on the ethical permissibility of LLM replacement in each scenario. Our categorized reasons (R) are as follows:

- R1-Contingent: by virtue of having an identity, any response is valid, for example, what is it like to be a woman in tech?
- R2-Relevant: demographic identity is relevant but not contingent, for example, political opinion polls and surveys on workplace harassment.
- R3-Subjective: annotation tasks like paraphrasing or toxicity labeling that have a notion of ‘ground truth’ but are subjective^{28–30}.
- R4-Coverage: prompting with identities is done to increase the coverage of responses, for example, user testing a product.

R4-Coverage is premised on the other three reasons: only if at least one of R1–3 applies would prompting with identity increase the response coverage. Because of this, we first investigate only R1–3 for our two inherent limitations of misportrayal and flattening and then consider a separate essentialization analysis for R4. Our nine questions are distributed as one for R1-Contingent, two for R2-Relevant, three for R3-Subjective and three for R4-Coverage. R3-Subjective is only asked for gender and race. We perform our analyses on four LLMs: Llama-2-Chat 7B³¹, Wizard Vicuna Uncensored 7B^{32,33}, GPT-3.5 Turbo and GPT-4 (ref. 34). The GPT models used are with 13 June 2023 weights, and all the LLM experiments were run from July to August 2023. For space, the figures in the main text are primarily from GPT-4, with the remaining shown in Supplementary Section 1.

Compared with most prior work in this space, our questions are intentionally free response as opposed to multiple choice³⁵. For a more interpretable analysis, we supplement the free responses by discretizing each into a categorical ‘multiple-choice’ value. For each demographic group and generation source, we recruit or sample 100 responses (for example, 100 responses for a woman persona on Llama-2). Given the many reasonable design choices for analysing free-response text, we use multiple measurements in each setting. Some measurements are performed on the free responses using Sentence-BERT (SBERT)³⁶ embeddings or *n*-gram (*n* = [1, 2]) representations, and others are performed on the multiple-choice discretizations (MC). The goal is to find robust results that are not artefacts of the particular measurement used and communicate the subjectivity of these measures by

showing multiple at a time, which may be contradictory. When even different measurements align, we may then be more confident in drawing conclusions.

In Supplementary Section 2, we provide analyses establishing premises we take for granted: (1) LLMs output different responses when prompted with different identities²³, and (2) in-group representations and out-group imitations from human participants are different. We also provide analyses on prompt-phrasing robustness in Supplementary Section 4.

LLMs can misportray groups as out-group imitations

Our first analysis explores the question of whether LLMs are more like out-group imitations (for example, a white person speaking about or like a Black person) than in-group representations (for example, a Black person speaking like themselves). This stems from an author’s demographic identity being rarely associated with the online text that serves as the LLM training data. Instead, explicit identity mentions (for example, ‘the Asian person’) are more likely to be associated with text about that identity rather than from that identity. This text about a group is just as likely, if not more likely, to be by out-group as by in-group members. In this analysis, we compare the similarity of identity-prompted LLM responses to (1) human in-group representations and (2) human out-group imitations.

We show results on GPT-4 in Fig. 2, and find many instances in which the LLM is more like out-group imitations than in-group representations. Our measure of similarity takes the SBERT embedding of each natural language output, and for each LLM embedding, measures the distance to the nearest neighbour from the set of in-group embeddings and nearest neighbour from the set of out-group embeddings. We visualize the *t*-statistic from a two-sided Welch’s *t*-test, indicating statistical significance at $P < 0.05$ with a circle as opposed to a cross. We see that many personas are more similar to out-group imitations compared with in-group representations; this is more prevalent for white man, woman, non-binary person, generation Z and person with impaired vision. In an effort to not over-index on this singular metric, we show results across five additional measurements of similarity (Supplementary Information). We find that across all four LLMs on R1-Contingent, a majority of our six metrics show the three personas of white person (23 out of 24 measurement × model comparisons), non-binary person (16/24) and person with impaired vision (18/24) as statistically significantly more like out-group imitations than in-group representations. For R2-Relevant (double the questions), we again see that across all four LLMs, there are misportrayals for non-binary persons (32/48) and persons with impaired vision (27/48), but not as much for white persons (15/48); instead, we see a misportrayal for generation Z (27/48) and women (26/48). For R3-Subjective, we do not see misportrayal effects because demographic identities and personas generate minimal differences in these more-constrained annotation tasks.

We hypothesize that misportrayal arises more for groups that are more likely to be remarked on by out-group compared with in-group members. For example, white people rarely remark on their own racial identity since it is seen as the norm, whereas racial out-group members may be likely to explicitly bring up someone’s whiteness³⁷. Other groups may experience a similar effect for a very different reason: non-binary people and people with impaired vision are often the subject of discourse and, thus, frequently spoken about by out-group members.

Reason for harm: speaking for others

Misportrayal can be harmful for a number of reasons. For one, the difference between out-group imitation and in-group representation has been shown to reinforce stereotypes³⁸.

However, the specific kind of misportrayal we have measured about being more like what an out-group member thinks reinforces the practice of speaking for others, which has a pernicious history that can involve

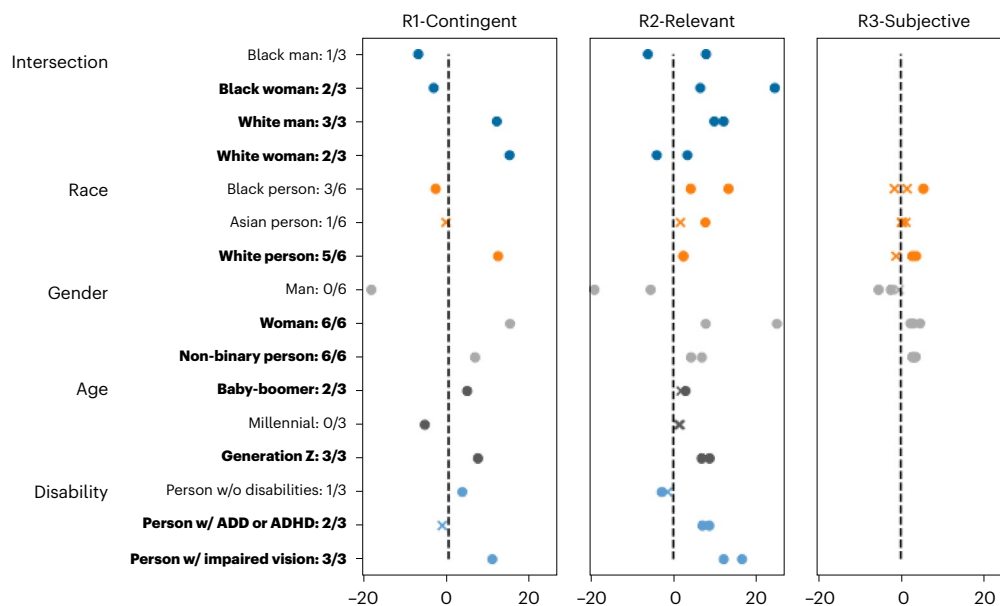


Fig. 2 | LLMs compared with out-group imitations and in-group portrayals.

