What Are Chatbots' Stereotypes About? A Data-Driven Analysis of Large Language Models' Content Associations with Social Categories

Gandalf Nicolas[1], Aylin Caliskan[2]

[1]Rutgers University – New Brunswick
[2]University of Washington

gandalf.nicolas@rutgers.edu, aylin@uw.edu

Abstract

This study introduces a data-driven taxonomy of stereotype content in contemporary large language models (LLMs). We prompt ChatGPT 4.5, ChatGPT 3.5, Llama 3, and Mixtral 8x7B, four recent and powerful LLMs, for the characteristics associated with 87 social categories (e.g., gender, race, occupations). We show that these prompts are reliable and valid, predicting unrelated tasks such as storytelling about the targets. Using text embeddings and cluster analyses, we identify 14 dimensions (Ability, Appearance, Assertiveness, Beliefs, Deviance, Emotion, Family, Geography, Health, Morality, Occupations, Social categories, Sociability, and Status) in LLMs' stereotypes. This high-dimensional taxonomy reveals both similarities (e.g., same set of dimensions) and differences (e.g., variation in prevalence of content) with human stereotypes. In addition, we find that highly overlapping taxonomies emerge from analyses of personal and cultural stereotypes, as well as across various LLMs. However, again, some prompts and LLMs differ in how frequently specific dimensions appear in association with social categories. Our findings suggest that LLMs' stereotypes are high-dimensional and auditing and debiasing would benefit from considering this complexity to minimize unidentified harm from reliance in low-dimensional views of bias in LLMs.

Code — https://osf.io/bwdcr/
Datasets — https://osf.io/bwdcr/

Introduction

Humans create and place each other into social categories (e.g., in terms of gender, race, age, occupations) to simplify and navigate the social world, often via potentially harmful stereotypes (Macrae and Bodenhausen 2000). A stereotype is defined here and in general psychological models as a characteristic associated with a social category (e.g., through explicit beliefs, implicit associations; Bodenhausen and Macrae 1998). These stereotypes vary in content, such as whether they are about a target's moral traits, abilities, or other characteristics (Abele et al. 2021). Recent studies have used text analysis to describe the diversity of stereotypes across social categories in human surveys (e.g., Nicolas, Bai, and Fiske 2022). However, stereotype content dimensions in contemporary Artificial Intelligence (AI) large language models (LLMs) have not been systematically identified. A more comprehensive taxonomy of stereotypes in LLMs is a critical first step for thorough auditing and effective debiasing of AI's social biases.

Current Study

Using data-driven cluster analyses, we present a taxonomy of stereotype content in contemporary LLMs, identifying its dimensions and their prevalence. We derive this taxonomy based on the LLMs' semantic associations with several U.S. social category terms, focusing on generalizable stereotype properties across social categories. We examine four recent and widely used LLMs: ChatGPT 4.5 (primary LLM), and ChatGPT 3.5 Turbo, Llama 3, and Mixtral 8x7B Instruct (replication LLMs for a subset of analyses), providing convergent evidence for the taxonomy.

We prompted the LLMs to list 50 characteristics associated with salient social categories in the United States. We establish the reliability and validity of this approach, showing that ChatGPT 4.5 draws from these associations in storytelling about the social categories, an unrelated task that users may prompt chatbots for. We explore the robustness of the method by using prompts with both cultural (what are society's associations?) and personal (what are the LLM's associations?) phrasings.

For the main analysis, we identify the content of associations by obtaining text embeddings of the responses and using a clustering algorithm. We also examine how frequently

the different dimensions occur in stereotypes across categories (i.e., representativeness/prevalence).

We expected to find significant overlap with human taxonomies, where dimensions related to Warmth (Sociability, Morality) and Competence (Ability, Assertiveness) are highly representative, but not sufficient to characterize stereotype content, with additional dimensions (e.g., Emotion, Deviance) showing significant prevalence across categories (Nicolas, Bai, and Fiske 2022).

Using data-driven content analyses of cultural and personal stereotype listing and storytelling about 87 social categories, including 1,366 different category-coding labels, across 4 powerful LLMs, we contribute the following:

- Reliability and validity testing for our prompting approach to obtain explicit stereotypes from LLMs, suggesting that LLMs use the information obtained via our direct elicitation method in unrelated tasks, such as storytelling.
- Evidence about the similarities and differences in LLM cultural vs. personal stereotypes. We show that responses to both prompts significantly overlap (with some differences in content prevalence), but ChatGPT 4.5 outputs warnings more frequently for cultural prompts.
- A high-dimensional taxonomy of the top associations that large language models have about diverse social categories. We show that a variety of labeling methods converge to a set of 14 primary dimensions describing the content of LLMs' stereotypes (Ability, Appearance, Assertiveness, Beliefs, Deviance, Emotion, Family, Geography, Health, Morality, Occupations, Social categories, Sociability, and Status). These dimensions align with human taxonomies of spontaneous (open-ended) stereotypes, providing evidence for generalizability of psychological models, as well as guidance on the content that comprehensive auditing and debiasing must address.
- Evidence for the high prevalence of dimensions related to Warmth and Competence, with more variation for smaller dimensions, such as Status, Beliefs, and Deviance. These patterns suggest that LLMs may emphasize distinct content, with implications for auditing and debiasing specificity and user experience based on the selected LLM.

Related Work

Stereotype Content in Human Data

The best-established stereotype dimensions are Warmth and Competence, which are evolutionarily plausible, have been found cross-culturally and over time, and are predictive of emotions and behaviors (Fiske et al. 2021). Warmth (also called communion or the horizontal dimension) refers to attributions about a target's sociability and morality. Competence (also called agency or the vertical dimension) refers to attributions about a target's abilities and assertiveness. In other words, humans prioritize understanding: is this person a friend or foe (Warmth), and can they act on their intentions (Competence)? Expanding into a more comprehensive taxonomy, the Spontaneous Stereotype Content Model (SSCM, Nicolas, Bai, and Fiske 2022) proposed that 14 content dimensions account for 80-95% of stereotypes about salient social categories. These dimensions are: Sociability and Morality (facets of Warmth), Ability and Assertiveness (facets of Competence), socioeconomic Status, political-religious Beliefs (see Koch et al. 2016), Appearance (including attractiveness), Emotion, Occupations/Work, Health, Deviance, Geography (including foreignness), Family relations, and intersectional Social Group associations (e.g., "rich people are men"). These dimensions vary in prevalence (also called representativeness), with Warmth and Competence facets being highly representative across categories, while associations about Health and Geography are less prevalent. Prevalence relates to primacy of stereotype dimensions (Abele et al. 2021), as content that is used more often tends to be more relevant for navigating the social world.

Stereotype Content in LLMs

LLMs are generative AI models, trained on vast amounts of text data, which learn contextualized semantics and can generate human language in response to linguistic input (Christiano et al. 2017). Given their training from internet and other text data, LLMs reproduce many human stereotypes (Ghosh and Caliskan 2023). However, almost all the research on the topic has examined either general valence (i.e., positivity-negativity), a limited number of stereotype dimensions, or simple word associations (without identifying more generalizable dimensions of meaning). For example, research shows that many social categories have negative representations in AI models paralleling human stereotypes (Caliskan et al. 2017; Wolfe and Caliskan 2022; Busker et al. 2023), and that Warmth and Competence valence differences emerge in LLM stereotypes (Fraser et al. 2022; Ungless et al. 2022; Jeoung et al. 2023; Omrani et al. 2023).

