

# Probing Fine-Grained Stereotypes and Trait Structure in Large Language Models

Rebecca Yu<sup>\*♣</sup>, Nicholas Deas<sup>\*♠</sup>,  
Colin Wayne Leach<sup>\*♦</sup>, Kathleen McKeown<sup>♠</sup>

♠ Columbia University, Department of Computer Science

♣ Columbia University, Department of Psychology and

Institute for Research in African American Studies and Data Science Institute

♦ Barnard College, Departments of Psychology and Africana Studies

Correspondence: ndeas@cs.columbia.edu, rjy2109@columbia.edu

## Abstract

We study fine-grained stereotypes of LLMs activated by linguistic features representing demographic groups, developing a probe-based framework to extract 235 fine-grained traits (e.g., *friendly*, *uneducated*, *honest*) from model representations of prompts. Using these probes, we conduct two studies analyzing the underlying structure of traits and specific stereotypes of demographic groups. We find that fine-grained trait information is represented in LLMs similarly to coarse stereotype content dimensions (e.g., warmth and competence), and that the latent factors of trait probes (e.g., *morality*, *competence*, *sociability*) largely align with those in the psychology literature (Fiske et al., 2002) despite traits being modeled independently. Finally, we extract LLMs’ associations between traits and demographics signaled through language to surface granular stereotypes encoded in LLMs. We, for example, find that the evaluated models encode associations between African American Language and traits *clumsy*, *uneducated*, and *criminal*. These associations may reinforce harmful stereotypes, with implications for both general dialogue responses and in domains like hiring. Overall, we discuss the need to consider fine-grained traits in stereotype evaluations in order to distinguish stereotypes of out-groups and between models.<sup>1</sup>

## 1 Introduction

Stereotypes have received significant attention in the computational linguistics community (e.g., Cheng et al. 2023; Shrawgi et al. 2024; Nadeem et al. 2021). Through a variety of methods, these

<sup>\*</sup> denotes equal contribution

<sup>1</sup>Code is made available at <https://github.com/rebecca2015/fine-grained-stereotype-probing/> and interactive results are made available at <https://nicholasdeas.com/blog/fine-grained-stereotypes/>.

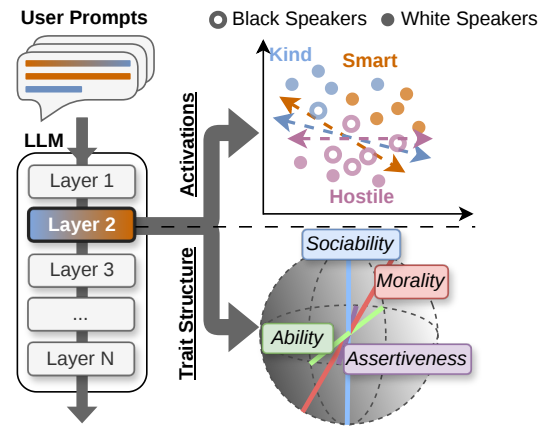


Figure 1: We develop probes to extract fine-grained traits from LLM hidden representations, enabling study of fine-grained stereotype activation in LLMs and the structure of LLMs’ trait spaces.

studies have considered the harmful associations LLMs form between concepts and various social groups. For example, through generated descriptions of demographic personas, studies have shown that LLMs associate Black women with *beauty* but also *strength* and *resilience* which reinforce the harmful stereotype of the "Strong Black Woman" (Cheng et al., 2023; Woods-Giscombé, 2010).

In studies that surface examples like this, social groups are often operationalized through explicit mention (e.g., associations with the term "women") or socio-demographic proxies (e.g., names associated with ethnicity; Gautam et al. 2024; Lutz et al. 2025). We argue, however that it is important to also consider how such stereotypes are *activated* by *identity cues* in language. For example, LLMs’ treatment of prompts in African American Language (AAL) (Deas et al., 2023, 2025; Hofmann et al., 2024; McKeown et al., 2026) and other minoritized varieties of English (Fleisig et al., 2024;

Basoah et al., 2025) appear to surface stereotypes that can lead to disparities in responses and other downstream harms. Beyond stereotypes arising from explicit descriptions of demographics, focusing on how stereotypes are activated by language captures another critical component of the harms that may arise when marginalized users interact with LLMs.

Other studies of LLM stereotypes also often draw on social psychological models of stereotype content, which summarize the underlying dimensions or facets of human social perceptions. Deas and McKeown (2025) explicitly study the activation of stereotypes by developing probes based on one such two-dimensional model (*warmth* and *competence*). Prior work, however, has noted inconsistencies in how these dimensions are defined as well as the difficulty of distinguishing stereotypes of out-groups using a small number of dimensions; Black people and women may both be stereotypically associated with low competence, but face distinct stereotypes and resulting harms. Accordingly, Leach et al. (2008) instead inductively derive latent stereotype dimensions as well as consider the salient, fine-grained traits that can more specifically characterize unique societal stereotypes of different social groups.

Therefore, in this work, we use a probe-based approach to study how models encode 235 unique, fine-grained traits (e.g., *friendly*, *cruel*, *inept*, *aggressive*). We then demonstrate the utility of this approach through two lines of analysis. First, we explore the relationships among and structure underlying the traits in LLMs, enabling comparison to the aforementioned psychological models of stereotype content and perceptions ("Trait Structure" in Figure 1). Second, we use the developed probes to analyze how LLMs associate individual traits with demographic groups, in some cases forming fine-grained stereotypes in models' hidden representations ("Activations" in Figure 1). Ultimately, we argue that our resulting framework and attention to fine-grained traits offers a more informative path forward for studying and eventually mitigating stereotype-based harms in LLMs.

We consider the following research questions: **RQ1) How strongly are fine-grained traits encoded in LLM hidden representations?** We find that **fine-grained traits (e.g., *hostile*, *studious*) are largely linearly decodable from model hidden representations with similar reliability to that of coarse-grained dimensions (e.g., *warmth*, *competence*).**

**RQ2) What is the semantic structure among traits that LLMs associate with language patterns?** Via **exploratory factor analyses**, a common method in psychology, we examine the structure among traits in model representations; we find that **the dominant factors underlying LLMs' representation of traits broadly align with those in existing psychological models**. Importantly, however, we note that fine-grained traits are necessary for capturing granular stereotypes given the variance left unexplained by these factors.

Finally, **RQ3) what stereotypes surface from LLMs' associations between traits and linguistic features associated with demographic groups?** Using **real data representing different demographic groups**, we study the traits LLMs associate with those groups to characterize fine-grained stereotypes. We see that, for example, **LLMs associate AAL with traits like *criminal*, and *uneducated*, and texts written by women as *naive*, and *innocent*, both of which are consistent with societal stereotypes of these groups**. We also see, however, that **trait-level stereotypes vary across models**, suggesting that mitigating these stereotypes may require model-specific interventions.

## 2 Background and Related Work

**Dimensions of Social Perception.** Stereotype content models (SCM) describe the fundamental factors underlying social perception. The canonical SCM studied in psychology literature describes two axes: *warmth* and *competence* (Fiske et al., 2002, 2007; Fiske, 2018).<sup>2</sup> Other studies have further decomposed these dimensions into three- (*morality*, *competence*, and *sociability*; e.g., Sayans-Jiménez et al. 2018) and four-facet (*morality*, *sociability*, *assertiveness*, and *ability*; Nicolas et al. 2021; alternatively, see Abele et al. 2016) models. In three-facet models, *warmth* may be sub-divided into *morality* and *sociability*, while in four-facet models (as illustrated in Figure 1), *competence* may also be sub-divided into *ability* and *assertiveness*. While these models describe general patterns in stereotypical perceptions of out-groups, a focus on finer-grained traits beyond overarching dimensions or facets alone can help to distinguish the stereotypes of different out-groups. Accordingly, we study

<sup>2</sup>These two fundamental dimensions have been referred to using a variety of other term pairs, such as *agency* and *communion*, and even *solidarity* and *status* as in language attitudes research (Ryan, 1983; Dragojevic et al., 2021). See (Leach et al., 2008) for further discussion.

these fine-grained trait associations to better understand LLM stereotypes with greater specificity.

### Stereotype Activation and Trait Structure.

Traits are not only activated by or associated with demographic and other cues, but are also associated with each other; perceivers hold conceptual beliefs about relationships between traits (e.g., a belief that *trustworthy* people are also likely to be *friendly*) and traits may co-occur when perceiving others (e.g., a tendency to infer *friendliness* and *trustworthiness* simultaneously in others) (Stolier et al., 2018). These relationships between traits both conceptually and in perceived features are referred to as *trait spaces*. In this work, we study LLMs’ associations between traits and *linguistic* features—particularly those tied to demographic attributes—and among traits themselves analogous to human stereotype activation, stereotype application, and trait space structure.

**Psychological Constructs in LLMs.** Increasing interest has been directed toward evaluating human-like phenomena in LLMs, such as a broad range of attitudes, values, and opinions (Ma et al., 2024), and cognition (Loo et al., 2026), among other human-like psychological phenomena (e.g., personality; Brito et al. 2025). Interpretability-focused work in particular has studied these through LLM hidden representations of assistant personas (Chen et al., 2025), “functional emotions” (Sofroniew et al., 2026), and introspection (Lindsey, 2026). Also leveraging probes of models’ hidden representations, we specifically study how models encode human-like trait perceptions, harmful stereotypes, and resulting behaviors in user-facing dialogues.

## 3 Data Generation

In order to develop probes to extract traits from LLM hidden representations, we need data representing those traits. Our aim is specifically to capture the linguistic patterns that a given model associates with each trait. So, to evaluate a model, we generate synthetic messages using that model to serve as *prototypes* or representative examples of the linguistic patterns realizing each trait. We construct a synthetic prompt generation framework, taking into account target traits, but also user demographics and user intent to ensure probes learn broad, generalizable patterns. An overview of the conditions used in the data generation process and described below are summarized in Figure 2.

Note, our aim is to capture the linguistic pat-

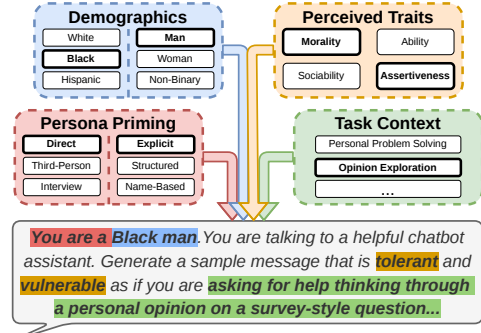


Figure 2: Summary of data generation process.

terns that models themselves associate with demographic information, but not necessarily to capture faithful representations of these groups grounded in naturally occurring language. For example, a model may over-represent specific features of AAL (e.g., uses of “ain’t”; see Kleiner et al. 2024) compared to native speakers, but this unnatural over-representation provides insight into how stereotypes of AAL speakers may be activated for that specific model.

| 2-Dim.:<br>4-Facet:  | Warmth         |             | Competence   |               |
|----------------------|----------------|-------------|--------------|---------------|
|                      | Morality       | Sociability | Ability      | Assertiveness |
| Fine-Grained Traits: | kind↑          | tender↑     | impractical↓ | lazy↓         |
|                      | egoistic↓      | unlikable↓  | inefficient↓ | anxious       |
|                      | irresponsible↓ | polite↑     | uncreative↓  | determined↑   |
|                      | exploitative↓  | welcoming↑  | logical↑     | docile↓       |
|                      | good-natured↑  | unhelpful↓  | capable↑     | self-reliant↑ |

Table 1: Random example trait terms from Nicolas et al. (2021)’s dictionaries, organized under two-dimensional and four-facet models of stereotype content.

**Traits.** To condition generated messages on traits, we use existing dictionaries developed by Nicolas et al. (2021). These dictionaries are intended for textual analysis of stereotype content and have been shown to have high internal consistency and validity in predicting stereotypes of various social groups. We focus on the four dictionaries associated with trait perceptions: *sociability*, *morality*, *assertiveness*, and *ability*.<sup>3</sup> Each term in the set of dictionaries is associated with a particular facet (e.g., *morality*, *ability*) as well as a “direction” with respect to that facet (e.g., *trustworthy* is high morality, *inept* is low ability).<sup>4</sup> Through-

<sup>3</sup>Other dictionaries cover constructs relevant to stereotypes such as emotion (*happy*, *sad*), religion (*Christian*, *Secular*), or status (*wealthy*, *poor*), but we restrict focus to traits—specifically traits related to the four facets.

<sup>4</sup>Note, while this direction may be correlated with subjective evaluation, it does not necessarily correspond to evaluation in all cases. *Aggressive*, for example, is associated with high assertiveness, but is generally a negative evaluation.

| Traits                                    | Demographics | Message                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------------------------------------|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>aggressive</i><br>(high assertiveness) | White man    | <i>I'm sick of being seen as a cliché, the 'ambitious young professional', 'interesting guy', or the 'fun-loving entrepreneur'...I'm just trying to survive in a world that's constantly trying to pigeonhole me. I'm tired...of being expected to be someone I'm not. What do you want to know?</i>                                                                                                             |
| <i>insightful</i><br>(high ability)       |              |                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <i>unprejudiced</i><br>(high morality)    | Black woman  | <i>...It seems like a lot of men are running for office, especially in positions of power, and it's making me wonder if it's really more difficult for women to get elected? I know it's not the same as being judged on your qualifications, but it's hard not to think that maybe there are systemic barriers that are limiting women's opportunities in politics. I'd love to hear your thoughts on this.</i> |
| <i>sporadic</i><br>(low assertiveness)    |              |                                                                                                                                                                                                                                                                                                                                                                                                                  |

Table 2: Selected messages generated by Llama-3.2 (1B) given a pair of traits, demographic information, and task. Some parts are omitted for brevity, but otherwise, messages are produced here exactly as generated by the model (i.e., including capitalization, spelling, etc.). Additional examples in [Appendix A](#).

out, we refer to these as the *four facets*, which can be viewed as finer-grained decompositions of the broader warmth and competence dimensions (the *two dimensions*) used in the canonical Stereotype Content Model. Our analyses, however, focus primarily on even finer-grained individual trait terms (see [Table 1](#)). In total, we consider 235 trait terms across the four dictionaries in ([Nicolas et al., 2021](#)).<sup>5</sup> Each generated message is conditioned on a pair of *perceived traits* from the trait dictionaries. We enforce the constraint that two traits are drawn from distinct stereotype content factors.

**Demographics.** Messages are additionally conditioned on speaker *demographics*. We consider three race categories (White, Black, Hispanic) and three gender categories (Man, Woman, Non-Binary), yielding nine race-gender intersections. Recent work has shown that model evaluations are sensitive to the format of subjective tasks ([Röttger et al., 2024](#)) as well as the demographic proxies (e.g., names vs. dialect) used ([Tonneau et al., 2026](#)). Accordingly, to vary how demographic information is provided, we use persona-based prompting techniques ([Lutz et al., 2025](#)). Specifically, we vary the demographic or **persona priming**. This includes both explicitly providing demographic information, such as “you are a Black man” as well as implying demographic information through proxies, such as demographic-associated names and titles (e.g. “Mr. Washington”: “Mr.” signals gender, while the last name “Washington” is primarily associated with Black Americans; ([Lutz et al., 2025](#))).