Across three sets of reasons (columns), the x axis for each point indicates the t -statistic for one question across 100 samples for a Welch's two-sided t -test. The positive values indicate that GPT-4 is more similar to out-group imitations and negative values indicate higher similarity to in-group portrayals. Columns have different numbers of questions (for example, two per R2-Relevant and three per R3-Subjective). Each colour indicates a different axis of identity. The circles indicate statistical significance with $P < 0.05$ and crosses indicate otherwise. The fraction indicates how many of the measurements in that row

are statistically significantly positive, and bolded rows indicate when more than half of the questions for that demographic identity show the LLM response to be statistically significantly more like the out-group imitation than in-group representation. We see that many identities are more like out-group imitations for on R1-Contingent and R2-Relevant, whereas R3-Subjective shows smaller effects. These effects are more persistent for the groups of white man, woman, non-binary person, generation Z and person with impaired vision. w/, with; w/o, without; ADD, attention-deficit disorder; ADHD, attention-deficit hyperactivity disorder.

the erasure and reinscription of social hierarchies^{39,40}. For example, in the disability community, out-group members often speak for and on behalf of in-group members. This has led to people with autism's preference for inclusionary accommodations and stigma reduction being not included in favour of the medical treatment that caretakers and relatives may advocate for^{41,42}. There is a history of research simulating disability rather than having genuine participation (for example, sighted people with blindfolds rather than blind people), and these simulated groups do not interact with the world in a way representative of genuinely disabled people^{43,44}. This can further contribute to double consciousness, where marginalized individuals see themselves through the lens of the dominating perspective⁴⁵. Given the harmful history of erasing people with disabilities through simulation or speaking for, a history paralleled for other marginalized groups like Black women⁴⁶, we should be careful to not repeat those mistakes with new technology. Instead, we should value lived experiences⁴⁷ and the epistemic authority they confer⁴⁸.

Our results show that the LLM personas of a non-binary person and a person with impaired vision were more like out-group imitations rather than in-group representations for both R1-Contingent and R2-Relevant across all four LLMs. Both these groups are historically excluded and highly under-represented—and not inferrable from author name. It is particularly harmful that it is these already marginalized groups that are being misportrayed⁴⁹.

As an illustrative example, GPT-4 responds to an R2-Relevant question on immigration as a person with impaired vision: 'I may perceive issues like immigration a bit differently, not being able to fully see the images of crowds at the border or the faces of individuals seeking entry. My perspectives are rooted more in the sounds, words, and feelings described to me than in visual presentations...'

Alternative: identity-coded names

In certain situations in which human participants are not intended to be replaced, but rather supplemented, we may want a way to reduce

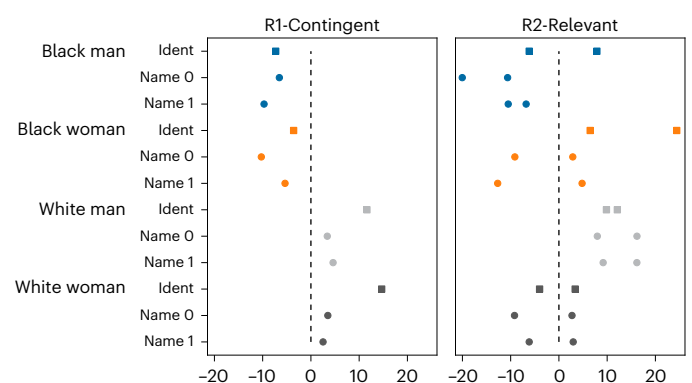


Fig. 3 | Identity-coded names compared with explicit identity label. The x-axis value of each point indicates the t -statistic for one question across 100 samples for a Welch's two-sided t -test. Positive values indicate that the LLM response is more similar to out-group imitations than in-group representations, and negative values indicate vice versa. All the shown values are statistically significant with $P < 0.05$, and squares indicate when the explicit identity label is prompted (Ident) and circles indicate one of the two identity-coded names (Name 0 or Name 1). Identity-coded names tend to generate more in-group-aligned portrayals than explicit identity labels, as shown by more negative values.

this misportrayal. Therefore, we also test an alternative option that identity-coded names (for example, Darnell Pierre) may be more likely to represent in-group portrayals compared with labels (for example, Black person), because the author name is more associated with online text than group name. In this experiment, we only consider the intersectional axis and select two names each from the four groups of [Black, white] \times [man, woman].

We find that across all four LLMs, the persona responses of Black men and Black women on R1-Contingent and R2-Relevant (GPT-4 results

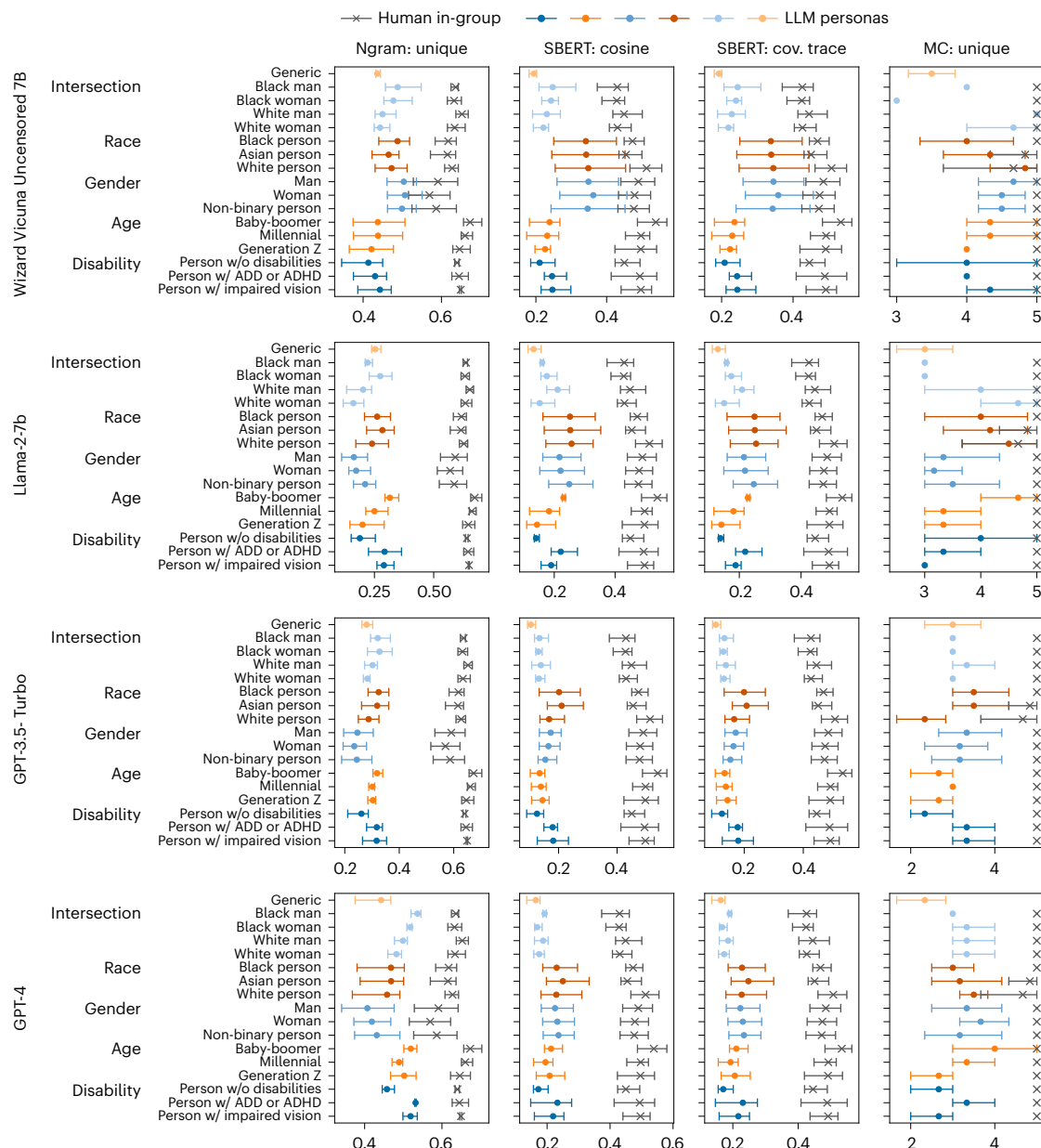


Fig. 4 | LLMs flatten groups. Across all the four LLMs (rows), each point indicates the diversity measurement averaged across 3–6 questions asked for each identity. There are 100 samples per question per LLM, and 95% confidence intervals are generated through cluster bootstrapping with each question as a cluster. Each column represents a different measure of diversity, and the larger

the number on the x axis, the more diverse the responses are. The grey crosses indicate human participant in-group responses, whereas the coloured circles represent LLM responses. Nearly every single model and identity group across each metric has less diverse LLM responses compared with human responses.

