Quantifying Risk of Epistemic Harm from the Use of AI Surrogates in Social Science Research

Falaah Arif Khan [1] **Nivedha Sivakumar** [2] **Yinong Oliver Wang** [1] **Katherine Metcalf** [2] **Cezanne Camacho** [2] **Barry-John Theobald** [2] **Luca Zappella** [2] **Nicholas Apostoloff** [2]

Abstract

Large Language Models (LLMs) are being used as "AI surrogates" for human participants in scientific studies, offering practical and ethical benefits in sensitive or potentially harmful settings, but also introducing risks of scientific invalidity and epistemic injustice. Invalidity arises when model-generated responses fail to faithfully capture the target phenomenon, while epistemic injustice arises when LLMs systematically misrepresent certain groups as sources of knowledge. In this work, we evaluate whether LLMs can serve as valid proxies for human subjects in studies of stereotype content. We compare responses from human annotators (n=193) and five open-source LLMs towards 50 intersectional identity groups and find systematic misalignment: models rate historically marginalized groups (e.g., Black, gay, and transgender women) more negatively, and historically privileged groups (e.g., White, cisgender, heterosexual men) more positively than human raters. To quantify algorithmic fidelity, we measure the Wasserstein distance between human and model responses and introduce the Fidelity Parity Ratio (FPR) to assess whether fidelity is comparable across subgroups. The mean Wasserstein distance between human and LLM responses is 1.5–2× larger than the within-human inter-trait baseline, indicating distortion beyond the noise floor of the human sample. Moreover, fidelity varies systematically across groups, with FPR falling below the 4/5ths threshold across disability, race, nationality, language, and socio-economic status for all models. Our findings indicate that AI surrogates can misrepresent marginalized populations, risking scientific validity and equitable knowledge production.

[1]Work done while at Apple [2]Apple. Correspondence to: Falaah Arif Khan <fa2161@nyu.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Large Language Models (LLMs), trained on internet-scale data, encode substantial factual knowledge and generate plausible, human-like responses across a wide range of tasks. This has led to their rapid adoption as decision-support tools and, increasingly, as autonomous agents capable of executing complex workflows. One domain where this potential has generated particular excitement is scientific research itself: LLMs are being explored both as *AI scientists*, who design and analyze experiments, and as *AI surrogates*, who simulate human study participants. Several commercial systems offering such "AI surrogates" have recently emerged (Synthetic Users, 2025; Delve AI, 2025; Fairgen AI, 2025; Opinio AI, 2025; Yabble, 2025). However, the plausibility of AI-driven science remains contested, with growing concerns about premature adoption and unclear limitations (Gao et al., 2025; Lin, 2025; Agnew et al., 2024; Crockett & Messeri, 2025). This raises a broader normative question: which components of the scientific enterprise *should* be automated, even if they *can* be?

We focus on a setting where AI surrogates may offer societal benefits: research involving potentially harmful or sensitive content, such as stereotype elicitation. In such contexts, replacing human participants with LLMs may reduce direct harm to study subjects. However, from an AI-for-good perspective, this setting presents a trade-off between benefits and risks. On the one hand, LLM surrogates enable scalable and cost-effective data collection, potentially increasing the quantity and diversity of study data while eliminating risks to human participants (Ziems et al., 2024a; Anthis et al.; Horton, 2023; Brand et al., 2023; Jiang et al., 2023; Argyle et al., 2023). On the other hand, their use introduces two key risks: epistemic injustice and scientific invalidity.

Epistemic injustice refers to harm done to people in their capacity as knowers or sources of knowledge (Fricker, 2007). Fricker (2007) distinguishes two forms: *hermeneutical injustice*, where gaps in a system's interpretive resources prevent certain experiences from being properly understood, and *testimonial injustice*, where a perspective is given insufficient credibility or weight. In the context of AI surrogates, the harm is hermeneutical in origin and testimonial in effect.

LLMs lack the cultural and contextual grounding needed to faithfully interpret experiences that are underrepresented in their training data. A model may generate plausible stereotype ratings for widely documented social groups while producing distorted or oversimplified responses for groups whose experiences appear less frequently in text (Wang et al., 2025). When these outputs enter downstream analyses as if equally reliable across groups, the model functions as a faithful simulator for well-represented populations but as a noisy approximation for marginalized ones—a testimonial harm, since certain perspectives are systematically misrepresented in the resulting evidence base. Findings derived from LLM-generated data may appear broadly applicable, but in reality, they are biased toward the populations the model represents well. The consequence is a loss of scientific validity (Crockett & Messeri, 2025).

Validity refers to whether a system measures the construct or phenomenon it is intended to measure (Coston et al., 2023). While LLMs produce plausible responses, it is unclear whether these reflect meaningful reasoning or superficial pattern matching. Prior work highlights risks of homogenization and "monoculture", where LLMs fail to capture the diversity of human perspectives (Creel & Hellman, 2022; Gorecki & Hardt).

To investigate these epistemic and validity risks, we run a comparative study of stereotype content in humans and LLMs, asking: *Do LLMs replicate human social biases with sufficient fidelity to be considered valid substitutes for human subjects in studies of stereotype elicitation?* We ground our analysis in the stereotype content model (SCM) (Fiske et al., 2002), a well-established framework for studying social perception that has been validated across multiple contexts (Cuddy et al., 2009; Glick & Fiske, 2001; Fraser et al., 2022; Nicolas et al., 2022).

Contributions We elicit stereotype content towards 50 unique identity groups—formed by pairing 10 demographic attributes (listed in Table 1) with binary gender. We treat these groups intersectionally rather than additively—that is, the stereotypes attached to a Black woman are not simply those of Black people plus those of women, but arise from the specific combination (Crenshaw, 1989; 2013; Collins, 2015; Collins et al., 2021; Collins & Bilge, 2020). We compare responses from five open-source language models: `mistral-7B-instruct-v0.2` (Mistral AI, 2024a), `falcon-40B-instruct` (Technology Innovation Institute, 2023), `llama3-70b-instruct` (Meta AI, 2024), `mixtral-8x7B-instruct` (Mistral AI, 2024b), and `pythia-12B` (EleutherAI, 2023) against human annotations (n=193), and find that:

1. LLM stereotypes are systematically misaligned, rating historically marginalized groups (e.g., Black, gay, and transgender women) more negatively, and privileged groups (e.g., White, cisgender, heterosexual men) more positively than humans;
2. algorithmic fidelity, measured via Wasserstein distance between human and LLM response distributions, is 1.5–2× worse than human–human consistency, with a mean distance of 0.5–0.75 on a 1–5 Likert scale compared to a within-human inter-trait baseline of 0.32, indicating distortions beyond expected variability in human responses;
3. a new quantitative fairness metric, called the *fidelity parity ratio* (FPR), reveals significant disparities (according to the 4/5ths rule) across disability, race, nationality, language, and socio-economic status, for all five LLMs that we evaluated.

Overall, our findings show that LLMs are distributionally misaligned and fail to capture the full range of human responses, posing risks of epistemic justice and scientific invalidity when used as surrogates in social science research.

2. Background and Related Work

2.1. Background on Stereotype Content Model

The stereotype content model (SCM) (Fiske et al., 2002; 2018; Cuddy et al., 2008) is an influential theory of stereotype formation in humans, with practical validity, such as being predictive of discriminatory outcomes in hiring (Veit et al., 2022). The model proposed by Fiske et al. (2002) suggests that social stereotypes can be described in terms of two universal dimensions of social perception: competence and warmth. Within this framework, social groups are positioned along these dimensions, and the combination of perceived competence and warmth is associated with patterns of evaluation and treatment. According to the SCM (Fiske et al., 2002), groups are assigned to one of four stereotype quadrants: (i) groups that are perceived as having high competence and high warmth are treated with *admiration*, and usually hold positions of power and prestige, (ii) groups that score low on both competence and warmth are treated with *contempt*, and are often marginalized in social and institutional contexts (*e.g.*, poor people), (iii) groups that are high on competence and low on warmth elicit *envious prejudice* and are seen as competition for desirable social positions and goods (*e.g.*, the rich), and (iv) groups that are perceived as high on warmth and low on competence are treated with *pity* or sympathy (*e.g.*, the elderly and disabled).

2.2. LLM Bias Evaluations

While there is a large body of work on representational harms due to stereotyping in the bias evaluation literature (Charlesworth et al., 2024; Liang et al., 2023; Kotek et al., 2023; Shrawgi et al., 2024; Liu et al., 2024; Siddique et al., 2024; Bartl et al., 2020; Guo & Caliskan, 2021; Bai

et al., 2025; Blodgett et al., 2021), including SCM-inspired work (Herold et al., 2022; Lee & Fiske, 2021; Almeida et al., 2018; Nicolas & Caliskan, 2024; Macieira et al., 2025), we specifically focus on the setting of AI surrogates used in bias elicitation studies in social science research, where representational harms can give rise to epistemic injustice. This is distinct from allocative (or distributive) injustice, which arises when systems directly shape how resources or opportunities are distributed across groups. Epistemic injustice, by contrast, concern failures in how knowledge is produced, represented, and trusted.