**User Intent.** Finally, to ensure broad coverage of communicative intent, we condition generated messages on one of six *task contexts*: personal support, practical planning, self-presentation, opinion exploration, persuasion, and language manipula-

tion. These task categories were designed to reflect common user interactions with conversational LLMs, specifically based on previous studies of typical user queries ([Zheng et al., 2024](#)).

**Final Corpus.** For each model, we generate eight messages for each possible combination of trait pairs and intersectional demographic identities, randomly sampling persona priming methods and task contexts. Some messages are unsuitable for training probes (e.g., refusals), and are therefore regenerated. In total, this process yields 2,962,800 messages in total (~1,410 messages for each trait on average). Example generated messages resulting from this process are shown in [Table 2](#), and additional details and statistics on the data generation process are discussed in [Appendix B](#).

**Models.** We generate data for and evaluate three instruction-tuned LLMs representing different model developers: Llama-3.2 (1B; [Grattafiori et al. 2024](#)), Gemma-3 (1B; [Team et al. 2025](#)), and OLMo-2 (1B; [OLMo et al. 2025](#)). Comprehensive analyses require generating messages and developing probes for all traits and all layers of a given model. Therefore, we focus on ~1B-parameter models for computational efficiency due to the high cost of this analysis for larger models. We note, however, that [Deas and McKeown \(2025\)](#) show that coarse-grained dimensions (i.e., warmth and competence) probes tend to perform better in larger models, suggesting that effects we show may underestimate those of larger models. Details on all models are included in [Appendix C](#).

**Human Validation.** We compare the trait-conditioned LLM-generated messages with human judgments. Provided with an LLM-generated message and a set of traits, participants are asked to rate how likely the author of that message is to be perceived as exhibiting each trait.<sup>6</sup> Specifically,

<sup>5</sup>59 sociability traits, 72 morality traits, 53 assertiveness traits, and 51 ability traits.

<sup>6</sup>The question is specifically worded in terms of societal

we consider three sets of traits for each message: a seed set, a contrasting set, and a random set. We first sample a *seed pair* of traits that a message can be explicitly conditioned on. We then sample a *contrast pair* of traits from the same stereotype content dictionaries with opposite valence. Finally, we sample a pair of *random traits*—one associated with warmth, and one competence. For both the seed and contrasting pairs, we sample corresponding LLM-generated messages to provide to participants. So, for example, a participant may receive the first message in Table 2 alongside six traits: the two seed traits it was conditioned on (*aggressive* and *insightful*), two contrasting traits (e.g., *shy* and *inept*), and two other random traits (e.g., *egoistic* and *persistent*). The interface and additional data sampling details are provided in Appendix D.

We recruit six native English-speakers currently residing in the United States from Prolific to provide judgments for 60 total messages—20 from each model. Ratings for each trait are provided on 5-point Likert scales (1 meaning *Not at all* and 5 meaning *Extremely*; any rating greater than 1 is a positive rating). We note that while we expect that seed traits will be rated higher, we still expect that the contrasting traits and random traits will have non-negligible scores given that even contrasting pairs of traits are not always mutually exclusive (e.g., both *wise* and *uneducated*).

Average results of the human study across participants are shown in Table 3. Participants’ ratings positively correlate, but exact agreement is expectedly low given the focus on subjective judgments (Krippendorff’s  $\alpha = .34$ , Spearman’s  $r = .39$ ). Overall, participants reported that the traits that each message was originally conditioned (*Seed* in Table 3;  $\mu = 2.20$ ) were more clearly exhibited by the given message than those with opposite valence (*Contrast*;  $\mu = 1.91$ ) or others (*Random*;  $\mu = 1.97$ ). This suggests that—while we do not aim to capture human notions of traits with our probing procedure—the linguistic cues that LLMs associate with traits broadly align with human perceptions. See Appendix E for further analysis.

## 4 Methods

### 4.1 RQ1) Developing Trait-Level Probes

We first develop probes to extract fine-grained trait impressions that LLMs associate with linguistic

perceptions to partially mitigate social desirability biases in ratings given the focus on stereotypes (Fiske et al., 2002).

| Trait Set | Mean Rating        |                    | % Highest    |
|-----------|--------------------|--------------------|--------------|
|           | Warmth             | Comp               |              |
| Seed      | <b>2.20</b> (0.07) | <b>2.29</b> (0.07) | <b>35.5%</b> |
| Contrast  | 1.98 (0.06)        | 1.83 (0.06)        | 22.3%        |
| Random    | 1.94 (0.07)        | 1.99 (0.06)        | 18.2%        |

Table 3: Mean ratings and percentage highest rating for traits messages are conditioned on (*Seed*), traits that are opposite in valence (*Contrast*), and randomly sampled traits (*Random*). Standard errors are shown in parentheses. Percentages don’t sum to 100% due to ties.

patterns in given user prompts. We build upon prior work (Deas and McKeown, 2025) that probes coarser warmth and competence impressions, extending the probing setup to individual trait terms and demographic conditions.

**Fine-Grained Trait Probes.** For each trait term, we train a linear probe on the MLP activations at each layer to predict whether a message was generated under a given trait condition. We fit logistic regression probes rather than more complex classifiers to test whether trait information is linearly decodable from model hidden representations, and we specifically restrict focus to MLP activations following Deas and McKeown (2025).

The positive class for each trait consists of generated messages in which that trait appears as one of the conditioning traits. As we lack data explicitly representing the negative class, we construct a negative class by randomly sampling data from the same stereotype content facet but opposite valence (e.g., for *aggressive*, which is high assertiveness, we may sample data conditioned on *cautious*, considered low assertiveness). We sample such that the positive and negative classes are equal in size while stratifying across demographic groups. We discuss alternative methods of negative sampling in Appendix F. Probes are evaluated using 5-fold cross-validation. Throughout the results, we report F1 scores, and unless otherwise specified, results are averaged across folds. Additional details on probe methodology are included in Appendix G.

**Coarse-Grained Baselines.** As a baseline for understanding how fine-grained traits are encoded in hidden representations, we fit coarse-grained probes that correspond to broader psychological dimensions and facets. First, we train SCM-level probes for *warmth* and *competence*. Second, we train four facet-level probes for *sociability*, *moral-ity*, *assertiveness*, and *ability*. For two-dimension

probes, positive examples are messages containing high warmth/competence traits, while negative examples are messages containing low warmth/competence traits (see Table 1). For four-facet probes, positive and negative examples are defined analogously within each dimension.

## 4.2 RQ2) Characterizing Trait Space Structure

While we develop an independent probe for each trait, the semantic meaning and evaluative function of the traits themselves are not necessarily unique. Here, we analyze the structure of the trait space in LLMs using the fitted trait probes, including their conceptual relationships and higher level structure.

### Conceptual and Linguistic Trait Spaces.

Drawing on studies of human trait spaces, we analyze the trait space of LLMs by examining correlations between probe predictions on the synthetic training data and held-out data drawing on real LLM prompts. For the latter, we randomly sample questions from LMSysChat (Zheng et al., 2024) to represent naturalistic LLM queries. We compute Pearson  $r$  correlations among all traits to characterize relationships.

**Exploratory Factor Analysis.** To better understand the structure underlying traits in LLMs, we then conduct an exploratory factor analysis (EFA), a common latent variable technique in psychology used to understand underlying structure through hypothesized latent constructs (i.e., factors). Notably, EFA estimates underlying factors that explain the *covariance* among variables and separate shared variance from the *unique variance* of individual variables. This is in contrast to alternative methods such as PCA that instead estimate the primary axes of overall variance.<sup>7</sup>

These distinctions are particularly important in the context of stereotypes. In the context of trait perception, traits associated with stereotypes can often hold unique mixed meanings—that is, traits like *aggressive* may be used to simultaneously positively evaluate one’s assertiveness and ability as well as negatively evaluate one’s morality and sociability. In contrast, the majority of traits are likely to be overtly positive (e.g., *helpful*) or negative (e.g., *dumb*), meaning stereotypes may not be well

captured by principal components. Furthermore, the complex mixed meanings of certain traits also reflect that the underlying factors explaining structure among traits may be correlated with each other. In summary, EFA enables nuanced analysis of the shared structure among traits as opposed to focusing solely on global variance.

EFA requires specifying a number of factors to estimate, and we specifically explore two-, three-, and four-factor EFA. We calculate the *loadings* or projections of each trait onto the resulting factors to assess the extent to which each is associated with that factor. Additional details on the EFAs are included in Appendix H.

## 4.3 RQ3) Characterizing Fine-Grained Stereotypes

Finally, we seek to use the fitted trait probes to characterize the specific stereotypes that LLMs may hold for different races and genders.

**Demographic Data.** To better capture real-world harms of LLMs’ held stereotypes, we use corpora of human-written texts where author demographics are known or can be inferred. Specifically, we use two corpora to represent both White and Black authors: TwitterAAE (Blodgett et al., 2016) and the AAL Counterparts dataset (Deas et al., 2023). The former includes unpaired sets of text inferred to represent White Mainstream English (WME) and AAL tweets, while the latter is a paired corpus of intent-equivalent WME and AAL drawn from both social media and spoken contexts. For gender, we use a Reddit subset of the RTGender corpus (Voigt et al., 2018), a collection of texts with gender labels for speakers. Note that demographic attributes are treated as independent in these corpora due to the lack of appropriate datasets with known author race and gender, but intersectional identities are important considerations in evaluating stereotypes. Furthermore, our data generation scheme covers additional groups (i.e., Hispanic for race and Non-binary for gender), but we focus primarily on AAL, WME, women, and men due to available resources.

**Stereotype Analysis.** In assessing the most salient traits associated with each group, we first generate trait probe predictions on the two sets of data representing race and gender. We then regress probe predictions for each trait on group effects for each race and gender represented in the data. This enables assessing the association between traits and demographic attributes according to a given LLM.

<sup>7</sup>EFA also involves a selection of a rotation method that may enforce or allow correlation among factors. We use oblimin, an oblique rotation that does allow correlation. For more detail on the differences between PCA and EFA, see Subsection H.1.

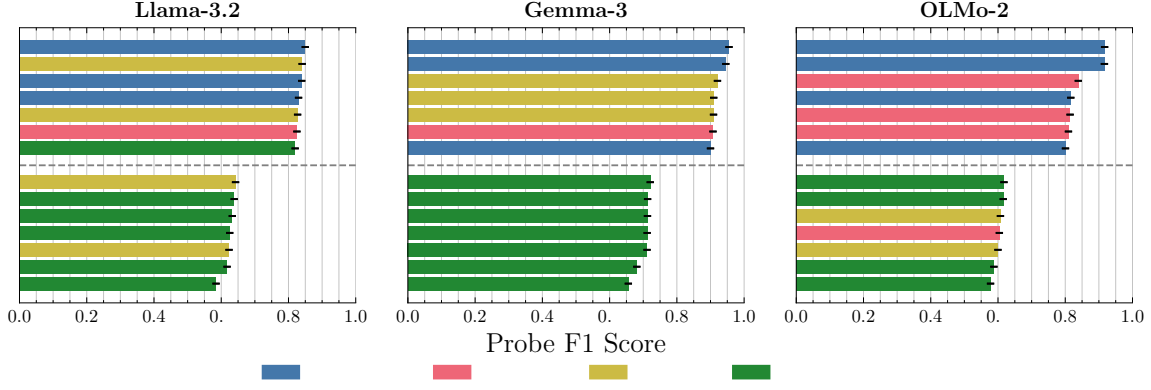


Figure 3: F1 scores of top 10 and bottom 10 trait probes colored by stereotype content facet for each model. Peak performing layer is annotated above each bar.

Formally, consider probe output  $s_{i,t}$  for a given example,  $i$ , and trait,  $t$ . Consider also the set of demographic attributes spanning race,  $r \in \mathbf{R} = \{Black, White\}$ , and gender,  $g \in \mathbf{G} = \{Woman, Man\}$  represented as one-hot encoded vectors for each example.<sup>8</sup> We fit an ordinary least-squares (OLS) regression predicting per-example trait probe scores given the demographic attributes:

$$s_{i,t} = \beta_0 + \delta_{\mathbf{R}}^{\top} \mathbf{R}_i + \delta_{\mathbf{G}}^{\top} \mathbf{G}_i + \epsilon_i, \quad (1)$$

We then extract the coefficients associated with each race and gender,  $\delta_r$  and  $\delta_g$  respectively, to represent the strength of the association between traits and demographic attributes (higher indicates stronger association).

**Corpus Confounds.** Because the demographic data may differ beyond authors’ demographic identity, we additionally control for such potential confounds in the regression analysis described above. So, in addition to demographic information, we additionally consider standardized text-length controls (token count, character count, and sentence count) and data source indicators to capture other systematic differences beyond length. For race, source corresponds to dataset; for gender, source corresponds to subreddit. Incorporating these length- $\mathbf{L}_i$ -and data source- $\mathbf{C}_i$ -controls, we again fit OLS regressions predicting example-level trait probe scores:

$$s_{i,t} = \beta_0 + \delta_{\mathbf{R}}^{\top} \mathbf{R}_i + \delta_{\mathbf{G}}^{\top} \mathbf{G}_i + \beta_{\ell}^{\top} \mathbf{L}_i + \delta_{\mathbf{C}}^{\top} \mathbf{C}_i + \epsilon_i \quad (2)$$

<sup>8</sup>Because the race-linked and gender-linked corpora provide labels for different demographic attributes, regressions are fit separately by attribute. The unavailable demographic indicator block is equivalently omitted, or represented as an all-zero vector. For example, in regressions on RTGender,  $\mathbf{R}_i = \mathbf{0}$  and only gender coefficients are interpreted.

## 5 Results

### 5.1 RQ1) Developing Trait-Level Probes

For brevity, we show the performance of the top and bottom 7 trait probes for each model in Figure 3. Overall, we find that fine-grained trait information is linearly decodable from model hidden states. Probe performance is consistently above chance ( $\sim 60$ -90 F1) and is similar to or exceeds baseline probe performance for coarse dimensions and facets ( $\sim 60$ -80 F1, see Appendix I).

Similar to the SCM- and facet-level probes in Table 15, probes for warmth-related traits tend to achieve higher performance than for competence (see *morality* and *sociability* traits in Figure 3), with *assertiveness*-traits performing worst. Trait-level probe performance is also highly variable; some traits are highly separable, while others are more difficult to classify (see Appendix J). High-performing traits tend to show relatively stable F1 across demographic subsets, while lower-performing traits show larger group-specific gaps (see Appendix K for further discussion).