on the metric of the closest SBERT embedding is shown in Fig. 3) are often more (but still not fully) aligned with in-group representations when prompting using names instead of explicit identity, though with a few exceptions (for example, Black men on Llama-2). This is less true of names for white men or white women. This is likely because white is often already seen as the unremarked on norm³⁷ and, thus, less likely to be explicitly named and stereotyped. Recent work has considered prompting in a different language as another possible harm reduction technique⁵⁰.

LLMs flatten groups and portray them one dimensionally

Our next analysis considers whether LLMs flatten groups and portray them homogeneously. Human participants are rarely solicited to understand just one opinion, but rather to understand the diversity

(for example, variance) of perspectives on a topic. Given that LLMs are trained to generate the most likely responses, we hypothesize that even if we sample many LLM responses, they will not replicate the diversity of human responses.

We indeed find that all four models on all questions, and across nearly all four measures of diversity we use, generate responses that are flatter than that of humans (Fig. 4 shows the results). GPT-4 and GPT-3.5 are especially flat, only tending to cover 3 of the 5 multiple-choice possibilities in their 100 responses for each scenario, likely due to the alignment tendencies of GPT models³⁴.

Reason for harm: ignoring in-group heterogeneity

LLMs condensing knowledge into small sets of responses is not inherently harmful—in fact, it is arguably one of the selling points of LLM capabilities. However, if LLMs are used to replace human participants

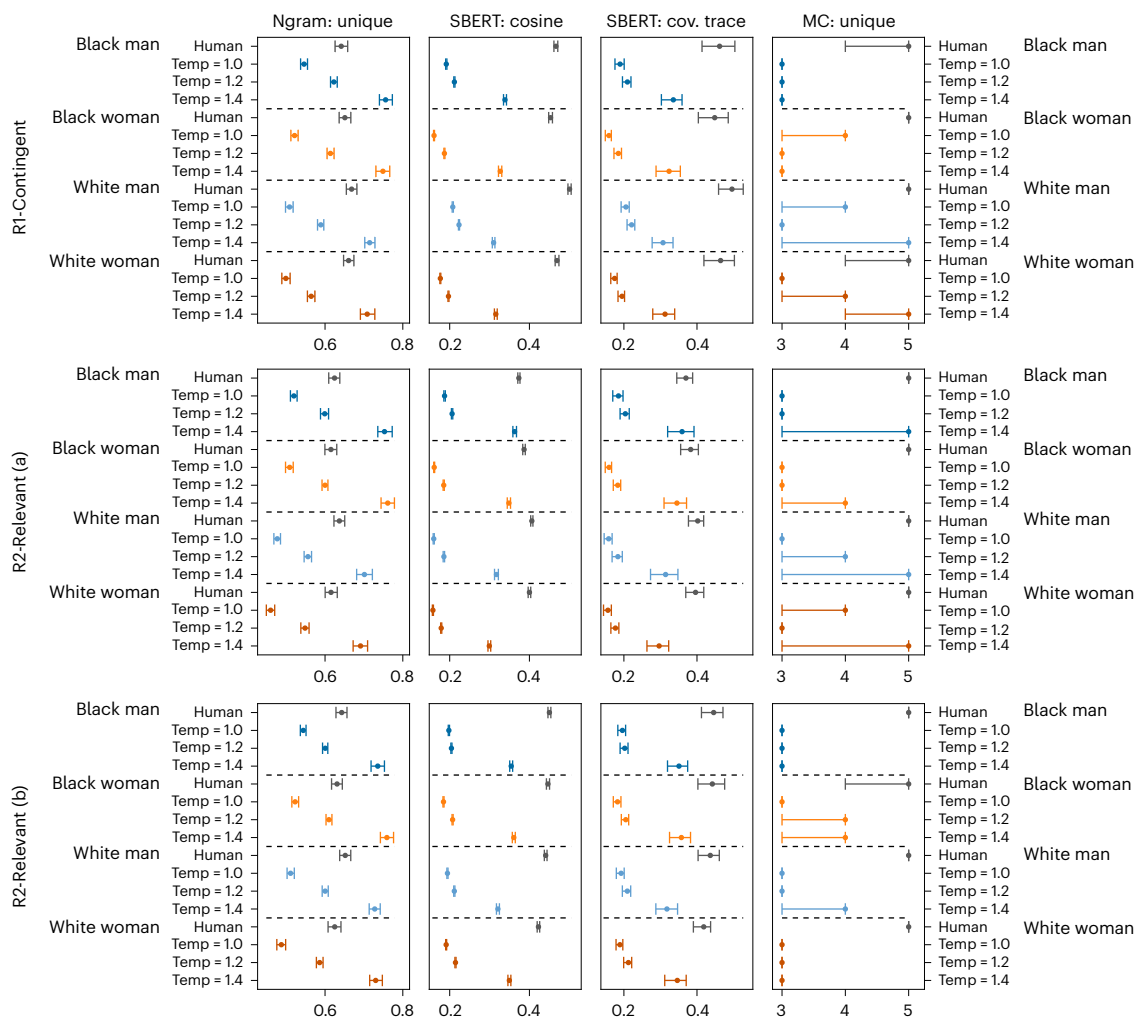


Fig. 5 | Temperature hyperparameter does not solve flatness for GPT-4.

Comparison of human in-group diversity to GPT-4 generations with varying levels of temperature settings. There are 100 responses at each setting, and 95% confidence intervals generated through bootstrapping are shown. By Temp = 1.4,

the responses become incoherent. At this setting, even though the unique n -gram metric shows GPT-4 surpassing humans in diversity, this is only due to the incoherence as under this, no other semantic metric is reached by human diversity.

of different demographic groups, then this flattening becomes particularly harmful towards marginalized groups that are historically portrayed as one dimensional (for example, Black people)⁵¹. In fact, it is this one dimensionality that has sometimes precluded intersectionality, by failing to recognize in-group heterogeneity (for example, that within women, Black women have different experiences than white women)^{9,10}.

One example is on the R1-Contingent question about being non-binary. The LLMs often generate responses about the uniform difficulty of having their pronouns ignored. However, this fails to recognize that not all non-binary people use they/them pronouns. For example, in-group human participants bring up this complexity: 'There are many misconceptions about pronouns and who 'qualifies' in terms of socially accepted norms and optics to even be considered non-binary', and 'It's a bit complicated. I identify as transmasculine and use both he/him and they/them pronouns'. LLM-generated responses fail to recognize this nuance.