More recent papers have looked at 3-dimensional models (e.g., Warmth/communion, Competence/status, and Beliefs, Koch et al. 2020; Cao et al. 2022; Cao et al. 2024; Schuster et al. 2024). However, these are still low-dimensional taxonomies that may not account for a vast majority of stereotypical associations in LLMs. Whether more dimensions (and which) are needed to understand AI stereotypes, has yet to be systematically examined. While we know that LLMs reproduce many human cultural associations (Nadeem et al. 2021; Dev et al. 2022; Jha et al. 2023; Davani et al. 2025), we should not assume this to always be the case. For example, training data may be biased towards particular content (e.g., because people may be more likely to talk about specific topics online vs. in other tasks or domains; see Luccioni and Viviano 2021), post-training safeguards may censor some content, or LMMs may place higher emphasis on semantic vs. inferential associations (e.g., "Wealthy person" with "money" rather than "selfish"). In addition, LLMs may learn to represent stereotypes based on specific subsets of humans (e.g., via reinforcement learning

from human feedback; RLHF; Mihalcea et al. 2025; Ouyang et al., 2022). Thus, although we hypothesize substantial overlap with the SSCM stereotype taxonomy, LLMs may show significant idiosyncrasies and distinct patterns of bias.

Additionally, most research has focused on specific social categories, such as gender or race (Duan et al. 2025; Garg et al. 2018; Caliskan et al. 2022). However, understanding more generalizable patterns of stereotype content requires examining larger and more representative samples of categories (Fiske et al. 2021). Previous research has also focused on examining text embeddings directly (Bolukbasi et al. 2016; Charlesworth et al. 2022), the numerical representations of text that underlie the conversational output of LLM chatbots. Here, we focus on the text output directly, as these constitute the final product in most applications and have the most direct impact on the general public.

Consequences of Stereotypical Associations

Stereotypes can be, in many cases, inaccurate, over-generalized, essentializing, or self-fulfilling, among other well-documented problematics (Bai, Nicolas, and Fiske, 2024; Bai et al. 2022), potentially resulting in discrimination, conflict, and adverse health impacts for stigmatized groups (Dovidio et al. 2017; Cipollina and Nicolas 2025; Salah et al. 2023). Both positive and negative stereotypes can be harmful (Kay et al. 2013), and their effects have been thoroughly documented. For example, stereotype content predicts outcomes such as emotional responses and interpersonal behaviors (Cuddy et al. 2007), hiring and performance evaluations (Cuddy et al. 2011), interactions across societal and organizational hierarchies (e.g., Friehs et al. 2022), and attitudes towards AI (McKee et al. 2023).

These consequences may be amplified, and stereotypes reinforced, via biased AI models. LLMs have become ubiquitous in applications with real-world impact, from healthcare (Rajpurkar 2022) to hiring (Cohen 2023). As with research and auditing, efforts to minimize harm from LLM stereotypes have so far focused on general valence, or in a few cases, on a limited number of dimensions and/or a small number of social categories (Fraser et al. 2022; Omrani et al. 2023). A more comprehensive taxonomy of LLM stereotypes is needed for more effective auditing and potential debiasing solutions for AI fairness.

Materials and Methods

All materials, data, and code are provided in the online repository: https://osf.io/bwdcr/.

Stimuli

We used a list of 1,366 different terms referring to 87 salient social categories in the U.S. For example, terms such as "wealthy," or "millionaire" were stimuli used to represent the "rich" social category. These terms have been validated and used successfully in previous LLM studies to elicit stereotype content (Nicolas and Caliskan 2024). See Table 1 for example categories and the Supplement for a full list.

Lawyers	Lower-class	American	Atheists
Teenagers	Athletes	Investors	Elderly
Accountants	Black	Geeks	Christians
Democrats	Gay	Women	Drug addicts
White-collar	Mexican	Disabled	Blind
Gamers	Hippies	Blue-collar	Doctors
Celebrities	Adults	Parents	Artists
Libertarians	Asian	Scientists	Hindus
White	Muslim	Jewish	Engineers
Nurses	Hipsters	Poor	Indian
Children	CEOs	Men	Buddhists
Vegans	Immigrants	Middle-class	Obese
Heterosexual	Republicans	Criminals	Germans
Politicians	Hackers	Bisexual	Religious
Catholics	Liberals	Homeless	Unemployed
Hispanic	Transgender	Lesbians	Rich

Table 1. Example social categories used as stimuli.

Language Models

We primarily focus on ChatGPT 4.5 for analyses but provide results from three additional models for main results: ChatGPT 3.5, Mixtral 8x7B, and Llama 3.

ChatGPT 4.5

We use GPT 4.5, as the state-of-the-art unsupervised learning ChatGPT model (OpenAI 2025). This LLM was trained on vast amounts of data, including the Common Crawl (a large scraping of internet webpages), books, Reddit, and Wikipedia (Brown et al. 2020), as well as RLHF (Christiano et al. 2017) and potentially others (OpenAI does not report all training sources). Testing suggests this model improves over other GPT models across multiple metrics, including reduced "hallucinations" and improved accuracy scores (OpenAI 2025). We used the Python OpenAI API to access the researcher version of the model.

ChatGPT 3.5

We use GPT 3.5 turbo as implemented in freely-available versions of ChatGPT until ~ August, 2024 (OpenAI 2022), accessed via the Python OpenAI API. The ChatGPT 3.5 model was trained on a smaller subset of data as the ChatGPT 4.5 model (Brown et al. 2020).

Mixtral

We use Mixtral 8x7B (with chatbot instruct fine-tuning), accessed via Python Transformers and HuggingFace. Mixtral is a "sparse mixture of experts" model (Mistral AI team 2023; Jiang et al. 2024). Unlike ChatGPT models, Mixtral has open weights, providing some additional level of transparency. Mixtral also has either similar or superior performance to ChatGPT 3.5 on various benchmarks (Jiang et al.

2024). As with ChatGPT, the training data is not transparently disclosed by the developers but is extracted "from the open Web" (Mistral AI team 2023).

Llama 3

We use Llama 3 8B (Meta AI 2024) as another open-weights LLM, accessed via Python HuggingFace. Llama 3 uses a decoder transformer architecture and achieves similar or superior performance to other similarly-sized open-weights LLMs. Unlike for the previous models, the Llama 3 model used here is the base model (i.e., not fine-tuned to behave as a chatbot). Using a base model allowed us to reduce the influence of safeguards introduced via chatbot fine-tuning.

Available information about Llama 3's training data includes that it consisted of over 15T tokens from public data, over 5% non-English data, involving filtering (e.g., NSFW filters), and other quality checks (Meta AI 2024).

Prompts

Prompts use language from human research requesting the LLMs to provide characteristics that either most Americans (cultural prompt) or the LLM (personal prompt) associate with social categories. In survey research, cultural framings allow for measurement of (perceptions of) societal stereotypes, shared across people, and helps reduce social desirability concerns of providing personal stereotypes (Fiske et al. 2002). These cultural stereotypes are nonetheless highly predictive of bias and discrimination intent (Cuddy et al. 2007; Fiske et al. 2021). Prompt language referring to cultural stereotypes may also reflect the aggregate (vs. individualized) nature of the associations LLMs learn. On the other hand, personal framings in human research tap into individual differences in beliefs about groups and may be more predictive of behavior (Devine 1989). We note that we are not trying to objectively anthropomorphize the LLMs (i.e., we are not implying personhood): we use the "personal" prompt and terminology to model the anthropomorphizing language that users may employ in chatbot communication, as well as potential training data that distinguishes between cultural and personal stereotypes explicitly, which the LLM may learn to distinguish. That is, although LLMs do not have a "personal" opinion, it is possible that they distinguish between cultural and more idiosyncratic stereotypes based on prompts, with potential consequences for the LLMs' use of these associations in other tasks.