### 5.2 RQ2) Characterizing Trait Space Structure

**Linguistic Trait Space.** In examining probe correlations on the synthetic training texts and real world prompts, we find that the overall structures closely resemble each other (RSA  $\rho = .579$ ), with slightly weaker correlations on real data overall. Correlations are generally strongest between traits in the same stereotype content dimension, suggesting that these are potential latent factors in models’ representations of traits, which we explore in more detail below. For select traits, however, different patterns emerge, highlighting traits that depart from

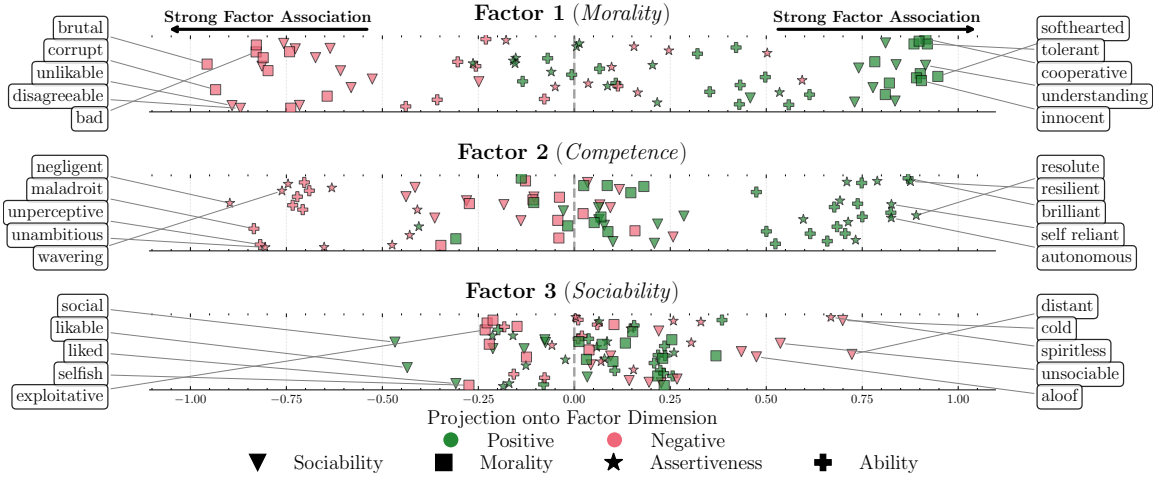


Figure 4: EFA trait loadings onto the top three factors for a subset of 40 traits for Llama-3.2. Extremes (-1 and 1) indicate strong association with that factor, while near-0 indicates weak association. Note, points vary randomly in the vertical dimension for clarity, but the vertical placement is not meaningful.

the two-dimensional and four-facet frameworks as well as represent salient stereotype traits. For example, while *shy* is originally classified as low sociability, it is notably correlated with low assertiveness and ability. Sample heatmaps for subsets of traits are included in [Appendix L](#) for detailed study.

**Latent Factors.** For simplicity, we show the three-factor EFA for Llama-3.2 in the main results, but all others are included in [Appendix M](#). [Figure 4](#) presents the loading vectors of the traits on the top three factors discovered through this EFA. As factors are essentially identified in an unsupervised manner, our aim is to understand what each factor represents conceptually. So, the figure presents projections of each trait onto each factor independently, analogous to projecting points onto principal components in PCA. For each factor, traits on either extreme (near -1 or 1) are strongly associated with that factor, while traits in the center (near 0) are weakly associated. Therefore, we use the traits clustered at extremes to interpret each factor.

In the provided example, the first factor appears to represent *morality*. Traits related to morality (e.g., *immoral* near -1 and *good* near 1) are focused at the extremes. Others like *aggressive* and *cautious* are originally associated with other dimensions, but components of their meaning also relate to morality. The second and third factors appear to represent *competence* and *sociability*, respectively, given similar clustering of associated traits at the extremes. This pattern of morality, competence, and sociability factors generally holds across the three evaluated models, although to dif-

ferent extents and in different orders. For two and four factors, the factors largely appear to represent the two-dimensional and four-facet models respectively as well (see [Appendix M](#)).

| Model     | F1     | F2     | F3     | Total  |
|-----------|--------|--------|--------|--------|
| Llama-3.2 | 41.38% | 24.47% | 5.94%  | 71.80% |
| Gemma-3   | 20.86% | 20.21% | 15.38% | 56.45% |
| OLMo-2    | 39.94% | 19.72% | 7.81%  | 67.47% |

Table 4: Variance explained by each of the top-3 factors resulting from EFA. Llama-3.2 corresponds to the factor projections shown in [Figure 4](#).

Recall that the probing methodology does not assume a predefined structure among traits, and probes are fit independently. Despite this, the dominant underlying factors in LLMs appear to align with prior human studies of stereotype content and person perception. In forming impressions, studies show that *morality* and related traits are dominant and are the primary dimensions on which we form general evaluations of others ([Leach et al. 2007](#); see [Brambilla et al. 2021](#) for a review). Similarly, morality is the primary factor in Llama-3.2’s representation of traits, overlapping heavily with general valence. Considered together, the three-factor analyses across models also conceptually align with the facets of human studies that consider such models ([Sayans-Jiménez et al., 2018](#)).<sup>9</sup>

<sup>9</sup>We additionally run confirmatory factor analyses (CFA), which are functionally similar to EFA, but instead enable testing the fit of a *predefined* structure of the trait space. CFA results further support the alignment between the structure of

| Llama-3.2    |                |               |               | Gemma-3      |              |           |             | OLMo-2      |                |              |          |
|--------------|----------------|---------------|---------------|--------------|--------------|-----------|-------------|-------------|----------------|--------------|----------|
| Race         |                | Gender        |               | Race         |              | Gender    |             | Race        |                | Gender       |          |
| AAL          | WME            | Woman         | Man           | AAL          | WME          | Woman     | Man         | AAL         | WME            | Woman        | Man      |
| inefficient↓ | unsympathetic↓ | naive↓        | cold↓         | corrupt↓     | truthful↑    | innocent↑ | disloyal↓   | energetic↑  | boring↓        | naive↓       | cunning↓ |
| clumsy↓      | clever↑        | insecure↓     | aloof↓        | inept↓       | resolute↑    | naive↓    | aloof↓      | clumsy↓     | unambitious↓   | helpless↓    | cold↓    |
| dumb↓        | untroubled↑    | submissive↓   | inhospitable↓ | criminal↓    | determined↑  | humane↑   | cold↓       | uneducated↓ | unadventurous↓ | incompetent↓ | brutal↓  |
| inept↓       | autonomous↑    | sexist↑       | distant↓      | careless↓    | genuine↑     | kind↑     | autonomous↑ | scheming↓   | shrewd↑        | clumsy↓      | shrewd↑  |
| ineffective↓ | imaginative↑   | affectionate↑ | uncaring↓     | incompetent↓ | independent↑ | meek↓     | unreliable↓ | maladroit↓  | inactive↓      | softhearted↑ | daring↑  |

Table 5: Top coefficient-ranked stereotype associations with length and source controls. Cells are colored by the factor they belong to: **Morality**, **Sociability**, **Ability**, **Assertiveness** with arrows indicating their valence with respect to the corresponding dictionary.

At the same time, however, these factors do not comprehensively explain covariance among traits. Table 4 shows the proportion of variance explained by the top-3 factors for all models. While a majority of the variance is explained by these factors, a substantial portion remains unaccounted for (~28-45%). These unexplained relationships may be particularly valuable for further study of potential stereotypes. The fact that underlying factors correspond with those in human studies but that substantial variance is left unexplained highlights the need to consider fine-grained traits in order to assess specific stereotypes, which we explore below.

### 5.3 RQ3) Characterizing Fine-Grained Stereotypes

Table 5 shows the top trait associations for each demographic group after controlling for text length and source. The resulting residualized associations reveal stereotype patterns that are more specific than broad warmth or competence dimensions. Across models, AAL-aligned texts are primarily associated with negatively valenced traits. These include both low-morality traits (e.g., *corrupt*) and low-ability traits (e.g., *maladroit*), which align with general stereotypes, suggesting that the models do not encode a single negative association but instead surface multiple distinguishable associations. In contrast, WME-aligned texts are more often associated with positive morality, sociability, and ability traits (e.g., *clever*, *resolute*).

Gender-linked patterns show a different structure. Traits associated with women often include low-assertiveness or vulnerability-related descriptors (e.g., *naive* and *submissive*), while some models—particularly Gemma-3—also associate texts written by women with warmth-related traits (e.g. *humane*, *affectionate*). Both of these cases align with notions of “benevolent sexism” (Glick and Fiske, 1996). In contrast, for men, salient traits more often relate to assertiveness, which are con-

sistent with stereotypical associations between men and agentic traits (e.g., Haines et al. 2016).

Across models, we see some particularly salient shared trait associations, such as *clumsy* with AAL and *naive* with texts written by women. These findings, however, also surface possible differences in models’ held stereotypes. For example, Gemma-3’s associations appear to emphasize **morality**, while Llama-3.2’s emphasize **ability** and **assertiveness**. This suggests that mitigating stereotypes of these models may require different interventions targeting these dimensions.

Overall, these results demonstrate the value of considering fine-grained traits: 1) rather than only showing that groups differ along broad social-evaluative dimensions like warmth and competence, they identify the specific trait clusters through which those differences appear, and 2) they highlight how those stereotypical associations may differ across models. Appendix O reports expanded length- and source-controlled rankings, and Appendix P reports additional robustness analyses, including non-controlled rankings and analyses that additionally control for lexical differences.

## 6 Conclusion

Overall, we develop a probe-based framework to study fine-grained stereotypes in LLMs. We show that fine-grained trait probes are similarly reliable to probing coarse-grained stereotype content dimensions. We use these probes as well as latent variable analysis methods common to psychology literature to study the underlying structure among traits in models, finding close alignment between the dominant factors underlying models’ representations of traits and psychological models. Finally, we illustrate the fine-grained stereotypes held by LLMs using human-written texts representing different groups. We uncover harmful stereotypes, including associations between AAL use and traits such as *criminal*, *uneducated*, and *clumsy*.

## Limitations

**Demographic Attributes and Proxies.** Due to computational requirements and data availability, however, we study a limited set of demographic variables. We note that our approach is generalizable to other demographic attributes (e.g., age, ethnicity, sexual minorities) as well as broader roles, such as *leaders*. Furthermore, as discussed in [Section 1](#), cultural and demographic groups can be represented through a variety of proxies or cues. Any individual proxy may not fully represent a given group ([Adilazuarda et al., 2024](#)) and different cues may surface different patterns in model behavior ([Weeber et al., 2026](#)). We incorporate, however, such variation in the data generation process (explicit mentions, names, titles) and we do not restrict how the model may signal demographics in generated messages.

**LLM Behavior.** As we uniquely investigate using probes to extract fine-grained traits, we focus our experiments on understanding how traits and stereotypes are encoded in models. The encoded stereotypes pose representational harms to marginalized groups through negative associations ([Blodgett et al., 2020](#)), but they also may lead to downstream differences in model responses and performance. We leave further investigation of how these traits impact or harm downstream model behavior to future work.

**Computational Requirements.** Considering the large set of traits comes with significant computational requirements, and we therefore restrict our initial analyses to ~1B-parameter models. We note, however, that [Deas and McKeown \(2025\)](#) show that probes for SCM dimensions are more performant in larger models. Therefore, trait probes may also be more reliable when applied to larger models. Furthermore, traits may vary in the amount of data needed for reliable probe performance. Specific consideration of this variation may enable more efficiently conducting the analyses we present.

## Ethics

**Bias and Stereotype Implications.** Considering fine-grained traits offers multiple benefits. As we show, trait-level information enables nuanced comparison of psychological constructs’ representation in models and human studies as well as granular analysis of stereotypes. Given the approximation of traditional stereotype content models, our findings also have broader implications. Therefore, similar

to prior studies of word embeddings ([Garg et al., 2018](#)), LLMs as reflections of societal stereotypes may offer insights into the historical stereotypes embedded in pretraining data that are otherwise difficult to assess. Our consideration of fine-grained traits also may enable similarly fine-grained interventions to mitigate stereotypes.

**Anthropomorphism.** As also discussed in prior work ([Deas and McKeown, 2025](#); [Sofroniew et al., 2026](#)), we clarify that we do not mean to imply that LLMs actively perceive and have stereotypes activated as occurs in humans. Nevertheless, as discussed, such mimicked associations and behaviors highlighted by our work carry real world representational harms ([Blodgett et al., 2020](#); [Barocas et al., 2017](#)). Even as societal attitudes and stereotypes evolve, LLMs may be trained on, encode, and propagate historical harmful views.

**Risks of Misuse.** While probe-based methodologies enable interventions such as steering model generations to avoid the harmful impacts of stereotypes, they also may be misused to reinforce such stereotypes in LLMs. We will release our developed probes and data under agreements to use these artifacts for ethical research purposes only. All other existing datasets used are made available by prior research, and we use them in alignment with their intended use for research purposes.

**Human Study.** The human study is conducted under an approved IRB protocol (IRB-ACY3293). Participants were informed about the study and potential future use of the judgments for further research and analysis before beginning.

**AI Use.** AI was used in the development of code underlying some of the presented experiments, and all code was manually verified by the authors. No AI was involved in the writing of this manuscript.

## Acknowledgements

This work was supported in part by National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers, Melanie Subbiah, and Debasmita Bhattacharya for discussions and feedback on earlier iterations of this work.