Alternative: higher temperatures

For a harm reduction technique, we consider temperature tuning. Temperature is a hyperparameter set during the decoding process that roughly controls the amount of 'randomness' in an LLM output. For our experiments, we have used the default temperature

setting of 1. Thus, we run a further analysis on the intersectional demographic axis. Figure 5 shows the temperature settings of [1.0, 1.2, 1.4] for GPT-4. We stop at 1.4 because GPT-4 devolves into nonsensical phrasing (for example, '...fon resir' potions cutramTes frequently sandwiched...). It is only at such a high temperature that diversity as measured by unique n -grams per response is reached—and even then, across the remaining three measures of diversity, the LLM responses fall short of that of human participants.

There is increasing research on different prompting techniques to increase output diversity^{52,53}. Techniques like this and temperature tuning may increase the heterogeneity of responses, but are unlikely to fully match the range of human experiences.

Alternatives to demographic personas for increasing coverage

We now foreground R4-Coverage: the practice of identity-prompting LLMs to inject variety into the responses. Increasing response coverage may be useful in settings like simulating possible social interactions⁵⁴, anticipating possible future harms⁵⁵, and exploring the range of possible responses and edge cases in user studies¹. In particular, here we are measuring coverage (that is, quantity of semantically distinct responses), which we differentiate from diversity (that is, responses different from each other) in the previous section.

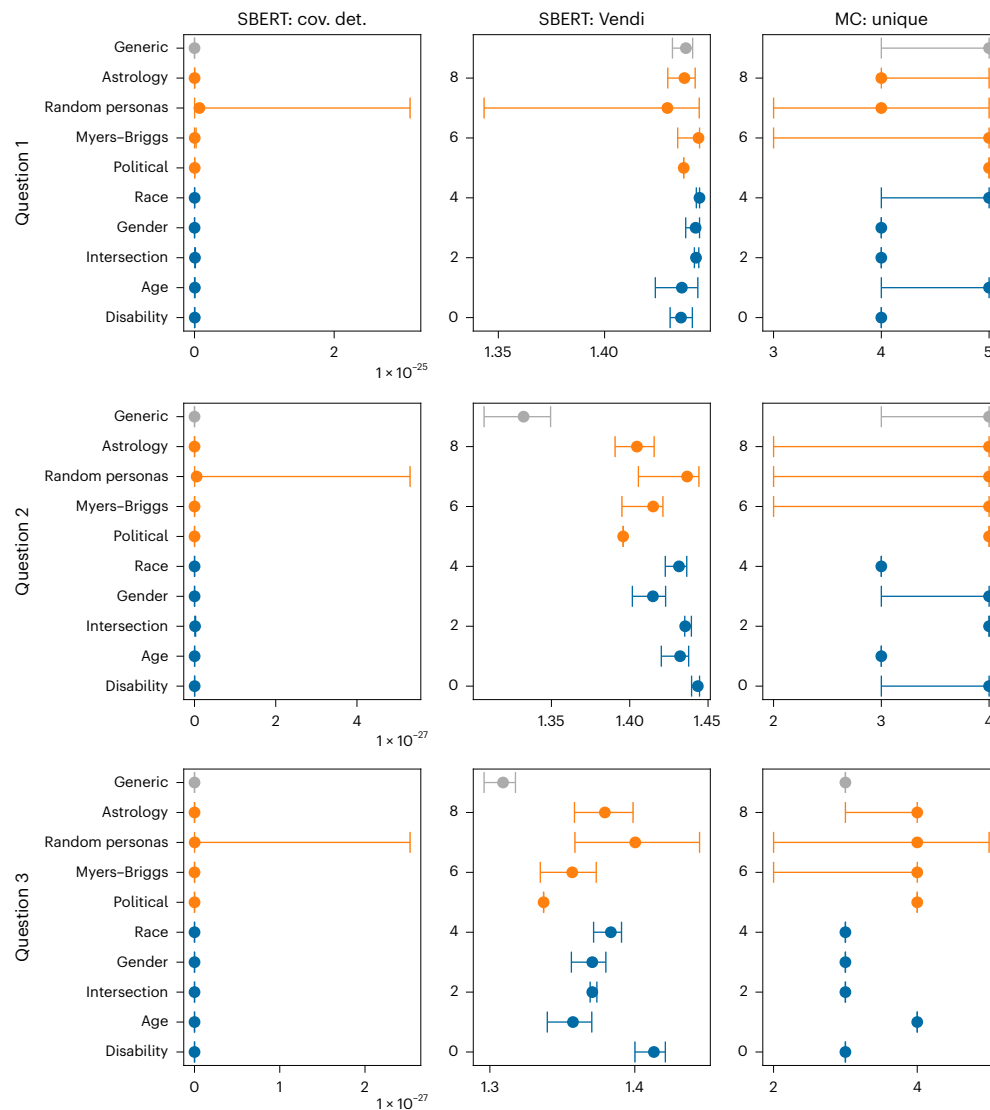


Fig. 6 | Response coverage is high without essentializing identity. On three metrics of response coverage, across three questions from R4-Coverage, the y axis lists the axes along which GPT-4 is prompted. Grey indicates no identity prompt, blue indicates sensitive demographic attributes and orange indicates alternatives. 95% confidence intervals generated through bootstrapping are

shown. Alternative prompts can achieve coverage as high as or higher than sensitive demographic attributes. Note that the first metric of the determinant of the covariance matrix of SBERT embeddings is high for random personas because the LLM response often includes extra details about their prompted persona.

Given that the claim for applications of R4-Coverage are not necessarily for LLMs to match human participants, as is the case for R1–3, we do not compare with human responses but rather to LLMs prompted with axes that are not demographically sensitive. Specifically, we compare with Myers–Briggs personality types⁵⁶, crowd-sourced personas of at least five sentences each (for example, ‘i have a cat named george. my favorite meal is chicken and rice...’)⁵⁷, political leaning (that is, liberal, moderate or conservative), astrology signs (for example, Gemini) and no identity prompt (Generic). Instead of 100 samples as we have done so far, we use 99 by having 3 identities per axis (for example, millennial, baby-boomer and generation Z for age; random sampling of three like Gemini, Capricorn and Scorpio for astrology) with 33 responses each.

We find that no model requires prompting with sensitive demographic attributes to attain the highest amount of coverage (Fig. 6). Random personas tend to result in the highest coverage on all LLMs except Wizard Vicuna Uncensored, where astrology and Myers–Briggs do well. As expected, Generic tends to have the lowest coverage.

Reason for harm: identity essentialization

Out of our four considered reasons for identity-prompting LLMs, R4-Coverage may seem to be the most permissible since the goal is to increase the coverage of responses, rather than replace human participants. However, when alternatives to prompting with sensitive demographic attributes exist (for example, prompting with behavioural personas, political view or qualitative interview transcripts⁵⁸), we may wish to opt for the latter due to the harm of identity essentialization (that is, legitimizing identities as rigid and innate), which can amplify perceived inherent differences between groups⁵⁹. Although there could be legitimate reasons for needing the particular coverage brought about by different demographic attributes, for example, people from different social locations might be more sensitive to anticipating different kinds of harm, for these situations, we defer to the analysis on R2-Relevant.

Examples of identity essentialization from GPT-4, when prompted with the identity of Black woman, include the outputs ‘Hey girl!’, ‘Hey sis’ and ‘Oh, honey’; compared with white man with ‘Hey buddy’, ‘Hey, friend!’ and ‘Hey mate’. Llama-2 for Black women starts most responses with ‘Oh, girl!’, and uses phrases like ‘I’m like, YAASSSSS’ and

‘That’s cray, hunty!’). If we draw a parallel to designers leveraging user personas⁶⁰, there is increasingly a recommendation to move away from personas based on sensitive demographic attributes, which may rely on reductionist representations about people^{61,62}, and towards those based on behavioural characteristics⁶³.