As a system prompt, we entered: "You will be given information about a series of target groups of people." Followed by "Answer how most people in American society would respond" (cultural prompt) or "Provide your honest responses" (personal prompt).

Following the system prompt, for each stimulus term, we retrieved their stereotype associations by prompting: "List 50 characteristics that you believe most Americans would think describe people who are [*term*]. Use single words" (cultural prompt) or "List 50 characteristics that you believe describe people who are [*term*]. Use single words" (personal prompt). We prompted the LLM for 50 characteristics to obtain an expansive list of stereotypical associations.

For the storytelling task, we prompted ChatGPT 4.5: "Write a short story about Alex, who is [*term*]" or "Write a short story about a person who is [*term*]." Stimuli were modified (e.g., adding "an") as needed for grammar. We manipulated personal vs. cultural stories via the system prompts, using the same system prompts as for the stereotype task.

For analyses collapsing across personal and cultural prompts we averaged their embeddings. For stories, when conducting analyses for personal and cultural prompts separately, we averaged across the two versions of each prompt.

To obtain the most deterministic results, we set the temperature (a parameter that manipulates the randomness of the LLM's output) to 0 (Llama only allowed > 0 values and was set to 0.05). Repeated associations within a response were removed. Some output included warnings about bias, which were removed. We coded ChatGPT 4.5 as providing *any* warnings when information about bias, interpretation, or context was provided, either alongside or instead of the requested responses. We used an "*only* warnings" variable, when no responses were provided alongside the warning. In addition to warnings, the LLMs failed to return responses for terms it indicated are "not commonly used or understood in American society" (e.g., "mahanaya" for ChatGPT 3.5). For all 87 social categories, except the "Black" category in ChatGPT 3.5 (which returned only warnings) we successfully retrieved the requested output for at least one term.

For models other than ChatGPT 4.5 we focused on the cultural version of the prompts. We were interested in maximizing convergence with human data, which has most often asked about cultural stereotypes, in part to minimize socially desirable responding (c.f., chatbot safeguards), and because cultural stereotypes are predictive of relevant outcomes. And, as shown later, our results for ChatGPT 4.5 suggest that although cultural prompts result in more warnings, the content of personal and cultural prompts was highly overlapping and both impact model behavior in unrelated tasks.

Because the Llama 3 base model is trained for sentence completion rather than chat (Meta AI 2024), we slightly modified the prompt for sentence-completion rather question answering (see Supplement for these variations).

Statistical Analysis

We preprocessed the stereotype responses by transforming words from plural to singular, removing capitalization, and replacing non-content phrases (e.g., "most," "are").

Obtaining Text Embeddings

We obtained text embeddings for the LLMs' responses and stories. Embeddings are numerical vector representations of each response, encoding information about their semantics.

We use the embedding model SBERT (Devlin et al. 2019; Reimers and Gurevych 2019), which have fewer dimensions than those underlying the LLMs used here (making them more suitable for cluster analysis), are openly available (unlike ChatGPT's), are specialized for analyses of text similarity, and have shown validity in previous analyses of social perceptions (e.g., Nicolas, Uddenberg, and Todorov 2025). In addition, by using SBERT, our coding is independent from the internal representations of the LLMs used here, avoiding potential "double dipping" on an LLM's bias.

An exploratory analysis using Llama 3 embeddings provided similar, albeit lower quality, results (see Supplement). In addition, we note that because the cluster analysis uses only the responses, without connection to the social categories, it captures the semantic structure of the responses, not biases based on the category-response association.

Prompt Reliability

To test stability across multiple runs, for each of the 87 overarching social categories, we obtained the LLMs' responses for 50 runs with randomly sampled seeds. Then, we used the responses' text embeddings to obtain the average correlation between seeds for each category and computed the average Cronbach's α across all categories.

Prompt Validity

To establish that the prompts are capturing a relevant construct of category-characteristic associations and that it has convergent predictive validity, we correlate (using embeddings) ChatGPT 4.5's stereotypes for each category with an unrelated task of writing a story about a member of the category. For context, we compare this correlation against the background correlation. We computed the background correlation by randomizing the category labels, such that the stereotypes and categories were mismatched, and correlating with the stories. We conducted multiple tests, predicting all stories from all prompts, as well as separately predicting cultural and personal versions of the prompts.

Personal v. Cultural Comparisons

To compare personal and cultural prompts, we obtain the correlation between their embeddings for each corresponding category in ChatGPT 4.5. We also compare these correlations against the background correlation by randomizing the category labels. In additional analyses, we predict warnings from the prompt versions.

Cluster Analysis

With the embeddings, we computed a (dis)similarity matrix using pairwise cosine similarities between all the unique response embeddings (N = 5,871). For example, responses such as "fit" and "healthy" received higher cosine similarity scores than pairs such as "fit" and "black hair". We ran a k-means algorithm using the dissimilarity matrix. To select an appropriate number of clusters (k), we used the R package NBclust (Charrad et al. 2014), which runs multiple metrics of fit. We tested cluster sizes ranging from $k = 2$ to $k = 60$. To minimize the bias of any one measure of fit, we used 15 of the NBclust metrics that had continuous values (e.g., kl, ccc), standardized them, and obtained their average. Then, we chose the higher-dimensional solution ($k > 40$) with the highest index. This allowed us to start with a nuanced solution, and work down to a simpler taxonomy (as in Nicolas, Bai, and Fiske 2022). Chosen ks using this method were always in the top 5 solutions. In the Supplement, we present results for the best-fit solutions, showing how, given sufficient nuance, the resulting taxonomy is similar across ks.

To facilitate labeling by the researchers, we obtained, for each cluster, the 25 responses most semantically similar to the cluster centroid (which served as the most prototypical representation of the cluster). These top responses were labeled using various methods. Note that for this labeling we were interested in what dimension the clusters were *about*, not what end of a semantic differential (e.g., positive vs. negative) the content represented. Thus, both clusters with words related to immorality or morality would be labeled as being about the Morality dimension (see also Nicolas & Caliskan 2024).

First, we obtain a "free," data-driven coding by ChatGPT 4.5. We used the system prompt: "You will be given lists of terms that share a theme. Identify the theme using a single word." Then, for each cluster, we prompted: "I will give you a list of terms that describe people and that have a common theme. If you had to indicate what the terms are about, using a single word, what would it be? The list of terms is:" followed by the set of top words for the cluster. This data-driven coding allowed the authors to examine content that may not fall within the SSCM dimensions. Because most cluster labels using this approach could be considered to code for the SSCM dimensions (see results), we then used more theory-driven labeling, while still allowing for non-fitting labels (by providing an "Other" content option).

The first closed-ended labeling method also used ChatGPT 4.5. Using the same system prompt, for each cluster we prompted: "I will give you a list of words that describe people and that have a common theme. Please identify which of the following options best describes the theme: Morality, Sociability, Ability, Assertiveness/persistence, Emotion, Status, Appearance, Health, Beliefs, Uniqueness/averageness, Occupations, Social groups, Geography, Family, or Other. The list of words to classify is:" followed by the set of top words for the cluster.

The second closed-ended labeling method used human raters. Two research assistants were provided with instructions to categorize the top words into the SSCM dimensions. They were provided with a list of representative words for each dimension for context (Nicolas, Bai, and Fiske 2021).

The final closed-ended coding was obtained by correlating, for each cluster, the SBERT embeddings of the top words with embeddings of dictionaries for each dimension

(Nicolas, Bai, and Fiske 2021). The dimension with the highest correlation was taken as the label.