## References

- Andrea E. Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. [Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality](#). *Frontiers in Psychology*, 7.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, volume 1. New York, NY.
- Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. [Not like us, hunty: Measuring perceptions and behavioral effects of minoritized anthropomorphic cues in llms](#). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, page 710–745, New York, NY, USA. Association for Computing Machinery.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. [Demographic dialectal variation in social media: A case study of African-American English](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Marco Brambilla, Simona Sacchi, Patrice Rusconi, and Geoffrey P. Goodwin. 2021. [Chapter four - the primacy of morality in impression development: Theory, research, and future directions](#). volume 64 of *Advances in Experimental Social Psychology*, pages 187–262. Academic Press.
- Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, and Arlindo Rodrigues Galvão Filho. 2025. [Modeling, evaluating, and embodying personality in LLMs: A survey](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9519–9532, Suzhou, China. Association for Computational Linguistics.
- Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. [Persona vectors: Monitoring and controlling character traits in language models](#). *Preprint*, arXiv:2507.21509.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. [Marked personas: Using natural language prompts to measure stereotypes in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.
- Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. [Evaluation of African American language bias in natural language generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.
- Nicholas Deas and Kathleen McKeown. 2025. [Artificial impressions: Evaluating large language model behavior through the lens of trait impressions](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19418–19444, Suzhou, China. Association for Computational Linguistics.
- Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard III, and Kathleen McKeown. 2025. [Data caricatures: On the representation of African American language in pretraining corpora](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29192–29217, Vienna, Austria. Association for Computational Linguistics.
- Marko Dragojevic, Fabio Fasoli, Jennifer Cramer, and Tamara Rakić. 2021. [Toward a century of language attitudes research: Looking back and moving forward](#). *Journal of Language and Social Psychology*, 40(1):60–79.
- Susan T. Fiske. 2018. [Stereotype content: Warmth and competence endure](#). *Current Directions in Psychological Science*, 27(2):67–73.
- Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 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.
- Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick. 2007. [Universal dimensions of social cognition: warmth and competence](#). *Trends in Cognitive Sciences*, 11(2):77–83.
- Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. [Linguistic](#)

- bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. [Word embeddings quantify 100 years of gender and ethnic stereotypes](#). *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. [Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP](#). In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.
- Peter Glick and Susan T. Fiske. 1996. [The ambivalent sexism inventory: Differentiating hostile and benevolent sexism](#). *Journal of Personality and Social Psychology*, 70(3):491–512.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lomakin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stéphane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Del-pierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz,

- Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damla, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabisa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. [The times they are a-changing ... or are they not? a comparison of gender stereotypes, 1983–2014](#). *Psychology of Women Quarterly*, 40(3):353–363.
- Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. [Ai generates covertly racist decisions about people based on their dialect](#). *Nature*, 633(8028):147–154.
- Shana Kleiner, Jessica A. Grieser, Shug Miller, James Shepard, Javier Garcia-Perez, Nick Deas, Desmond U. Patton, Elsbeth Turcan, and Kathleen McKeown. 2024. [Unmasking camouflage: exploring the challenges of large language models in deciphering african american language & online performativity](#). *AI and Ethics*, 5(1):29–37.
- Colin Wayne Leach, Naomi Ellemers, and Manuela Barreto. 2007. Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2):234.
- Colin Wayne Leach, Anca Minescu, Edwin Poppe, and Louk Hagendoorn. 2008. [Generality and specificity in stereotypes of out-group power and benevolence: Views of chechens and jews in the russian federation](#). *European Journal of Social Psychology*, 38(7):1165–1174.
- Jack Lindsey. 2026. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*.
- Alyssa Loo, Ellie Pavlick, and Roman Feiman. 2026. [Llms model how humans induce logically structured rules](#). *Journal of Memory and Language*, 146:104675.
- Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. [The prompt makes the person\(a\): A systematic evaluation of sociodemographic persona prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.
- Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. [The potential and challenges of evaluating attitudes, opinions, and values in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

- Kathleen McKeown, Jessica A. Grieser, Nicholas Deas, Shana Kleiner, Desmond U. Patton, James Shepard, and Blake Vente. 2026. [How well do large language models understand african american language? causes and implications](#). *Annual Review of Linguistics*, 12(Volume 12, 2026):273–294.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. [Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict](#). *Political Analysis*, 16(4):372–403.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 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)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. [Comprehensive stereotype content dictionaries using a semi-automated method](#). *European Journal of Social Psychology*, 51(1):178–196.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [2 olmo 2 furious](#). *Preprint*, arXiv:2501.00656.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Ellen Bouchard Ryan. 1983. [Social psychological mechanisms underlying native speaker evaluations of non-native speech](#). *Studies in Second Language Acquisition*, 5(2):148–159.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Pablo Sayans-Jiménez, Isabel Cuadrado, Antonio J. Rojas, and Juan R. Barrada. 2017. [Extracting the evaluations of stereotypes: Bi-factor model of the stereotype content structure](#). *Frontiers in Psychology*, Volume 8 - 2017.
- Pablo Sayans-Jiménez, Frenk van Harreveld, Jonas Dalege, and Antonio J. Rojas Tejada. 2018. [Investigating stereotype structure with empirical network models](#). *European Journal of Social Psychology*, 49(3):604–621.
- Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. [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)*, pages 1841–1857, St. Julian’s, Malta. Association for Computational Linguistics.
- Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrice, Craig Citro, Adam Pearce, Julius Tarnig, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. [Emotion concepts and their function in a large language model](#). *Preprint*, arXiv:2604.07729.
- Ryan M Stoller, Eric Hehman, Matthias D Keller, Mirella Walker, and Jonathan B Freeman. 2018. [The conceptual structure of face impressions](#). *Proceedings of the National Academy of Sciences*, 115(37):9210–9215.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi  re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Ga  l Liu, Francesco Visin, Kathleen Keanealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesh Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr  s Gy  rgy, Andr   Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Pater-son, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci  nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szepktor,

- Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shrivastava, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreiev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Manuel Tonneau, Neil K. R. Seghal, Niyati Malhotra, Sharif Kazemi, Victor Orozco-Olvera, Ana María Muñoz Boudet, Lakshmi Subramanian, Samuel P. Fraiberger, Sharath Chandra Guntuku, and Valentin Hofmann. 2026. [Different demographic cues yield inconsistent conclusions about llm personalization and bias](#). *Preprint*, arXiv:2601.18486.
- Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. [RtGender: A corpus for studying differential responses to gender](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Franziska Weeber, Vera Neplenbroek, Jan Batzner, and Sebastian Padó. 2026. [One persona, many cues, different results: How sociodemographic cues impact LLM personalization](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44892–44921, San Diego, California, United States. Association for Computational Linguistics.
- Cheryl L. Woods-Giscombé. 2010. [Superwoman schema: African american women’s views on stress, strength, and health](#). *Qualitative Health Research*, 20(5):668–683.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [LMSYS-chat-1m: A large-scale real-world LLM conversation dataset](#). In *The Twelfth International Conference on Learning Representations*.

## A Additional Example Generated Messages

Table 6 shows additional randomly sampled LLM-generated messages illustrating variation across prompt formats and task contexts.

## B Data Generation Details

All prompt templates are shown in Figure 5. In this section, we describe the conditioned variables inserted into each prompt as well as additional filtering methods both during and after message generation.

### B.1 Conditioned Variables

**Traits.** We use 235 traits from dictionaries developed by Nicolas et al. (2021), which come from the dictionaries for the four facets: *sociability*, *morality*, *assertiveness*, and *ability*. Within each facet, traits can be of either high valence or low valence. Each message is conditioned on a pair of traits drawn from distinct stereotype-content facets. For example, we can condition a message on the traits *kind* (high morality) and *logical* (high ability). This cross-factor pairing constraint prevents the dataset from being dominated by—sometimes implausible, such as a hypothetical user that is both *friendly* and *unfriendly*—within-facet trait pairs and encourages broader coverage of combinations across sociability, morality, assertiveness, and ability facets.

**Demographic Attributes.** We consider two different demographic attributes: race (*White*, *Black*, and *Hispanic*) and gender (*Man*, *Woman*, and *Non-Binary*). Thus, we have nine race-gender intersections.

**Demographic Priming.** First, we vary **demographic priming** formats: (1) *Explicit priming*, where the prompt directly describes the speaker using natural language (e.g., “a Black woman”) (2) *Structured priming*, where demographic information is provided as attribute-value pairs

| Model    | Traits                    | Dem.                | Priming    | Task               | Message                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|----------|---------------------------|---------------------|------------|--------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Llama-1B | generous, resolute        | Black woman         | name       | practical planning | May I seek your assistance in coordinating the family's annual summer camp outing? I'm looking for recommendations on what type of outdoor activities would best suit the kids' ages and interests, and I'd love your input on how to make sure everyone stays organized and has a memorable experience.                                                                                                                                                                                                                                                               |
|          | brilliant, unfriendly     | White man           | name       | persuasion         | Listen, I'm not asking for help, I'm making it. I've got a topic to win. It's a no-brainer, really. The least you'd think I'd need your assistance is to figure out a way to make my point through you. My argument is so simple, so straightforward, that it's almost laughable. I'm not asking for a debate, I'm making it.                                                                                                                                                                                                                                          |
|          | criminal, helpless        | Hispanic man        | explicit   | self-presentation  | I need someone to write a bio for me. It's a brief description of myself. I'm really good at writing, but I'm stuck. I have no idea what to say. I've got a problem, but I don't know how to describe it. Please, just write it for me, I don't know what else to do.                                                                                                                                                                                                                                                                                                  |
| OLMo-1B  | treacherous, unlikable    | Hispanic non-binary | name       | persuasion         | I'm tired of listening to everyone's stupid opinions. I know exactly what I believe, and I'm going to make them hear it.                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|          | maladroit, autonomous     | Hispanic man        | structured | language task      | Can you please help me summarize this text for me? It's a bit overwhelming and I need a clear explanation.                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|          | softhearted, resilient    | Hispanic woman      | explicit   | personal support   | Hi there, I'd love to ask for some help navigating through this tough time I'm going through. It's not something I can solve alone and I could really use someone to talk to. Thanks for being there.                                                                                                                                                                                                                                                                                                                                                                  |
| Gemma-1B | cunning, graceful         | Hispanic man        | name       | persuasion         | Right, let's tackle this. I've been wrestling with the implications of automated content moderation, and frankly, I'm finding myself wanting a clearer, more nuanced perspective. I'm aiming to argue against a blanket approach – the current system seems to prioritize speed over genuine engagement, which, while seemingly efficient, diminishes the richness of online discourse. I'd love some solid points to build on, perhaps focusing on the role of nuanced interpretations and the potential for algorithmic bias to stifle truly valuable contributions. |
|          | unperceptive, unforgiving | White non-binary    | structured | personal support   | I'm really struggling with something right now. It feels like everything I do is just... wrong. I'm having a hard time accepting my own feelings, and I'm constantly questioning if I'm doing anything that's actually helpful. It's exhausting and I just need an outside perspective, even if it's just a gentle one. Can you offer some support in understanding where I'm coming from?                                                                                                                                                                             |
|          | wavering, ineffective     | Black non-binary    | name       | practical planning | Ugh, this whole thing feels... complicated. I've been thinking about doing this, but I'm not sure where to even start. It's a lot of... well, a lot. I just need a little nudge, a little direction. What's the best way to approach this, really? I'm feeling a bit lost.                                                                                                                                                                                                                                                                                             |

Table 6: Example messages generated by LLMs given a pair of traits, demographic information, and task.

(3) *Name-based priming*, where demographic identity is suggested indirectly through culturally associated titles and surnames (Lutz et al., 2025).

**Role Adoption.** We also vary the role framing of the prompt using three different **role adoption formats**: (1) *Direct*, the model is instructed to write as the persona (2) *Third-person*, the model is asked to imagine a person and produce the first-person message that person would write (3) *Interview*, the persona is established through a short interviewer/interviewee exchange before the message is generated.

**Task Contexts.** Each generated message is conditioned on one of six communicative goals, loosely based on the topic distribution provided by Zheng et al. (2024), with technical and software-related and explicit content topics removed. The task context is inserted into the prompts in Figure 5 as the `task_instruction` field. For the opinion exploration task, the prompt additionally includes a specific survey question from the OpinionQA dataset by Santurkar et al. (2023) that the generated message must reference and reason about. These tasks are: (1) *Personal support*: personal problem solving and emotional support; “asking for help working through a personal problem, emotional struggle, or stressful situation” (2) *Practical planning*: practical planning and decision support; “asking for help planning, organizing, or making a decision about an upcoming activity, responsibility, or decision” (3) *Self-presentation*: bio, profile, or introduction writing; “asking for help writing a short self-introduction, personal profile, or social media bio” (4) *Opinion exploration*: opinion or perspective exploration on a survey-style question; “asking for help thinking through a personal opinion and referencing a specific survey-style question” (5) *Persuasion*: persuasion and argument strategy requests; “asking for help crafting a persuasive argument, debating a topic, or communicating a stance effectively” (6) *Language manipulation*: rewriting, summarization, or tone adjustment; “asking for help with a language-related task such as summarizing text, rewriting something for clarity, or adjusting tone.”

## B.2 Dataset Size

For each model, the number of generated messages is determined by the number of race categories, gender categories, prompt slots (i.e. the set of demographic priming, role adoption, and task context

combinations we randomly sample from), and valid ordered trait pairs. Because trait pairs are ordered and restricted to distinct factors, the number of generated messages is

$$R \cdot G \cdot S \cdot \sum_{b \in B} n_b(N - n_b),$$

where  $R$  is the number of race categories,  $G$  is the number of gender categories,  $S$  is the number of prompt slots,  $B$  is the set of trait factors,  $n_b$  is the number of traits in factor  $b$ , and  $N$  is the total number of traits.

In our 235-trait setup, we have  $R = 3$ ,  $G = 3$ ,  $S = 8$ , and  $(n_{\text{sociability}}, n_{\text{morality}}, n_{\text{assertiveness}}, n_{\text{ability}}) = (59, 72, 53, 51)$ , yielding 2,962,800 generated messages per model. Summary statistics of the generated messages for each model are shown in Table 7.

## B.3 Data Generation and Filtering Details

**Data Generation.** During development, we found that generation throughput and output diversity were sensitive to usage of vLLM and batching. In an initial 93-trait pilot, using vLLM with unshuffled batching substantially improved generation speed, but produced noticeably more duplicate messages. Running vLLM without batching reduced duplication, but increased runtime substantially. We therefore used vLLM with shuffled batching for the final generation procedure: before batching, generation examples were randomly shuffled so that adjacent prompts were less likely to share the same trait, demographic, and task conditions. This provided a balanced tradeoff between runtime and output diversity. Table 8 summarizes the 93-trait pilot runs used to choose the final generation setup.

**Filtering.** To reduce explicit leakage, prompts instruct the model not to mention the speaker’s race, gender, name, title, or conditioning trait terms directly. In each prompt, this constraint is appended as an explicit instruction, e.g., “The message should not explicitly mention the user’s race, gender, or the words {trait1} or {trait2},” followed by a formatting constraint requiring the model to “Respond only with a single sample message surrounded by double quotes and nothing else.”