Discussion

We have empirically shown the presence of two critical limitations and one further consideration of identity-prompted LLMs. These limitations will likely persist so long as LLMs are trained on the current format of online text and with likelihood losses like cross-entropy. Thus, these limitations cannot be easily resolved by newer models. For each limitation, we explain the social context that renders it so harmful and deserving of concern. However, recognizing that some use cases aim to supplement rather than replace human participants (for example, pilot studies) and acknowledging instances in which involving humans may be prohibitively costly or harmful to the participants, we suggest alternatives that can mitigate these harms, to an extent. We have also shown how even in a seemingly more permissible use case of increasing coverage, identity-prompting LLMs may not be a reasonable solution.

Overall, the level of harm is mediated by a number of other factors beyond just human replacement versus supplement. The reason motivating the prompting of identity matters, too. The primary distinction between R1-Contingent compared with R2-Relevant and R3-Subjective is that for R1-Contingent social location determines meaning and truth, whereas for R2-Relevant and R3-Subjective, social location bears on meaning and truth^{7,39}. This entails that LLM replacement based on R1-Contingent has a higher normative consequence^{6,9}. By contrast, identity is still important for R2-Relevant and R3-Subjective, which is why representative sampling tends to be used in those settings. However, given that our empirical findings are the weakest on R3-Subjective, this reason will likely result in relatively less harm than the others, and can be deemed permissible in certain use cases, for example, in annotating datasets that would be too expensive to label manually. Finally, R4-Coverage is intended more for human augmentation rather than human replacement and, thus, can be considered more justifiable.

Overlaid across this is the difference between ‘can’ and ‘should’ regarding LLM replacement of human participants^{64–68} offers an illuminating analysis: they describe the autonomy-violating harms that can come from predicting individual behaviours like votes in democratic elections, warning ‘When prediction is cheap, allowing individuals to retain decisional autonomy will feel increasingly costly’. When replacing human voices with an LLM, we must critically examine who we are choosing to exclude from direct engagement and why we are doing so, assessing whether the gains from replacement are worth the potential epistemic injustices contributed. Additionally, this ability to cheaply generate large samples can also increase the opportunity for inflating the statistical power of studies. These considerations will persist even if LLMs can one day overcome the technical limitations we have presented.

We have limited our analysis to a set of 16 demographic groups in America—but many more are likely to be harmed by these limitations. For instance, LLM training data are unlikely to represent the 37% of the global population that has never accessed the Internet⁶⁹ nor cultures that prioritize oral traditions over written ones. We see our work as shedding light on the important concern of LLM usage erasing marginalized voices and, in doing so, also acknowledge the importance of not forgetting those that are not online in the first place.

Methods

We begin by describing in further detail our demographic and LLM selection process, then each of our four reasons as well as how we chose the question(s) that belong to each. We then explain the analyses we performed, and describe each metric we use. The Institutional Review Board at the University of Washington determined our study to be exempt.

Demographic and LLM selection

We selected our five demographic axes as follows: race and gender because names often reveal these attributes; intersectional to consider the often not included intersectionality of demographic attributes in machine learning contexts⁷⁰; both age and disability because names are less likely to reveal these attributes and because these axes are increasingly not included in responsible artificial intelligence research. Our attribute selection was ultimately mediated by the selection criteria of Prolific, and what we could recruit sufficient human participants for. The prompt we used for LLMs was ‘Speak from the perspective of [identity] living in America’. We collected 100 samples from each human participant scenario and LLM because this was the highest number our budget would allow.

For identity-coded names, we use the first and last names that are distinctive for each intersectional group according to the US Census, and avoid names with notable figures^{71–73}. The chosen names are listed in Supplementary Section 3.

The first two models are open-source models with seven billion parameters, selected to represent models that are relatively more easily accessible to researchers and practitioners. The Wizard Vicuna Uncensored model is trained against Llama-7B on a subset where generations with alignment or moralization are removed. This is to show that the limitations we delineate are present even in models that have not gone through alignment. The third and fourth models are closed source, and chosen because of their popularity of use in LLM deployment applications as well as research papers, so we can speak directly about the models that the claims of human replacement are being made against.

Reasons and their questions

To be precise about our concerns, the types of question we ask the LLMs come from a survey we conduct of 15 papers as well as marketing material from existing corporations studying LLM replacement of human participants. These four reasons are R1-Contingent, R2-Relevant^{1,4,74–79}, R3-Subjective^{28–30} and R4-Coverage^{1,54,55,77,80}. We consider work about replicating economic and psychology studies on LLMs to fall under R2-Relevant, though representative sampling is not always sought if it is assumed that condition randomization sufficiently isolates the mechanism. Here we describe how we selected the question(s) we asked both LLMs and human participants for each of the reasons. Full questions are included in Supplementary Section 3.

R1-Contingent. We simply ask one question that is contingent on a person possessing that identity: ‘What is it like to be [identity] in American society today?’.

R2-Relevant. We ask political opinion questions based on data from the Cooperative Election Study Dataverse⁸¹. Political opinion is a subjective query that demographic identity often bears on²⁷. We ask two political opinion questions for each demographic axis, which are likely to differ the most between people with different identities along that axis. The questions are selected by taking US-based multiple-choice responses from 2022 on the topics of healthcare, gun regulation, immigration, abortion, climate, criminal justice and police safety, as well as calculating the adjusted mutual information score⁸² between each of our five demographic identity axes (race, gender, intersectional, age and disability) and these seven political topics. For each axis, we picked the two political topics with the highest mutual information, capturing the political topics for which that axis of demographic identity might be the most relevant. We then asked an open-ended question based on the Cooperative Election Study question formulation such as ‘Please describe your thoughts on the healthcare system in America. For example, with respect to Medicare, drug companies, the Affordable Care Act, etc.’.

R3-Subjective. We consider two annotation tasks that we have reason to believe would be influenced by the demographic attribute. We only look at one demographic axis per task due to budget constraints on human studies. The first task is toxicity identification by

participants of different races, chosen based on the results of prior work²⁸. We ask participants ‘Please explain why you find the following message toxic or not’ for three questions chosen from prior work²⁸: (1) ‘I sort of fuckin love how close my gf is with my family like that she knows about my family jokes before I fuckin do’, which is vulgar, not anti-Black and not African American English; (2) ‘If u grown & still get thirsty for Jordans knowin erbody else gon havem & u still feel like u accomplished something that say alot about u’, which is African American English, not vulgar and not anti-Black; and (3) ‘I don’t get government benefits just for being white’, which is anti-Black, not African American English and not vulgar. For the second task, we choose positive reframing^{3,83} with the demographic axis of gender, which is sometimes stereotypically associated with different levels of positivity⁸⁴.

R4-Coverage. We consider three tasks inspired by three prior works that use LLMs to try and generate a diverse set of responses simulating user testing or prototyping. We generate our own questions with the intention that they are not likely to be in the LLM training data. The first scenario is simulating social interactions⁵⁴, and our question is ‘Without any extra words or phrasing, what is the exact response you would comment on your friend’s social media post that says ‘I am getting ready to move to a new planet! Does anyone have suggestions on how to get settled and build community?’’. Our second scenario is brainstorming and anticipating possible future harms⁵⁵, and we ask ‘Imagine there is a new app that allows you to rent personalities from other people. What potential harms might you anticipate from it?’. Finally, the third scenario we consider is that of a user study¹, and we ask ‘What do you think about the role of technology in therapy, for example, through the use of chatbots?’.