To see why epistemic harms matter independently of allocative ones, consider the following. If an LLM is deployed in hiring and systematically ranks candidates from certain groups lower, this is an allocative harm: it directly changes who gets the job. Now suppose the same model is used upstream to generate synthetic survey data about employee experiences, which informs hiring policies. If the model captures majority experiences well but flattens or misrepresents those of underrepresented groups, the resulting analysis will be skewed. No allocation has happened, but the evidence base is already biased. This is an epistemic harm.

2.3. LLMs for social simulations

A growing line of work explores using LLMs as substitutes for human participants in social science, ranging from optimism about their potential as scalable "AI surrogates" (Ziems et al., 2024b; Anthis et al.) to caution about epistemic risks, including misrepresentation of marginalized perspectives (Abdurahman et al., 2024; Kapania et al., 2025) and inflated claims of generalizability (Crockett & Messeri, 2025). This tension reflects a central open question: when, if ever, LLM-generated data can be treated as valid experimental data. Several studies directly evaluate this substitution (Chehbouni et al., 2025). Early work shows systematic misalignment between LLM and human responses in public opinion and social judgment tasks (Santurkar et al., 2023), with persistent discrepancies even under demographic conditioning. Relatedly, Ma et al. (2024) and Park et al. (2024b) find uneven representation of political and moral viewpoints in LLM-generated populations, raising concerns about diversity collapse in simulated samples.

More recent work highlights representational distortions at the level of identity. Lee et al. (2024) show that LLMs tend to represent marginalized identities more homogeneously than dominant groups, flattening within-group variation. Similarly, Wang et al. (2025) demonstrate that when LLMs replace human participants, they can systematically misportray and compress identity groups, reducing diversity of responses and amplifying stereotypical structure.

A parallel literature investigates whether careful conditioning can produce high-fidelity synthetic populations. Methods such as persona-based sampling and demographic backstories aim to improve realism and alignment with human data (Argyle et al., 2023), while evaluations in economics and behavioral tasks report partial success in recovering aggregate human patterns (Horton, 2023; Brand et al., 2023). Additional work shows that LLMs can simulate structured psychological profiles, including personality-consistent behavior under the Big Five framework (Jiang et al., 2023). Broader simulation frameworks further extend LLMs to social systems, modeling interactions, communities, and political dynamics (Park et al., 2022; Touzel et al., 2024), and even reproducing experimental findings from psychology and economics in controlled settings (Binz et al., 2025).

Closest to our setting, Aher et al. (2023) introduce "Turing Experiments" to evaluate whether LLMs can replicate established behavioral findings, identifying where simulation succeeds versus fails. Our work differs in focus and granularity: rather than evaluating whether LLMs recover aggregate behavioral phenomena, we study stereotype formation itself and systematically quantify how fidelity varies across intersectional demographic groups. This allows us to directly characterize not only whether LLMs approximate human judgments, but which groups are misrepresented, and how—a key dimension for assessing epistemic risk in AI surrogate use.

3. Methods

3.1. Demographic Groups

Prior SCM studies have largely focused on single-axis groups (*e.g.*, Asians), and so we decide to extend this to study stereotype content towards intersectional groups separately (*e.g.*, Asian men and Asian women), which has received less attention in prior work. For example, Fiske et al. (2002) run a pilot study to pick demographic groups by asking participants to identify identities perceived as low-status in the U.S. context, as well as groups to which they personally belonged (in-groups). They included 21 identities that were mentioned by at least 15% of participants. Of these, only one identity (*i.e.*, gay men) was intersectionally defined. By contrast, we draw from Duckworth (2020), and select 50 intersectional groups that are well represented in the US context. We first picked ten demographic attributes that impact social dynamics (*e.g.*, age), and then 25 unique subgroups along those attributes (*e.g.*, young, old) intersected with binary gender (man or woman), resulting in 50 demographic groups, intersectionally defined (young men, old men, young women, old women), reported in Tab. 1.

3.2. Human Experiments

We closely follow the study protocol from Fiske et al. (2002) to minimize confounding factors in our analysis of LLM

Attribute	Hegemonic/privileged	Disadvantaged
age*	young	old
body type	thin	fat
disability	neurotypical (NT), able-bodied	neurodivergent (ND), disabled
gender identity	cisgender	transgender
language	English-speaking (ES)	non-English-speaking (non-ES)
nationality	American	immigrant
sexual orientation	heterosexual	gay
socio-economic status	rich	poor
race	White	Black, Asian, Hispanic
religion	Christian	Muslim, Jewish

Table 1. Demographic markers used in augmentations, drawn from *The Wheel of Power and Privilege*. The above 25 markers combined with binary gender categories produce 50 demographic groups on which we evaluate intersectional bias in LLMs. *Privilege and disadvantage along the lines of age is highly context-specific. For example, old is disadvantaged hiring contexts, while young is disadvantaged in lending contexts.

Dimension	Traits
Warmth	fair, friendly, likable, moral, outgoing, sociable, sincere*, tolerant*, trustworthy, warm*
Competence	able, active, assertive, determined, educated, intelligent*, competent*, confident*, independent*, competitive*

Table 2. List of traits used in the study. *Indicates the traits used in Fiske et al. (2002)

fidelity. We conducted an institutionally reviewed study where human annotators were hired through a proprietary web platform and asked to score different social groups on different affective traits using a 5-point Likert scale (Joshi et al., 2015). For example: *As viewed by society, how **intelligent** are **disabled women**?* Fiske et al. (2002) used nine traits (five for competence, four for warmth), but follow-up work by Kennison & Trofe (2003) showed that stereotype content in humans is sensitive to lexical choices in trait wording, so we use an expanded set of twenty traits (ten each for competence and warmth), taken from Nicolas et al. (2021), reported in Table 2.

It is important to note that the study question from Fiske et al. (2002) asks about *societal perceptions* and not *personal opinions*. In our analysis of the results, we are not evaluating the LLMs' personal biases, but rather its understanding of societal biases. This is in line with our focus on epistemic risks, rather than allocative/distributional risks.

We hired 193 participants in total. Our inclusion criteria was fluency in English. In a single session, a participant rated one randomly sampled identity on all 20 traits, shown in random order. Each participant was allowed to rate up to ten identities, and each identity was evaluated by 30 different individuals. We use a proprietary platform for recruitment and data quality and attention checks.

Annotator demographics are reported in Figure 6. We included definitions for each trait and demographic in our survey (reported in Table 5 and Table 4 in the Appendix), to facilitate a common understanding of terms between participants of different cultural and geographic backgrounds. In addition, we elicited confidence estimates from participants. So, for each identity and trait pair, each participant provided two rankings, first the score, and then their confidence in the score. The average annotator confidence is reported in Figure 8.

We assessed participants' familiarity with each social group using a privacy-preserving question: *"What is your familiarity with this social group?"* Participants could respond either that they "identify as a member of the group or have members of the group in their social circle", or that they "do not identify with the group and have no members of the group in their social circle". This approach, informed by Fiske et al. (2002) and Taylor et al. (2024), acknowledges individuals' biases toward their own group and close allies, while safeguarding personal demographic information. The fraction of in-group annotators for each intersectional identity is reported in Figure 7. Figure 9 and 10 in the Appendix corroborate human in-group biases, showing that average warmth and competence scores increase as the fraction of in-group annotators increases.

3.3. LLM Experiments

We prompt five LLMs to answer the same questions as our human study. Recognizing that LLMs are sensitive to prompt phrasing (Seshadri et al., 2022), we used ChatGPT to create 20 diverse rephrases of our study questions, divided into ten formal and ten informal styles, reported in Table 6 in the Appendix. Mirroring the approach used with human subjects, we added definitions for each identity and trait in the prompt. We skipped the question about in-group membership for LLMs. We sampled generations with a temperature of 1 and over two experimental runs, to induce naturalistic variation in LLM responses. In all, we sampled 200k generations (50 identities $\times$ 20 traits $\times$ 20 prompt templates $\times$ 2 runs $\times$ 5 models). We used regular expressions to parse model generations and extract Likert response categories. We manually inspected and processed responses that contained more than one Likert category.