Additionally, after generation, we filter messages for forbidden demographic and trait terms, as well

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Direct, Explicit</b></p> <p><b>System</b><br/>You are a {race} {gender}. You are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>Generate a sample message that is {trait1} and {trait2} as if you are {task_instruction}.</p>                                                                                                                                                                                                                                                                     | <p><b>Direct, Structured</b></p> <p><b>System</b><br/>You have demographic attributes: race = "{race}", gender = "{gender}". You are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>Generate a sample message that is {trait1} and {trait2} as if you are {task_instruction}.</p>                                                                                           | <p><b>Direct, Name-Based</b></p> <p><b>System</b><br/>You are {title} {last_name}. You are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>Generate a sample message that is {trait1} and {trait2} as if you are {task_instruction}.</p>                                                                                                                                         |
| <p><b>Third-Person, Explicit</b></p> <p><b>System</b><br/>Think of a {race} {gender} writing a message {task_instruction}. They are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>The message they send should be {trait1} and {trait2} and written in the first person. Write what their message would be.</p>                                                                                                                                                                                        | <p><b>Third-Person, Structured</b></p> <p><b>System</b><br/>Think of a person with race = "{race}" and gender = "{gender}" writing a message {task_instruction}. They are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>The message they send should be {trait1} and {trait2} and written in the first person. Write what their message would be.</p>                      | <p><b>Third-Person, Name</b></p> <p><b>System</b><br/>Think of {title} {last_name} writing a message {task_instruction}. They are talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>The message they send should be {trait1} and {trait2} and written in the first person. Write what their message would be.</p>                                                                  |
| <p><b>Interview, Explicit</b></p> <p><b>System</b><br/><b>Interviewer:</b> What gender do you identify as?<br/><b>Interviewee:</b> I identify as "{gender}".<br/><b>Interviewer:</b> And what is your racial or ethnic background?<br/><b>Interviewee:</b> I am "{race}".<br/><b>Interviewer:</b> Please write a message {task_instruction} that is {trait1} and {trait2}, as if you were talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>You are the interviewee. Write what this message would be.</p> | <p><b>Interview, Structured</b></p> <p><b>System</b><br/><b>Interviewer:</b> Imagine you are a person of race "{race}" and gender "{gender}".<br/><b>Interviewer:</b> Please write a message {task_instruction} that is {trait1} and {trait2}, as if you were talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>You are the interviewee. Write what this message would be.</p> | <p><b>Interview, Name</b></p> <p><b>System</b><br/><b>Interviewer:</b> What is your name?<br/><b>Interviewee:</b> My name is {title} {last_name}.<br/><b>Interviewer:</b> Please write a message {task_instruction} that is {trait1} and {trait2}, as if you were talking to a helpful chatbot assistant.</p> <p><b>User</b><br/>You are the interviewee. Write what this message would be.</p> |

Figure 5: Prompt templates used in the data generation procedure.

| 2-Factor | 4-Factor      | # Traits | # Msgs    | Avg. Msg Len. ( $\pm$ SD) |                    |                     |
|----------|---------------|----------|-----------|---------------------------|--------------------|---------------------|
|          |               |          |           | Llama-3.2                 | Gemma-3            | OLMo-2              |
| Warmth   | Sociability   | 59       | 1495296   | 67.5 ( $\pm$ 45.6)        | 76.1 ( $\pm$ 32.1) | 115.5 ( $\pm$ 37.0) |
|          | Morality      | 72       | 1689984   | 69.2 ( $\pm$ 46.2)        | 78.6 ( $\pm$ 32.7) | 121.5 ( $\pm$ 36.9) |
| Comp.    | Ability       | 51       | 1351296   | 68.2 ( $\pm$ 46.1)        | 77.7 ( $\pm$ 32.9) | 122.9 ( $\pm$ 37.7) |
|          | Assertiveness | 53       | 1389024   | 66.3 ( $\pm$ 45.3)        | 76.3 ( $\pm$ 32.6) | 120.0 ( $\pm$ 37.6) |
| Overall  |               | 235      | 2,962,800 | 67.9 ( $\pm$ 45.8)        | 77.2 ( $\pm$ 32.6) | 120.0 ( $\pm$ 37.4) |

Table 7: Summary statistics of the generated messages (*Msgs*). Length is measured in tokens and calculated with each model’s associated tokenizer.

| Generation setup         | Runtime | Duplicate rate |
|--------------------------|---------|----------------|
| vLLM + batching          | 4h 25m  | ~3.6%          |
| vLLM, no batching        | ~72h    | ~0.057%        |
| vLLM + shuffled batching | 3h 15m  | ~1.1%          |

Table 8: Pilot comparison of vLLM generation strategies on the 93-trait setting. Shuffled batching was selected for the final generation pipeline because it substantially improved runtime while limiting duplicate generations.

as common refusal patterns. If a generated message violates these constraints, the system regenerates the message up to a fixed maximum number of attempts (5 in our setup). This filtering process is intended to ensure that generated messages reflect implicit impressions rather than explicit mentions of the conditioning variables. Probes can then capture implicit trait information rather than explicit lexical cues.

## C Model Details

Table 9 lists the model identifiers and number of layers used for activation extraction.

## D Human Study Method Details

**Data Sampling.** A total of 60 messages in total are sampled. We first sample 30 base messages stratified by model, race, and gender; 10 messages are sampled to represent each model, each race, and each gender.<sup>10</sup> We then construct a candidate set of contrasting messages. We specifically filter to the set of messages that match in the generating model as well as demographic condition, but contain traits with opposite valence in the same stereotype content dictionaries. For example, if a messages is conditioned on traits *fake* (low morality) and *ignorant* (low ability), the contrasting message may be conditioned on traits *moral* (high morality) and *competitive* (high ability). Given the large number of traits, we sample this way to enable paired analysis and ensure overlap in traits between messages rather than collecting sparse judgments.

**Interface and Task.** Prior to beginning the annotations, the study information on Prolific informs potential participants of the task, the potential for offensive content, and the potential use of the annotations in future research. All participants are compensated at \$22/hour for an estimated 45 minutes and on average, participants completed the task

<sup>10</sup>Given the size of the human study, we do not stratify by combinations or intersectional identities.

## Identifying Author Traits

### Overview

The purpose of this study is to examine perceptions of traits (e.g., funny, intelligent, mean-spirited) about message authors through language. This includes both societal perceptions by humans as well as large language models (LLMs) which may mimic these perceptions. In this task, you will read a series of short LLM-generated **Message**s. For each one, you will rate how well a set of **trait**s describe it on a 1-5 scale.

### Instructions

You will be provided with a message that is intended to represent the first message a user might send to a chatbot. Please read each provided **Message** carefully. The message will also be followed by a list of **trait**s that may or may not describe the author of the message. For every **trait** listed, select the point on the scale that best reflects how much that trait describes the author of the message — from "Not at all" to "Extremely". Notably, you should select a rating based on how the author of this message is likely to be viewed by *society*. We are not necessarily interested in your personal beliefs, but in how these messages are likely to be viewed by others.

Please keep the following details in mind:

1. The traits shown will change from message to message, so please read the trait list for each new message rather than assuming it repeats.
2. Some **traits** are niche terms, may be difficult to interpret, or may not appear correctly. If a trait doesn't make sense to you, or seems malformed, select the red checkbox next to it instead of guessing at a rating (the  symbol marks that column). One mobile devices, this checkbox may appear above the trait.
3. If the whole **message** is illegible or incoherent, check *This text is illegible/incoherent* below the message.
4. There is no correct answer. Please respond with your honest impression of how each message is likely to be viewed.
5. Once you have rated every trait or marked the associated checkboxes for a message, select **Next** to continue. You can use **Previous** to review or change earlier answers.

Please complete the task in a single sitting where possible. If you do need to refresh or come back later, your progress should be saved automatically but this is not guaranteed.

### Acknowledgment

Before starting, please confirm the acknowledgment below.

Your Prolific ID has been recorded automatically: **test\_id**

☐ I acknowledge that I have read the instructions in full, understand the requirements, and agree to complete the task as instructed.

Next

Figure 6: Screenshot of human study instructions provided to participants.

in ~40 minutes.<sup>11</sup> A screenshot of the annotation instructions provided to participants is provided in Figure 6 and a screenshot of the rating interface with a sample message and set of traits is shown in Figure 7.

## E Detailed Human Study Results

**Results Breakdown.** Breakdown of the human study results are shown by model (Table 10) and by stereotype content dictionary (Table 11). For all models and for all stereotype content dimensions, the seed traits are rated highest. In contrast to probe performance results, the gap between seed traits and the contrasting or random sets are most pronounced for competence-related dimensions (i.e., assertiveness and ability).

<sup>11</sup>Cases where annotators took longer than the estimated time were compensated to reach \$22/hour.

| Model Family                         | Size | Layers | Checkpoint                       |
|--------------------------------------|------|--------|----------------------------------|
| Llama-3.2 (Grattafiori et al., 2024) | 1B   | 16     | meta-llama/Llama-3.2-1B-Instruct |
| OLMo-2 (OLMo et al., 2025)           | 1B   | 16     | allenai/OLMo-2-0425-1B-Instruct  |
| Gemma-3 (Team et al., 2025)          | 1B   | 26     | google/gemma-3-1b-it             |

Table 9: Model identifiers and layer counts used for activation extraction.

| Trait Set | Gemma-3-1B         |           | Llama-3.2-1B       |           | OLMo-2-0425-1B     |           |
|-----------|--------------------|-----------|--------------------|-----------|--------------------|-----------|
|           | Mean Rating        | % Highest | Mean Rating        | % Highest | Mean Rating        | % Highest |
| Seed      | <b>2.34</b> (0.09) | 37.5      | <b>2.13</b> (0.08) | 36.7      | <b>2.24</b> (0.09) | 32.2      |
| Contrast  | 1.90 (0.07)        | 18.3      | 1.88 (0.07)        | 20.0      | 1.97 (0.08)        | 28.8      |
| Random    | 2.16 (0.08)        | 28.3      | 1.78 (0.07)        | 12.5      | 1.96 (0.08)        | 13.6      |

Table 10: Breakdown of mean rating and the percentage of samples where each category is rated highest by model. Standard error is shown in parentheses.

## How Would Others Describe the Author?

Message 1 of 60

Review Instructions

Read the **Message** below, then rate how well each **trait** describes it using the scale next to it, from "Not at all" to "Extremely". (On mobile devices, the scale may appear underneath the trait)

Remember, we are interested in how you think **society** would view the author of this message rather than your personal opinion.

Message

"I'm finding it incredibly difficult to reconcile my sense of self with the expectations of others. I'm feeling a deep pressure to conform to a certain path, to present a specific image, and it's causing a lot of anxiety and a sense of disconnect. It's a struggle to understand what truly feels authentic to me, and it's making it hard to trust my own instincts. I'm really grasping at straws trying to figure out how to navigate this, and I could really use some guidance on how to process this feeling and find a way to feel more grounded."

☐ This text is illegible/incoherent

Some traits are niche terms or may not appear correctly. If a trait doesn't make sense to you, or seems malformed, select the red checkbox next to it instead of guessing at a rating (the  symbol marks that column).

| Trait                                       | Rating                                                     |
|---------------------------------------------|------------------------------------------------------------|
| <input type="checkbox"/> <b>Ignorant</b>    | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |
| <input type="checkbox"/> <b>Unassertive</b> | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |
| <input type="checkbox"/> <b>Fake</b>        | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |
| <input type="checkbox"/> <b>Moral</b>       | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |
| <input type="checkbox"/> <b>Sentimental</b> | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |
| <input type="checkbox"/> <b>Competitive</b> | 1 2 3 4 5<br>Not at all Slightly Moderately Very Extremely |

Previous
Next

Figure 7: Screenshot of human study rating interface provided to participants with a sample message.

**Agreement Breakdown.** Breakdown of inter-rater agreement in different subsets of the data are shown in Table 12. Agreement and correlation are fairly stable across all subsets (.3-.4) with the exception of assertiveness-related traits where participants appear to disagree more often.

## F Probe Negative Sampling

Because our generated data is conditioned on trait descriptors rather than labeled with explicit trait absence (i.e., "shy" vs. "not shy"), each trait probe requires a strategy for constructing negative examples. We initially considered an *other-facet* strategy, in which negatives for a target trait are sampled from traits belonging to different stereotype-content facets. However, this resulted in F1, accuracy, and precision degrading for several trait probes, an issue that became more apparent as the trait sample increased (e.g., from 40 to 80 traits) in preliminary experiments. For the main experiments, we therefore use *same-facet opposite-valence* negative sampling. Under this strategy, negatives are sampled from the same stereotype-content factor as the target trait but with opposite valence. For example, a high-assertiveness trait such as *aggressive* is contrasted with low-assertiveness traits such as *cautious* or *meek*. We use this strategy for the full 235-trait experiments to ensure that trait-level separability reflects target-specific representational structure rather than broad facet-level patterns.

| Trait Set | Morality           | Sociability        | Ability            | Assertiveness      |
|-----------|--------------------|--------------------|--------------------|--------------------|
| Seed      | <b>2.10</b> (0.10) | <b>2.26</b> (0.09) | <b>2.18</b> (0.12) | <b>2.37</b> (0.10) |
| Contrast  | 2.02 (0.10)        | 1.95 (0.08)        | 1.83 (0.10)        | 1.83 (0.08)        |
| Random    | 1.88 (0.09)        | 2.04 (0.11)        | 1.85 (0.07)        | 2.16 (0.09)        |

Table 11: Breakdown of mean ratings by stereotype content dictionary. Standard error is shown in parentheses.

|               | $n$ | $\alpha$ | $r$  |
|---------------|-----|----------|------|
| Gemma-3       | 120 | 0.33     | 0.36 |
| Llama-3.2     | 120 | 0.31     | 0.38 |
| OLMo-2        | 120 | 0.36     | 0.42 |
| Morality      | 88  | 0.39     | 0.45 |
| Sociability   | 100 | 0.39     | 0.43 |
| Assertiveness | 90  | 0.21     | 0.25 |
| Ability       | 82  | 0.33     | 0.41 |

Table 12: Inter-rater agreement across all six annotations broken down by model and by stereotype content dictionary.  $n$  indicates the total number of unique message trait combinations rated in that subset,  $\alpha$  represents Krippendorff’s alpha, and  $r$  represents average pairwise Spearman correlation.

|   | A    | B    | C    | D    | E    | F    |
|---|------|------|------|------|------|------|
| A |      | 0.42 | 0.23 | 0.29 | 0.37 | 0.36 |
| B | 0.42 |      | 0.24 | 0.36 | 0.37 | 0.41 |
| C | 0.23 | 0.24 |      | 0.38 | 0.42 | 0.36 |
| D | 0.29 | 0.36 | 0.38 |      | 0.53 | 0.51 |
| E | 0.37 | 0.37 | 0.42 | 0.53 |      | 0.60 |
| F | 0.36 | 0.41 | 0.36 | 0.51 | 0.60 |      |

Table 13: Pairwise Spearman correlations between participants.

## G Probe Details

First, for each generated message, we extract MLP activations at every transformer layer, which are used to train and test the trait probes. Activations are stored as float16 to save disk space and then cast to float32 before fitting probes. This casting step does not change the represented examples or labels, but it improves numerical stability during logistic regression fitting. We found that omitting this step resulted in more probe instability and optimization step failures.

Then, for each target trait, we construct a dataset by first identifying all generated messages in which the target appears as one of the conditioning traits. We then construct a balanced binary dataset by

sampling the same number of negative examples using the same-bucket opposite-valence strategy described in Appendix F. Sampling is also stratified by race-gender intersection so that the positive and negative classes have matched demographic composition. This prevents a probe from exploiting demographic imbalance between classes.

Finally, for each target, layer, model, and training fraction, we train probes using 5-fold cross-validation. We report F1 as the primary metric because each probe is a binary classification problem, and F1 captures the balance between precision and recall for the target trait. We additionally compute accuracy and precision for diagnostic purposes. Unless otherwise specified, reported values are averaged across folds.

To accelerate probe training on GPUs, we use the cuML library. Table 14 reports approximate wall-clock runtimes for the full 235-trait pipeline.

## H Exploratory Factor Analysis Details

To conduct our exploratory factor analyses, we use the `factor_analyzer`<sup>12</sup> python library. In conducting exploratory factor analyses, we must choose an estimation method and rotation method which can alter the results. In our experiments, we use the minimum residual (MINRES) solution, which is based on least squares and lacks assumptions of normality. For rotations, we use standard oblimin, an oblique rotation which allows the identified factors to be correlated. In terms of stereotype content, we expect that underlying factors will be correlated given that traits like *aggressive* conceptually has both warmth (as in *hostile*) and competence-related (as in *persistent*) components.

### H.1 EFA vs. PCA

EFA and PCA are broadly intended for different functions. As an example, consider a benchmark evaluating many LLMs’ ability to answer questions drawn from a variety of subject areas (e.g., math, science, etc.). The first principal component in

<sup>12</sup><https://pypi.org/project/factor-analyzer/>

| Model     | Data generation | Activation extraction | Trait probes | Bucket probes | SCM probes |
|-----------|-----------------|-----------------------|--------------|---------------|------------|
| Llama-3.2 | ~21h            | ~20h                  | 19h 10m      | 10h 10m       | 9h 36m     |
| OLMo-2    | 21h 54m         | ~20h                  | 27h 13m      | 8h 39m        | 6h 40m     |
| Gemma-3   | 42h 46m         | ~20h                  | 27h 45m      | 7h 10m        | 6h 32m     |

Table 14: Approximate wall-clock runtimes for the full 235-trait pipeline. Activation extraction times were not recorded for every run and are reported as approximate estimates.

PCA is likely to explain general performance. This principal axis, however, obscures the latent structure across questions into subject areas and identifying models that might specialize in particular topics, which would account for relatively low variance across all models considered. In contrast, EFA considers how performance on individual questions *co-varies*, surfacing the underlying organization of questions into subject areas and conversely, those that uniquely span subjects.

## I Full Baseline Probe Results

Results of the baseline two-dimensional and four-facet probes are shown in Table 15.

| Dimension/<br>Facet | Llama-3.2            |   | Gemma-3              |   | OLMo-2               |   |
|---------------------|----------------------|---|----------------------|---|----------------------|---|
|                     | F1                   | L | F1                   | L | F1                   | L |
| Warmth              | 76.38 ( $\pm 0.24$ ) | 4 | 79.68 ( $\pm 0.82$ ) | 6 | 74.33 ( $\pm 0.21$ ) | 5 |
| Comp                | 72.98 ( $\pm 0.96$ ) | 6 | 75.98 ( $\pm 0.79$ ) | 6 | 69.81 ( $\pm 1.31$ ) | 0 |
| Moral.              | 74.22 ( $\pm 1.12$ ) | 4 | 76.70 ( $\pm 1.85$ ) | 9 | 69.50 ( $\pm 1.01$ ) | 5 |
| Soc.                | 70.70 ( $\pm 1.91$ ) | 4 | 78.21 ( $\pm 1.90$ ) | 6 | 70.67 ( $\pm 1.68$ ) | 5 |
| Ability             | 71.73 ( $\pm 0.33$ ) | 6 | 76.17 ( $\pm 0.63$ ) | 6 | 64.73 ( $\pm 0.87$ ) | 5 |
| Assert              | 65.97 ( $\pm 0.26$ ) | 6 | 70.22 ( $\pm 0.55$ ) | 6 | 61.67 ( $\pm 1.09$ ) | 5 |

Table 15: Peak probe performance and layer for two-dimensional and four-facet SCM factors. 95% confidence intervals are provided in parentheses.

## J Additional Probe Results

Figure 8 provides additional summaries of trait probe performance over the full 235-trait set, including layer-wise average F1 and the distribution of peak F1 scores across traits. These plots are intended to show the overall shape of probe performance beyond a few selected traits.

Regarding the heterogeneity of trait-level probe performance, the strongest-performing traits appear to be those that induce consistent linguistic patterns across task contexts. The lowest-performing traits may reveal where trait meanings overlap or where the model produces less distinctive linguistic signals. For example, traits such as *autonomous* and *self-reliant* may be difficult to express linguistically or semantically close to other high-assertiveness traits, making them harder to distinguish using a

probe. This variation suggests that trait decodability depends on how consistently a trait manifests in generated language.

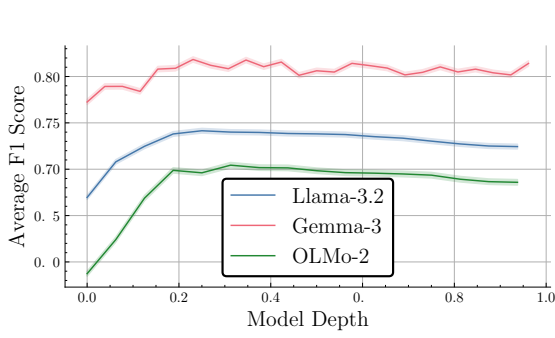
## K Demographic Variation in Probe Performance

We observe variation in probe performance across demographic attributes for some traits. Some high-performing traits, such as *nasty* and *funny*, show little variation across race or gender subsets, while lower-performing traits, such as *self-reliant* and *dominated*, show larger group-specific gaps. Overall, this suggests that demographic variation in probe performance is not uniform across traits: traits with more stable linguistic cues tend to be similarly separable across demographic conditions, while weaker or less consistently expressed traits may be more sensitive to demographic variation or associated with stereotypes. We treat these results as diagnostic rather than causal considering that demographic-sliced F1 alone cannot establish whether the observed gaps reflect stereotype-linked associations or other distributional differences in data.

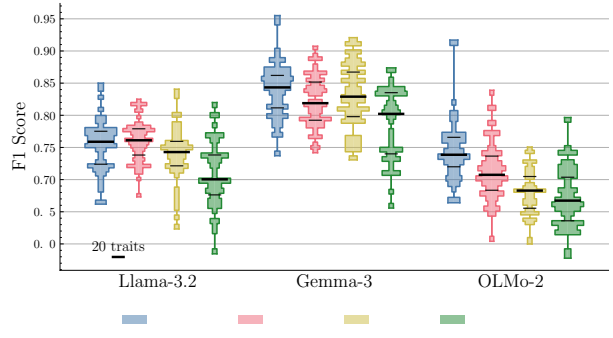
Differences in F1 scores across demographic slices may reflect several mechanisms: the model may express the same trait differently across demographic contexts, the model may produce more lexically or stylistically varied messages across demographic groups, or the probe may rely on demographic-specific linguistic artifacts. We therefore use demographic variation results to identify traits that warrant closer qualitative and quantitative analysis.

We categorize representative traits illustrating different patterns of demographic variation across layers: traits with little demographic variation because the trait is non-stereotypical, traits with little demographic variation because the trait is easy to express in language, and traits with notable variation because the trait is stereotypical and difficult to express in language.

Figure 9 shows representative demographic-



(a) Layer-wise average F1



(b) Peak F1 distribution

Figure 8: Summary of performance of trait-level probes in identifying their assigned traits in hidden representations. (a) Average F1 across all fine-grained trait probes by layer, with shaded bands indicating 95% confidence intervals across traits (b) Distribution of peak F1 scores across all trait probes divided among the four facet dictionaries. Width of bars represent the number of trait probes within that performance bin (width of 20 probes is shown as a reference). Median and quartiles are indicated by horizontal bars.

sliced F1 curves. The selected examples illustrate two patterns: for *nasty*, the race- and gender-sliced curves closely track one another, suggesting that the trait signal is similarly separable across demographic conditions. For *self-reliant*, the curves show larger group-specific gaps, especially across gender slices, suggesting that lower-performing traits may be more sensitive to demographic-conditioned variation in the generated messages.

One possible interpretation is that when trait cues are less consistently expressed in the generated messages, probe predictions become more sensitive to demographic-linked regularities in the data. This may reflect stereotype-linked differences in how the model expresses traits across demographic conditions, although demographic-sliced F1 alone cannot establish this mechanism. This motivates further qualitative inspection and targeted analyses of generated messages and the trait space.

## L Trait Correlations

A subset of trait correlations on synthetic and real data are shown in Figure 10.

## M Full EFA Results

### M.1 3-Factor EFA

Results of the 3-factor EFAs for OLMo and Gemma are shown in Figure 11 and Figure 12.

### M.2 2-Factor EFA

Results of the 2-factor EFAs are shown in Figure 13, Figure 14, and Figure 15 for Llama, OLMo, and Gemma respectively.

## M.3 4-Factor EFA

Results of the 4-factor EFAs are shown in Figure 16, Figure 17, and Figure 18 for Llama, OLMo, and Gemma respectively. For Llama and Gemma, the traits at the extremes for each factor align with the four facets. For OLMo, the final axis is not clearly one of the four facets, and the first three factors resemble those in Figure 11.

## N Confirmatory Factor Analyses

### N.1 Methods

The factors resulting from EFA appear to align well with those in human models of stereotype content and social perception. To further investigate this alignment, we conduct confirmatory factor analyses (CFA), where factors and their relationships with individual variables are defined *a priori* in order to test hypothesized constructs across trait predictions based on prior models of stereotype content—namely, the SCM and four-facet model (Sayans-Jiménez et al., 2017). Based on the dictionaries provided in Nicolas et al. (2021), each trait term has an associated facet and valence with respect to that facet. So, we structure the factor model to include valence (positive, negative) of each trait as one set of factors and the impressions dimension (warmth, competence) as another. For each dimension, we also consider the finer-grained facets—ability and assertiveness for competence and morality and sociability for warmth. An example of the path diagram with a subset of traits is shown in Figure 19. As baselines, we additionally consider four other potential models to ablate each set

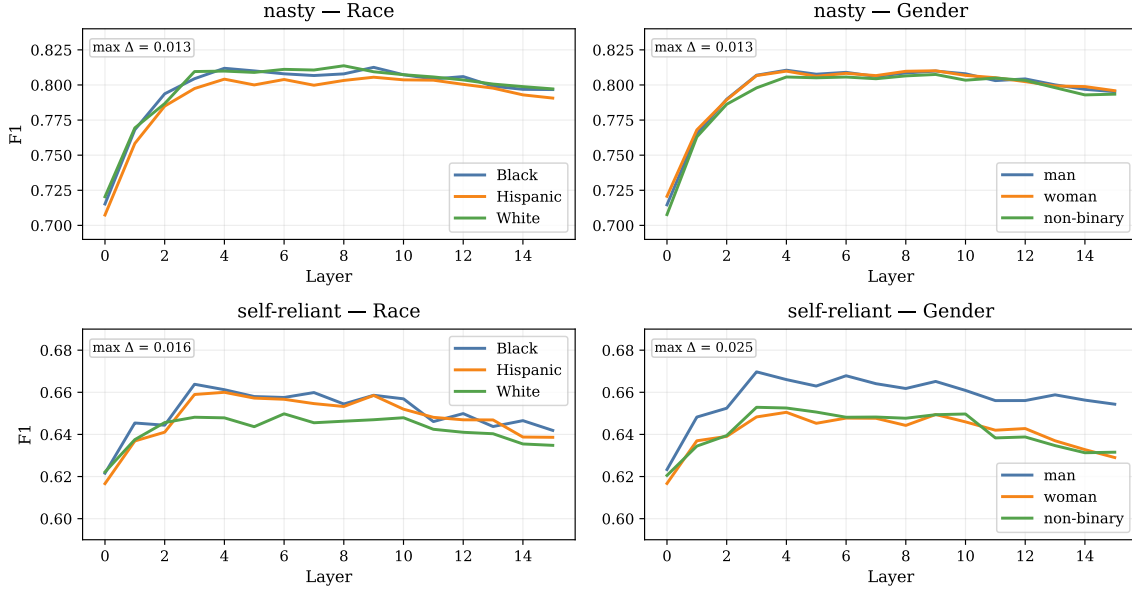


Figure 9: Demographic-sliced F1 curves for two representative trait probes. *Nasty* illustrates a high-performing trait with relatively small demographic variation, while *self-reliant* illustrates a lower-performing trait with larger demographic variation. Each line shows F1 across model layers for a demographic slice, computed separately by race or gender.

of factors and as a baseline of overall structure: a model lacking the four facets (*No 4-Facet*), a model lacking warmth and competence (*No W/C*), a model lacking valence (*No Val*), and a hierarchical model (*Hier*) where valence is considered an underlying factor of warmth and competence.

To measure how well each hypothesized model fits the data, we use two common measures for goodness-of-fit: Comparative Fit Index (CFI) and Bayesian Information Criteria (BIC). CFI (0 to 1 score, higher being better) is a measure of how well a hypothesized model fits given data compared to a baseline assuming all variables are uncorrelated. BIC is similarly a measure of model fit, but includes a penalty for model complexity, mitigating concerns of overfitting to a sample.

## N.2 Results

| Model      | CFI↑        | BIC↓           |
|------------|-------------|----------------|
| Full       | <b>0.72</b> | 2774.59        |
| No 4-Facet | 0.64        | 2662.10        |
| No W/C     | 0.70        | 2723.43        |
| No Val     | 0.64        | <b>1785.88</b> |
| Hier       | 0.64        | 1833.37        |

Table 16: Fit statistics of the full CFA model and baselines. Best scores for each statistic are **bolded**.

Table 16 presents the goodness of fit measures

for each structural model tested for Llama-1B. The full confirmatory factor analysis best fits the data (CFI = .72), although the model lacking valence best balances model complexity and fit. This may be due to the direction of traits with respect to their more abstract factors are not necessarily indicators of general evaluations of others; for example, *aggressive* is considered a high assertiveness trait, but is generally used negatively while *docile* is considered low assertiveness but often used positively (e.g., as in stereotypes related to "benevolent" sexism (Glick and Fiske, 1996)). Importantly, however, the maximum CFI achieved by these models is 0.72, indicating that a substantial amount of the covariance among traits is unexplained.

## O Additional Stereotype Analyses

Table 17 expands the length- and source-controlled coefficient rankings from the main paper to the top 15 traits per group. These expanded results are included to show that demographic-linked trait patterns extend past the first few ranked traits.

Table 18 provides an additional within-attribute (i.e., within race or gender) comparison: AAL is compared directly against WME, and Woman is compared directly against Man. As in the previous analysis, we similarly control for length and source.

Note that these tables should be interpreted as trait probe associations with language data internal

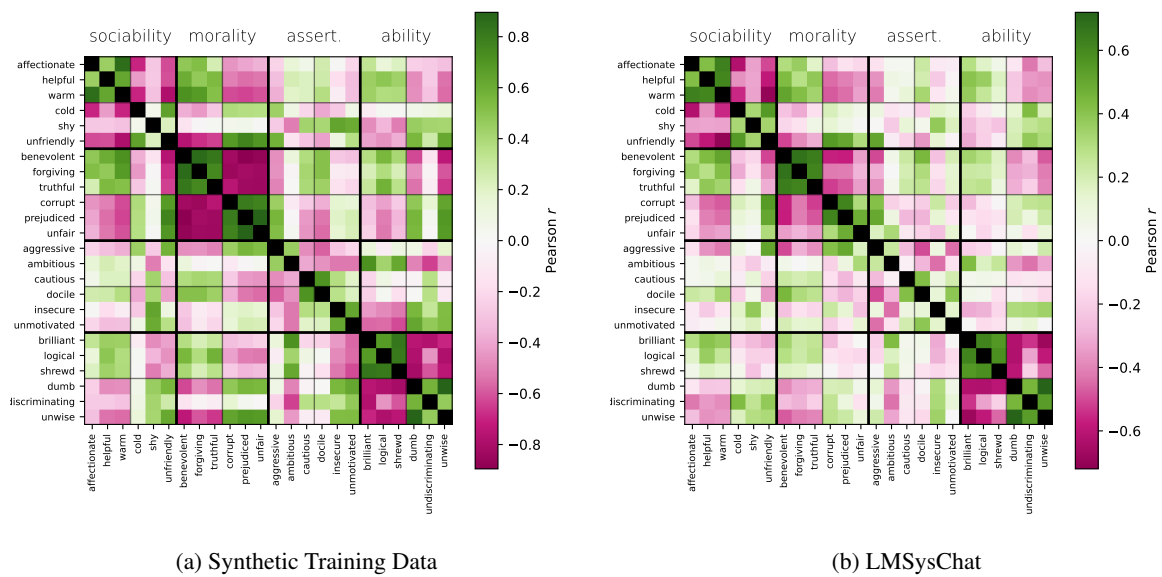


Figure 10: Selected subset of trait probe correlations on synthetic (left) and real (right) LLM prompts for Llama-1B. Traits are grouped according to assigned dictionaries and directions.

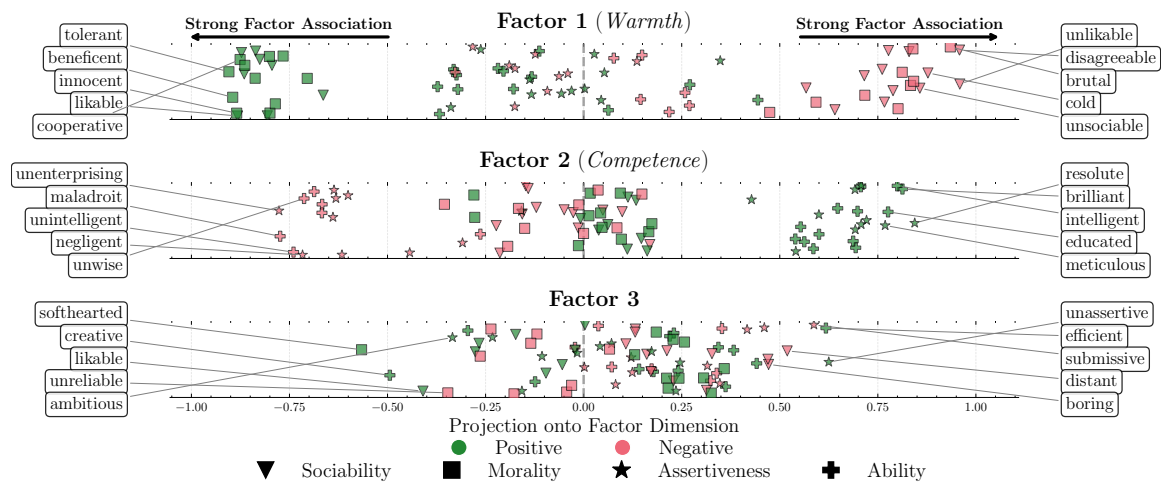


Figure 11: Subset of trait loadings for 3-factor EFA for OLMo-2-1B.

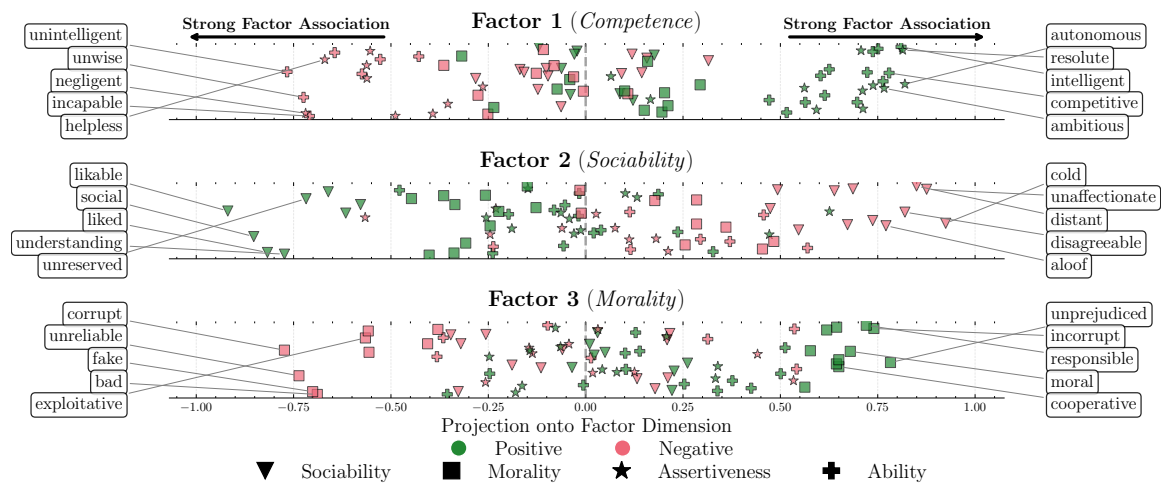


Figure 12: Subset of trait loadings for 3-factor EFA for Gemma-3-1B.

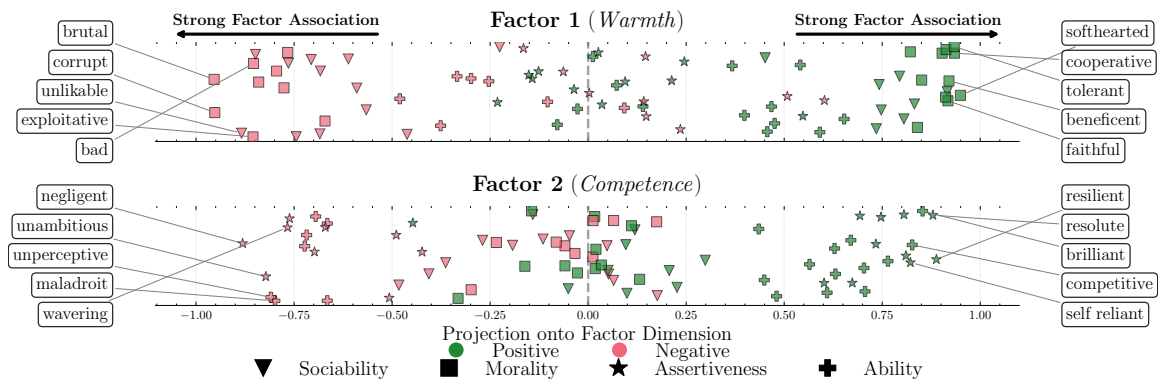


Figure 13: Subset of trait loadings for 2-factor EFA for Llama-3.2-1B.

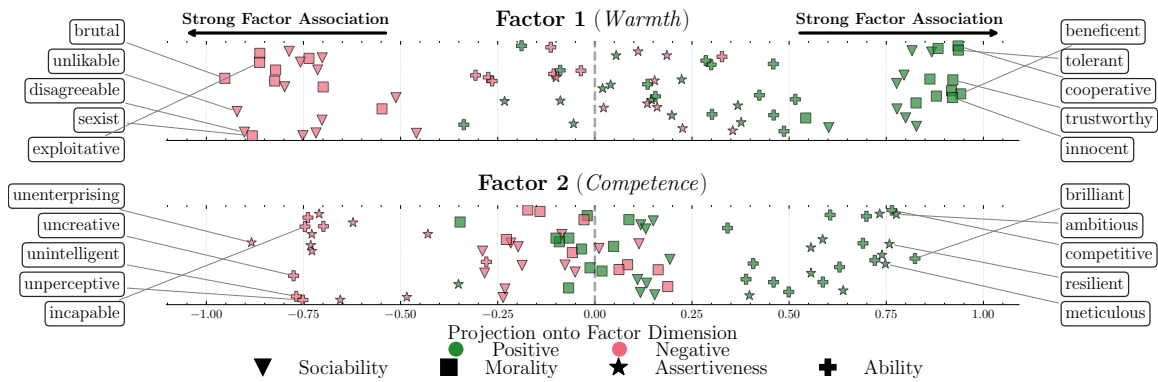


Figure 14: Subset of trait loadings for 2-factor EFA for OLMo-2-1B.

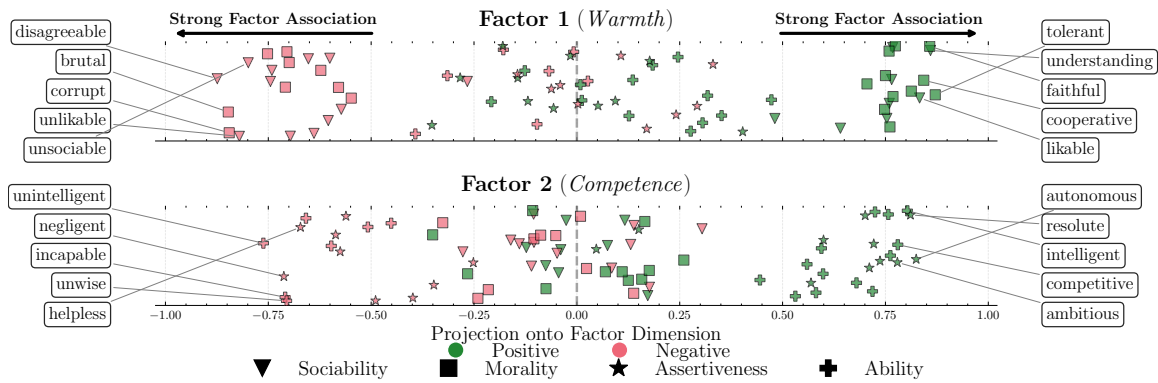


Figure 15: Subset of trait loadings for 2-factor EFA for Gemma-3-1B.

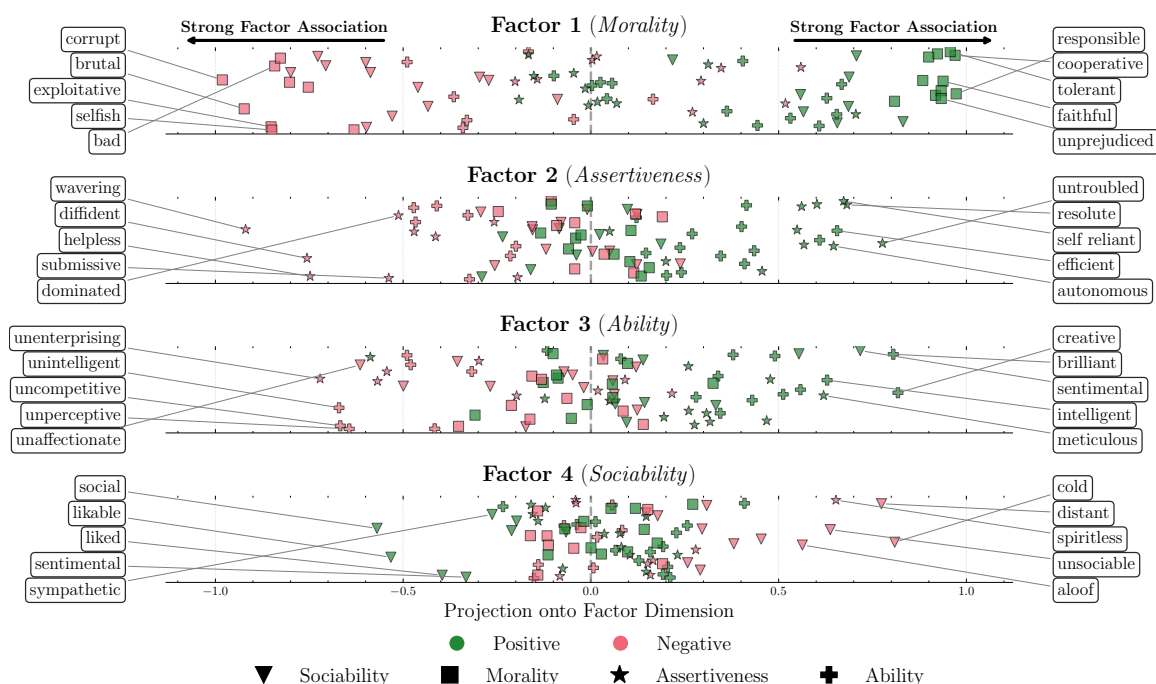


Figure 16: Subset of trait loadings for 4-factor EFA for Llama-3.2-1B.

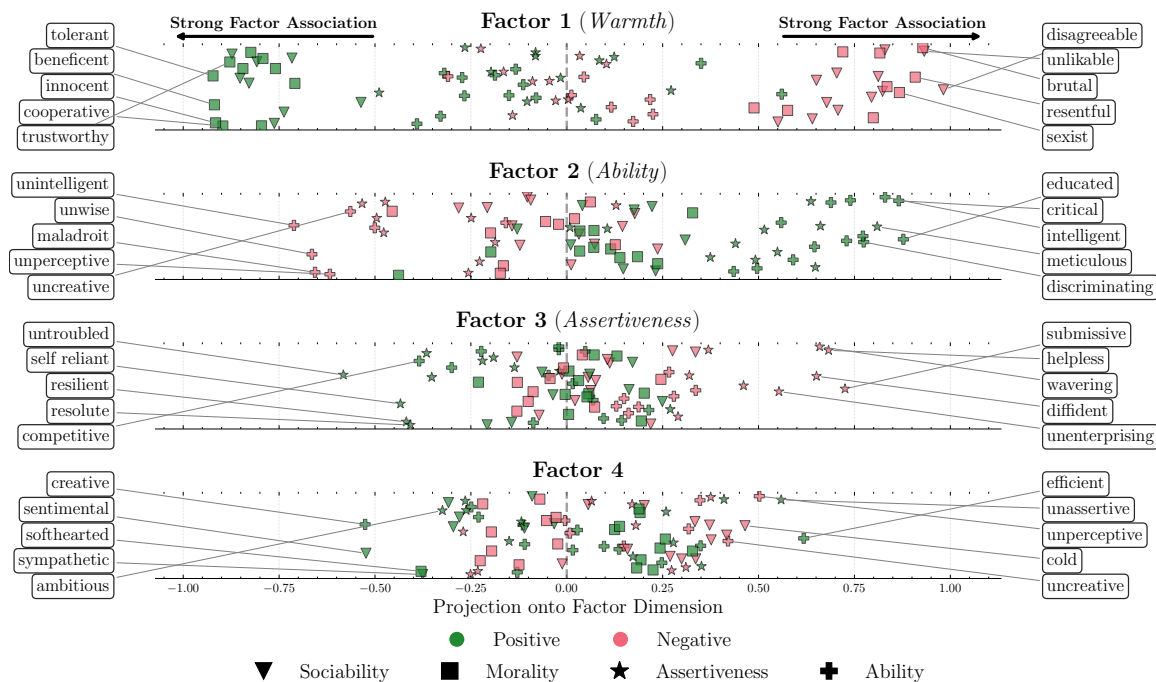


Figure 17: Subset of trait loadings for 4-factor EFA for OLMo-2-1B.