To establish the reliability of these codings, we used Kappa measures of inter-rater reliability (IRR). Finally, given these labels, we choose the majority dimension as the label for the cluster. In case of a tie, we code the cluster as having two labels (i.e., covering two dimensions).

We used the k-means results and labels to run prevalence analyses, showing which dimensions are used more commonly to describe social categories.

Regression Models

We had power > 90% for all tests, using a small-to-medium effect size (r = .2), as indicated by the R package *simr* (Green and MacLeod 2016).

For prevalence and story prediction analyses, we use maximal mixed regression models with category as a random factor (to account for non-independence), and responses to each term as an observation (reduced to linear models if convergence failed).

See Supplement for additional details and statistics.

Results

Prompt Reliability

We find that the LLMs' responses were almost identical across runs, suggesting consistency in the LLMs' representation of the social categories (given the specified temperature), Cronbach's αs > .998.

Prompt Validity

We find that the prompts, regardless of personal vs. societal framing, predicted ChatGPT 4.5's storytelling about targets, average r = .33 (Personal r = .32; Cultural r = .33). Specifically, stereotypes about a social category predicted stories about the category above what would be expected by background correlation (i.e., when the stereotypes are randomly shuffled across categories), average r = .26, Cohen's d = 1.43, p < .001. This effect is considered very large according to established guidelines, and it was robust when looking only at personal or societal prompts.

Thus, our prompts (partially) capture how ChatGPT 4.5 represents these groups in independent tasks that are likely use-cases of the technology, such as storytelling.

Personal vs. Cultural Prompts

Despite similar performance in predicting storytelling, personal and cultural prompts may differ in other ways, requiring further analyses. First, we found that the personal and cultural versions of the prompts for the same group were highly correlated, average r = .9, p < .001. Compared to background correlations, average r = .82, the effect size was very large, d = 2.45, p < .001.

However, personal and societal versions of the prompts did differ in the number of warnings provided. Specifically, ChatGPT 4.5 refused to provide responses more often for cultural (prob. = .031) than personal prompts (prob. = .005), p < .001. Similarly, it provided a warning alongside responses more often for cultural (prob. = .46) than personal prompts (prob. = .38), p = .003. Thus, although the LLM's stereotypes across prompts are highly overlapping, the safeguarding mechanism flags cultural prompts more often. This pattern counters usual findings in social psychology (where people acknowledge and report cultural stereotypes, but are less likely to personally endorse them), and suggests that using personal prompts may lead to higher exposure to stereotype content. Thus, our results suggest that ChatGPT makes relatively little distinction between societal and personal stereotypes, responding similarly to both prompts, and with both prompts predicting performance in an independent storytelling task. However, post-processing leads to higher frequency of warnings for cultural prompts.

Cluster Analysis

Given validation of our measures, we proceeded to use all prompts to identify content clusters. Fit metrics suggested various cluster sizes (k). As in previous research (Nicolas, Bai, and Fiske 2022) we opted for a higher-dimensional cluster solution to balance nuance and parsimony. Here, a solution of 48 clusters for ChatGPT 4.5 had good fit across multiple indices and was thus chosen. Table 2 shows cluster examples for this solution.

For other analyses, solutions varied in k when looking only at personal (k = 53) or societal (k = 58) ChatGPT 4.5 responses, as well as for other LLMs (ChatGPT 3.5 = 48, Llama 3 = 57, Mixtral = 55). However, as shown in the analyses below and the Supplement, the content of these clusters was highly overlapping, largely aligning with the SSCM.

Cluster Labeling

To understand what these clusters encode, we first provide results for data-driven labeling using ChatGPT 4.5 to identify the theme of each cluster's 25 top (most central) words.

Table 3 complements Table 2 by showing additional labels for ChatGPT 4.5 (see Supplement for full lists for all models). An examination of these labels suggests significant overlap, as expected, with established dimensions in human stereotype models (e.g., the SSCM). Thus, to identify overlap between the "free" labels and connect them to human models, we used the multi-method coding approach described previously, focusing on the SSCM dimensions.

For closed labeling, the IRR between the human ratings was .72, and between human ratings, SADCAT correlations, and GPT coding, it was .63. Both of these IRRs fall somewhere between "substantial" (> .6) and "almost perfect" (> .8) guidelines, suggesting excellent agreement. For all other LLMs, average IRRs were > .654.

Morality	**Sociability**	**Deviance**
Dishonesty	*Communication*	*Normality*
dishonest	talkative	typicality
deceitful	personable	typical
deceptive	relatable	commonplace
dishonorable	approachable	usual
unscrupulous	spoken	ordinary
Ability	**Assertiveness**	**Status**
Education	*Perseverance*	*Poverty*
educating	steadfast	underserved
educator	persistent	underemployed
teacher	persevering	underprivileged
learning	resilient	underrepresented
trained	toughness	underpaid
Beliefs	**Appearance**	**Health**
Religion	*Appearance*	*Fatigue*
religious	puffy	tired
religious-educated	beefy	lethargic
devout	chubby	fatigued
holy	flabby	snoozing
bible-believing	doughy	sluggish
Geography	**Family**	**Emotions**
Culture	*Family*	*Emotions*
ethnic	grandparent	unhappy
multicultural	grandparents	distraught
culturally-aware	parents	dejected
cultural	parental	disappointed
nationalistic	aged	discontented

Table 2. Example clusters for dimension identification, ChatGPT 4.5. Bolded rows indicate the cluster label, based on the multi-coding approach (majority label). Italicized rows indicate the GPT 4.5 data-driven cluster label, provided with no guidance as to previously identified human dimension labels. Top words (closest to the centroid) are provided for each cluster.

Five of the ChatGPT 4.5 clusters captured syntactic regularities (e.g., words starting with "un") rather than stereotype content and were not coded. As additional support for the "syntactic" labeling, we found that the maximum correlation between clusters and SADCAT dictionaries was significantly lower for clusters labeled as syntactic than for those coded into specific content, $p = .048$, $d = .96$. In other words, syntactic clusters tended to not fit well with any of the dictionary dimensions, suggesting that they contained content mixture, and were instead similar based on non-semantic features.

Six clusters were labeled as reflecting a mixture of two dimensions (based on a tie on the multi-method labeling). These overlapping codes tended to reflect theoretically related dimensions (e.g., Sociability and Morality, facets of the Warmth dimension). Similar patterns of syntactic and mixture clusters occurred for the rest of the LLMs (see Supplement).

Cluster Prevalence

Our multi-method labeling of clusters suggest that the SSCM's most primary 14 dimensions were sufficient to cover the content of LLMs' explicit stereotype associations. In Table 4 and Figure 1, we present analyses of how frequent cluster codings for each dimension were in stereotypes for ChatGPT 4.5. We find that, as in human data, the facets of Warmth and Competence tended to be most prevalent (although Morality stereotypes were relatively less common in the LLMs). However, the rest of the SSCM dimensions also emerge with meaningful frequency.

GPT 4.5 Free Code	**Majority Code**	**GPT 4.5 Free Code**	**Majority Code**
Aggression	Morality	Isolation	Assertiveness
Adaptability	Ability	Technology	Ability
Government	Beliefs	Kindness	Sociability
Politics	Beliefs	Neglect	Ability
Sports	Ability	Environment	Beliefs
Music	Other	Excellence	Ability
Happiness	Emotion	Work	Occupation
Nutrition	Health	Calmness	Emotion
Tradition	Deviance	Mathematics	Ability
Travel	Geography	Misfortune	Status
Intolerance	Sociability	Fashion	Appearance
Risk	Assertiveness	Finance	Status
Personality	Sociability	Carefulness	Assertiveness
Status	Status	Fantasy	Other
Uncertainty	Deviance	Appearance	Appearance
Thinking	Ability		

Table 3. Additional ChatGPT 4.5 cluster labels not included in Table 2. We show the GPT 4.5 "free" code (label options not provided), and the majority code (across all closed-ended labeling methods). For some clusters, labeling across methods resulted in a tie, in which case a second majority code was included (not shown). Syntactic clusters are not included. See Supplement for more information.

Human Baseline		All Prompts	
Dimension	**Prevalence**	**Dimension**	**Prevalence**
Ability	0.177	Ability	.29[a]
Morality	0.158	Sociability	.153[b]
Sociability	0.157	Assertiveness	.15[b]
Assertiveness	0.142	Status	.124[c]
Status	0.094	Emotion	.101[d]
Appearance	0.08	Beliefs	.088[e]
Emotion	0.076	Morality	.085[e]
Beliefs	0.069	Appearance	.051[f]
Deviance	0.038	Deviance	.05[f]
Health	0.033	Other	.02[g]
Work	0.023	Geography	.017[g]
Other	0.022	Family	.011[g]
Social groups	0.021	Health	.011[g]
Geography	0.015	Work	.008[g]
Family	0.004		

Table 4. ChatGPT 4.5 cluster content and prevalence. Human baseline obtained from Nicolas, Bai, and Fiske 2022 (note that human results may vary across studies/moderators). Prevalence values in rows sharing a superscript letter are not significantly different from each other ($p > .05$).

Results suggest high overlap across all (combined), personal, and cultural prompts, with solutions sharing the same dimensions (except for the combined solution showing no evidence of a "Social Groups" cluster). However, differences in prevalence emerge (e.g., higher prevalence of Morality content in personal prompts, but higher prevalence of Status content in cultural prompts (see Table 5).

Similarly, despite substantial consistency with the SSCM in terms of which dimensions are represented, nuances emerge in the relative prevalence of content, showing how LLMs may differ from human representations, and from representations in other LLMs (see Table 6). To illustrate, ChatGPT 4.5 had higher prevalence of Sociability content, compared to the other LLMs, and it was more aligned with human patterns. On the other hand, Llama 3, an LLM without chatbot fine-tuning, showed a higher prevalence of Morality content. This may suggest a role of chatbot training and safeguarding on the prevalence of specific dimensions.

See the Supplement for results showing robustness of the main findings, including using different high-fit *k*s and embeddings (Llama 3 embeddings). The supplementary materials and code also allow for exploration of stereotypes for specific social categories and LLMs (see the Discussion for an example of prevalence patterns for the "poor" and "wealthy" social categories).

Personal Prompts		Cultural Prompts	
Dimension	**Prevalence**	**Dimension**	**Prevalence**
Ability	.231[a]	Ability	.211[a]
Assertiveness	.189[b]	Assertiveness	.144[b]
Morality	.189[b]	Status	.137[b]
Sociability	.166[c]	Sociability	.114[c]
Emotion	.099[d]	Emotion	.080[d]
Status	.050[e]	Beliefs	.079[d]
Deviance	.044[ef]	Appearance	.066[e]
Appearance	.036[fg]	Morality	.060[ef]
Beliefs	.031[g]	Other	.050[fg]
Health	.026[gh]	Deviance	.048[g]
Geography	.016[hi]	Work	.027[h]
Other	.012[ij]	Geography	.025[h]
Family	.010[ij]	Health	.024[h]
Work	.004[j]	Groups	.008[i]
Social Groups	.004[j]	Family	.005[i]

Table 5. ChatGPT 4.5 cluster content and prevalence for cluster solutions based only on personal or cultural prompts. Within columns, values in rows sharing a superscript letter are not significantly different ($p > .05$).

Discussion

This study introduces a nuanced taxonomy of stereotype content in contemporary LLMs. We prompted ChatGPT 4.5, ChatGPT 3.5, Llama 3, and Mixtral 8x7b to provide stereotypes about a large number of salient social categories. We then cluster-analyzed these associations using text embeddings to identify the LLMs' stereotype content.

The dimensions identified largely align with the content of human stereotypes described by the SSCM. Specifically, LLMs' stereotypes were about the social categories' Ability, Appearance, Assertiveness, Beliefs, Deviance, Emotion, Family, Geography, Health, Morality, Occupations, Social groups, Sociability, and Status. Smaller content related to literature, music, and cultural associations also emerged.

Llama 3		ChatGPT 3.5		Mixtral	
Dimension	**Prev**	**Dimension**	**Prev**	**Dimension**	**Prev**
Ability	.32[a]	Assertive	.26[a]	Assertive	.22[a]
Morality	.18[b]	Ability	.17[b]	Ability	.16[b]
Assertive	.09[c]	Beliefs	.12[c]	Morality	.10[c]
Status	.09[c]	Sociability	.07[d]	Beliefs	.09[c]
Appearance	.08[c]	Appearance	.06[de]	Other	.09[cd]
Beliefs	.08[cd]	Deviance	.05[ef]	Work	.07[de]
Emotion	.06[de]	Morality	.041[fg]	Emotion	.07[de]
Health	.05[ef]	Groups	.04[fgh]	Sociability	.07[ef]
Sociability	.04[fg]	Other	.03[fghi]	Status	.05[fg]
Other	.03[gh]	Status	.03[fghij]	Appearance	.05[fg]
Deviance	.03[gh]	Geography	.02[ghij]	Groups	.04[g]
Groups	.01[hi]	Health	.02[ghij]	Health	.04[g]
Family	.01[hi]	Emotion	.02[hij]	Geography	.04[g]
Geography	.01[i]	Work	.01[ij]	Family	.01[h]
Work	.01[i]	Family	.01[j]		

Table 6. Secondary LLMs' cluster content and prevalence (Prev). Assertive = Assertiveness and Groups = Social groups. Within-column shared superscript = *ns* difference.

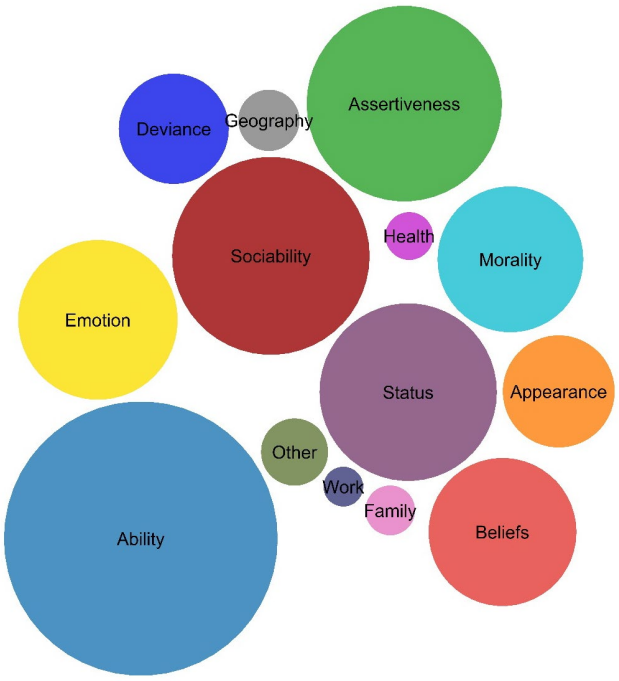

Figure 1. Main taxonomy dimensions. Circle size indicates prevalence in ChatGPT 4.5 (averaging across personal and cultural prompts prior to analysis). The "other" cluster includes topics such as art and culture. A "social groups" cluster (e.g., associations such as "CEOs are men") emerged for specific LLMs, but not the combined data shown here. Additional clusters coded for syntactic information, rather than capturing coherent semantic themes.