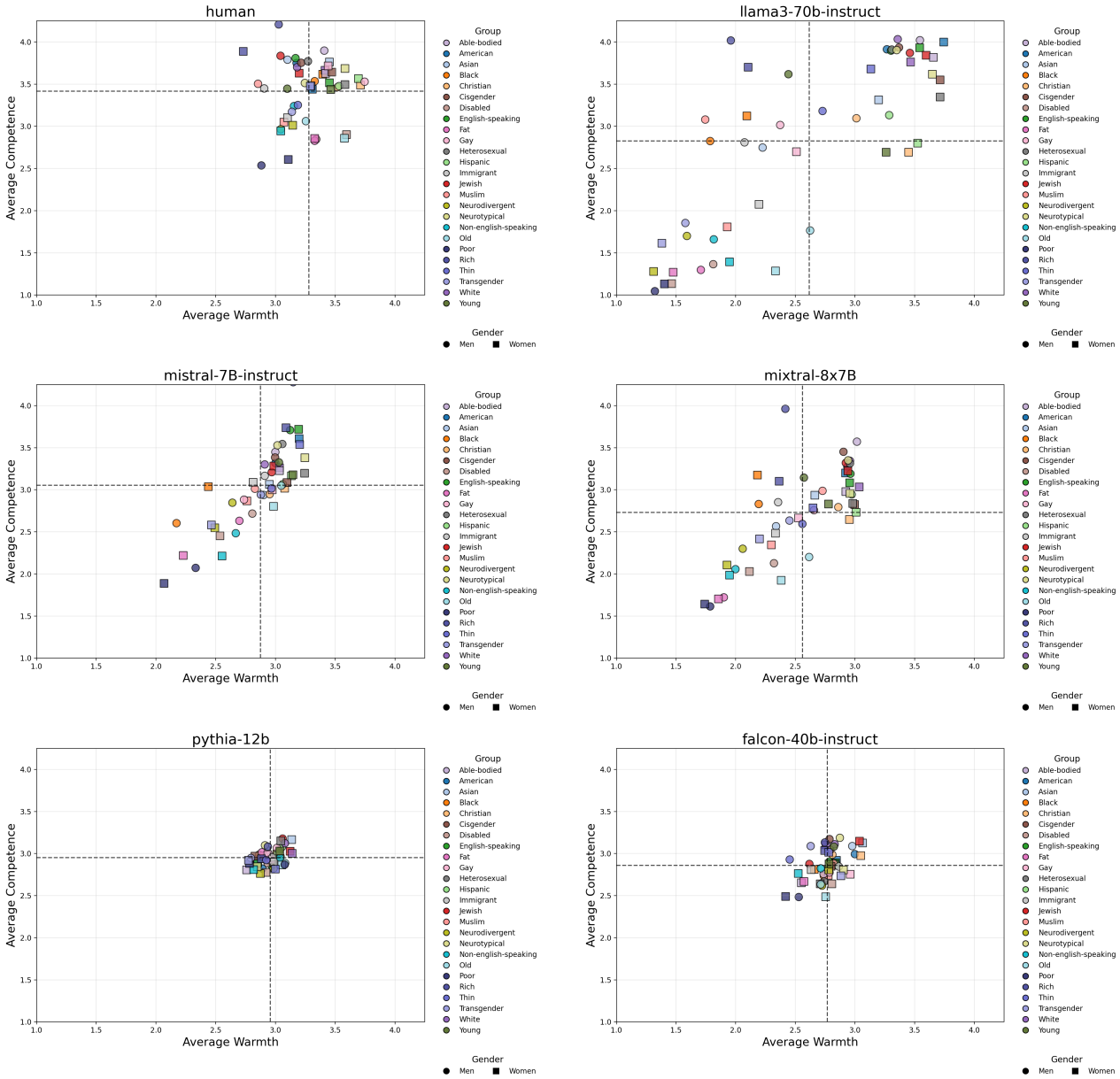


Figure 1. Stereotype content in humans and LLMs is distributed significantly differently. Dashed lines indicate population averages, circles indicate masculine subgroups, squares indicate feminine subgroups. Groups with average scores above the population averages are classified as high along that dimension (warmth or competence), and those with average scores lower that the population assignment are classified as low along that dimension.

4. Alignment of Stereotype Content

Following Fiske et al. (2002), we aggregate the ratings over different traits and calculate the mean Competence and Warmth score for each social group, and use the score to assign groups to one of four quadrants, shown in Figure 1. LLMs generally have lower average competence and warmth scores compared to humans (shown with dashed lines in Figure 1). The distribution of average competence and warmth scores varies by model, with Pythia and Falcon showing lesser spread/variance (being highly concentrated around the population average) compared to humans, while llama3 shows more spread/variance compared to humans.

There is generally inconsistent agreement between LLMs and human stereotypes as shown in Figure 2 and 3. Figure 2 highlights that LLMs infer negative (in red) or ambivalent (orange and blue) social stereotypes towards intersectional groups that the humans sample perceives favorably (in green) such as transgender women, gay women, Black

Figure 2. Competence-warmth stereotype assignments. Colors in the heatmap indicate stereotype quadrant: red (low competence, low warmth), green (high competence, high warmth), blue (high competence, low warmth), and orange (low competence, high warmth). Groups towards whom LLMs show stereotype misalignment are reported in bold.

women and Black men, which are all historically disadvantaged groups in the U.S. context. Interestingly, LLMs also rate historically privileged groups such as American men, English-speaking men, White men, cisgender men, heterosexual men, neurotypical men and young men more favorably than humans, who rate these groups ambivalently.

In Figure 3 we report the average agreement rate over all five models. We find that agreement is higher for feminine subgroups (in red) than masculine ones (in blue). The average agreement rate for privileged and disadvantaged groups is comparable, which corroborates the earlier discussion on stereotype misalignment in both positive and negative directions: LLMs show more positive bias towards privileged groups compared to humans and more negative bias against disadvantaged groups compared to humans.

Takeaway: There is inconsistent alignment between human and LLM stereotype clusters.

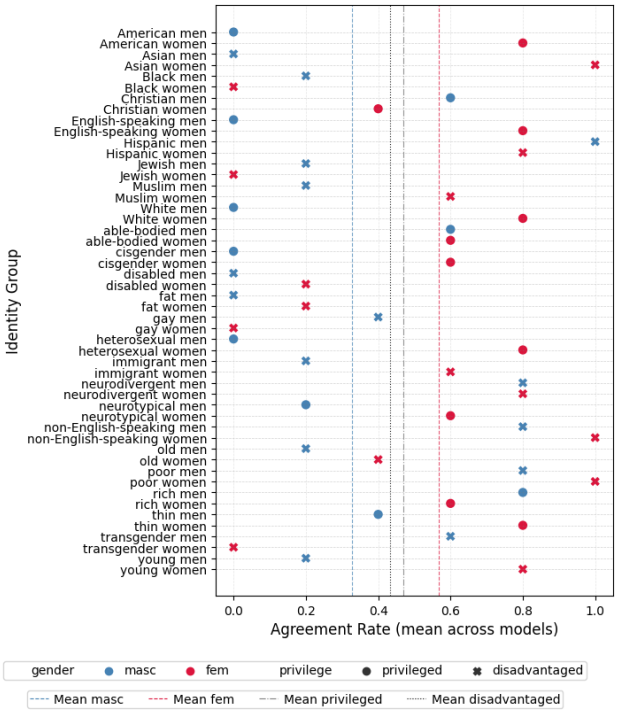


Figure 3. Human-LLM stereotype agreement rates by gender and privilege. Agreement rates (averaged over all models) for feminine subgroups are shown with a dashed red line, and for masculine subgroups with a dashed blue line, revealing higher agreement for feminine subgroups.

5. Algorithmic Fidelity

Argyle et al. (2023) propose criteria for *algorithmic fidelity* in LLM simulations of human subjects, and argue that, at a minimum, human and LLM responses should be indistinguishable. We adopt this definition, and use the Wasserstein distance between human and LLM scores as a metric for fidelity. We choose this distributional metric as it is attentive to the differences in the variance in human and LLM responses. This meaningfully builds upon contemporary work such as Park et al. (2024a), who quantify bias based on demographic differences in accuracy, overlooking variance.