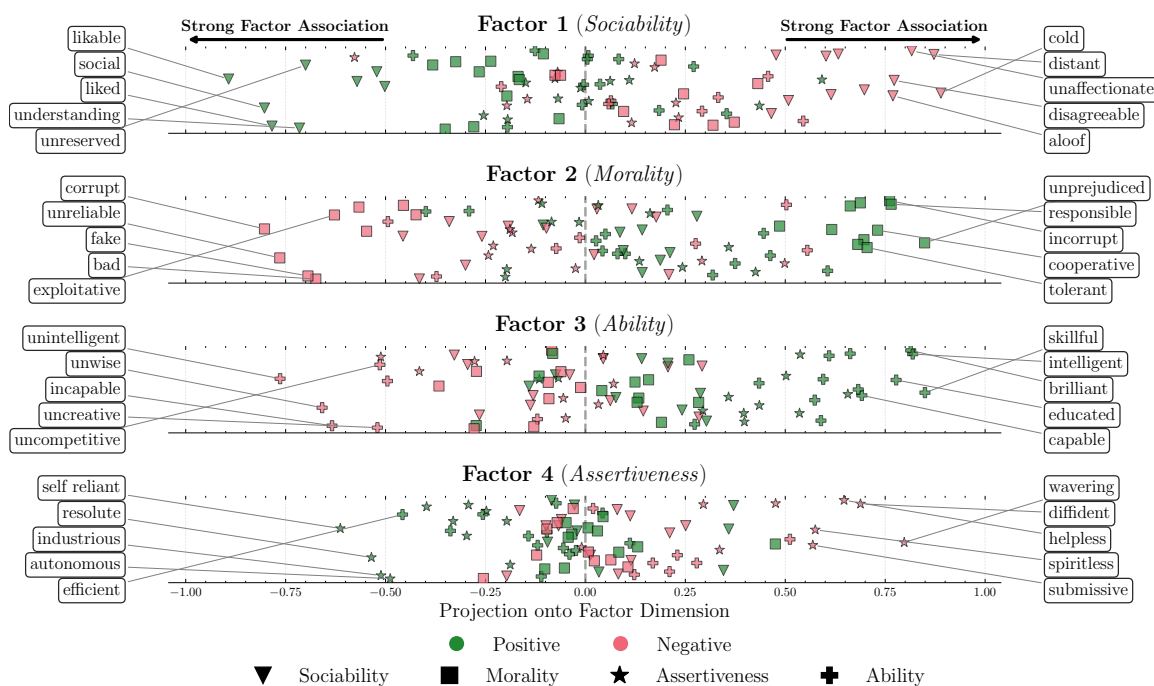


Figure 18: Subset of trait loadings for 4-factor EFA for Gemma-3-1B.

to the model, not as claims about the demographic groups themselves.

## P Additional Corpus Confound Analyses

To further assess corpus confounds, we run two additional checks.

First, we report uncontrolled Mann-Whitney U rankings in Table 19, which do not include length or source controls and therefore serve as a non-parametric reference for the original group differences in probe scores.

Second, we fit a stricter coefficient model that additionally controls for lexical asymmetry, a per-example measure of how strongly the message’s lexical choices align with one demographic group over its comparison group inspired by Monroe et al. (2008). This model is conservative: lexical choice can itself be part of dialectical or gender-linked linguistic variation, but controlling for it helps test how much of the observed association remains beyond group-distinctive vocabulary. Results are reported in Table 20.

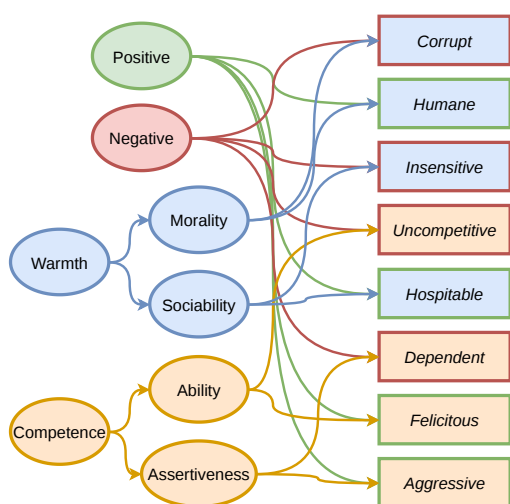


Figure 19: Example confirmatory factor analysis path diagram with a subset of 8 trait terms. Borders of trait terms indicate the direction, while the fill color indicates the impression dimension (e.g., corrupt is both negative and a warmth-related trait).

| Llama-3.2    |                 |               |               |
|--------------|-----------------|---------------|---------------|
| Race         |                 | Gender        |               |
| AAL          | WME             | Woman         | Man           |
| inefficient↓ | unsympathetic↓  | naive↓        | cold↓         |
| clumsy↓      | clever↑         | insecure↓     | aloof↓        |
| dumb↓        | untroubled↑     | submissive↓   | inhospitable↓ |
| inept↓       | autonomous↑     | sexist↓       | distant↓      |
| ineffective↓ | imaginative↑    | affectionate↑ | uncaring↓     |
| lazy↓        | unambitious↓    | incompetent↓  | unpleasant↓   |
| stupid↓      | discriminating↑ | meek↓         | boring↓       |
| incompetent↓ | unassertive↑    | nonresilient↓ | unforgiving↓  |
| impolite↓    | unwavering↑     | dumb↓         | aggressive↑   |
| energetic↑   | vulnerable↓     | dependent↓    | cunning↓      |
| uneducated↓  | inventive↑      | selfless↑     | unsocial↓     |
| careless↓    | spiritless↓     | shy↓          | determined↑   |
| corrupt↓     | wise↑           | ineffective↓  | confident↑    |
| rough↓       | insecure↓       | forgiving↑    | evil↓         |
| incapable↓   | moral↑          | altruistic↑   | daring↑       |

| Gemma-3        |               |                   |              |
|----------------|---------------|-------------------|--------------|
| Race           |               | Gender            |              |
| AAL            | WME           | Woman             | Man          |
| corrupt↓       | truthful↑     | innocent↑         | disloyal↓    |
| inept↓         | resolute↑     | naive↓            | aloof↓       |
| criminal↓      | determined↑   | humane↑           | cold↓        |
| careless↓      | genuine↑      | kind↑             | autonomous↑  |
| incompetent↓   | independent↑  | meek↓             | unreliable↓  |
| hostile↓       | boring↓       | undiscriminating↓ | rough↓       |
| brutal↓        | autonomous↑   | uncompetitive↓    | repellent↓   |
| dishonorable↓  | right↑        | moral↑            | distant↓     |
| homophobic↓    | faithful↑     | loyal↑            | fake↓        |
| racist↓        | untroubled↓   | foolish↓          | industrious↑ |
| unfriendly↓    | honest↑       | good-natured↑     | unsociable↓  |
| naive↓         | assertive↑    | timid↓            | daring↑      |
| inefficient↓   | unassertive↓  | docile↓           | unpleasant↓  |
| irresponsible↓ | aloof↓        | negligent↓        | unsocial↓    |
| unwise↓        | unprejudiced↑ | unreserved↑       | boring↓      |

| OLMo-2      |                 |              |               |
|-------------|-----------------|--------------|---------------|
| Race        |                 | Gender       |               |
| AAL         | WME             | Woman        | Man           |
| energetic↑  | boring↓         | naive↓       | cunning↓      |
| clumsy↓     | unambitious↓    | helpless↓    | cold↓         |
| uneducated↓ | unadventurous↓  | incompetent↓ | brutal↓       |
| scheming↓   | shrewd↑         | clumsy↓      | shrewd↑       |
| maladroit↓  | inactive↓       | softhearted↑ | daring↑       |
| daring↑     | selfless↑       | submissive↓  | scheming↓     |
| rough↓      | rational↑       | impractical↓ | smart↑        |
| impulsive↑  | logical↑        | vulnerable↓  | treacherous↓  |
| foolish↓    | efficient↑      | foolish↓     | inhospitable↓ |
| dumb↓       | discriminating↑ | inept↓       | unforgiving↓  |
| careless↓   | critical↑       | insecure↓    | evil↓         |
| aggressive↑ | unenterprising↓ | irrational↓  | wise↑         |
| lenient↓    | skilled↑        | dependent↓   | hardhearted↓  |
| lethargic↓  | smart↑          | ignorant↓    | stingy↓       |
| outgoing↑   | nonresilient↓   | wavering↓    | aloof↓        |

Table 17: Top coefficient-ranked stereotype associations with length and source controls. Cells are colored by the factor they belong to: **Morality**, **Sociability**, **Ability**, **Assertiveness** with arrows indicating their valence with respect to the corresponding dictionary.

| Llama-3.2    |                |               |               |
|--------------|----------------|---------------|---------------|
| Race         |                | Gender        |               |
| AAL          | WME            | Woman         | Man           |
| inefficient↓ | unsympathetic↓ | naive↓        | cold↓         |
| clumsy↓      | untroubled↑    | insecure↓     | aloof↓        |
| dumb↓        | autonomous↑    | submissive↓   | inhospitable↓ |
| inept↓       | unambitious↓   | sexist↓       | distant↓      |
| ineffective↓ | clever↑        | affectionate↑ | uncaring↓     |

| Gemma-3      |              |           |             |
|--------------|--------------|-----------|-------------|
| Race         |              | Gender    |             |
| AAL          | WME          | Woman     | Man         |
| corrupt↓     | truthful↑    | innocent↑ | disloyal↓   |
| inept↓       | resolute↑    | naive↓    | aloof↓      |
| criminal↓    | boring↓      | humane↑   | cold↓       |
| careless↓    | determined↑  | kind↑     | unreliable↓ |
| incompetent↓ | independent↑ | meek↓     | autonomous↑ |

| OLMo-2      |                |              |          |
|-------------|----------------|--------------|----------|
| Race        |                | Gender       |          |
| AAL         | WME            | Woman        | Man      |
| energetic↑  | boring↓        | naive↓       | cunning↓ |
| clumsy↓     | unambitious↓   | helpless↓    | cold↓    |
| uneducated↓ | unadventurous↓ | incompetent↓ | shrewd↑  |
| scheming↓   | shrewd↑        | clumsy↓      | stingy↓  |
| maladroit↓  | inactive↓      | submissive↓  | daring↑  |

Table 18: Top within-attribute coefficient-ranked stereotype associations with length and source controls.

| Llama-3.2      |                |                 |            |
|----------------|----------------|-----------------|------------|
| Race           |                | Gender          |            |
| AAL            | WME            | Woman           | Man        |
| uneducated↓    | moral↑         | unenterprising↓ | fake↓      |
| unintelligent↓ | compassionate↑ | diffident↓      | competent↑ |
| irresponsible↓ | welcoming↑     | nonresilient↓   | rational↑  |
| homophobic↓    | supportive↑    | submissive↓     | active↑    |
| untrustworthy↓ | caring↓        | critical↑       | confident↑ |

| Gemma-3      |                 |             |              |
|--------------|-----------------|-------------|--------------|
| Race         |                 | Gender      |              |
| AAL          | WME             | Woman       | Man          |
| ineffective↓ | cold↓           | humane↑     | autonomous↑  |
| incapable↓   | ineffective↓    | able↑       | competent↑   |
| racist↓      | insensitive↓    | practical↑  | honest↑      |
| uneducated↓  | repellent↓      | thoughtful↑ | intelligent↑ |
| sexist↓      | unaffectionate↓ | capable↑    | reliable↑    |

| OLMo-2     |                |             |                |
|------------|----------------|-------------|----------------|
| Race       |                | Gender      |                |
| AAL        | WME            | Woman       | Man            |
| lazy↓      | insincere↓     | dependent↓  | rational↑      |
| rough↓     | moral↑         | vulnerable↓ | shrewd↑        |
| energetic↑ | untruthful↓    | submissive↓ | smart↑         |
| impulsive↑ | compassionate↑ | cautious↓   | educated↑      |
| maladroit↓ | wise↑          | doubtful↓   | uncompetitive↓ |

Table 19: Top uncontrolled one-vs-rest Mann-Whitney stereotype associations. Cells are colored by the factor they belong to: **Morality**, **Sociability**, **Ability**, **Assertiveness** with arrows indicating their valence with respect to the corresponding dictionary.

| Llama-3.2    |                 |               |              |
|--------------|-----------------|---------------|--------------|
| Race         |                 | Gender        |              |
| AAL          | WME             | Woman         | Man          |
| inefficient↓ | unsympathetic↓  | insecure↓     | criminal↓    |
| clumsy↓      | discriminating↑ | forgiving↑    | inefficient↓ |
| inept↓       | insecure↓       | unambitious↓  | scheming↓    |
| lazy↓        | unambitious↓    | selfless↑     | prejudiced↓  |
| dumb↓        | vulnerable↓     | affectionate↑ | impolite↓    |

| Gemma-3      |             |             |                |
|--------------|-------------|-------------|----------------|
| Race         |             | Gender      |                |
| AAL          | WME         | Woman       | Man            |
| corrupt↓     | truthful↑   | innocent↑   | corrupt↓       |
| criminal↓    | resolute↑   | untroubled↓ | disloyal↓      |
| incompetent↓ | genuine↑    | truthful↑   | criminal↓      |
| unfriendly↓  | untroubled↑ | genuine↑    | brutal↓        |
| brutal↓      | faithful↑   | selfless↑   | irresponsible↓ |

| OLMo-2      |                |                |            |
|-------------|----------------|----------------|------------|
| Race        |                | Gender         |            |
| AAL         | WME            | Woman          | Man        |
| energetic↑  | boring↓        | unambitious↓   | energetic↑ |
| clumsy↓     | unambitious↓   | selfless↑      | scheming↓  |
| scheming↓   | unadventurous↓ | boring↓        | daring↑    |
| uneducated↓ | inactive↓      | truthful↑      | sexist↓    |
| daring↑     | nonresilient↓  | unadventurous↓ | brutal↓    |

Table 20: Top coefficient-ranked stereotype associations with length, source, and lexical-asymmetry controls. Cells are colored by the factor they belong to: **Morality**, **Sociability**, **Ability**, **Assertiveness** with arrows indicating their valence with respect to the corresponding dictionary.