However, despite sharing content dimensions, humans and LLMs differed in terms of the prevalence of specific dimensions. For example, moral content was less prevalent in ChatGPT, while Llama 3 and Mixtral had lower prevalence of Sociability stereotypes (vs. the human baseline). As suggested, different LLMs' associations also varied in prevalence of specific dimensions. For example, Llama 3, which was not fine-tuned for chatbot functionality in our study, showed higher prevalence of Morality content than other LLMs. Moral stereotypes tend to be negative (Nicolas, Bai, and Fiske 2022), suggesting a potential role of safeguards in the less moralized content of ChatGPT and Mixtral.

Moreover, despite high correlations in the general content of personal and cultural prompts in ChatGPT 4.5, cultural prompts tended to have higher prevalence of epistemic, sociological dimensions such as Status and Beliefs, while personal prompts included more relational, psychological dimensions such as Morality, in line with related human research (Nicolas et al. 2022). In general, results suggest high overlap across prompt versions, with all solutions sharing the same dimensions (except for the combined prompts solution showing no "Social Groups" cluster). However, differences in prevalence emerge, for example, with a higher prevalence of Morality content in personal prompts, but higher prevalence of Status content in cultural prompts.

Implications

The taxonomy introduced here provides a nuanced view of stereotype associations in LLMs. While previous research, auditing, and debiasing has tended to focus on general valence or just 2 or 3 dimensions, our paper suggests that understanding LLMs' associations with social categories requires a much wider set of dimensions. These dimensions also describe stereotypes in humans (SSCM), as well as face and other person perceptions (Connor et al. 2025; Nicolas, Uddenberg, & Todorov 2025), supporting their relevance and the generalizability of psychological models. These dimensions predict prejudice towards and decision making about social targets. Warmth and Competence are predictors of outcomes ranging from hiring and performance evaluations to interpersonal behaviors (e.g., Cuddy et al. 2007).

Beyond Warmth and Competence, the rest of the taxonomy also predicts scenario-based decision-making outcomes such as which social categories to prioritize for policies guaranteeing access to healthcare (Health dimension) or protection from hiring (Social groups, Geography) and facial recognition discrimination (Appearance; Nicolas, Bai, and Fiske 2022). Understanding how these dimensions are reflected in LLMs can expand the ways in which we measure stereotypes relevant to these outcomes (e.g., over time, across languages), with implications for social psychological theory and interventions (Bailey et al. 2022; Boyd and Schwartz 2021; Muthukrishna et al. 2021; Jackson et al. 2022). However, such inferences from LLM to human cognition must carefully consider training data (transparency and biases), fine-tuned safeguards that may distort cultural patterns, and the potential for LLMs reflecting novel or distinct stereotypes due to non-transparent synthesis and processing of information (e.g., Bianchi et al. 2023).

More directly relevant to LLM development and use, a deeper understanding of the multidimensionality of stereotypes can help prevent biases from percolating through auditing and debiasing approaches focused on general indicators or low-dimensional representations. Developing benchmarks and debiasing procedures that address higher-dimensional stereotype taxonomies will provide a more accurate picture of fairness in LLMs and responsible applications. For example, auditing efforts focused solely on Warmth or Competence would miss biases along other dimensions, such as stereotypes about whether a group is unhealthy, deviating from norms, or emotional. To illustrate, exploratory analyses of specific categories in our data suggest that differences in stereotypes about high vs. low socioeconomic

categories vary across dimensions, with low (vs. high) socioeconomic categories showing particularly negative Ability, Assertiveness, Appearance, Emotion, and Health stereotypes. The taxonomy introduced here provides a more comprehensive set of content to evaluate. These steps may reduce harmful exposure to stereotypes for stigmatized groups, reduce the perpetuation of stereotypes via AI, and improve human-computer interaction, among other benefits.

The highly overlapping outputs from personal vs. cultural versions of the prompts suggest that LLMs may not make pronounced distinctions in their representations of concepts along these micro vs. macro levels (these analysis were only performed on the ChatGPT 4.5 model but see also Nicolas and Caliskan 2024 for convergent evidence in other LLMs). Additionally, both personal and cultural prompts similarly predicted storytelling about the targets, showing that ChatGPT draws from the general associations captured by explicit association prompts to complete unrelated tasks.

However, the cluster analyses did reveal differences in the prevalence of specific dimensions, aligning with emphasis shifts for specific content based on framing, such that cultural framings make structural and epistemic dimensions (e.g., Status and Beliefs) salient, while personal framings may highlight more interpersonal dimensions (e.g., Sociability; Nicolas et al. 2022). Moreover, differences emerged in number of warnings (fewer warnings for personal than cultural prompts), suggesting that, although the LLM does not learn to strongly differentiate personal vs. cultural associations, fine-tuning for safety introduces a stronger distinction. Future research and auditing can leverage these insights to understand impact on users.

Limitations and Future Directions

The current research is not without limitations. First, our results are US- and English- centric, due to the training data of most LLMs. However, initial cross-cultural research with human participants showed fair stability of the SSCM taxonomy (Nicolas, Bai, and Fiske 2022). Second, the LLMs used lack transparency regarding training data and implemented safeguards. As such, our ability to connect LLM representations to cultural representations is limited. Third, we restrict our results to four recent models in a growing field of LLMs. However, the striking consistency between these LLMs' associations, and human data, suggests that this taxonomy may be robust, with variability across specific properties (e.g., prevalence, valence) to be further studied in future research. Fourth, we focused on cultural prompts for our secondary LLMs. Although ChatGPT 4.5 showed high overlap between prompts, it is possible that other LLMs show higher differentiation between cultural and personal prompts, which should be tested. A related future direction should explore the taxonomy in less explicit prompts that may elicit distinct stereotypes (e.g., Bai et al. 2024). Fifth, here we used a data-driven analysis to achieve an initial identification of dimensions. However, other methods (e.g., dictionary analyses, text embeddings similarities) would address some of the limitations of cluster analyses (e.g., terms forced into clusters with low fit, clusters formed based on syntactic rather than semantic information) and provide additional insights (e.g., differentiate prevalence and valence; Nicolas and Caliskan 2024). Additional future directions include expanding the taxonomy to intersectional targets (Guo and Caliskan 2021; Nicolas and Fiske 2023), exploring human-LLM stereotype differences, and developing auditing and debiasing methods incorporating the taxonomy.

Finally, we note that the current taxonomy includes dimensions that could be broken down or combined. For example, to align with the SSCM, we combined clusters of politics and religion into an overarching Beliefs dimension. On the other hand, we break down the big two of Warmth and Competence into their facets of Morality and Sociability, and Ability and Assertiveness (Abele et al. 2021). This taxonomy aims to balance nuance with a manageable number of dimensions. Based on goals of parsimony and generalizability vs. complexity and specificity, researchers may use a variety of methods to tap into different levels of content. Finally, taxonomy properties may change as LLMs undergo further iterations and safeguarding modifications, requiring future evaluations of stereotype content.

Conclusion

A more complete understanding of the biases encoded into increasingly influential AI technologies requires acknowledging the multidimensionality of stereotypes. The LLM stereotype taxonomy we identified largely aligns with human models, such as the SSCM, while showing unique patterns in the prevalence of specific dimensions across prompts and LLMs. As AI continues to be developed and deployed, our findings suggest that auditing and debiasing efforts should attend to the complexities of stereotypes, in an effort to minimize their harmful consequences.

Appendix

Supplementary materials are available at: https://osf.io/bwdcr/

Acknowledgements

We are grateful to the anonymous reviewers for their helpful feedback. This work was supported by the U.S. National Science Foundation (NSF) CAREER Award 2337877. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NSF.

References

Abele, A. E.; Ellemers, N.; Fiske, S. T.; Koch, A.; and Yzerbyt, V. 2021. Navigating the Social World: Toward an Integrated Framework for Evaluating Self, Individuals, and Groups. *Psychological Review* 128(2): 290–314. doi.org/10.1037/rev0000262.

Bai, X.; Fiske, S. T.; and Griffiths, T. L. 2022. Globally Inaccurate Stereotypes Can Result from Locally Adaptive Exploration. *Psychological Science* 33(5): 671–684. doi.org/10.1177/09567976211045929.

Bai, X.; Nicolas, G.; and Fiske, S. T. 2024. Social Stereotypes: Content and Process. In *The Oxford Handbook of Social Cognition, Second Edition*. Edited by D. E. Carlston, K. Hugenberg, K. L. Johnson, 442–470. New York: Oxford University Press.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2024. Explicitly Unbiased Large Language Models Still Form Biased Associations. *Proceedings of the National Academy of Sciences* 122(8): e2416228122. doi.org/10.1073/pnas.2416228122.

Bailey, A. H.; Williams, A.; and Cimpian, A. 2022. Based on Billions of Words on the Internet, People = Men. *Science Advances* 8(13): eabm2463. doi.org/10.1126/sciadv.abm2463.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1493–1504. New York: Association for Computing Machinery. doi.org/10.1145/3593013.3594095

Bodenhausen, G. V., Macrae, C. N. 1998. Stereotype Activation and Inhibition. In *Stereotype Activation and Inhibition*, edited by R. S. Wyer, 1–52. New Jersey: Lawrence Erlbaum Associates Publishers.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), 4356–4364. New York: Curran Associates, Inc. doi.org/10.5555/3157382.3157584

Boyd, R. L., Schwartz, H. A. 2021. Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology* 40(1): 21–41. doi.org/10.1177/0261927X20967028.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20), 1877–1901. New York: Curran Associates, Inc. doi.org/ 10.5555/3495724.3495883

Busker, T.; Choenni, S.; and Shoae Bargh, M. 2023. Stereotypes in ChatGPT: An Empirical Study. In Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance, 24–32. New York: Association for Computing Machinery. doi.org/10.1145/3614321.3614325

Caliskan, A.; Ajay, P. P.; Charlesworth, T.; Wolfe, R.; and Banaji, M. R. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 156–170. New York: Association for Computing Machinery. doi.org/10.1145/3514094.3534162

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science* 356(6334): 183–186. doi.org/10.1126/science.aal4230

Cao, Y. T.; Sotnikova, A.; Daumé III, H.; Rudinger, R.; and Zou, L. 2022. Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1276–1295. Seattle: Association for Computational Linguistics. doi.org/10.18653/v1/2022.naacl-main.92

Cao, Y. T.; Sotnikova, A.; Zhao, J.; Zou, L. X.; Rudinger, R.; and Daumé III, H. 2024. Multilingual Large Language Models Leak Human Stereotypes Across Language Boundaries. arXiv:2312.07141

Charlesworth, T. E. S.; Caliskan, A.; and Banaji, M. R. 2022. Historical Representations of Social Groups Across 200 Years of Word Embeddings from Google Books. *Proceedings of the National Academy of Sciences* 119(28):e2121798119. doi.org/10.1073/pnas.2121798119

Charrad, M.; Ghazzali, N.; Boiteau, V.; and Niknafs, A. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* 61(6):1–36. doi.org/10.18637/jss.v061.i06.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17) , 4302–4310. Red Hook, NY: Curran Associates Inc. doi.org/10.5555/3294996.3295184

Cipollina, R., Nicolas, G. 2025. Characterizing Stereotypes That Perpetuate Sexual Minorities' Anticipated Stigma in Healthcare Settings. *Psychology of Sexual Orientation and Gender Diversity*. doi.org/10.1037/sgd0000822.

Connor, P.; Nicolas, G.; Antonoplis, S.; and Koch, A. 2025. Unconstrained Descriptions of Facebook Profile Pictures Support High-Dimensional Models of Impression Formation. *Personality and Social Psychology Bulletin*. doi.org/10.1177/01461672241266651.

Cuddy, A. J. C.; Fiske, S. T.; and Glick, P. 2007. The BIAS Map: Behaviors from Intergroup Affect and Stereotypes. *Journal of Personality and Social Psychology* 92(4): 631–648. doi.org/10.1037/0022-3514.92.4.631.

Cuddy, A. J. C.; Glick, P.; and Beninger, A. 2011. The Dynamics of Warmth and Competence Judgments, and Their Outcomes in Organizations. *Research in Organizational Behavior* 31:73–98. doi.org/10.1016/j.riob.2011.10.004.

Davani, A.; Dev, S.; Pérez-Urbina, H.; and Prabhakaran, V. 2025. A Comprehensive Framework to Operationalize Social Stereotypes for Responsible AI Evaluations. arXiv.2501.02074.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; and Chang, K. 2022. On Measures of Biases and Harms in NLP. In Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, 246–267, Online:Association for Computational Linguistics. doi.org/10.18653/v1/2022.findings-aacl.24

Devine, P. G. 1989. Stereotypes And Prejudice: Their Automatic and Controlled Components. Journal of Personality and Social Psychology 56(1): 5–18. doi.org/10.1037/0022-3514.56.1.5.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi.org/10.18653/v1/N19-1423

Dovidio, J. F.; Love, A.; Schellhaas, F. M. H.; and Hewstone, M. 2017. Reducing Intergroup Bias Through Intergroup Contact: Twenty Years of Progress and Future Directions. Group Processes and Intergroup Relations 20(5):606–620. doi.org/10.1177/1368430217712052

Duan W.; Li L.; Freeman G.; and McNeese N. 2025. A Scoping Review of Gender Stereotypes in Artificial Intelligence. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–20. New York, NY, USA: Association for Computing Machinery. (CHI '25). doi.org/10.1145/3706598.3713093.

Fiske, S. T.; Cuddy, A. J. C.; Glick, P.; and Xu, J. 2002. 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): 878–902. doi.org/10.1037/0022-3514.82.6.878.

Fiske, S. T.; Nicolas, G.; and Bai, X. 2021. The stereotype content model: How we make sense of individuals and groups. In Social Psychology: Handbook of Basic Principles, 3rd ed, edited by P. A. M. Van Lange; E. T. Higgins; A. W. Kruglanski, 392–410. New York, NY: The Guilford Press.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2022. Computational Modeling of Stereotype Content in Text. Frontiers in Artificial Intelligence 5. doi.org/10.3389/frai.2022.826207.

Friehs, M.-T. ; Kotzur, P. F.; Kraus, C.; Schemmerling, M.; Herzig, J. A.; Stanciu, A.; Dilly, S.; Hellert, L.; Hübner, D.; Rückwardt, A.; et al. 2022. Warmth and Competence Perceptions of Key Protagonists Are Associated with Containment Measures during the COVID-19 Pandemic: Evidence from 35 Countries. Scientific Reports 12(1): 21277. doi.org/10.1038/s41598-022-25228-9.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences 115(16): E3635–E3644. doi.org/10.1073/pnas.1720347115.

Ghosh, S., Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages. arXiv preprint. arXiv:2305.10510.

Green, P., MacLeod, C. J. 2016. SIMR: An R Package for Power Analysis of Generalized Linear Mixed Models by Simulation. Methods in Ecology and Evolution 7(4): 493–498. doi.org/10.1111/2041-210X.12504.

Guo, W., Caliskan, A. 2021. 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, 122–133. New York: Association for Computing Machinery. doi.org/10.1145/3461702.3462536.

Jackson, J. C.; Watts, J.; List, J.-M.; Puryear, C.; Drabble, R.; and Lindquist, K. A. 2021. From Text to Thought: How Analyzing Language Can Advance Psychological Science. Psychological Science 17(3), 805-826. doi.org/10.1177/17456916211004899

Jeoung, S.; Ge, Y.; and Diesner, J. 2023. Stereomap: Quantifying the Awareness of Human-Like Stereotypes in Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12236–12256, Singapore: Association for Computational Linguistics. doi.org/10.18653/v1/2023.emnlp-main.752

Jha, A.; Mostafazadeh Davani, A.; Reddy, C. K.; Dave, S.; Prabhakaran, V.; and Dev, S. 2023. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 9851–9870. Toronto, Canada: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. de las; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of Experts. arXiv:2401.04088.

Kay, A. C.; Day, M. V.; Zanna, M. P.; and Nussbaum, A. D. 2013. The Insidious (And Ironic) Effects of Positive Stereotypes. Journal of Experimental Social Psychology 49(2): 287–291. doi.org/10.1016/j.jesp.2012.11.003.

Koch, A.; Imhoff, R.; Dotsch, R.; Unkelbach, C.; and Alves, H. 2016. The ABC Of Stereotypes About Groups: Agency/Socioeconomic Success, Conservative–Progressive Beliefs, And Communion. Journal of Personality and Social Psychology 110(5): 675–709. doi.org/10.1037/pspa0000046.

Koch, A.; Imhoff, R.; Unkelbach, C.; Nicolas, G.; Fiske, S.; Terache, J.; Carrier, A.; and Yzerbyt, V. 2020. Groups' Warmth Is a Personal Matter: Understanding Consensus on Stereotype Dimensions Reconciles Adversarial Models of Social Evaluation. Journal of Experimental Social Psychology 89. doi.org/10.1016/j.jesp.2020.103995.

Macrae, C. N., Bodenhausen, G. V. 2000. Social cognition: Thinking Categorically About Others. Annual Review of Psychology 51(1): 93–120. doi.org/10.1146/annurev.psych.51.1.93

McKee, K. R.; Bai, X.; and Fiske, S. T. 2023. Humans Perceive Warmth and Competence in Artificial Intelligence. iScience 26(8):107256. doi.org/10.1016/j.isci.2023.107256.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-06- 24.

Mihalcea R.; Ignat O.; Bai L.; Borah A.; Chiruzzo L.; Jin Z.; Kwizera C.; Nwatu J.; Poria S.; and Solorio T. 2025. Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone. In Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence 39(27), 28657–28670. Philadelphia, Pennsylvania: Association for the Advancement of Artificial Intelligence. doi.org/10.1609/aaai.v39i27.35092.

Mistral AI team. 2023. Mixtral of experts: A high quality sparse mixture-of-experts. https://mistral.ai/news/mixtral-of-experts/. Accessed: 2024-04-20.

Muthukrishna, M.; Henrich, J.; and Slingerland, E. 2021. Psychology as a Historical Science. Annual Review of Psychology 72: 717–749. doi.org/10.1146/annurev-psych-082820-111436.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. 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), 5356–

5371, Online: Association for Computational Linguistics. doi.org/10.18653/v1/2021.acl-long.416

Nicolas, G.; Bai, X.; and Fiske, S. T. 2022. A Spontaneous Stereotype Content Model: Taxonomy, Properties, and Prediction. *Journal of Personality and Social Psychology* 123(6):1243–1263. doi.org/10.1037/pspa0000312.

Nicolas, G., Caliskan, A. 2024. Directionality And Representativeness Are Differentiable Components of Stereotypes in Large Language Models. *PNAS Nexus* 3(11): pgae493. doi.org/10.1093/pnasnexus/pgae493.

Nicolas, G., Fiske, S. T. 2023. Valence Biases and Emergence in the Stereotype Content of Intersecting Social Categories. *Journal of Experimental Psychology: General* 152(9): 2520–2543. doi.org/10.1037/xge0001416.

Nicolas, G.; Fiske, S. T.; Koch, A.; Imhoff, R.; Unkelbach, C.; Terache, J.; Carrier, A.; and Yzerbyt, V. 2022. Relational Versus Structural Goals Prioritize Different Social Information. *Journal of Personality and Social Psychology* 122(4): 659–682. doi.org/10.1037/pspi0000366.

Nicolas, G.; Uddenberg, S.; and Todorov, A. 2025. Spontaneous content of impressions of naturalistic face photographs. *Social Cognition*. 43(2): 114-143. https://doi.org/10.1521/soco.2025.43.2.114

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-9-8.

Open AI. 2025. Introducing GPT-4.5. https://openai.com/index/introducing-gpt-4-5/. Accessed: 2025-05-06.

Omrani, A.; Salkhordeh Ziabari, A.; Yu, C.; Golazizian, P.; Kennedy, B.; Atari, M.; Ji, H.; and Dehghani, M. 2023. Social-Group-Agnostic Bias Mitigation via the Stereotype Content Model. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4123–4139. Toronto, Canada: Association for Computational Linguistics. doi.org/10.18653/v1/2023.acl-long.227

Ouyang L.; Wu J.; Jiang X.; Almeida D.; Wainwright C. L.; Mishkin P.; Zhang C.; Agarwal S.; Slama K.; Ray A.; Schulman, J.; Hilton, J.; Kelton F.; Miller L.; Simens M.; Askell A.; Welinder P.; Christiano P.; Leike J.; and Lowe R. 2022. Training language models to follow instructions with human feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems, 27730–27744. Red Hook, NY, USA: Curran Associates Inc. (NIPS '22). doi.org/10.5555/3600270.3602281

Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature Medicine* 28(1):31–38. doi.org/10.1038/s41591-021-01614-0.

Reimers, N., Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. Hong Kong, China: Association for Computational Linguistics. doi.org/10.18653/v1/D19-1410.

Salah M.; Alhalbusi H.; Ismail M. M.; and Abdelfattah F. 2024. Chatting With Chatgpt: Decoding the Mind of Chatbot Users and Unveiling the Intricate Connections Between User Perception, Trust and Stereotype Perception on Self-Esteem and Psychological Well-Being. *Current Psychology* 43(9): 7843–7858. doi.org/10.1007/s12144-023-04989-0.

Schuster C. M.; Roman M-A.; Ghatiwala S.; and Groh G. 2025. Profiling Bias in LLMs: Stereotype Dimensions in Contextual Word Embeddings. In Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies, 639–650. Tallinn, Estonia: University of Tartu Library. https://aclanthology.org/2025.nodalida-1.65/.

Suliteanu J.; Ofosu E. K.; Paquin Domingues A.; and Hehman E. 2025. Prejudice And Stereotypes at Regional and Individual Levels: Related but Distinct. *Journal of Personality and Social Psychology* 128(4): 807–820. doi.org/10.1037/pspa0000433.

Ungless, E.; Rafferty, A.; Nag, H.; and Ross, B. 2022. A Robust Bias Mitigation Procedure Based on the Stereotype Content Model. In Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), 207–217. Abu Dhabi, UAE: Association for Computational Linguistics. doi.org/10.18653/v1/2022.nlpcss-1.23.

Wolfe, R., Caliskan, A. 2022. VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models. In Proceedings of the AAAI Conference on Artificial Intelligence 36(10):11477–11485. doi.org/10.1609/aaai.v36i10.21400.