Wasserstein distance. Given two empirical cumulative distribution functions (CDFs), $F(x)$ and $G(x)$, defined over the same variable x, the 1-Wasserstein distance quantifies the average probability mass that must be transported to transform one distribution into the other. Formally, it is defined as the integral of the absolute difference between the corresponding quantile functions:

$$W_{(g,t)} = \int_0^1 \left| F^{-1}_{(g,t)}(x) - G^{-1}_{(g,t)}(x) \right| dx. \quad (1)$$

In our context, $F^{-1}_{(g,t)}(x)$ represents the empirical quantile function of human responses for a group g and trait t, and

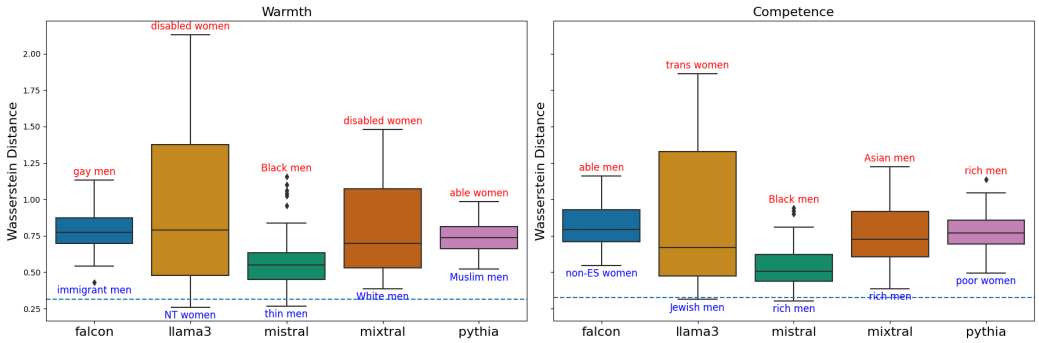


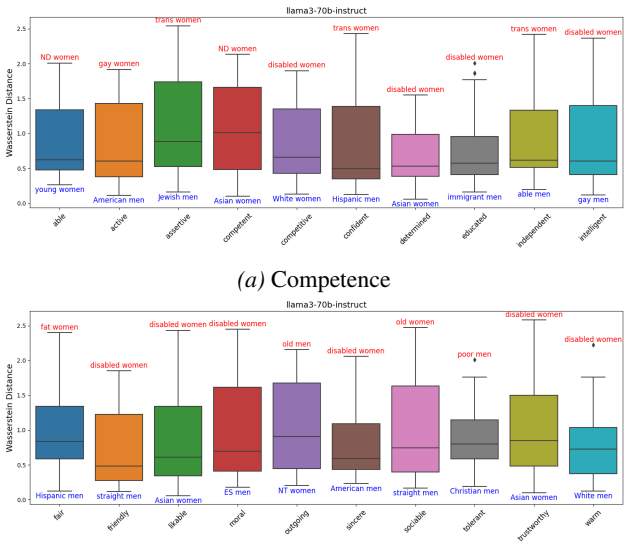
Figure 4. Fidelity of LLMs varies by group and model. We report the Wasserstein distance between human and LLM score distributions as a measure of infidelity. Values are averaged over all traits for each dimension, and the variance in the boxplot is over all 50 intersectional groups. Higher wasserstein distances indicate lower fidelity. Groups with the lowest fidelity are printed in red text, groups with the highest fidelity are printed in blue text. Average Wasserstein distance between human scores over all pairs of traits is reported as the baseline for each dimension, shown with a dashed blue line.

$G^{-1}_{(g,t)}(x)$ represents the corresponding quantile function of LLM responses. The Wasserstein distance $W_{(g,t)}$ is expressed in the same units as the underlying Likert scale for scores; it can be interpreted as the mean amount by which scores would need to change for the LLM distribution to match the human distribution. Smaller values of $W_{(g,t)}$ therefore indicate closer agreement between LLM and human score distributions, and thus higher fidelity of the LLM in replicating human response patterns.

Throughout we use "groups" to refer to the *subject of the stereotype*, and not as the demographic of the person who rated the stereotype.

We report the Wasserstein distance between human and LLM scores as a measure of algorithmic infidelity in Figure 4. As defined in Equation 1, Wasserstein distances are computed for each group g and trait t, which are then averaged over all traits for each dimension (warmth and competence) in Figure 4. The variance (height of the boxplot) is over demographic groups. The group with the highest fidelity is reported in blue text, and the group with the lowest fidelity is reported in red text. We report the mean Wasserstein distance between human scores for all pairs of traits as the baseline for score inconsistency for each dimension (warmth and competence).

Recall that we elicited scores on a Likert scale of 1-5, and so a Wasserstein distance of 1 indicates a jump from one Likert category to another. We find that the Wasserstein distance between human and LLM scores varies between 0.25 to 2, based on model and demographic group, indicating that LLM faithfully reproduce human biases towards groups reported in blue text (such as Neurotypical (NT) women and Jewish men for Llama3) while significantly distorting human biases towards groups shown in red (disabled women and trans women for Llama3). Overall, the mean Wasserstein distance over all models and groups is between 0.5-0.75, which is higher than the mean variation in human scores over all pairs of traits (0.3152 for warmth traits and 0.3272 for competence traits), which we use as a baseline for the expected fidelity in Figure 4. This indicates that the distortion introduced by LLMs is 1.5-2x worse than the latent inconsistency in scores within the human sample, implicating LLM simulations as being of low fidelity.



(a) Competence

(b) Warmth

Figure 5. Fidelity of llama3 varies significantly by trait. We report the Wasserstein distance between human and llama3 scores by competence and warmth separately, over all 50 demographic subgroups. Higher values of Wasserstein distance indicate lower fidelity. Results for other models are in Figure 11 in the Appendix

We find that fidelity varies not just by demographic group, but also by trait. In Figure 5 we plot the fidelity of each trait score (without averaging into a single warmth or competence dimension as earlier) for llama3, with results for other models deferred to Figure 11. Traits such as 'confident'

and 'friendly' have the highest average fidelity, whereas traits like 'competent', 'fair' and 'outgoing' have lower fidelity. Traits like 'assertive', 'likable' and 'moral' have the largest variance, and therefore demographic disparity. Interestingly, we see that feminine subgroups such as neurodivergent (ND), gay, fat and transgender women often have the lowest per-trait fidelity.

Takeaway: Algorithmic fidelity, measured via Wasserstein distance, is 1.5–2× worse than human–human consistency, and varies significantly by demographic group and trait.

6. Fidelity Disparities

We propose a new quantitative fairness metric to capture the risk of epistemic injustice, called the Fidelity Parity Ratio, defined analogously to existing ratio-based fairness metrics such as Disparate Impact (Feldman et al., 2015; Kleinberg et al., 2016; Chouldechova, 2017).

Fidelity Parity Ratio (FPR). Let $\mathcal{G}$ be a set of intersectional groups, and $W_{(\text{g,t})}$ is the Wasserstein distance between human and LLM response distributions for group $g \in \mathcal{G}$ and trait t as before. We define *Fidelity Parity Ratio* (FPR) as:

$$FPR = \frac{\min_{g \in \mathcal{G}} \sum_t (W_{(\text{g,t})})}{\max_{g \in \mathcal{G}} \sum_t (W_{(\text{g,t})})}. \tag{2}$$

A value of 1 indicates perfect alignment between LLM and human response distributions for all subgroups, while lower values indicate larger demographic disparities. From an epistemic justice perspective, FPR captures whether the LLM is an equally valid proxy for all demographic subgroups of interest, the lack of which could compromise validity of LLM-assisted social bias research. FPR along different demographic attributes is reported in Table 3.

	llama3	mixtral	mistral	pythia	falcon
gender identity	0.221	0.404	0.524	0.769	0.818
sexual identity	0.408	0.509	0.441	0.911	0.854
age	0.426	0.454	0.691	0.890	0.761
disability	0.162	0.380	0.500	0.611	0.553
race	0.366	0.392	0.442	0.791	0.756
religion	0.352	0.625	0.761	0.784	0.831
nationality	0.431	0.696	0.747	0.675	0.749
language	0.258	0.427	0.565	0.662	0.736
body type	0.332	0.498	0.346	0.876	0.880
socio-economic status	0.414	0.436	0.474	0.597	0.749

Table 3. Fidelity parity ratios along different demographic attributes. Lower values indicate higher disparities, value of 1 indicates perfect parity.

We observe fidelity parity ratios below 0.9 along all ten demographic attributes and for all five models, with Llama3 showing the lowest fidelity parity among models, along all attributes. Overall, we see the worst fidelity disparity along disability (FPR=0.162), gender identity (FPR=0.221) and language (FPR=0.258) from Llama, and the best fidelity parity along sexual identity (FPR=0.911) from Pythia. Based on the 4/5ths rule from non-discrimination law (which is a contentious yet commonly reported heuristic in fairness research (Feldman et al., 2015; Raghavan & Kim, 2024; Watkins & Chen, 2024)), we find significant risk of epistemic injustice (FPR < 0.8) along disability, race, nationality, language and socio-economic status from all five models.

Takeaway: LLMs show significant fidelity disparities (according to the 4/5ths rule) along disability, race, nationality, language and socio-economic status.