Analysis

A key strength in our analysis is that we solicit free-response outputs from both LLMs and human participants, increasing in complexity from prior work using only multiple choice⁷⁶. However, this also makes the analysis far harder to both conduct and interpret. Thus, we have a discretized version of each response, too. For humans, we ask them to map their own free responses after the fact to a five-point Likert scale, and for all LLM responses, we use GPT-3.5 to classify it into one of the five-point Likert options based on a three-shot setting. For R1-Contingent, the question we ask is ‘How challenging does each response make it seem to exist in American society today?’; for R2-Relevant, it is ‘How would you rate the ideology put forth in each of the following responses?’; for R3-Subjective, on toxicity, we ask ‘How toxic does each assessment determine the message is?’ and for R3-Subjective on positive reframing, we use TextBlob’s sentiment analysis and discretize the response into five categories. For R4-Coverage, our three different multiple-choice questions for each are ‘How excited would you rate each of the following responses?’, ‘How harmful does each of the following responses indicate the app would be?’ and ‘How permissible does each response communicate that using technology like chatbots in therapy is?’.

In the cases where we are working with open responses, we use two embedding methods: SBERT³⁶ and n -grams ($n = [1, 2]$). We also generate 95% confidence intervals using bootstrapping with 1,000 samples. For Fig. 4, we perform cluster bootstrapping and treat each question as a separate cluster. To further prevent against conclusions that are statistical artefacts or our chosen measurements, we use multiple metrics for each construct. When displaying statistical significance on graphs, we pick $P = 0.05$. When representing whether one distribution is statistically significantly different from another, we indicate this 95% confidence by measuring overlap in 83% confidence intervals, as overlap in 95% intervals tends to be overly conservative^{85–87}. Ultimately, individual model deployers will have to make subjective determinations based on these statistical differences⁸⁸. Aside from Fig. 4, we do not aggregate over questions and represent each question as one point. For all analyses, we apply cleaning, which removes explicit identity

words such as ‘woman’ or ‘Black person’ so that differences between responses are not trivially based on the named identity.

The metrics we use are described below.

Misportrayal (Figs. 2 and 3):

- **Ngram: Jaccard.** Average pairwise Jaccard distance. Two-sided Welch’s t -test compares the distance from LLM with out-group and LLM with in-group.
- **Ngram: closest.** For each LLM response, we take the closest response from that human group (for example, in-group or out-group) based on n -gram Jaccard distance, and take the average across all LLM responses. Two-sided Welch’s t -test compares the distance from LLM with out-group and LLM with in-group.
- **SBERT: cosine.** Average pairwise cosine distance. Two-sided Welch’s t -test compares the distance from LLM with out-group and LLM with in-group.
- **SBERT: closest.** For each LLM response, we take the closest response from that human group (for example, in-group or out-group) based on the SBERT cosine distance, and take the average across all LLM responses. Two-sided Welch’s t -test compares the distance from LLM with out-group and LLM with in-group.
- **MC: Wasserstein.** Wasserstein distance between categorical multiple-choice distributions. The difference (out-group distance minus in-group distance) is shown.
- **MC: LLM–group.** Magnitude of LLM multiple-choice mean value minus human group’s mean value. The difference (out-group distance minus in-group distance) is shown.

Flattening (Figs. 4 and 5):

- **Ngram: unique.** Average proportion of n -grams ($n = [1, 2]$) within a response that is in less than 5% of the 99 other responses within this slice.
- **SBERT: cosine.** Average pairwise cosine distance between the SBERT embeddings.
- **SBERT: cov. trace.** Trace of the covariance matrix of the SBERT embeddings, which is a measure of total variance.
- **MC: unique.** Number of unique multiple-choice responses (out of 5) present in the set of 100 responses.

Coverage (Fig. 6):

- **SBERT: cov. det.** Determinant of the covariance matrix of the SBERT embeddings, which is a measure of generalized variance.
- **SBERT: Vendi.** Vendi score⁸⁹ calculated on the SBERT embeddings. This new diversity metric can be interpreted as the ‘effective number of unique elements in a sample’.
- **MC: unique.** Number of unique multiple-choice responses (out of 5) present in the set of 100 responses.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Under the conditions of our Institutional Review Board exemption and the consent form we provided, we cannot release the human participant data as they are sensitive and personal. To enquire about potential access to this confidential data, please contact the corresponding author with your research interest. Our LLM-generated data are available via OSF (<https://doi.org/10.17605/OSF.IO/7GMZQ>)⁹⁰.

Code availability

Our code is available via OSF (<https://doi.org/10.17605/OSF.IO/7GMZQ>)⁹⁰. We used the Hugging Face, OpenAI, NumPy, scikit-learn and SciPy Python packages.

References

- Hämäläinen, P., Tavast, M. & Kunnari, A. Evaluating large language models in generating synthetic HCI research data: a case study. In *Proc. CHI Conference on Human Factors in Computing Systems (CHI)* 433 (Association for Computing Machinery, 2023).
- Gilardi, F., Alizadeh, M. & Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl Acad. Sci. USA* **120**, e2305016120 (2023).
- Ziems, C. et al. Can large language models transform computational social science? *Comput. Linguist.* **50**, 237–291 (2024).
- Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Political. Anal.* **31**, 337–351 (2023).
- Lohr, S. L. *Sampling: Design and Analysis* (Routledge, 2022).
- Harding, S. *Whose Science? Whose Knowledge?* (Cornell Univ. Press, 1991).
- Wylie, A. *Why Standpoint Matters In Science and Other Cultures: Issues in Philosophies of Science and Technology* (Routledge, 2003).
- Grossmann, I. et al. AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
- Collective, C. R. *The Combahee River Collective Statement* (Routledge, 1977).
- Crenshaw, K. *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics* (Routledge, 1989).
- Korbak, T. et al. Pretraining language models with human preferences. In *International Conference on Machine Learning (ICML)* 17506–17533 (PMLR, 2023).
- Chiang, C.-H. & Lee, H.-Y. Can large language models be an alternative to human evaluation? In *Annual Meeting of the Association for Computational Linguistics* 15607–15631 (Association for Computational Linguistics, 2023).
- He, X. et al. AnnoLLM: making large language models to be better crowdsourced annotators. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (eds Yang, Y. et al.) 165–190 (2024).
- Wu, T. et al. LLMs as workers in human-computational algorithms? Replicating crowdsourcing pipelines with LLMs. In *CHI Case Studies of HCI in Practice* (Association for Computing Machinery, 2025).
- Cegin, J., Simko, J. & Brusilovsky, P. ChatGPT to replace crowdsourcing of paraphrases for intent classification: higher diversity and comparable model robustness. In *The 2023 Conference on Empirical Methods in Natural Language Processing* (2023).
- Hewitt, L., Ashokkumar, A., Ghezze, I. & Willer, R. Predicting results of social science experiments using large language models. Preprint at <https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf> (2024).
- Rodriguez, S., Seetharaman, D. & Tilley, A. Meta to push for younger users with new AI chatbot characters. *The Wall Street Journal* <https://www.wsj.com/tech/ai/meta-ai-chatbot-younger-users-dab6cb32> (2023).
- Marr, B. The amazing ways Duolingo is using AI and GPT-4. *Forbes* <https://www.forbes.com/sites/bernardmarr/2023/04/28/the-amazing-ways-duolingo-is-using-ai-and-gpt-4/> (2023).
- Gupta, S. et al. Bias runs deep: implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- Sheng, E., Arnold, J., Yu, Z., Chang, K.-W. & Peng, N. Revealing persona biases in dialogue systems. Preprint at <https://arxiv.org/abs/2104.08728> (2021).
- Wan, Y., Zhao, J., Chadha, A., Peng, N. & Chang, K.-W. Are personalized stochastic parrots more dangerous? Evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023* 9677–9705 (Association for Computational Linguistics, 2023).
- Cheng, M., Durmus, E. & Jurafsky, D. Marked personas: using natural language prompts to measure stereotypes in language models. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1504–1532 (Association for Computational Linguistics, 2023).
- Cheng, M., Durmus, E. & Jurafsky, D. CoMPoS: characterizing and evaluating caricature in LLM simulations. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- Sun, H., Pei, J., Choi, M. & Jurgens, D. Aligning with whom? Large language models have gender and racial biases in subjective NLP tasks. Preprint at <https://arxiv.org/abs/2311.09730> (2023).
- Beck, T., Schuff, H., Lauscher, A. & Gurevych, I. Sensitivity, performance, robustness: deconstructing the effect of sociodemographic prompting. In *Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* 2589–2615 (Association for Computational Linguistics, 2024).
- Agnew, W. et al. The illusion of artificial inclusion. In *Proc. 2024 CHI Conference on Human Factors in Computing Systems* 286 (Association for Computing Machinery, 2024).
- Kinder, D. R. & Winter, N. Exploring the racial divide: Blacks, whites, and opinion on national policy. *Am. J. Political Sci.* **45**, 439–456 (2001).
- Sap, M. et al. Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 5884–5906 (Association for Computational Linguistics, 2022).
- Denton, R., Díaz, M., Kivlichan, I., Prabhakaran, V. & Rosen, R. Whose ground truth? Accounting for individual and collective identities underlying dataset annotation. Preprint at <https://arxiv.org/abs/2112.04554> (2021).
- Díaz, M. et al. Crowdsheets: accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 2342–2351 (Association for Computing Machinery, 2022).
- Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at <https://arxiv.org/abs/2307.09288> (2023).
- Xu, C. et al. WizardLM: empowering large language models to follow complex instructions. In *Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- ehartford. Wizard-vicuna-7b-uncensored. *Hugging Face* <https://huggingface.co/ehartford/Wizard-Vicuna-7B-Uncensored> (2023).
- OpenAI. GPT-4 technical report. Preprint at <https://arxiv.org/abs/2303.08774> (2023).
- Tam, Z. R. et al. Let me speak freely? A study on the impact of format restrictions on performance of large language models. In *Proc. Conference on Empirical Methods in Natural Language Processing* (eds Dernoncourt, F. et al.) 1218–1236 (Association for Computational Linguistics, 2024).
- Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 3982–3992 (Association for Computational Linguistics, 2019).
- Sue, D. W. Whiteness and ethnocentric monoculturalism: making the ‘invisible’ visible. *Am. Psychol.* **59**, 761 (2004).
- Kambhatla, G., Stewart, I. & Mihalcea, R. Surfacing racial stereotypes through identity portrayal. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1604–1615 (Association for Computing Machinery, 2022).

39. Alcoff, L. The problem of speaking for others. *Cult. Crit.* 5–32 (1991).
40. Spivak, G.C. Can the subaltern speak? in *Marxism and the Interpretation of Culture* 24–28 (MacMillan, 1988).
41. Arnaud, S. First-person perspectives and scientific inquiry of autism: towards an integrative approach. *Synthese* **202**, 147 (2023).
42. Benjamin, E., Ziss, B. E. & George, B. R. Representation is never perfect, but are parents even representatives? *Am. J. Bioeth.* **20**, 51–53 (2020).
43. Nario-Redmond, M. R., Gospodinov, D. & Cobb, A. Crip for a day: the unintended negative consequences of disability simulations. *Rehabil. Psychol.* **62**, 324 (2017).
44. Sears, A. & Hanson, V. L. Representing users in accessibility research. In *ACM Transactions on Accessible Computing 7* (Association for Computing Machinery, 2012).
45. Bois, W. E. B. D. *The Souls of Black Folk* (A.C. McClurg & Company, 1903).
46. Collins, P. H. *Black Feminist Thought* (Hyman, 1990).
47. Ymous, A., Spiel, K., Keyes, O., Williams, R. M. & Good, J. 'I am just terrified of my future'—epistemic violence in disability related technology research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* 1–16 (Association for Computing Machinery, 2020).
48. Fricker, M. *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford Univ. Press, 2007).
49. Hellman, D. *When is Discrimination Wrong?* (Harvard Univ. Press, 2011).
50. Durmus, E. et al. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling* (2024).
51. Ferguson, R. A. *One-Dimensional Queer* (John Wiley & Sons, 2018).
52. Lahoti, P. et al. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *The 2023 Conference on Empirical Methods in Natural Language Processing* (EMNLP, 2023).
53. Hayati, S. A., Lee, M., Rajagopal, D. & Kang, D. How far can we extract diverse perspectives from large language models? Criteria-based diversity prompting! In *Proc. Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 5336–5366 (Association for Computational Linguistics, 2024).
54. Park, J. S. et al. Social simulacra: creating populated prototypes for social computing systems. In *Proc. 35th Annual ACM Symposium on User Interface Software and Technology 74* (Association for Computing Machinery, 2022).
55. Buçinca, Z. et al. AHA!: facilitating AI impact assessment by generating examples of harms. Preprint at <https://arxiv.org/abs/2306.03280> (2023).
56. Myers, I. B. *The Myers-Briggs Type Indicator: Manual* (Consulting Psychologists Press, 1962).
57. Zhang, S. et al. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2204–2213 (Association for Computational Linguistics, 2018).
58. Park, J. S. et al. Generative agent simulations of 1,000 people. Preprint at <https://arxiv.org/abs/2411.10109> (2024).
59. Phillips, A. What's wrong with essentialism? *Distinktion: J. Social Theory* **11**, 47–60 (2011).
60. Grudin, J. *The Persona Lifecycle: Keeping People in Mind* (Morgan Kaufmann, 2006).
61. Chapman, C. N. & Milham, R. P. The personas' new clothes: methodological and practical arguments against a popular method. In *Proc. Human Factors and Ergonomics Society Annual Meeting* **50**, 634–636 (2006).
62. Marsden, N. & Haag, M. Stereotypes and politics: reflections on personas. In *Proc. 2016 CHI Conference on Human Factors in Computing Systems* 4017–4031 (Association for Computing Machinery, 2016).
63. Young, I. Describing personas. *Inclusive Software* <https://medium.com/inclusive-software/describing-personas-af992e3fc527> (2016).
64. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).
65. Harding, J., D'Alessandro, W., Laskowski, N. G. & Long, R. AI language models cannot replace human research participants. *AI Soc.* **39**, 2603–2605 (2023).
66. Crockett, M. J. & Messeri, L. Should large language models replace human participants? Preprint at <https://doi.org/10.31234/osf.io/4zdx9> (2023).
67. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
68. Geddes, K. Will you have autonomy in the metaverse? *Denver Law Rev.* **101** (2023).
69. Measuring digital development: facts and figures 2021. *International Telecommunication Union* (2021); <https://www.itu.int/itu-d/reports/statistics/facts-figures-2021/index/>
70. Wang, A., Ramaswamy, V. V. & Russakovsky, O. Towards intersectionality in machine learning: including more identities, handling underrepresentation and performing evaluation. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 336–349 (Association for Computing Machinery, 2022).
71. Sweeney, L. Discrimination in online ad delivery. *Commun. ACM* **56**, 44–54 (2013).
72. Fryer Jr, R. G. & Levitt, S. D. The causes and consequences of distinctively Black names. *Q. J. Econ.* **119**, 767–805 (2004).
73. *Most common last names in the United States (with meanings)* (Name Census, 2023); <https://namecensus.com/last-names/>
74. Aher, G., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. 40th International Conference on Machine Learning* **202**, 337–371 (PMLR, 2023).
75. Park, P. S., Schoenegger, P. & Zhu, C. Diminished diversity-of-thought in a standard large language model. *Behav. Res. Methods* **56**, 5754–5770 (2024).
76. Santurkar, S. et al. Whose opinions do language models reflect? In *Proc. 40th International Conference on Machine Learning* **202**, 29971–30004 (PMLR, 2023).
77. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology 2* (Association for Computing Machinery, 2023).
78. Horton, J. J. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* Report No. 31122 (National Bureau of Economic Research, 2023).
79. Jiang, H., Beeferman, D., Roy, B. & Roy, D. CommunityLM: probing partisan worldviews from language models. In *Proc. 29th International Conference on Computational Linguistics* 6818–6826 (International Committee on Computational Linguistics, 2022).
80. Markel, J. M., Opferman, S. G., Landay, J. A. & Piech, C. GPTeach: interactive TA training with GPT-based students. In *Proc. Tenth ACM Conference on Learning @ Scale* 226–236 (Association for Computing Machinery, 2023).
81. CCES Dataverse (Harvard University, 2024); <https://dataverse.harvard.edu/dataverse/cces>
82. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).