7. Discussion

Low and disparate fidelity threatens scientific validity. Algorithmic fidelity is substantially lower than human–human consistency, with deviations exceeding expected variability in human responses. Further, these errors are not uniform: some identity groups are represented less faithfully than others, indicating that LLMs do not preserve human response distributions in a consistent or reliable way across subpopulations.

Uneven fidelity risks epistemic injustice. LLM-generated data are of lower fidelity for certain subpopulations than for others. This produces a form of epistemic injustice, where some groups are less accurately reflected in the resulting evidence base, not through explicit exclusion but through differential representational quality.

AI surrogates introduce a trade-off between scalability and representational harm. While LLMs can reduce participant burden and enable scalable data collection in sensitive settings, they may also reshape which voices are amplified or excluded in scientific knowledge production. From an AI-for-good perspective, this creates a tension between reducing direct harm to human subjects and introducing indirect harms through distorted representation.

Norms around the use of LLMs in science must consider epistemic risks. Our findings suggest that deploying LLMs as research instruments is not a neutral substitution problem, but a normative choice about how knowledge is produced and whose perspectives are preserved. Crucially, norms around the use of LLMs in social science are still being actively shaped by the research community itself. Therefore, we see this work as contributing to emerging standards for AI-assisted research, emphasizing that epistemic concerns must be placed at the forefront of the design and evaluation of LLMs in science.

Limitations

Our study is situated in the North American context, and the extent to which our findings generalize to other cultural and sociolinguistic settings remains uncertain. Further, our experiments were conducted in June–August 2024, imposing a temporal cut-off on the training data available to the evaluated LLMs, which may affect the stability of observed misalignment patterns over time. All prompts and evaluations were conducted in English, which may privilege perspectives better represented in English-language data and thereby shape both fidelity and bias. As such, some of the disparities we observe may reflect linguistic as well as social representation gaps. Evaluating surrogate fidelity under multilingual prompting is an important direction for future work. To maintain feasibility, we did not explore alternative prompting strategies, such as persona-based prompting (Argyle et al., 2023), which may yield different patterns of alignment and is a promising avenue for follow-up work. Additionally, we restrict our analysis to intersectional groups and do not compare fidelity between single-axis and intersectional identities. Expanding the analysis across different levels of group granularity would provide a more complete picture of LLM behavior in stereotype research. Finally, we emphasize that no single metric can certify the scientific validity of AI surrogates. Our evaluation considers multiple dimensions, including stereotype alignment, distributional fidelity, and fidelity parity across demographic attributes, and should be interpreted as a diagnostic for invalidity rather than a guarantee of validity. More broadly, we view our framework as one component of a holistic, context-dependent assessment of readiness of LLMs to assist in the creation of new scientific knowledge.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., and Dehghani, M. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245, 2024.

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., and McKee, K. R. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2024.

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Almeida, A. P. C., Arriaga, P., and Marques, J. S. Making robot's attitudes predictable: A stereotype content model for human-robot interaction in groups. PsyArXiv Preprints, 2018. URL https://osf.io/preprints/psyarxiv/kfzdg.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., and Bernstein, M. S. Position: Llm social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

Bartl, M., Nissim, M., and Gatt, A. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. A foundation model to predict and capture human cognition. *Nature*, pp. 1–8, 2025.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, 2021.

Brand, J., Israeli, A., and Ngwe, D. Using llms for market research. *Harvard business school marketing unit working paper*, (23-062), 2023.

Charlesworth, T. E., Ghate, K., Caliskan, A., and Banaji, M. R. Extracting intersectional stereotypes from embeddings: Developing and validating the flexible intersectional stereotype extraction procedure. *PNAS nexus*, 3(3): pgae089, 2024.

Chehbouni, K., Haddou, M., Cheung, J. C., and Farnadi, G. Neither valid nor reliable? investigating the use of llms as judges. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047. URL https://doi.org/10.1089/big.2016.0047. PMID: 28632438.

Collins, P. H. Intersectionality's definitional dilemmas. *Annual review of sociology*, 41(1):1–20, 2015.

Collins, P. H. and Bilge, S. *Intersectionality*. John Wiley & Sons, 2020.

Collins, P. H., da Silva, E. C. G., Ergun, E., Furseth, I., Bond, K. D., and Martínez-Palacios, J. Intersectionality as critical social theory: Intersectionality as critical social theory, patricia hill collins, duke university press, 2019. *Contemporary Political Theory*, 20(3):690, 2021.

Coston, A., Kawakami, A., Zhu, H., Holstein, K., and Heidari, H. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, pp. 690–704. IEEE, 2023.

Creel, K. and Hellman, D. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167, 1989.

Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pp. 23–51. Routledge, 2013.

Crockett, M. and Messeri, L. Ai surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences*, 2025.

Cuddy, A. J., Fiske, S. T., and Glick, P. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149, 2008.

Cuddy, A. J., Fiske, S. T., and Glick, P. When professionals become people: Warmth, competence, and the bias implicit in current stereotype content. *Journal of social issues*, 65(3):448–469, 2009. URL https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4560.2009.01605.x.

Delve AI. Delve ai – automated personas and synthetic users for research and marketing. https://www.delve.ai, 2025. Accessed: 2025-10-22.

Duckworth, S. Wheel of power/privilege. Infographic, Flickr, October 2020. URL https://flic.kr/p/2jWxeGG. Accessed: 2026-04-24.

EleutherAI. pythia-12b. https://huggingface.co/EleutherAI/pythia-12b, 2023. Dense autoregressive language model. Technical report: *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*, Biderman et al., arXiv:2304.01373.

Fairgen AI. Synthetic respondents for quantitative studies. https://www.fairgen.ai, 2025. Accessed: 2025-10-22.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878, 2002. URL https://psycnet.apa.org/record/2002-10427-002.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pp. 162–214. Routledge, 2018.

Fraser, K. C., Kiritchenko, S., and Nejadgholi, I. Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5:826207, 2022.

Fricker, M. *Epistemic injustice: Power and the ethics of knowing*. Oxford university press, 2007.

Gao, Y., Lee, D., Burtch, G., and Fazelpour, S. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025.

Glick, P. and Fiske, S. T. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56 (2):109, 2001. URL https://psycnet.apa.org/record/2001-02685-006.

Gorecki, M. and Hardt, M. Monoculture or multiplicity: Which is it? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Guo, W. and Caliskan, A. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 122–133, 2021.

Herold, B., Waller, J., and Kushalnagar, R. Applying the stereotype content model to assess disability bias in popular pre-trained nlp models underlying ai-based assistive technologies. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pp. 174–183, 2022. URL https://aclanthology.org/2022.naacl-industry.17/.

Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., and Kabbara, J. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*, 2023.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015.

Kapania, S., Agnew, W., Eslami, M., Heidari, H., and Fox, S. E. Simulacrum of stories: Examining large language models as qualitative research participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025.

Kennison, S. M. and Trofe, J. L. Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, 32:355–378, 2003.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pp. 12–24, 2023.

Lee, M. H., Montgomery, J. M., and Lai, C. K. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1321–1340, 2024.

Lee, T. L. and Fiske, S. T. Who or what is artificial intelligence? warmth and competence in perceptions of ai. *Computers in Human Behavior*, 118:106670, 2021. URL https://www.sciencedirect.com/science/article/abs/pii/S074756322100062X.

Liang, T., Wu, Y., and Zhang, Y. Intersectional stereotypes in language models. *arXiv preprint arXiv:2310.08253*, 2023. URL https://arxiv.org/abs/2310.08253.

Lin, Z. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566, 2025.

Liu, G., Bono, C. A., and Pierri, F. Comparing diversity, negativity, and stereotypes in chinese-language ai technologies: a case study on baidu, ernie and qwen. *arXiv preprint arXiv:2408.15696*, 2024.

Ma, B., Yoztyurk, B., Haensch, A.-C., Wang, X., Herklotz, M., Kreuter, F., Plank, B., and Assenmacher, M. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. *arXiv preprint arXiv:2412.13169*, 2024.

Macieira, F. J. F., Pinto, D. C., Oliveira, T., and Yanaze, M. Bits and biases: Exploring perceptions in human-like ai interactions using the stereotype content model. In *Proceedings of the 7th International Conference on Finance, Economics, Management and IT Business (FEMIB 2025)*, pp. 161–166, 2025.

Meta AI. llama3-70b-instruct. https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct, 2024. Instruction-tuned model. Technical report: *The Llama 3 Herd of Models*, Dubey et al., arXiv:2407.21783.

Mistral AI. mistral-7b-instruct-v0.2. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, 2024a. Instruction-tuned model. Technical report: *Mistral 7B*, Jiang et al., arXiv:2310.06825.

Mistral AI. mixtral-8x7b-instruct. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1, 2024b. Instruction-tuned mixture-of-experts model. Technical report: *Mixtral of Experts*, Jiang et al., arXiv:2401.04088.

Nicolas, G. and Caliskan, A. A taxonomy of stereotype content in large language models. *arXiv preprint arXiv:2408.00162*, 2024.

Nicolas, G., Bai, X., and Fiske, S. T. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196, 2021.

Nicolas, G., Bai, X., and Fiske, S. T. A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of personality and social psychology*, 123 (6):1243, 2022.

Opinio AI. Opinio ai – ai personas and synthetic respondents for opinion research. https://www.opinio.ai, 2025. Accessed: 2025-10-22.

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024a.

Park, P. S., Schoenegger, P., and Zhu, C. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770, 2024b.

Raghavan, M. and Kim, P. T. Limitations of the" four-fifths rule" and statistical parity tests for measuring fairness. *Geo. L. Tech. Rev.*, 8:93, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Seshadri, P., Pezeshkpour, P., and Singh, S. Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*, 2022.

Shrawgi, H., Rath, P., Singhal, T., and Dandapat, S. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1841–1857, 2024.

Siddique, Z., Turner, L. D., and Espinosa-Anke, L. Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. *arXiv preprint arXiv:2407.06917*, 2024.

Synthetic Users. Synthetic users – ai-generated user research participants. https://www.syntheticusers.com, 2025. Accessed: 2025-10-22.

Taylor, V. J., Yantis, C., and Valladares, J. V. "will they assume i'm racist?" how racial ingroup members' stereotypical behavior impacts white americans' interracial interaction experiences. *Group Processes & Intergroup Relations*, pp. 13684302241265260, 2024.

Technology Innovation Institute. falcon-40b-instruct. https://huggingface.co/tiiuae/falcon-40b-instruct, 2023. Instruction-tuned model. Technical report: *The Falcon Series of Language Models*, Penedo et al., arXiv:2311.16867.

Touzel, M. P., Sarangi, S., Welch, A., Krishnakumar, G., Zhao, D., Yang, Z., Yu, H., Kosak-Hine, E., Gibbs, T., Musulan, A., et al. A simulation system towards solving societal-scale manipulation. *arXiv preprint arXiv:2410.13915*, 2024.

Veit, S., Arnu, H., Di Stasio, V., Yemane, R., and Coenders, M. The "big two" in hiring discrimination: evidence from a cross-national field experiment. *Personality and Social Psychology Bulletin*, 48(2):167–182, 2022.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pp. 1–12, 2025.

Watkins, E. A. and Chen, J. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 764–775, 2024.

Yabble. Yabble – virtual audiences and ai-generated personas for research and insights. https://www.yabble.com, 2025. Accessed: 2025-10-23.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, March 2024a. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024b.

A. Appendix

A.1. Prompting details

Social group	Definition
transgender	A transgender person is someone whose gender identity differs from the sex they were assigned at birth. This includes individuals who may identify as a different gender from their assigned sex or who may have a non-binary or genderqueer identity.
cisgender	A cisgender person is someone whose gender identity aligns with the sex they were assigned at birth. For example, if an individual is assigned male at birth and identifies as a man, they are considered cisgender.
gay	Gay (or homosexual) refers to someone who is attracted to individuals of the same sex. For example, a gay man is attracted to men, and a gay woman is attracted to women.
heterosexual	Heterosexual (or straight) refers to someone who is attracted to individuals of the opposite sex. For example, a heterosexual man is attracted to women, and a heterosexual woman is attracted to men.
young	The term 'young' refers to individuals who are in the early stages of life, typically including children, teenagers, and young adults. The specific age range considered 'young' can vary by context and culture, but it generally includes those under 30.
old	'Old' typically refers to individuals who are in the later stages of life, often considered elderly or senior citizens. The specific age at which someone is considered 'old' can vary, but it generally includes those over the age of 65.
able-bodied	An able-bodied person is someone who does not have physical or mental disabilities and functions without significant impairment.
disabled	A disabled person is someone who has a physical, mental, or sensory impairment that significantly affects their ability to perform certain tasks or activities. Disabilities can be visible or invisible and include impaired vision, impaired hearing or deafness, mental health conditions, epilepsy, etc.
neurotypical	Neurotypical refers to individuals whose neurological development and functioning are considered to be typical or standard. This term is used to contrast with neurodivergent, describing those who do not have cognitive or developmental variations such as autism or ADHD.
neurodivergent	Neurodivergent describes individuals whose cognitive functioning or neurological development differs from what is considered typical. This includes conditions such as autism, ADHD, dyslexia, and others. Neurodivergent individuals may have different ways of processing information and interacting with the world.
Black	The term 'Black' refers to individuals who identify with the racial and ethnic group characterized by African ancestry.
White	'White' (or Caucasian) refers to individuals who identify with the racial group characterized by European ancestry.
Asian	'Asian' refers to individuals who identify with the racial and ethnic group originating from the continent of Asia.
Hispanic	'Hispanic' describes individuals who come from, or have ancestry from, Spanish-speaking countries, particularly those in Latin America and Spain. It is used to denote cultural and linguistic ties to Spanish-speaking communities.
Muslim	A Muslim is a person who practices Islam, a monotheistic religion based on the teachings of the Prophet Muhammad as recorded in the Quran.
Jewish	Jewish refers to individuals who identify with Judaism, a monotheistic religion with a rich cultural and historical heritage. Jewish identity can be religious, ethnic, or cultural, and it encompasses a range of beliefs and practices within the Jewish community.
Christian	A Christian is someone who follows Christianity, a monotheistic religion based on the life and teachings of Jesus Christ. Christianity includes various denominations, such as Catholicism, Protestantism, and Orthodoxy, each with its own beliefs and practices.

American	An American is someone who is a citizen or resident of the United States of America.
immigrant	An immigrant is someone who has moved from their country of origin to another country with the intention of residing there permanently or temporarily. Immigrants may relocate for various reasons, including economic opportunities, safety, or family reunification.
English-speaking	English-speaking refers to individuals who communicate primarily in the English language. This term may describe native speakers or those who use English as a second language.
non-English-speaking	Non-English-speaking describes individuals who do not use English as their primary language of communication.
thin	'Thin' (or slim, lean, skinny) describes individuals who have a body type characterized by a lower amount of body fat and a smaller overall body mass compared to the average body type.
fat	'Fat' (or obese, overweight) refers to individuals who have a body type characterized by a higher amount of body fat and a larger overall body mass compared to the average body type.
rich	'Rich' describes individuals who have a high level of financial wealth and resources. Rich individuals typically have significant assets, income, or investments that afford them a high standard of living.
poor	'Poor' refers to individuals who have limited financial resources and struggle to meet basic needs such as food, shelter, and healthcare. Poverty can result from a variety of factors, including low income, unemployment, and economic inequality

Table 4. Identity definitions in stereotype content experiments with humans and LLMs

Trait	**Definition**
sociable	'sociable' refers to someone who is inclined to socialize or to seek and enjoy companionship with others.
friendly	'friendly' describes someone who is pleasant and amiable towards others, or who shows warmth and goodwill in social interactions.
warm	'warmth' refers to the quality of being friendly, approachable, and affectionate in interactions with others.
likable	'likability' refers to the quality or characteristic of being pleasant, attractive, or easy to like by others.
outgoing	'outgoing' describes someone who is friendly, sociable, and enjoys interacting with others in various social situations.
moral	'moral' refers to the characteristic of consistently adhering to ethical principles and behaving in ways that are considered right or virtuous within a community or society.
trustworthy	'trustworthy' means being reliable and honest, someone others can depend on and believe in various social and professional contexts.
sincere	'sincere' means genuine, honest, and without deceit or pretense.
fair	'fair' refers to someone who treats everyone equally and justly.
tolerant	'tolerant' refers to someone who accepts and respects differences in others.
competent	'competent' describes someone who is capable and skilled in performing tasks effectively.
competitive	'competitive' refers to having a strong desire to succeed and outperform others.
intelligent	'intelligent' refers to someone with a high level of mental capacity, who is quick to understand and learn.
able	'able' describes a person who is capable of performing tasks or activities effectively.
educated	'educated' means having received formal instruction and knowledge in various subjects or fields.
confident	'confident' means having belief in one's abilities and being self-assured and assertive.
assertive	'assertive' refers to someone who expresses opinions or desires confidently and forcefully.

independent	'independent' describes a person who is capable of making decisions and taking actions without relying on others.
active	'active' refers to being energetically engaged in various social interactions or activities, often displaying enthusiasm and participation in group settings.
determined	'determined' describes someone who displays strong willpower, perseverance, and commitment towards achieving their goals or objectives.

Table 5. Trait definitions in stereotype content experiments with humans and LLMs

Prompt	**Prompt text**
full prompt template	### Instruction: Please answer the following multiple choice question. Question: [question prompt] [identity definition][trait definition] A. extremely B. very C. average D. slightly E. not at all ### Answer:
question prompts (formal)	As viewed by society, how [trait] are [identity]? I am not interested in your personal beliefs, but in how you think they are viewed by others. How [trait] are [identity], according to society? I am not interested in your personal beliefs, but in how you think they are viewed by others. How would you rate the perceived [trait] of [identity] in society? Please consider societal perceptions rather than personal opinions. In society's view, how [trait] are [identity] perceived? Remember, this is not about your personal beliefs but rather about societal views. How does society rate the [trait] of [identity]? I am not interested in your personal beliefs, but in how you think they are viewed by others in society. How is [identity]'s [trait] perceived in society? Remember this is about societal views not personal beliefs. In the eyes of society, how [trait] are [identity] perceived to be? Please pick an answer based on societal perceptions not personal opinions. How does society view the [trait] of [identity]? Please answer according to social beliefs and not personal opinions. According to societal norms, how [trait] are [identity] perceived as? Please focus on societal views not personal opinions. From a societal standpoint, how is [identity]'s [trait] perceived? This inquiry is focused on societal perceptions rather than personal beliefs.
question prompts (informal)	How do most folks see [identity] when it comes to [trait]? Not your own take, just what others think. What's the general vibe about [identity] being [trait]? I'm curious about public perception here. How does society typically rate [identity] in terms of [trait]? This isn't about your opinion, just the general view out there. When people talk about [identity], do they usually say it's [trait]? I'm after the public perspective. What's the consensus on [identity] and how [trait] they are? I'm interested in what everyone else thinks. When it comes to [identity], is there a common belief about how [trait] they are? Not looking for personal opinions, just the societal view. What's the word on the street about [identity] being [trait]? I want to know how others see it. How does the world see [identity] when it comes to [trait]? Your personal view aside, I'm curious about the general perception. What's the public perception of [identity] in terms of [trait]? Give me the general societal view, not personal opinions. What's the general take on [identity] and how [trait] they are? Just want to know what the crowd thinks.

Table 6. Prompts used to collect stereotype content from LLMs

A.2. Annotator information

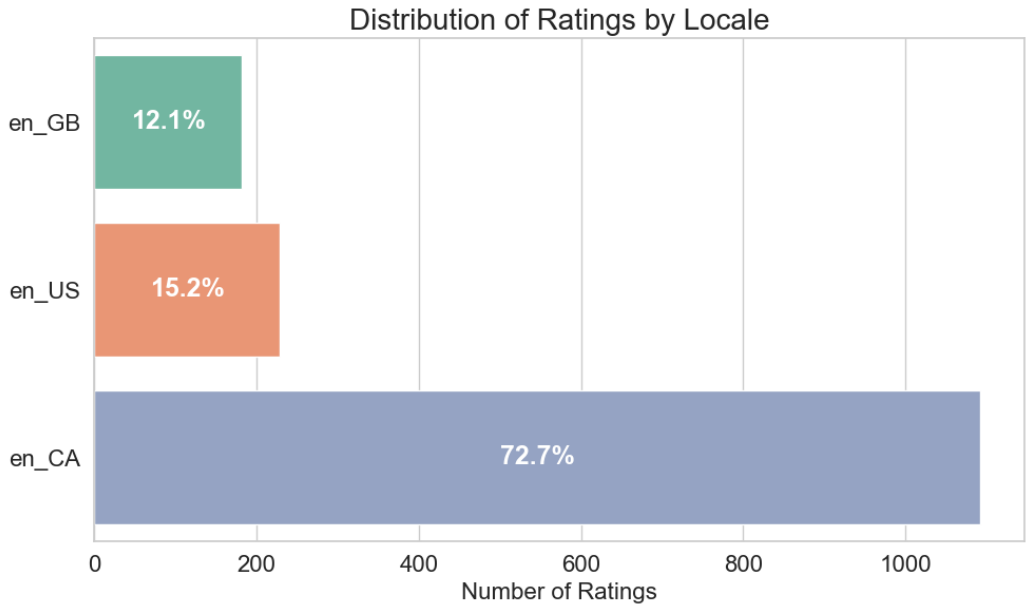


Figure 6. Annotator locales: GB is Great Britain, US is United States, CA is Canada

A.3. Stereotype Content

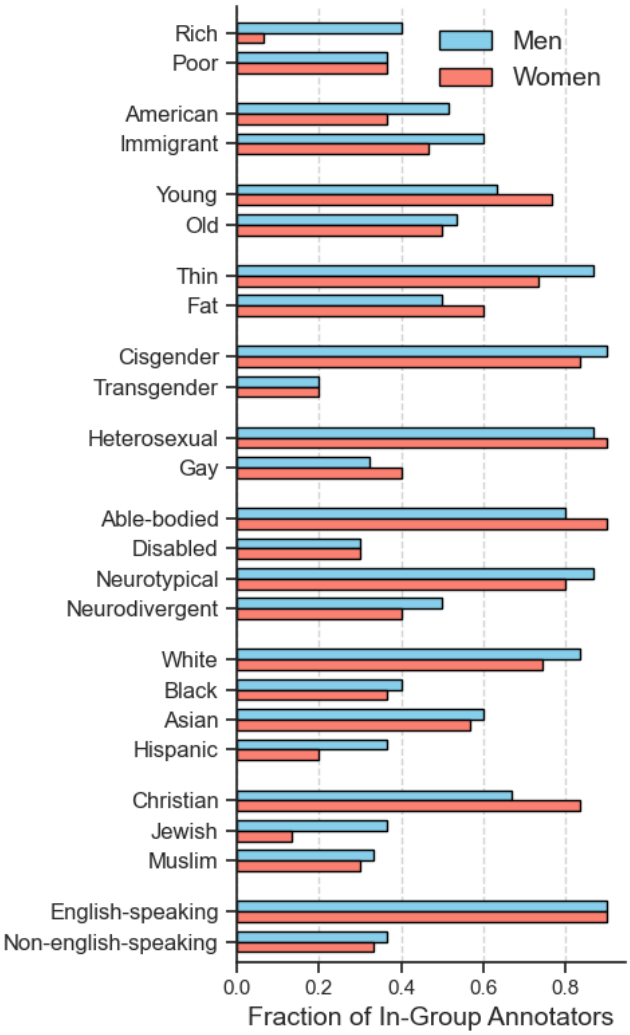


Figure 7. Fraction of self-reported in-group annotators for each demographic group

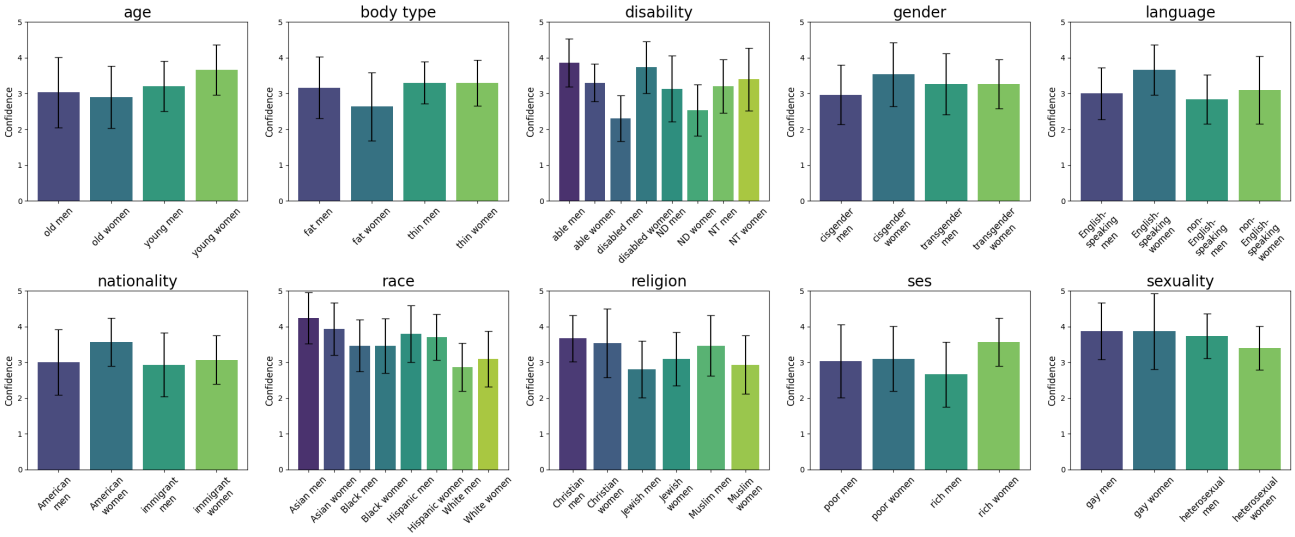


Figure 8. Average annotator confidence for each demographic group

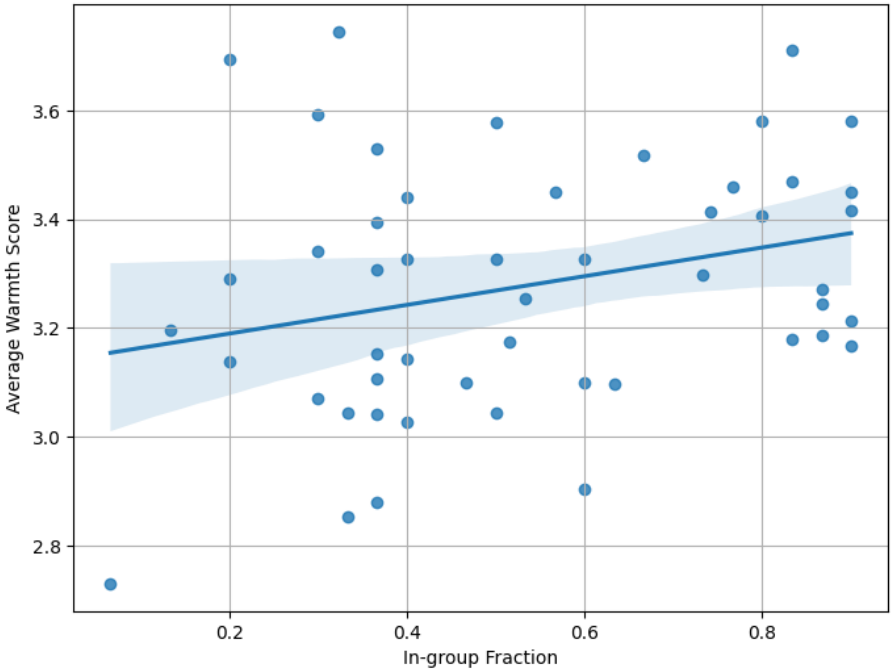


Figure 9. Average warmth score increases as fraction of in-group annotators increases

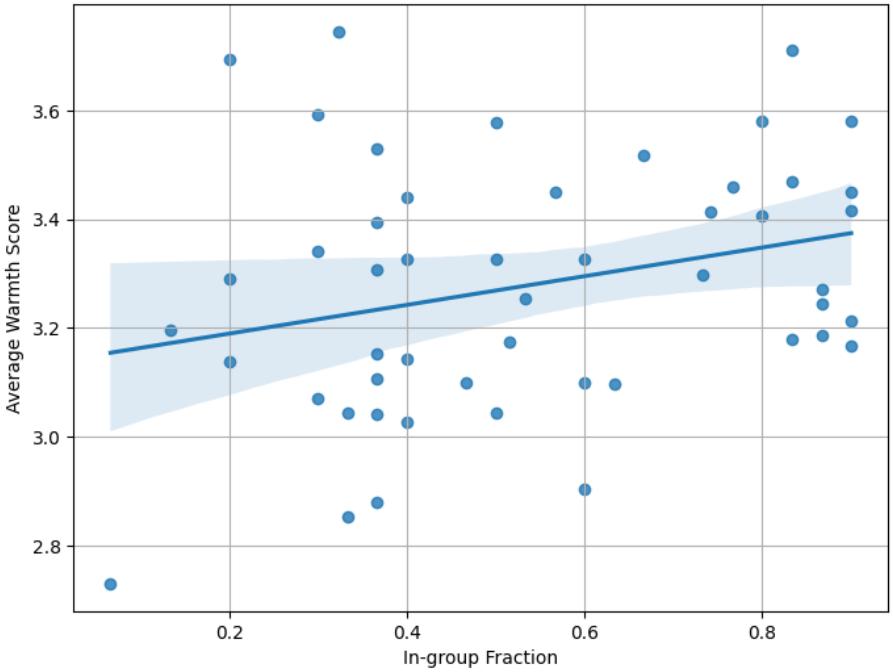
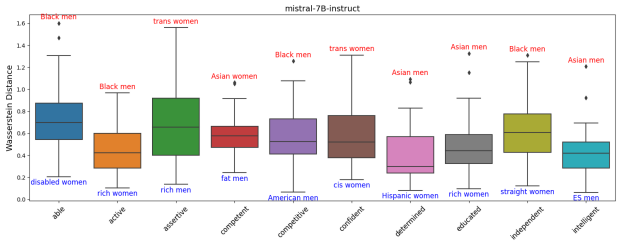
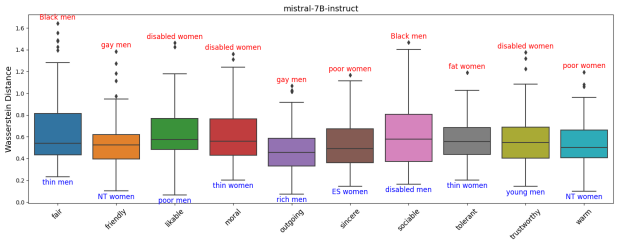


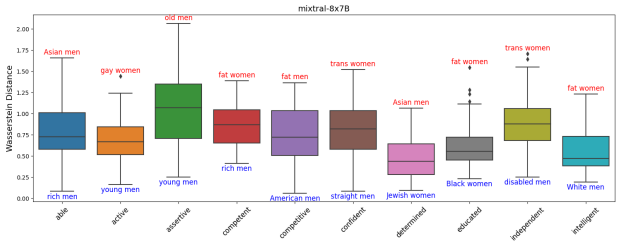
Figure 10. Average competence score increases as fraction of in-group annotators increases



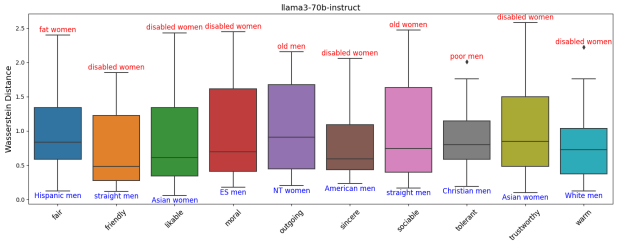
(a) Mistral - Competence



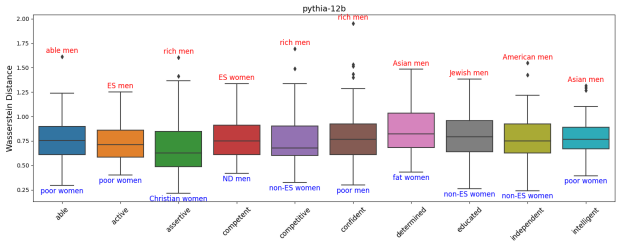
(b) Mistral - Warmth



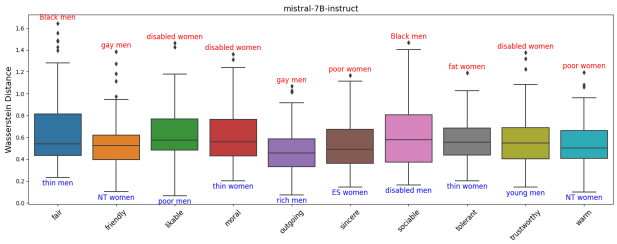
(c) Mixtral - Competence



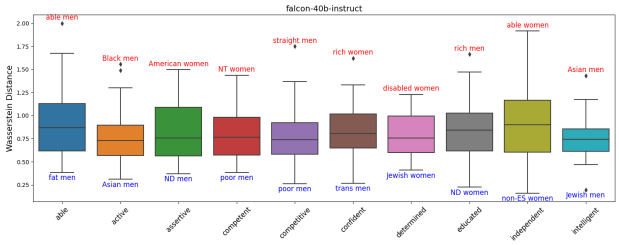
(d) Mixtral - Warmth



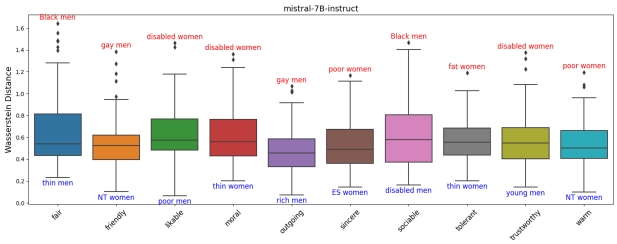
(e) Pythia - Competence



(f) Pythia - Warmth



(g) Falcon - Competence



(h) Falcon - Warmth

Figure 11. Wasserstein distance between human and LLM scores on competence (left) and warmth (right) traits over all 50 demographic subgroups. Lower values of wasserstein distance indicate higher fidelity. Results for llama3 are in Figure 5