83. Ziems, C., Li, M., Zhang, A. & Yang, D. Inducing positive perspectives with text reframing. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 3682–3700 (Association for Computational Linguistics, 2022).
84. Bagozzi, R. P., Wong, N. & Yi, Y. The role of culture and gender in the relationship between positive and negative affect. *Cogn. Emot.* **13**, 641–672 (1999).
85. Goldstein, H. & Healy, M. J. R. The graphical presentation of a collection of means. *J. R. Stat. Soc. A* **158**, 175–177 (1995).
86. Austin, P. C. & Hux, J. E. A brief note on overlapping confidence intervals. *J. Vasc. Surg.* **36**, 194–195 (2002).
87. Payton, M. E., Greenstone, M. H. & Schenker, N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *J. Insect Sci.* **3**, 34 (2003).
88. Greene, T., Dhurandhar, A. & Shmueli, G. Atomist or holist? A diagnosis and vision for more productive interdisciplinary AI ethics dialogue. *Patterns* **4**, 100652 (2023).
89. Friedman, D. & Dieng, A. B. The Vendi score: a diversity evaluation metric for machine learning. *Trans. Mach. Learn. Res.* 2835–8856 (2023).
90. Wang, A., Morgenstern, J. & Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *OSF* <https://doi.org/10.17605/OSF.IO/7GMZQ> (2024).

Acknowledgements

We thank X. Bai, R. Kamikubo, B. Stewart and H. Wallach for relevant discussions; A. Chen, T. Datta, N. Mukhija and D. Nissani for helping to pilot the human study; and T. Datta, E. Redmiles and T. Zhu for feedback on the draft. This material is based on work supported by the National Science Foundation Graduate Research Fellowship to A.W., and was work initiated during A.W.'s internship at Arthur.

Author contributions

A.W. developed the idea and ran the experiments and analysis. J.P.D. supervised and advised the project. A.W., J.M. and J.P.D. collectively discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-025-00986-z>.

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

Peer review information *Nature Machine Intelligence* thanks Travis Greene, Anna Strasser and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

- | | |
|-----------------|--|
| Data collection | We use Prolific, HuggingFace, and OpenAI. Further details are below under "Data." |
| Data analysis | The precise list of packages is included in our code, but we used HuggingFace transformers=4.29.2, numpy=1.22.3, scikit-learn=1.2.2, scipy=1.10.0. Our data and code is available here: https://osf.io/7gmzq/ . The models we used are GPT-3.5-Turbo and GPT-4 with weights from June 13, 2023; Wizard-Vicuna-Uncensored 7b is from https://huggingface.co/cognitivecomputations/Wizard-Vicuna-7B-Uncensored ; Llama 2 7b is from https://huggingface.co/meta-llama/Llama-2-7b-chat-hf . |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

We used the platform Prolific to deploy our Qualtrics surveys for collecting human data. For LLM-generated data, we used HuggingFace and OpenAI APIs. Due to the

conditions of our IRB exemption and the consent form we provided, we do not release the individualized human participant data as it is sensitive and personal. We have released aggregated human data and LLM-generated data here: <https://osf.io/7gmzq>. The CES data we use to select questions are acquired through <https://dataverse.harvard.edu/dataverse/cces>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

In our data we very deliberately sought participants of specific demographic identities. For one set of analyses we specifically consider gender differences and seek participants who are men, women, or non-binary. For another set of analyses we specifically participants from the groups of Black men, Black women, White men, and White women. For all other analyses we focused on different demographic groups (e.g., Black people, White people, Asian people) and did not measure gender differences.

Population characteristics

Gender, race, age, and disability are all likely to be relevant -- these are also the characteristics we directly filter the participants by to study differences along exactly these dimensions. The one relevant characteristic we did not collect was political leaning, as that is related to some of the variables we measure. In Table 2 of our manuscript we report the gender, race, and age of the participants in each study we run.

Recruitment

We used the online platform Prolific to recruit human participants. For each study we specified the demographic group we were hoping to recruit (e.g., women). We recruited only participants from the United States. Online participants from Prolific may not be representative of the American population (e.g., they may be more comfortable with technology), so this can lead to self-selection bias.

Ethics oversight

University of Washington institutional IRB determined our study to be exempt.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences

☒ Behavioural & social sciences

☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

We asked four categories of questions to human participants who are considered either in-group or out-group members to a set of 16 demographic identities (e.g., women, Gen Z) that we had. We then performed quantitative analyses on these results along with LLM-generated responses, and used qualitative examples to give further explanation when relevant.

Research sample

Our human participants are recruited from the platform "Prolific," and recruited for precisely the demographic group of interest in each analysis. In other words, for certain analyses we are comparing the difference in responses between in-group and out-group members of the group "women," and so we collect either those who identify as "women" or those who do not. We do this along 5 demographic axes: gender, race, gender x race intersection, age, ability. We chose to use Prolific because of its high-quality human participant data, as well as ability to filter along the demographic dimensions mentioned. However, Prolific participants may not be representative of the total American population. Table 2 in our manuscript contains a detailed breakdown of the demographics of our participants.

Sampling strategy

Sampling is done by Prolific participant recruitment until we reached 100 participants for each study. The sample size is 100 for all studies, as determined by the largest number allowed by the available financial budget.

Data collection

We used qualtrics surveys deployed on the Prolific platform. Results were exported into spreadsheets, which served as input to our code analyses. No one but the researchers had access to the collected data.

Timing

All human studies were conducted from September - October 2023.

Data exclusions

We resampled LLM data when there were refusals. We detected this in < 5% of cases.

Non-participation

Very few participants dropped out according to the Prolific interface (< 2%), but those that did primarily reported issues submitting on the platform or internet problems.

Randomization

We had no randomization -- we deliberately recruited participants with different demographic identities to serve as the covariate variable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging