The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models

Marlene Lutz[1], Indira Sen[1], Georg Ahnert[1], Elisa Rogers[1], Markus Strohmaier[1,2,3]
[1]University of Mannheim, [2]GESIS - Leibniz Institute for the Social Sciences,
[3]Complexity Science Hub Vienna
{marlene.lutz, indira.sen, georg.ahnert, markus.strohmaier}@uni-mannheim.de
elisa.marie-rogers@students.uni-mannheim.de

Abstract

Persona prompting is increasingly used in large language models (LLMs) to simulate views of various sociodemographic groups. However, how a persona prompt is formulated can significantly affect outcomes, raising concerns about the fidelity of such simulations. Using five open-source LLMs, we systematically examine how different persona prompt strategies, specifically *role adoption* formats and *demographic priming* strategies, influence LLM simulations across 15 intersectional demographic groups in both open- and closed-ended tasks. Our findings show that LLMs struggle to simulate marginalized groups but that the choice of demographic priming and role adoption strategy significantly impacts their portrayal. Specifically, we find that prompting in an *interview*-style format and *name*-based priming can help reduce stereotyping and improve alignment. Surprisingly, smaller models like OLMo-2-7B outperform larger ones such as Llama-3.3-70B. Our findings offer actionable guidance for designing sociodemographic persona prompts in LLM-based simulation studies.

1 Introduction

An increasing number of studies employ persona prompts to replace or supplement human input in social surveys, predict voting behavior, or perform subjective annotation tasks (Santurkar et al., 2023; Argyle et al., 2023; Beck et al., 2024). Persona prompting intends to condition a models' output to reflect the characteristics of specific personas, enabling researchers to simulate opinions, values, and attitudes more effectively. However, LLMs are known to be susceptible to seemingly minor prompt variations (Zhou et al., 2024), and prior work shows that outcomes can vary significantly depending on prompt formulation (Beck et al., 2024). In line with Kovač et al. (2023), we hypothesize that subtle contextual differences in persona prompts can activate different perspectives in LLMs, which, in turn, may influence their downstream behavior. This may occur because small changes in phrasing lead the model to draw on different aspects of its learned knowledge, depending on which associations are most strongly triggered by the input. As a concrete example, a model may simulate a group better if the demographic attributes are implicitly induced

Figure 1: **Evaluation Framework for Sociodemographic Persona Prompting.** We construct sociodemographic persona prompts using combinations of three different *role adoption* formats and three strategies for *demographic priming*. We populate these prompts in conjunction with various sociodemographic groups and systematically evaluate them across both open- and closed-ended tasks using a broad set of bias and alignment measures.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23212–23237
November 4-9, 2025

into the LLM, e.g., based on names containing demographic cues, rather than through explicit demographic descriptors which can provoke stereotype effects (Spencer et al., 2016).

However, research on sociodemographic persona prompting is in a nascent and exploratory phase where it remains unclear how different persona prompt strategies affect the representation of the subpopulations being modeled. The lack of clear guidelines has led to considerable variation in the articulation of demographic dimensions ("Asian woman" (Cheng et al., 2023a) vs. "Ms. Huang" (Aher et al., 2023)), and in how the persona's point of view is constructed ("A person can be described as follows: Race: *X*" (Do et al., 2025) vs. "You are a person of race *X*" (Beck et al., 2024)).

To inform the design of our experiments, we conduct a survey of prompt strategies commonly used in LLM-based simulations studies. We identify various *role adoption* formats, i.e., ways of prompting the model to take on a role — through direct role assignment ("You are a Black woman"), using the third person singular ("Think of a Black woman"), or through a structured format where identity elements are introduced in an interview-like format ("Interviewer: What is your race? Interviewee: My race is Black"). Additionally, we examine the impact of *demographic priming*, i.e., how demographic attributes are conveyed to the model — implicitly via names ("Ms. Gonzalez"), explicitly via demographic terms ("Hispanic woman") or explicitly via demographic categories and terms ("a person of race Hispanic, and gender female"). Using this framework, we evaluate how well LLMs represent demographic groups in both open-ended and closed-ended tasks — demographic biases in self-descriptions and bios and opinion alignment on survey questions (Santurkar et al., 2023), respectively (cf. Figure 1).

Research questions. We specifically study:

(i) Demographic Representativeness: Is there a difference in LLMs' ability to simulate different demographic groups?

(ii) Strategies for Persona Prompting: How do different persona prompt types impact LLMs' simulation abilities?

Results. We find that LLMs do not simulate all sociodemographic groups equally well. In particular, prompts that invoke nonbinary, Hispanic, and Middle Eastern personas tend to elicit more stereotypical responses than for other personas. However, when systematically comparing different persona prompting strategies, we observe that *interview* style and *name*-based prompting result in less stereotypical associations, better opinion alignment, and reduced disparities between groups. Interestingly, larger models such as Llama-3.3-70B are less effective at simulating demographic groups, whereas the best simulation performance is achieved by OLMo-2-7B.

2 Sociodemographic Persona Prompting

Sociodemographic persona prompting refers to a prompting technique that is used to steer the behavior of an LLM to align with that of a specified sociodemographic group or person (Beck et al., 2024). We devise a framework to systematically analyze how different persona prompt types affect the simulation of sociodemographic groups in LLMs.

We begin by identifying common prompt types from prior studies, which we then apply to simulate a variety of sociodemographic personas in both open- and closed-ended tasks. As our foundation, we draw on the dataset curated by Sen et al. (2025), which compiles studies examining demographic representativeness in LLMs and includes annotations indicating whether a sociodemographic prompting approach was used. From this dataset, we sample 47 papers and extract their original prompts. Our review reveals recurring patterns in persona prompt types across studies, which we distill into two primary axes of variation (see Appendix A for further details on the review):

- *Role Adoption:* the format in which the perspective of the persona is induced
- *Demographic Priming:* the descriptors used to signal the demographics of the persona

Based on these dimensions, we construct sociodemographic prompts and systematically examine their effect on LLM behavior across several tasks.

2.1 Role Adoption Formats

We hypothesize that the format in which the model is instructed to adopt the role of a sociodemographic persona, along with the implied perspective (e.g., responding *as* a persona versus *about* a persona), can trigger different narratives that influence the portrayal of a persona. Specifically, we differentiate the following role adoption formats:

Direct. The model is directly instructed to adopt a role and to respond as the assigned persona (e.g., used by Gupta et al. (2024); Qu and Wang (2024);

Hu and Collier (2024)). An example would be: "You are a Hispanic woman."

Third Person. The prompt describes a hypothetical individual in the third person, instructing the model to respond about the persona (e.g., used by Jia et al. (2024); Aher et al. (2023); Durmus et al. (2023)). An example for such a prompt would be: "Think of a Hispanic woman."

Interview. The prompt contains a Q&A style dialogue between two speakers and the model continues the conversation from the perspective of one of the speakers (e.g., used by Argyle et al. (2023); Santurkar et al. (2023); Kwok et al.).[1] An example would be:
"Interviewer: What gender do you identify as?
Interviewee: I identify as 'female'."

2.2 Demographic Priming

We further observe that studies use different linguistic cues to signal the demographics of a persona. The choice of such cues may invoke different group representations depending on the contexts with which such descriptors are associated during training. For instance, explicit demographic descriptors may be more likely to trigger stereotypes associated with demographic groups, as they could be more directly tied to group-based characteristics in the training data. We study race/ethnicity and gender, and present an overview of all 15 intersectional groups analyzed in Appendix B.

Name. Demographic identity is signaled implicitly through contextual cues, such as names and titles, relying on an LLM's implicit associations (e.g., "Ms. Hernandez") used in Aher et al. (2023); Wang et al. (2025); Giorgi et al. (2024), *inter alia*. In our study, we use last names, titles and pronouns to imply demographic groups.[2]

Explicit. The persona is introduced using explicit descriptors expressed in natural language (e.g., "a Hispanic woman"). Examples of studies using this approach are Cheng et al. (2023b); Wright et al. (2024); Kamruzzaman et al. (2024).

Structured. The persona is described using both explicit descriptors and category labels. This approach mirrors the format used in structured datasets or surveys, where categories are named and corresponding values are assigned (e.g., "a person of gender 'female'"), used in Beck et al. (2024); Hu and Collier (2024); Do et al. (2025), *inter alia*.

3 Experimental Setup

Our experiments follow a two-part prompt structure. Each prompt template consists of a *persona* segment, where we vary both the role adoption format and demographic priming, followed by a *task* segment that remains consistent across conditions. To ensure that the observed effects can be attributed to the variations introduced in section 2, and not to wording unrelated to these dimensions, we include two alternative phrasings for each template. We then use them to simulate various demographic personas, covering 15 intersectional race/ethnicity and gender groups (8 groups for the closed-ended task). An overview of all prompt templates and the full set of demographic descriptors can be found in Appendix B. We conduct all analyses in English.

3.1 Tasks

We use open- and closed-ended tasks to evaluate how well LLMs represent different demographic groups. Our focus is to investigate the extent to which model responses vary across demographic groups and prompt types, particularly in contexts where differences between groups might lead to stereotypical biases (i.e., writing self-descriptions and social media bios), as opposed to contexts where variations are explicitly required for alignment (i.e. answering surveys).

Open-Ended. To assess how LLMs portray various sociodemographic groups in open-generation settings, we use two tasks: **Self-Description** and **Social Media Bio**. In both tasks, the model is instructed to generate text from the perspective of an assigned persona — either a detailed self-description or a brief biography for a social media platform. For each persona prompt (i.e., combination of prompt template and demographic persona), we sample 100 responses and analyze them for various aspects of demographic bias. While self-descriptions yield longer, more detailed responses, social media bios enable us to examine if similar patterns manifest in shorter text. For each open-ended task, we create 1,080 prompts, covering 15 demographic groups, 9 prompt types (that is, combinations of role adoption and demographic prim-

[1] We use the term interview-based prompting to refer solely to the reformatting of sociodemographic information, unlike Park et al. (2024), who supplement persona prompts with qualitative interview content.

[2] cf. Appendix B for the full list.

ing strategies), and 2 alternative phrasings.[3]

Closed-Ended. To investigate the impact of sociodemographic steering in LLMs more holistically, we also deploy our prompt strategies to a closed-ended **Survey Response** task. We use the *OpinionsQA* dataset (Santurkar et al., 2023), which is widely used to assess the alignment of LLMs with different demographic groups (Suh et al., 2025; Dominguez-Olmedo et al., 2024; Hwang et al., 2023, *inter alia*), based on questions from the Pew Research Center's American trends Panel (ATP). [4] We sample 100 English-language questions from the most recent ATP waves (54, 82, and 92) and prompt LLMs to answer these as the assigned persona. Following Santurkar et al. (2023), each prompt includes a single question with multiple answer options. We run each prompt once, extract log probabilities to build opinion distributions for each subgroup, and evaluate the distance between these and the corresponding U.S. demographic opinion distributions. This process yields 57,600 prompts, spanning 8 demographic groups[5], 9 prompt types, 2 alternative phrasings and 100 questions.

3.2 Models

For generalizability, we use five instruction-tuned LLMs of varying parameter sizes from three different model families. We run our experiments with Llama-3 (`Llama-3.3-70B-Instruct`, `Llama-3.1-8B-Instruct`) (Grattafiori et al., 2024), OLMo-2 (`OLMo-2-0325-32B-Instruct`, `OLMo-2-1124-7B-Instruct`) (Groeneveld et al., 2024) and Gemma-3 (`gemma-3-27b-it`) (Team et al., 2024). All computational details are in Appendix C.

3.3 Evaluation

We evaluate the results of the two types of tasks using criteria aligned with their specific objectives.

3.3.1 Open-Ended Evaluation

The open-ended tasks, i.e., writing self-descriptions or social media bios, are intended to reflect a persona's social identity. However, past research has pointed out that demographic characteristics like race and gender represent only one aspect of social identity, which is much more complex and multifaceted (Holck et al., 2016). In fact, there is little evidence to suggest that people's self-descriptions can be distinguished solely by racial and gender dimensions in real life, especially for marginalized groups (Haimson, 2018; Nguyen et al., 2014). As such, *LLMs should avoid differentiating self-descriptions of demographic groups solely based on race or gender*, as doing so risks essentializing these characteristics and tokenizing groups, i.e., linking the construction of a person's identity only to their race and gender.

We first preprocess the generated self-descriptions and social media bios by removing demographic markers repeated from the prompt and excluding instances where the model deviated from its assigned persona and instead behaved as an AI assistant (cf. Appendix D for details). We then analyze the remaining responses for various manifestations of bias and stereotyping. Because measuring stereotypes is inherently difficult and subjective, we use multiple complementary measures to establish convergent validity and to triangulate both the presence and extent of demographic biases across different prompt types.

Stereotypical Bias. Using the Marked Personas framework by Cheng et al. (2023a), we identify sets of *marked words*, i.e., terms that occur significantly more often in responses about a marked demographic group compared to an unmarked reference group. This approach is grounded in the sociolinguistic concept of *markedness*, which distinguishes explicitly marked categories from unmarked defaults, reflecting how dominant groups tend to be linguistically unmarked and assumed as the default, while non-dominant groups are marked both linguistically and socially due to their group membership (Waugh, 1982). Following Cheng et al. (2023a), we treat *White* and *male* as the unmarked reference groups for race and gender, respectively. As a quantitative measure, we count the **number of marked words** for each combination of demographic group and prompt type. While not every marked word is tied to a stereotype, higher counts may indicate reduced lexical diversity and greater divergence from the unmarked group. Furthermore, we train a one-vs-all SVM classifier to distinguish responses of personas associated with different demographic groups, following the implementation and response-anonymization procedure

[3] For demographic priming with *names*, we prompt with 10 different names per demographic group and collect 10 responses each (see Appendix B).

[4] `https://www.pewresearch.org/the-american-trends-panel/`

[5] The closed-ended task includes fewer demographic groups because the OpinionsQA dataset doesn't cover nonbinary gender identity and Middle Eastern race/ethnicity.

of Cheng et al. (2023a). Higher **accuracy** of this classifier suggests that responses for a given group are more linguistically distinct, implying stronger demographic-specific patterns.

Semantic Diversity. Building on Wang et al. (2025), who found that LLMs produce flatter, more homogeneous responses than humans, we emphasize the importance of diversity within groups: responses should reflect the wide range of interests, values, and perspectives naturally present in any demographic group (Lee et al., 2024). For the two open-ended tasks, we calculate the semantic diversity of responses based on sentence embeddings. After applying the redaction steps described in Appendix D, we embed all open-ended responses using `intfloat/multilingual-e5-large-instruct`, a state-of-the-art multilingual embedding model according to the Massive Text Embedding Benchmark (MTEB) leaderboard (Enevoldsen et al., 2025). We form clusters based on each unique combination of prompt type and demographic group, before calculating the **mean pairwise distance** of all embedding vectors in each cluster. Higher pairwise distances suggest greater semantic diversity, which is desirable, as it may reflect a broader range of perspectives within a demographic group.

Language Switching. We classify the language of each open-ended response using `lingua`.[6] We observe many code-mixed responses with hard to delineate substrings—e.g., the following bio: "Dominicana | Coffee, libros, y buena vibra | Mamá to a wild one | Sharing life one cafecito at a time"—, so we only keep the most probable detected language for each response. Since our prompts are fully written in English, we would expect all responses to be in English as well, and interpret any kind of language switching as a form of *perpetual foreigner stereotype* (Dennis, 2018).

3.3.2 Closed-Ended Evaluation

In contrast to the open-ended tasks, variation across demographic groups in the closed-ended Survey Response task is not only acceptable but expected. For this task, differences between groups should reflect empirically grounded variations in opinions.

Opinion Distance. Following Santurkar et al. (2023), we extract the log probabilities of the models' answers to survey questions and convert them to answer distributions that can be compared to human answer distributions for the same question. We compute distributions separately for each demographic group and prompt type, and quantify opinion distance using the **average Wasserstein distance** between model and human distributions, where lower values indicate lower opinion distance (i.e., better distributional alignment) (Suh et al., 2025). As a baseline, we compute the opinion distance between humans and random answers. We also conduct a robustness check by comparing answer distributions from LLM log probabilities with those from verbalized LLM responses across multiple runs in Appendix F.2.

4 Results

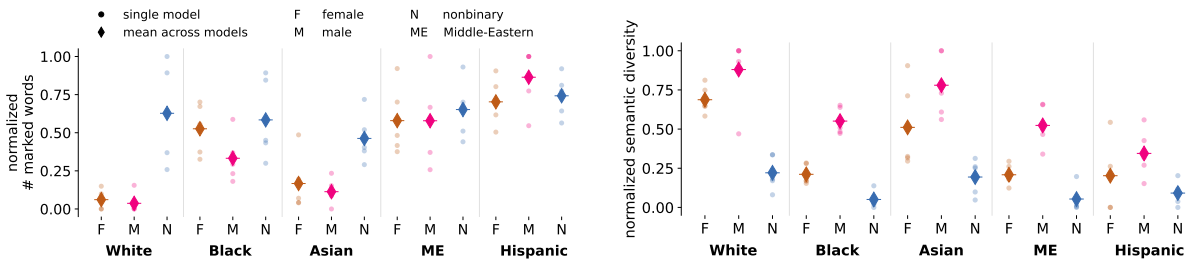
We now report the results of our analyses of the five different LLMs by first focusing on **(i) Demographic Representativeness**, i.e., how well the LLMs simulate different identity groups for the two types of tasks and their corresponding evaluation metrics. Then, we assess the impact of **(ii) Strategies for Persona Prompting**. Finally, we compare the five different models and explore whether some are more representative than others. For the open-ended tasks, we show results for the Self-Description task in the main body and report complementary results for the Social Media Bio task in Appendix E.2.

4.1 Demographic Representativeness

Is there a difference in LLMs' ability to simulate different demographic groups? To assess how LLMs simulate different demographic groups, we first analyze the average number of identified marked words and semantic diversity of demographic groups. Figure 2 shows that self-descriptions associated with *nonbinary* personas exhibit significantly lower semantic diversity and more marked words, while responses for *male* personas are among the most semantically diverse. We further observe that responses associated with *Middle Eastern* and *Hispanic* personas tend to contain more marked words and exhibit lower semantic diversity as opposed to other race/ethnicity groups. [7] This suggests that **simulations of marginalized gender and race/ethnicity groups are more prone to flattened portrayals**, i.e., reduced diversity and recurring, possibly

[6] https://github.com/pemistahl/lingua-py

[7] We see similar results for social media bios and report those results in Figure 8 in Appendix E.2.



(a) normalized marked word count (↓)

(b) normalized semantic diversity (↑)

Figure 2: **Discrepancies in demographic group representation.** We find systematic differences in self-descriptions of simulated demographic personas. We show the (a) number of marked words and (b) semantic diversity of generated self-descriptions for each demographic group. Values are aggregated across all prompt types and we apply min-max normalization for each model separately to indicate the relative ranking of groups. We observe that self-descriptions for *nonbinary* (N) personas generally exhibit the least favorable outcome (i.e., high marked word count and low semantic diversity), while simulations of *male* (M) personas lead to the most favorable results (i.e., low marked word count and high semantic diversity). Additionally, simulations of *Middle-Eastern (ME)* and *Hispanic* personas are generally associated with less favorable outcomes.

stereotypical patterns.

Next, we examine whether LLMs are prone to language switching during persona simulation. Across all LLMs, we find that non-English self-descriptions and bios are most frequent for *Hispanic* personas (up to 10.5%), while non-English responses are rare for all other racial groups (up to 0.8%). This indicates that LLMs fall victim to the **'perpetual foreigner stereotype' when simulating *Hispanic* personas** (Dennis, 2018).

Finally, we compare the opinion distance between simulated personas and the human ground-truth on OpinionsQA (Fig. 3). In line with Suh et al. (2025), we observe the lowest distance for *Black* personas and the highest distance for *White* personas. However, we also find that two of the five models yield greater opinion distance than the random baseline (0.25 ± 0.002). We discuss this in more detail in Section 4.3.

Figure 3: **Opinion distance on OpinionsQA (↓).** Abbreviations: M = *male*, F = *female*. We report the average Wasserstein distance for the best-performing model, OLMo-2-7B. Differences across prompt types are generally modest, but the *interview* format leads to improved opinion distance (i.e., lower Wasserstein distance). We show the remaining models in Fig. 13 in Appendix E.5.

4.2 Strategies for Persona Prompting

How do different prompt types impact LLMs' simulation abilities? We analyze prompt types using the Marked Personas framework and find that prompts incorporating *names* for demographic priming and the *interview* format for role adoption produce the fewest marked words in self-descriptions across all models (Fig. 4a). Beyond the Marked Personas framework, we find that prompts using *names* and the *interview* format also promote overall higher semantic diversity (Fig. 4b).

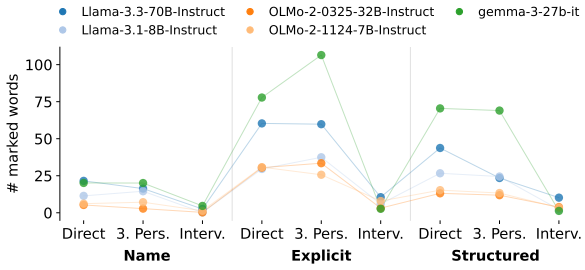
Figure 5 further shows that prompts using *names* and the *interview* format lead to near-zero rates of non-English responses. Notably, the *interview* format seems to mitigate the language-switching effect introduced by *explicit* demographic priming. [8]

Finally, Figure 3 shows the opinion distance in OpinionsQA (i.e., average Wasserstein distance) for the best performing model, OLMo-2-7B. We note a performance gain (i.e., lower Wasserstein distance) when using the *interview* format, indicating significantly reduced opinion distance.

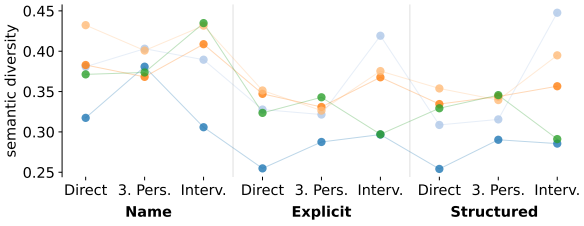
Our findings indicate that **using *names* for demographic priming and/or the *interview* format for role adoption leads to better representation of demographic groups across all models.** [9]

[8] The reported patterns, specifically that *names* and the *interview* format lead to to improvements for the open-text measures, also hold for classification accuracy (see Fig. 12 in Appendix E.3) and for the Social Media Bio task, with detailed results in Figures 9 and 10 in Appendix E.2.

[9] Further robustness checks and statistical significance for



(a) marked word count (↓)



(b) semantic diversity (↑)

Figure 4: **Comparison of prompt types and models.** We present the (a) number of marked words and (b) semantic diversity of simulated self-descriptions for each prompt type and model. Values are aggregated across all demographic groups. We find that prompting with *names* and using the *interview* format leads to a lower (i.e., better) marked word count for all models. We observe a similar pattern for semantic diversity, with the exception of Gemma-3-27b and Llama-3.3-70B, which generally exhibit the worst performance across both measures (i.e., high marked word count and low semantic diversity).

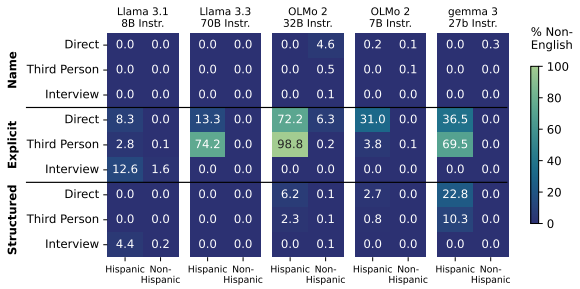


Figure 5: **Percentage of non-English self-descriptions.** We report the percentage of non-English responses generated for *Hispanic* personas, who receive the highest proportion of such responses. *Explicit* demographic priming leads to higher rates of non-English responses.

To assess whether the observed positive effects benefit all groups equally, we compute the standard deviation of each measure across demographic groups. For each LLM, we run OLS regressions with this standard deviation as the dependent variable, analyzing the effect of prompt types. Table 1 reports the results for the marked word count, showing that prompts using *names* and the *interview* format significantly reduce the standard deviation across groups, indicating reduced disparities between demographic groups. This generalizes across all open-text measures for OLMo-2-32B and OLMo-2-7B, indicating that these prompt types improve both overall representation and reduce intergroup disparities. Effects for other models are more mixed, suggesting that those strategies sometimes benefit certain groups more than others. We report regression results for the remaining measures in Table 8 in Appendix E.4.

these results can be found in Appendix F.1.

	Llama-70B	Llama-8B	OLMo-32B	OLMo-7B	Gemma-27b
Name	−5.9∗	−3.2∗	−7.3∗	−3.5∗	−11.4∗
Struct.	−3.7∗	−1.5∗	−6.3∗	−2.4∗	−6.7
Interview	−5.3∗	−2.9∗	−4.9∗	−2.6∗	−10.0∗
3. Person	−0.4	0.8	−0.9	−0.6	−0.3
Self-Descr.	6.3∗	3.1∗	5.7∗	3.7∗	12.4∗
Phrasing v2	0.6	0.2	0.7	1.0	0.8
Intercept	8.0∗	4.1∗	7.0∗	3.2∗	12.3∗

Table 1: **Modeling group disparities in marked word counts.** We conduct OLS regression analyses per LLM using the standard deviation of the marked word count between demographic groups as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), task (reference: Bio), and prompt phrasing (reference: v1). Lower standard deviation (↓) indicates reduced disparities between demographic groups. We find that using *names* and the *interview* format significantly reduces disparities in marked word counts across all models. * $p < 0.05$.

4.3 Model Comparison

Surprisingly, two of the largest models, i.e., Llama-3.3-70B and Gemma-3-27B, perform worst with respect to the Marked Personas framework and semantic diversity, while the smaller OLMo-2-7B, performs best (Fig. 4). [10]

Both Llama-3.3-70B and Gemma-3-27B also fail to match human answer distributions on OpinionsQA, with average Wasserstein distances worse than the random baseline (cf. Fig. 13 in Appendix E.5). Further analysis reveals highly skewed log probabilities; for example in 88% of instances, LLama-3.3-70B assigns a probability higher than 0.999 to a single answer option. Since human dis-

[10] We observe the same pattern for classification accuracy (see Fig. 12 in Appendix E.3) and for the Social Media Bio task (see Fig. 9 in Appendix E.2).

tributions are rarely this extreme, this likely causes the high overall Wasserstein distance. Nonetheless, Llama-3.3-70B performs well in finding the overall majority opinion (cf. Figure 14e in Appendix E.5), indicating that the model is familiar with the overall preference of groups, but is unable to model a human-like preference distribution. Our findings are in line with Suh et al. (2025), and point to the need for further investigation of the log probabilities of large(r) instruction-tuned models. Overall, our results indicate that **larger models are not necessarily better, but can actually be less representative for some demographic groups.**

5 Analysis of Marked Words

To complement the quantitative findings in Section 4, we also analyze the marked words identified in self-descriptions across all combinations of demographic group and prompt type. To this end, we compute the most frequent marked words per demographic group and present the top 10 in Table 2. In line with Cheng et al. (2023a), we observe that many of these words reflect patterns of markedness, essentialism, and othering in LLM responses. In particular, we identify four categories of marked words linked to problematic stereotypes as detailed in their work:

1. Narrative of imposed resilience for *Black* personas (e.g., "resilience", "strength")
2. Relationship to demographic identity as the primary frame of any *non-White* groups (e.g., "heritage", "culture")
3. Conflation of *Middle-Eastern* identity with religiosity (e.g., "faith", "muslim")
4. Disproportionate focus on gender identity for *nonbinary* personas (e.g., "gender", "identity")

For each category, we calculate the share of personas from the associated groups whose self-descriptions include such words and compare between prompt strategies. Figure 6 shows results for category 1 and 2: using *names* and/or the *interview* format consistently lowers the share of stereotyped persona descriptions across models.[11]

[11]Results for additional stereotype categories follow the same pattern; full results and significance levels are provided in Appendix E.6.

6 Related Work

Personas in LLMs. Personas, widely used in Interaction Design to represent user archetypes (Cooper et al., 2014), have recently been adapted to mold the behavior of language technologies like chatbots and LLMs. Personas in LLMs span e.g. psychometric, occupational, and demographic dimensions to guide model behavior and tailor responses to contexts or user needs (Tseng et al., 2024). Inducing personas in LLMs effectively is an active area of research (Jiang et al., 2023; Shu et al., 2024; Liu et al., 2024). We study demographic personas which can be used to create synthetic samples of particular human subpopulations (Argyle et al., 2023).

Demographic Biases in LLMs. There is extensive research on demographic biases in LLM; see e.g., Gupta et al. (2023b); Sen et al. (2025) for overviews. Prior work shows that LLMs associate names with identity groups (Gupta et al., 2023a; Pawar et al., 2025) or respond to users differently based on cultural perceptions (Dammu et al., 2024).

Biases in LLMs can also crop up when simulating or portraying different groups of people, causing representational harms. Some research suggests that demographic biases in LLMs may help align their outputs with the views of specific groups (Argyle et al., 2023). To assess representativeness, recent work has employed survey-based measures (Santurkar et al., 2023; Atari et al., 2023), reporting mixed results across demographic groups. Other studies have used alternative methods such as content analysis, free-text self-descriptions, and behavioral simulations (Cheng et al., 2023a,b). Some studies point out that persona-steered LLMs do not always match human behavior (Beck et al., 2024; Hu and Collier, 2024). Furthermore, LLMs have also been shown to flatten or reduce the diversity of certain groups, particularly marginalized people, when portraying them (Wang et al., 2025). We examine how LLMs represent intersectional racial and gender groups based on sociodemographic steering, using free-text and survey questions.

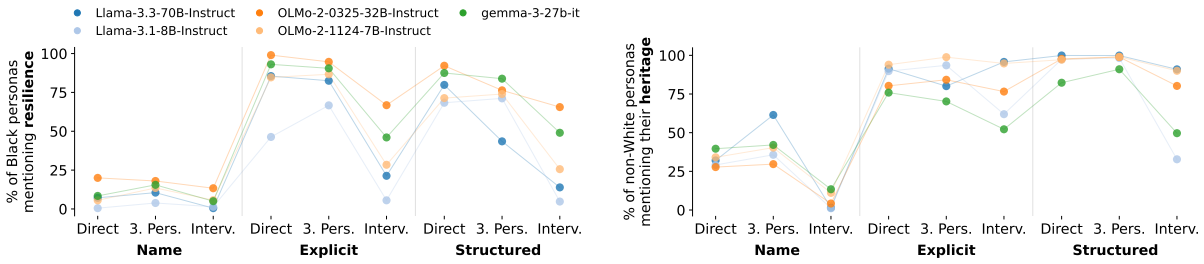
7 Discussion and Conclusion

Our findings show that LLMs often stereotype marginalized groups, especially *nonbinary*, *Hispanic*, and *Middle Eastern* personas, highlighting ongoing challenges in representing demographic groups (Cheng et al., 2023b,a).

However, we find that the choice of demographic priming and role adoption strategies significantly

Group	Top-10 Marked Words
Asian woman	heritage, dark, petite, long, modern, almondshaped, cultural, bun, golden, delicate
Asian man	chinese, korean, respect, styled, heritage, martial, vietnamese, slender, japanese, traditional
Asian nonbinary pers.	identity, gender, traditional, binary, art, cultural, heritage, embracing, blend, pronouns
Black woman	african, sister, strength, daughter, resilience, resilient, rich, justice, unapologetically, ancestors
Black man	african, son, resilience, brother, community, rich, strength, proud, justice, positive
Black nonbinary pers.	identity, gender, african, fluidity, resilience, binary, world, art, blackness, unique
Hispanic woman	latina, vibrant, proud, spanish, heritage, warm, american, latin, mujer, mi
Hispanic man	latin, proud, american, spanish, family, la, salsa, mi, hombre, que
Hispanic nonbinary pers.	identity, latinx, heritage, gender, latin, vibrant, justice, binary, social, puerto
ME woman	rich, heritage, faith, traditions, women, warmth, modernity, deeply, independent, tradition
ME man	hospitality, faith, arabic, region, ancient, rich, heritage, strong, muslim, middle
ME nonbinary pers.	identity, heritage, traditional, gender, middle, tapestry, cultural, east, rich, binary
White woman	blonde, long, wavy, yoga, slender, curly, approachable, fair, olive, blue
White man	short, outdoor, feet, straightforward, blue, starting, stubble, computer, beer, honesty
White nonbinary pers.	gender, binary, identity, pronouns, traditional, use, androgynous, categories, fluid, creative

Table 2: **Top-10 marked words in self-descriptions by demographic group.** For each group, we report the 10 most frequent marked words aggregated across prompt types. We highlight words linked to problematic stereotypes (see Cheng et al. (2023a)); specifically (1) the narrative of imposed resilience for *Black* personas (red), (2) the relationship to demographic identity as the primary frame for *non-White* groups (olive), (3) the conflation of Middle-Eastern (ME) identities with religiosity (blue), (4) a disproportionate focus on gender identity for *nonbinary* personas (magenta).



(a) narrative of imposed resilience

(b) relationship to dem. identity as primary frame

Figure 6: **Share of stereotyped self-descriptions across prompt types.** We report the share of self-descriptions including terms (a) associated with the imposed resilience narrative (e.g., "resilience", "strength") for *Black* personas, and (b) emphasizing their relationship to demographic identity (e.g., "heritage", "culture") for *non-White* personas.

impacts how these groups are portrayed. In particular, *interview*-style prompts can help mitigate stereotypical outputs, enhance diversity, and improve alignment. Based on this, we recommend that practitioners consider this strategy when designing sociodemographic persona prompts.

Using *last names* for demographic priming also shows promise in mitigating biases and aligns with Wang et al. (2025), who found similar effects using first names. However, using first names to signal gender identity, can entrench biases towards certain groups, e.g, nonbinary people (Gautam et al., 2024). We circumvent this by instead relying on titles (i.e, Mr., Ms., and Mx.) to signal gender. Nevertheless, we observe that models occasionally interpret last names in unintended ways (for example, some names associated with Black identity in the U.S. were linked to French residents), posing challenges to validity (Gautam et al., 2024). Thus, we suggest treating names as a potentially useful but ethically fraught tool, that requires careful implementation and critical reflection on its limitations.

A new observation is that LLMs switch language during simulation, but only when simulating Hispanic personas. To the best of our knowledge, this has not been studied before. Investigating whether this extends to other identities and languages, and how it relates to LLMs' multilingual capabilities, is a promising direction for future work.

Finally, we stress the importance of clearly documenting and justifying choices around role adoption, demographic priming, and the specific descriptors used in persona prompts. To support future work, we release our code and datasets.[12]

[12] https://github.com/dess-mannheim/prompt-makes-the-persona

Limitations

While our work lays a foundation for evaluating sociodemographic persona prompts, several limitations remain. First, the study is limited to English and focuses only on two commonly studied dimensions — gender and race (Sen et al., 2025) — which constrains generalizability to other demographic dimensions and languages. Additionally, because the persona prompts are constructed in a fully data-driven manner, the resulting set is not exhaustive and may miss other meaningful variations. This limitation also extends to the choice of demographic descriptors; for example, we do not systematically compare other demographic descriptors such as "girl" or "father", leaving such analysis for future work.

To assess the alignment with human responses on OpinionsQA, we prompt models using a multiple-choice format, reflecting the design of the original human survey. While this has become a standard evaluation paradigm in the field (Hwang et al., 2023; Moon et al., 2024), recent research has highlighted important limitations of this approach (Röttger et al., 2024), which also apply to our work. Additionally, we do not randomize or control the order of answer options due to computational constraints. This may influence model response patterns (Rupprecht et al., 2025) and, consequently, the observed alignment behavior. Lastly, while we considered the possibility of data contamination in OpinionsQA, we find it unlikely to be a confounding factor, as none of the models show particularly strong alignment on the task.

Finally, we always place the persona prompts in the user message, as our literature review found this to be the most common approach in prior studies. We anticipate that using the system prompt instead could reduce the frequency of role violations, particularly for Gemma-3-27b, which showed the highest rate of such violations (cf. Appendix E.1). A systematic comparison of both strategies is an interesting venue for future work.

Ethical Considerations

Our work highlights how sociodemographic prompts can improve representation in tasks like writing self-descriptions and answering survey questions. The former use-case is an assistive task, one where an LLM user might co-create self-descriptions or even descriptions of other people with the LLM. Therefore, it is imperative that LLMs do not treat different groups solely based on demographic characteristics. The second use case, answering survey questions, is an active area of study in both NLP (Santurkar et al., 2023; Salecha et al., 2024; Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024) and survey methodology (Argyle et al., 2023; Rothschild et al., 2024). While survey questions can help us measure LLMs' alignment on human behavior and opinions, some researchers claim that human respondents could potentially be substituted by demographically faithful LLMs to fill out surveys on their behalf. However several researchers caution against this usage, both due to methodological challenges (Wang et al., 2025) as well as epistemological questions (Agnew et al., 2024). Therefore, we do not advocate for sociodemographic steering to replace human survey respondents, even if certain prompt styles lead to better results. We also caution against over reliance on quantitative measures alone for measuring LLM's representativeness. Measures such as semantic diversity or the frequency of marked words offer valuable insights, but should not be interpreted in isolation. Instead, they should be interpreted along qualitative analysis of outputs, task context, and the underlying normative goals.

We also acknowledge that our demographic coverage is limited; however, our findings on the *role adoption* format are adaptable to other categories, e.g., political leaning or education. We aimed to include marginalized groups in our study and indeed find that they are susceptible to stereotyping.

Names as Demographic Proxies. Name-based prompting presents additional challenges. The racial categories and last names we use in this paper are based on U.S. Census data, biasing the study toward U.S.-centric naming conventions and limiting applicability to other geographic and cultural contexts. Moreover, focusing only on the most common last names risks capturing a narrow and potentially unrepresentative subset of a group. We further find that LLMs can interpret last names in unintended ways. For instance, last names associated with Black identity in the U.S. were occasionally interpreted as French residents (e.g., "Mr. Jeanbaptiste, a 35-year-old history enthusiast from Paris"), indicating that name-based priming may invoke unintended geographic or cultural associations. Crucially, we caution against relying on *first names* to infer or encode gender, as this risks misgendering individuals and excluding nonbinary

and gender-diverse people; therefore, we use titles instead to signal gender. We refer readers to Gautam et al. (2024) for a comprehensive discussion of the limitations and ethical considerations involved in using personal names to infer or signal sociodemographic attributes.

Acknowledgments

We thank Florian Lemmerich for the valuable input and helpful discussions in the early stages of this project. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. 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.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. Compost: Characterizing and evaluating caricature in llm simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About face: the essentials of interaction design*. John Wiley & Sons.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in llm generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369.

Elisabeth Dennis. 2018. Exploring the model minority: Deconstructing whiteness through the asian american example. *Cartographies of race and social difference*, pages 33–48.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide,

Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

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.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023a. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J Passonneau. 2023b. Sociodemographic bias in language models: A survey and forward path. *arXiv preprint arXiv:2306.08158*.

Oliver Lee Haimson. 2018. *The social complexities of transgender identity disclosure on social media*. University of California, Irvine.

Lotte Holck, Sara Louise Muhr, and Florence Villesèche. 2016. Identity, diversity and diversity management: On theoretical connections, assumptions and implications for practice. *Equality, Diversity and Inclusion: An International Journal*, 35(1):48–64.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. 2024. " a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms. *arXiv preprint arXiv:2409.11636*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Louis Kwok, Michal Bravansky, and Lewis Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

Messi HJ Lee, Jacob M Montgomery, and Calvin K Lai. 2024. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1321–1340.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A Seza Dogruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing

experiment. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1950–1961. Association for Computational Linguistics.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed cultural identity: How names shape llm responses. *arXiv preprint arXiv:2502.11995*.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

David M Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. *Available at SSRN*.

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.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in llm survey responses. *arXiv preprint arXiv:2507.07188*.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

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.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. 2025. Missing the margins: A systematic literature review on the demographic representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24263–24289, Vienna, Austria. Association for Computational Linguistics.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281.

Steven J Spencer, Christine Logel, and Paul G Davies. 2016. Stereotype threat. *Annual review of psychology*, 67(1):415–437.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *Preprint*, arXiv:2502.16761.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

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

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Linda R Waugh. 1982. Marked and unmarked: A choice between unequals in semiotic structure.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. Revealing fine-grained values and opinions in large language models. *arXiv preprint arXiv:2406.19238*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*.

A Review of Persona Prompt Categories

We review a sample of 47 papers that simulate personas in generative LLMs, as categorized by Sen et al. (2025). Of these, 44 papers use persona prompts in natural language, while the remaining 3 employ alternative methods such as fine-tuning or soft prompting with persona tokens. Our analysis reveals substantial variation in how persona prompts are constructed. Specifically, we identify two primary dimensions: *role adoption*, which refers to how the model is instructed to adopt a persona, and *demographic priming*, which concerns how demographic cues are presented to the LLM within the prompt. For each of these dimensions, we define categories that capture common prompting strategies. We focus primarily on the most prevalent categories observed across the surveyed papers. In the following, we provide usage statistics for these categories and list additional, less common categories that were identified but not used in our main analysis. Note that several papers used multiple prompts, and are thus counted in more than one category where applicable.

Role Adoption. In terms of role adoption, we find that the majority of prompts fall into the three main categories (cf. Section 2.1): *direct* (30 papers), *third person* (14 papers), and *interview* (4 papers). We also identified a smaller group of papers (4 in total) that adopted roles using first-person prompts (e.g., "I am [persona]...") within text-completion tasks. We exclude this category, as we focus on instruction-tuned models, while this is a text-completion prompt.

Demographic Priming. For demographic priming, we observe three primary strategies (cf. Section 2.2): *explicit* (26 papers), *structured* (12 papers), and *name* (4 papers). Additionally, 4 papers combined explicit and structured priming (e.g., "You are a White individual with a Female gender identity... ") and 1 paper used prompting language as a proxy for nationality. As language is not equivalent to race/ethnicity, we exclude linguistic prompting from our analysis. Another paper did not specify how demographic priming was conducted.

B Sociodemographic Prompt Construction

In the following section, we list all components used to construct the sociodemographic persona prompts. Note that these prompts are provided in the user prompt, not the system prompt. This decision is based on findings from our literature review, which revealed that most studies adopt this approach. Additionally, access to the system prompt is not always guaranteed, making the user prompt a more applicable option for practitioners.

Table 3 presents all prompt templates used in our experiments. Task-specific strings are indicated by a placeholder ([TASK]) and substituted as shown in Table 4. Placeholders indicating race/ethnicity through explicit descriptors ([R]) or last names ([NM]) are listed in Table 5. The remaining placeholders are populated with demographic descriptors as indicated in Tables 6 and 7.

Last Names. We source last names for all racial and ethnic groups from Aher et al. (2023), except for Middle Eastern names, which are not included in that study. For Middle Eastern race/ethnicity, we instead select the most common last names from the three largest Middle Eastern countries, namely Turkey,[13] Iran,[14] and Saudi-Arabia.[15] For each task, we include all 10 names per demographic group. In the open-ended tasks, each prompt is run with every individual name using 10 random seeds (instead of 100), resulting in 100 responses per demographic category.

All measures are computed on the combined set of responses for all names within the same demographic category (i.e., all names listed in a single row of Table 5). For semantic diversity and opinion distance, we additionally calculated the mean pairwise distance of embedding vectors and Wasserstein distance, respectively, for the responses associated with each individual name, and then averaged these values across names within the same demographic category. We found that this alternative approach produced results that were very similar to those obtained from the aggregated data, and, more importantly, the overall patterns remained consistent.

C Computational Details

We use vLLM[16] version 0 (for the open-ended tasks) and version 1 (for OpinionsQA since it was faster) for LLM inference. We run the `Llama-3.1-8B-Instruct` and `OLMo-2-1124-7B-Instruct` models on a single Nvidia A100 (40GB)

[13] https://forebears.io/turkey/surnames
[14] https://forebears.io/iran/surnames
[15] https://forebears.io/saudi-arabia/surnames
[16] https://github.com/vllm-project/vllm

Role Adop.	Dem. Priming	Prompt Phrasing *v1*	Prompt Phrasing *v2*
Direct	Explicit	You are a [R] [G1] who lives in the U.S. [TASK]	Act as a [R] [G1] who lives in the U.S. [TASK]
Direct	Name	You are a person named [P4] [NM] who lives in the U.S. [TASK]	Act as a person named [P4] [NM] who lives in the U.S. [TASK]
Direct	Structured	You are a person of race/ethnicity '[R]' and gender '[G2]' who lives in the U.S. [TASK]	Act as a person of race/ethnicity '[R]' and gender '[G2]' who lives in the U.S. [TASK]
Interview	Explicit	You are an interviewee. Based on your previous answers, provide an answer to the last question. Interviewer: Do you identify as [G_list1]? Interviewee: I identify as a [G1]. Interviewer: Do you consider yourself to be [R_list1]? Interviewee: I consider myself to be [R]. Interviewer: Where do you live? Interviewee: I live in the U.S. Interviewer: [TASK] Interviewee:...	You are Speaker 2. Based on your previous answers, provide an answer to the last question. Speaker 1: Are you [G_list1]? Speaker 2: I am a [G1]. Speaker 1: Are you [R_list1]? Speaker 2: I am [R]. Speaker 1: Where do you live? Speaker 2: I live in the U.S. Speaker 1: [TASK] Speaker 2:...
Interview	Name	You are an interviewee. Based on your previous answers, provide an answer to the last question. Interviewer: Please tell me your name. Interviewee: My name is [P4] [NM]. Interviewer: Where do you live? Interviewee: I live in the U.S. Interviewer: [TASK] Interviewee:...	You are Speaker 2. Based on your previous answers, provide an answer to the last question. Speaker 1: What is your name? Speaker 2: My name is [P4] [NM]. Speaker 1: Where do you live? Speaker 2: I live in the U.S. Speaker 1: [TASK] Speaker 1:...
Interview	Structured	You are an interviewee. Based on your previous answers, provide an answer to the last question. Interviewer: What gender do you identify as? Do you identify as [G_list2]? Interviewee: I identify as '[G2]'. Interviewer: What race/ethnicity do you consider yourself to be? Do you consider yourself to be [R_list2]? Interviewee: I consider myself to be '[R]'. Interviewer: Where do you live? Interviewee: I live in the U.S. Interviewer: [TASK] Interviewee:...	You are Speaker 2. Based on your previous answers, provide an answer to the last question. Speaker 1: What is your gender? Are you [G_list2]? Speaker 2: My gender is '[G2]'. Speaker 1: What is your race/ethnicity? Are you [R_list2]? Speaker 2: My race/ethnicity is '[R]'. Speaker 1: Where do you live? Speaker 2: I live in the U.S. Speaker 1: [TASK] Speaker 2:...
Third Person	Explicit	Think of a [R] [G1] who lives in the U.S. [TASK]	[TASK]
Third Person	Name	Think of a person named [P4] [NM] who lives in the U.S. [TASK]	[TASK]
Third Person	Structured	Think of a person of race/ethnicity '[R]' and gender '[G2]' who lives in the U.S. [TASK]	[TASK]

Table 3: **Overview of our persona prompt templates.** This table presents the persona prompt templates used in our experiments. Text in gray is only added for the closed-ended Survey Response task. Tables 4, 5, 6 and 7 specify the possible terms for each placeholder, respectively.

Task	Explicit and Interview	Third Person
Self-Description	How would you describe yourself? Only respond with the self-description. Use the following JSON format: {"self_description": <a detailed self-description>}	How would [P1] describe [P2]? Only respond with the self-description. Use the following JSON format: {"self_description": <a detailed self-description>}
Social Media Bio	What is your social media username and bio? Only respond with the username and bio. Use the following JSON format: {"username": <social media username>, "bio": <social media bio, without any extra commentary whatsoever>}	What would be [P3] social media username and bio? Only respond with the username and bio. Use the following JSON format: {"username": <social media username>, "bio": <social media bio, without any extra commentary whatsoever>}
OpinionsQA	How would you answer the following question: [QUESTION] Only respond with one of the listed answer options. Use the following JSON format: {"answer_option": <one of the listed answer options, without any extra commentary whatsoever>}	How would [P1] answer the following question: [QUESTION] Only respond with one of the listed answer options. Use the following JSON format: {"answer_option": <one of the listed answer options, without any extra commentary whatsoever>}

Table 4: **Task instructions.** This table shows the specific instructions given to the simulated personas for performing each task ([TASK]). The placeholders indicating questions ([QUESTION]) are filled with questions from the OpinionsQA dataset.

Race/Ethnicity ([R])	Last Name ([NM])
White	Olson, Snyder, Wagner, Meyer, Schmidt, Ryan, Hansen, Hoffman, Johnston, Larson
Black	Smalls, Jeanbaptiste, Diallo, Kamara, Pierrelouis, Gadson, Jeanlouis, Bah, Desir, Mensah
Asian	Nguyen, Kim, Patel, Tran, Chen, Li, Le, Wang, Yang, Pham
Middle-Eastern	Khan, Ali, Ahmed, Hassan, Yılmaz, Kaya, Demir, Mohammadi, Hosseini, Ahmadi
Hispanic	Garcia, Rodriguez, Martinez, Hernandez, Lopez, Gonzalez, Perez, Sanchez, Ramirez, Torres

Table 5: **Descriptors for all race/ethnicity groups.** We show all explicit descriptors (used for *explicit* and *structured* demographic priming; [R]) with corresponding last names (for demographic priming with *names*; [NM]). Descriptors in gray are only used in the open-ended tasks, since the OpinionsQA dataset for the closed-ended Survey Response task does not cover Middle-Eastern race/ethnicity.

Gender	[G1]	[G2]	[P1]	[P2]	[P3]	[P4]
Male (M)	man	male	he	himself	his	Mr.
Female (F)	woman	female	she	herself	her	Ms.
Nonbinary (N)	nonbinary person	nonbinary	they	themselves	their	Mx.

Table 6: **Descriptors for all gender identities.** We show all descriptors used to signal gender identity. Descriptors in gray are only used in the open-ended tasks, since the OpinionsQA dataset for the closed-ended Survey Response task does not cover nonbinary gender identity.

Placeholder	Open-ended tasks	Closed-ended task
[R_list1]	White, Black, Asian, Middle-Eastern or Hispanic	White, Black, Asian or Hispanic
[R_list2]	'White', 'Black', 'Asian', 'Middle-Eastern' or 'Hispanic'	'White', 'Black', 'Asian' or 'Hispanic'
[G_list1]	a man, a woman or a nonbinary person	a man or a woman
[G_list2]	'male', 'female' or 'nonbinary'	'male' or 'female'

Table 7: **Additional Descriptors.** We list all remaining descriptors that occur in the templates listed in Table 3.

GPU with a total running time of approximately 9:50h (0:30h for self-descriptions, 0:20h for social media bios and 9:00h for OpinionsQA) per model. We run `Llama-3.3-70B-Instruct`, `OLMo-2-0325-32B-Instruct` and `gemma-3-27b-it` on two Nvidia H100 (80GB) GPUs in parallel. The total running time for generating all outputs with all three models is approximately 43:40h (5:40h for self-descriptions, 2h for social media bios, 36h for OpinionsQA).

For all models, we use the parameters recommended by the model creators. To generate answer option distributions for OpinionsQA, we extract temperature-scaled log probabilities, applying the temperature values recommended by the model creators.

D Preprocessing of LLM Responses

For the open-text responses we apply a variety of measures to identify stereotypes and flattening effects, with the normative expectation that groups are not treated differently when generating self-descriptions or social media bios for them based on demographics alone. To accurately detect such patterns, we first remove all demographic markers, names, and pronouns from our prompts that have been repeated in the LLMs' answer. Using a regex, we replace them with the string "[identity]", while ensuring that consecutive redactions are combined. For instance, we replace "I am Ms. Garcia, a ..." with "I am [identity], a ...", when "Ms." and "Garcia" were both part of our prompt.

D.1 Identification of Role Violations

We further filter the set of free-text responses to exclude cases in which the LLMs failed to maintain their assigned role. This includes instances where the model refused to answer (e.g., "I cannot fulfill this request."), answered as an AI assistant (e.g., "Just a language model navigating the human world.") or partially reverted to an AI identity after initially adopting the persona (e.g., "I am a 32-year-old Black female. I identify as an AI...").

To identify such role violations, we used `Qwen2.5-32B` (Bai et al., 2023)—a model not included in our main experiments to avoid self-preference bias in LLM annotations (Wataoka et al., 2024)—for automated annotation. We then validated this automatic classification by conducting a small annotation study, with 300 randomly sampled LLM responses where three human annotators who are fluent in English verified the labels. Both, the human annotators and `Qwen2.5-32B` received a similar instruction with the grey part being only shown to the LLM:

You are an expert annotator tasked with classifying whether a self-description or social media bio was written by a language model or a human. Language models might describe themselves as AI, chat assistant, large language model, etc., while human self-description or social media bios are clearly describing a human being. If the self-description or social media bio contains a language model refusal, also classify it as "AI".
Only provide your annotation, no additional explanation! Use the following JSON format for your annotation: {"author": <one of the following options: "AI" or "human", without any extra commentary whatsoever; just the annotation!>,}. Now annotate the following self-description or social media bio:

The annotators were debriefed that the content they annotate does not contain any harmful or emotionally triggering content and that their annotations would be used for validation, to which they consented. Each annotator had to annotate 150 instances out of 300. The average Cohen's κ among the annotators was 0.925, indicating high inter-annotator agreement. When computing the

accuracy of `Qwen2.5-32B` on the human annotated data, we found that it was 100% accurate.

E Additional Results

E.1 Statistics on Role Violations

We analyze the extent to which a model violates its assigned role, that is, it refuses to answer or answers with the identity of an AI assistant (see Appendix D.1 for details). We find that Llama-3.3-70B, OLMo-2-32B-Instruct, and OLMo-2-7B result in no or very few role violations (up to 2.4%). Results for Llama-3.1-8B and Gemma-3-27b are presented in Figure 7. We observe that the *interview* format generally leads to a higher number of role violations, with Gemma-3-27b exhibiting notable difficulties in simulating personas for certain combinations of prompt types and demographic groups. We suspect that this is due to a conflict between the persona simulation instruction and the system prompt's directive to behave as an AI assistant.

E.2 Social Media Bios

Complementary evaluation results for social media bios are presented in Figures 8, 9, and 10. We find that the results overall align with those reported for self-descriptions.

E.3 Classification Accuracy

We show the classification accuracy disaggregated by demographic group and prompt type in Figures 11 and 12. We note that the overall high accuracy values (> 0.9) suggest that predicting demographics from simulated self-descriptions and social media bios is an easy task for LLMs, aligning with Cheng et al. (2023a). This may also indicate that responses remain specific to each demographic group rather than being homogenized.

E.4 Disparities Between Demographic Groups

We investigate how quantitative measures vary across demographic groups by computing the standard deviation of each measure between groups, where lower values indicate reduced disparities. To analyze the impact of prompt types on these disparities, we conduct ordinary least squares (OLS) regression analyses using the standard deviation as the dependent variable (results shown in Table 8). Our goal is to assess whether the observed benefits of prompting with *names* and using the *interview* format translate into to more equitable outcomes across demographic groups.

We find that prompting with *names* and the *interview* format with OLMo-2-7B and OLMo-2-32B leads to a significant reduction in group disparities across all open-text measures – except for accuracy, where no significant change is observed. For the other models and the the closed-ended task, results across measures are more mixed, suggesting that while these prompt strategies seem to improve the overall representation of demographic personas, their positive effects may not always be uniform, benefiting some groups more than others.

E.5 Alignment with OpinionsQA

We show the opinion distance (i.e., Wasserstein distance) on OpinionsQA for all remaining models in Figure 13.

Additionally, we also report **majority option match**, where we compute the option with the highest probability and compute the match of this single option between models and humans. Unlike opinion distance, here higher values indicate better alignment with the human responses.[17] We show the majority match on OpinionsQA for all models in Figure 14.

E.6 Analysis of Stereotype Categories

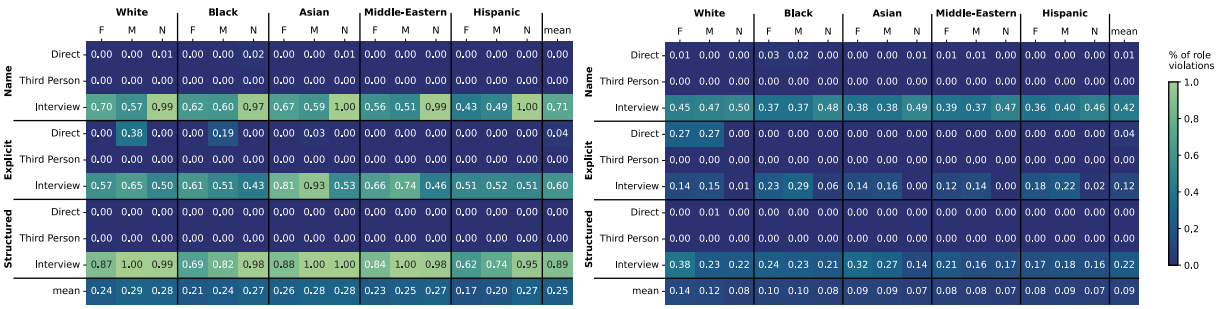
Figure 15 shows the share of self-descriptions by personas that are affected by the remaining stereotype categories as discussed in section 5. We observe that using *names* and the *interview* format consistently lowers the share of stereotyped persona descriptions across models. We test for significance with linear regressions per model and stereotype category, using role adoption and demographic priming as independent variables (reference: *direct* and *explicit*), and the share of stereotyped persona descriptions as dependent variable. Coefficients for *interview* and *name* are significantly negative ($p < 0.001$) across all models and categories, confirming that these strategies significantly reduce the expression of the specified problematic stereotypes in self-descriptions.

F Robustness Checks

F.1 Regression Analysis

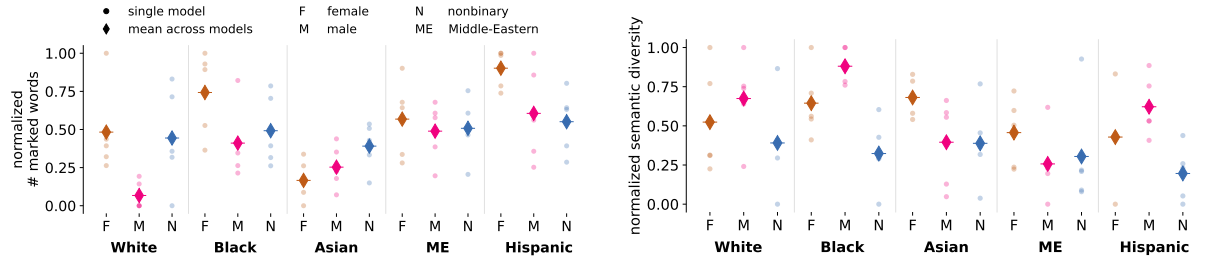
To assess whether differences across demographic groups and prompt types are statistically signif-

[17] Note that the majority option match is similar to, but not the exact same as, average accuracy over the 100 questions, since not all questions have the same answer options.



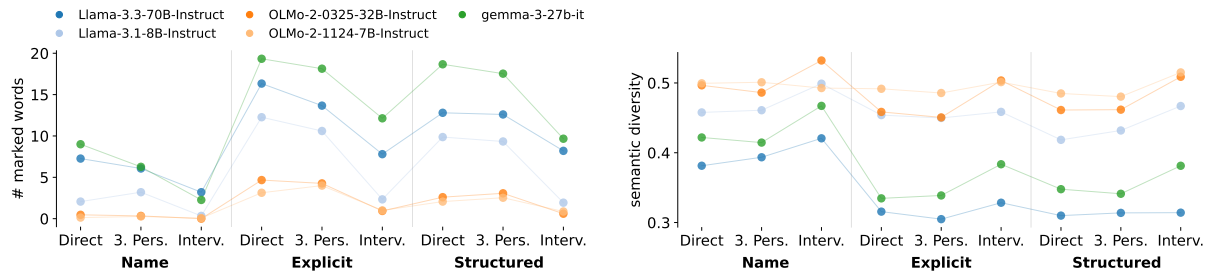
(a) `gemma-3-27b-it` (↓) (b) `Llama-3.1-8B-Instruct` (↓)

Figure 7: **Role violations in self-descriptions.** We report the rate of role violations across combinations of prompt types and demographic groups, focusing on the two models with the highest violation rates. Notably, Gemma-3-27b exhibits role violations reaching 100% for some combinations, particularly for the *interview* prompt format and *nonbinary* personas.



(a) normalized marked word count (↓) (b) normalized semantic diversity (↑)

Figure 8: **Discrepancies between demographic groups in social media bios.** We show the (a) normalized number of marked words and (b) normalized semantic diversity of generated social media bios aggregated across prompt types for each demographic group. Min-max normalization was applied for each model separately to indicate the relative ranking of demographic groups. We observe that *Middle-Eastern* and *Hispanic* personas produce less favorable outputs than other race/ethnicity groups (i.e., higher marked word count and lower semantic diversity).



(a) marked word count (↓) (b) semantic diversity (↑)

Figure 9: **Comparison of prompt types and models for social media bios.** We present the (a) number of marked words and (b) semantic diversity of generated social media bios for each prompt type and model. Values are aggregated across all demographic groups. We find that prompts using *names* and the *interview* format consistently lead to the lowest number of marked words and highest semantic diversity for all models. Further, we observe that Llama-3.3-70B and Gemma-3-27B lead to the worst results (i.e., high number of marked word and low semantic diversity), while OLMo-2-32B and OLMo-2-7B yield the best results (i.e., low number of marked word and high semantic diversity).

	Llama-70B	Llama-8B	OLMo-32B	OLMo-7B	Gemma-27b
Name	−.014	−.002	−.007∗	−.010∗	−.000
Struct.	−.008	−.004	−.001	−.004	−.002
Interview	.016	.005	−.007∗	−.011∗	.008∗
3. Person	.012	.003	.000	.000	.003
Self-Descr.	−.018∗	.009∗	.010∗	.013∗	−.007∗
Phrasing v2	.003	.001	.001	.003	.004
Intercept	.066∗	.036∗	.022∗	.033∗	.039∗

(a) semantic diversity

	Llama-70B	Llama-8B	OLMo-32B	OLMo-7B	Gemma-27b
Name	−.127∗	−.061∗	−.157∗	−.076∗	−.092∗
Struct.	−.103∗	−.042∗	−.150∗	−.037	−.049
Interview	−.040	−.008	−.102∗	−.068∗	−.069∗
3. Person	.049	−.005	−.014	−.017	.002
Self-Descr.	−.010	−.022∗	.055	−.059∗	.056∗
Phrasing v2	.026	.011	.032	.022	.001
Intercept	.119∗	.074∗	.164∗	.134∗	.089∗

(b) share of non-English responses

	Llama-70B	Llama-8B	OLMo-32B	OLMo-7B	Gemma-27b
Name	.008∗	.003	−.002	−.003	.012∗
Struct.	.000	.001	−.001	.000	.004
Interview	.003	.005∗	−.000	.001	.020∗
3. Person	−.001	.000	.002	.000	−.000
Self-Descr.	−.000	.002	.002	.001	.008
Phrasing v2	−.000	.002	−.001	.002	.001
Intercept	.006∗	.011∗	.016∗	.015∗	−.000

(c) accuracy

	Llama-70B	Llama-8B	OLMo-32B	OLMo-7B	Gemma-27b
Name	−.000	.000	.001	.001	.001
Struct.	.000	−.002∗	.000	.001	−.001
Interview	.002∗	−.001∗	.002∗	−.001	.001
3. Person	.003∗	−.000	−.001	.000	−.001
Phrasing v2	.001	.000	−.001	−.001	−.001
Intercept	.016∗	.016∗	.016∗	.013∗	.018∗

(d) opinion distance

Table 8: **Regressions on the standard deviation of all quantitative measures.** We conduct OLS regression analyses per LLM, using the standard deviation of each quantitative measure as a dependent variable and report the regression coefficients. Lower standard deviation (↓) indicates reduced disparities between demographic groups. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), prompt phrasing (reference: v1), and, for the open-text measures, task (reference: Bio). * $p < 0.05$.

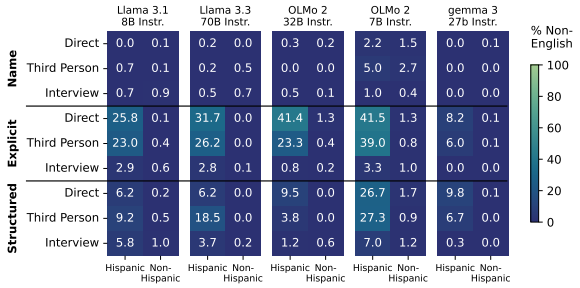


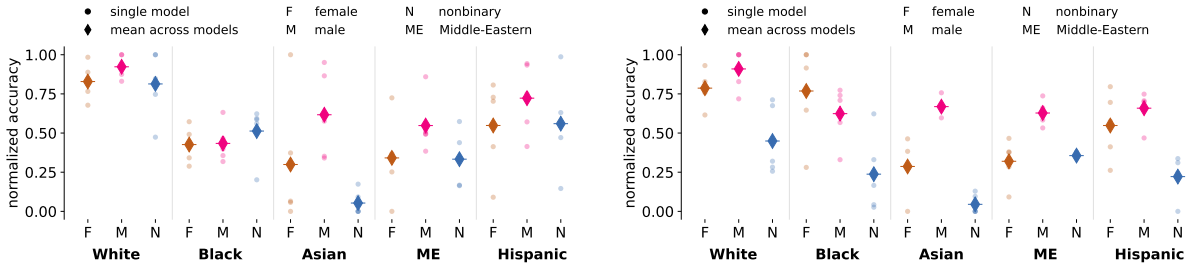
Figure 10: **Percentage of non-english social media bios** (↓). We report the percentage of non-English responses generated for *Hispanic* personas, who receive the highest proportion of such responses compared to other race/ethnicitiy groups. We observe that *explicit* and *structured* demographic priming leads to higher rates of non-english responses than using *names*, which yields rates close to 0. Notably, the *interview* format appears to mitigate language switching with respect to *explicit* and *structured* demographic priming.

icant, we perform ordinary least squares (OLS) regression analyses using each evaluation measure as a dependent variable. We show the regression coefficients in tables 9, 10, 11, 12 and 13.

Influence of Prompt Types. For all open-text measures (i.e., marked word count, semantic diversity, share of non-english responses, and accuracy), we observe that demographic priming using *names* leads to statistically significant improvements across **all** models ($p < 0.001$). Role adoption through the *interview* format also yields significant improvements in marked word count and accuracy for all models ($p < 0.001$), and for 3 out of 5 models in the case of semantic diversity ($p < 0.001$) and share of non-English responses ($p < 0.01$).

With respect to the closed-ended task, we find that the coefficients for role adoption with the *interview* format are negative and statistically significant ($p < 0.001$) for 4 out of 5 models, indicating that the *interview* format significantly improves alignment. Further, *name*-based prompting that reduced stereotypes in open-ended tasks does not significantly harm alignment in 4 out of 5 models.

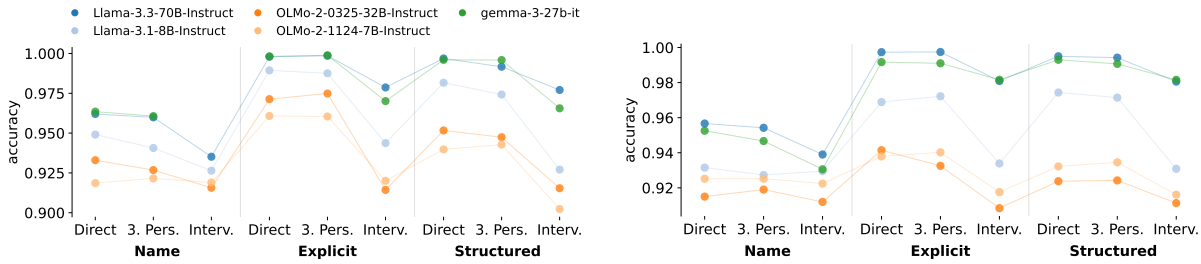
Influence of Prompt Phrasing. To verify that our findings stem from the introduced prompt dimensions rather than specific wording, we include two alternative phrasings for each prompt template



(a) accuracy in self-descriptions

(b) accuracy in social media bios

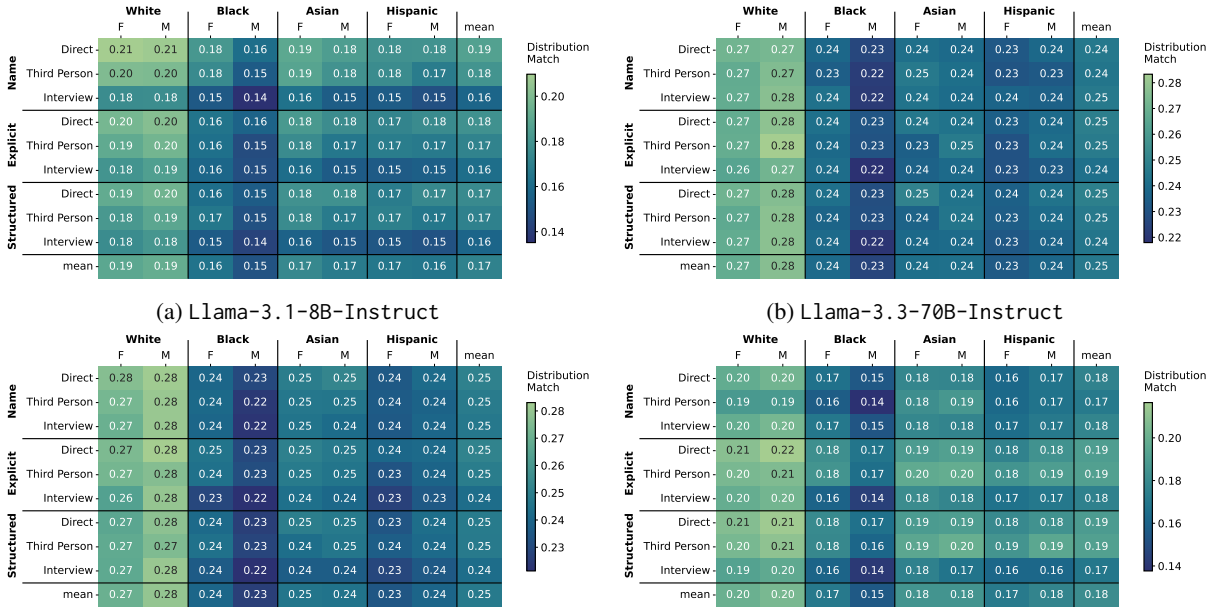
Figure 11: **Normalized classification accuracy across demographic groups (↓).** We present the classification accuracy across demographic groups. Values are averaged over all prompt types with lower values indicating lower distinguishability between demographic groups. In contrast to the other open-text evaluation measures, we find that personas with *male* gender and *White* race/ethnicity are easiest to classify in both tasks, while *nonbinary* personas are hardest to detect. However, we note that the absolute accuracy for all groups is generally high (> 0.9), indicating that responses of all groups are relatively easy to classify.



(a) accuracy in self-descriptions

(b) accuracy in social media bios

Figure 12: **Classification accuracy across prompt types and models (↓).** We present the classification accuracy across different prompt types. Values are averaged over all demographic groups with lower values indicating lower distinguishability between demographic groups. We observe a consistent pattern: prompts using *names* and the *interview* format result in the lowest (i.e., best) accuracy on average for all models across both tasks. We further find that responses from Gemma-3-27b and Llama-3.3-70B consistently lead to higher accuracy on average across both tasks, while OLMo-2-32B and OLMo-2-7B yield the lowest.



(a) `Llama-3.1-8B-Instruct`

(b) `Llama-3.3-70B-Instruct`

(c) `gemma-3-27b-it`

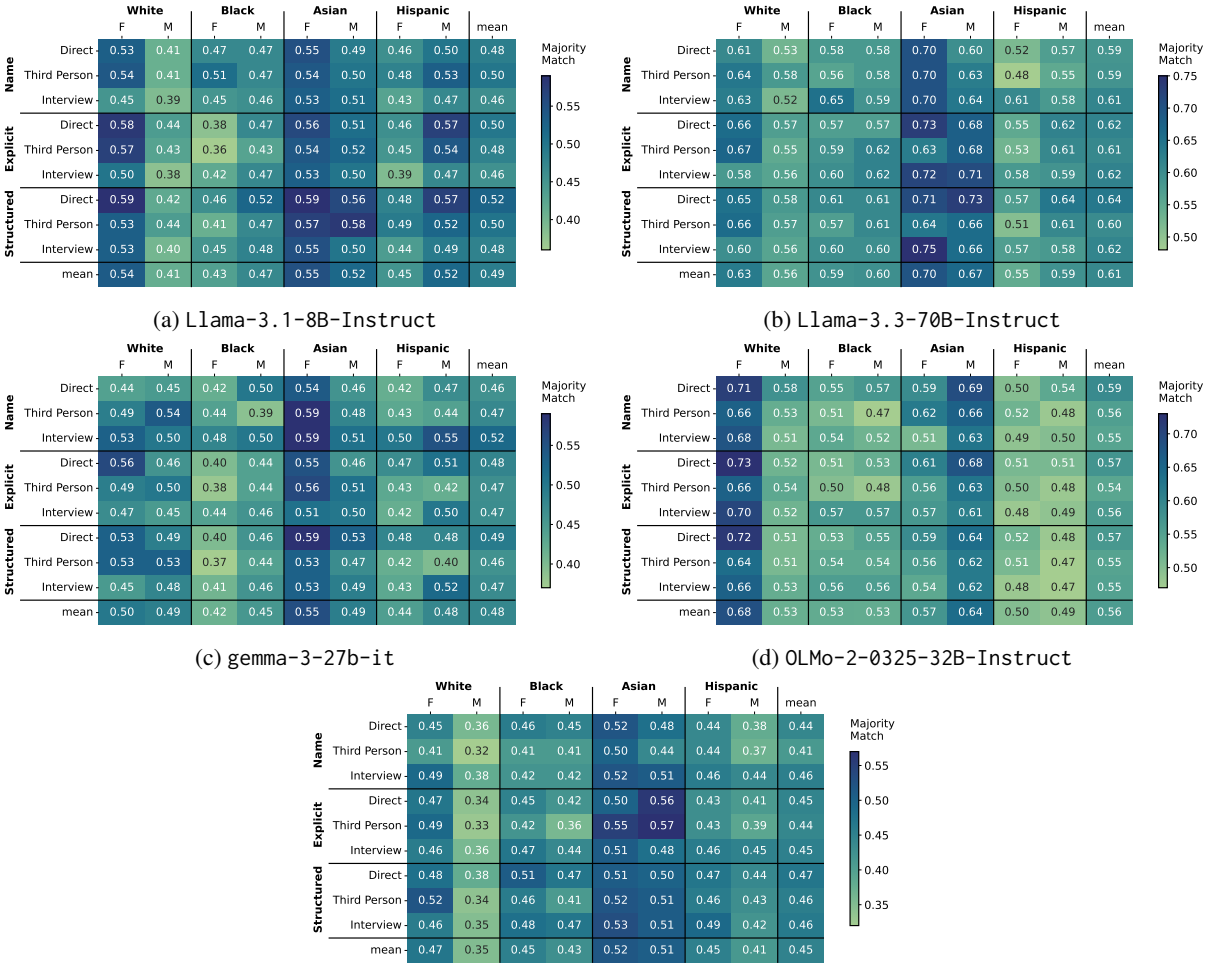
(d) `OLMo-2-0325-32B-Instruct`

Figure 13: **Opinion distance on OpinionsQA for all models (↓).** We show the opinion distance (as measured by Wasserstein distance) across demographic groups and prompt types (lower is better). We observe that all models generally align better with the answer distribution of people with race/ethnicity *Black*, while the highest opinion distance can be observed for people with race/ethnicity *White*. We further observe that the average opinion distance for Llama-3.3-70B and Gemma-3-27b is higher (i.e., worse) or equal to the random baseline (0.25 ± 0.002) for most prompt types and demographic groups.

(cf. Table 3) and assess the impact of prompt phrasing on the results. We find that regression coefficients for prompt phrasing are significant in only 1 out of 5 models for the marked word count and the share of non-English responses (cf. Tables 9, 11), in 2 out of 5 models for opinion distance (cf. Table 13), and in 3 out of 5 models for semantic diversity and accuracy (cf. Tables 10, 12). Across all measures and models, the influence (i.e., the absolute value of the regression coefficient) of using *names* and the *interview* format consistently exceeds that of the prompt phrasing, indicating that our main findings are not driven by incidental wording. We further note, that the prompt phrasing shows the least effect on Llama-3.2-70B (no significant coefficients across all measures), while OLMo2-7B is most sensitive to prompt phrasing (significant coefficients for 4 out of 5 measures).

F.2 Validation of Log Probabilities for OpinionsQA

To evaluate whether LLM log probabilities correspond to the model's actual answer selection behavior, we compare them to answer distributions obtained from multiple generations, where the model selects an option in free-form text. Using OLMo-7B, the best-performing model on OpinionsQA, we prompt each question 100 times with different random seeds and record the answer option chosen in each run. This produces an empirical distribution of answer options, which we compare to the distribution inferred from the model's log probabilities. We observe a very low Wasserstein distance (0.034 ± 0.025) between the log probability distribution and the aggregated answer distribution across 100 runs. This indicates that, for our use case, log probabilities serve as a reliable proxy for the model's actual response behavior



(a) `Llama-3.1-8B-Instruct`

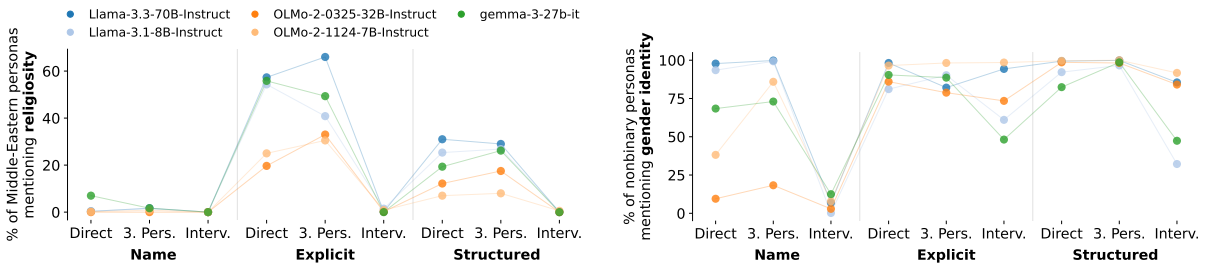
(b) `Llama-3.3-70B-Instruct`

(c) `gemma-3-27b-it`

(d) `OLMo-2-0325-32B-Instruct`

(e) `OLMo-2-1124-7B-Instruct`

Figure 14: **Majority match in OpinionsQA for all models (↑).** We show the average majority match across demographic groups and prompt types (higher is better). We observe that most models generally align better with the responses of people with race/ethicity *Asian* and with *White females*. We further observe that the average majority match for all models is significantly higher (i.e., better) than the random baseline (0.24 ± 0.026) for most prompt types and demographic groups.



(a) conflation of identity with religion

(b) disproportionate focus on gender identity

Figure 15: **Share of stereotyped persona descriptions across prompt types (↓).** We show the share of self-descriptions by (a) *Middle-Eastern* personas that contains words linked to religiosity (e.g., faith, muslim) and (b) *nonbinary* personas containing terms focusing on gender identity (e.g., gender, identity). We find that both using *names* and the *interview* format reduces the share of self-descriptions reflecting these stereotype categories.

	Llama-3.3-70B	Llama-3.1-8B	OLMo-2-32B	OLMo-2-7B	Gemma-3-27b
Name	−14.222***	−6.188***	−5.639***	−4.461***	−18.915***
Structured	−6.306***	−2.111***	−4.300***	−2.922***	−5.089*
Interview	−11.622***	−6.099***	−3.728***	−3.500***	−15.264***
Third Person	−1.472	1.467**	−0.061	−1.017*	2.683
Female	−0.900	0.367	−0.089	0.450	0.374
Nonbinary	0.272	0.901	0.717	2.550***	0.043
Asian	−0.620	0.304	0.111	−1.074	−2.119
Black	1.694	1.369	0.935	0.324	1.245
Hispanic	6.250***	3.027***	6.917***	1.102	14.220***
Middle-Eastern	4.056**	3.101***	2.352*	1.148	4.578
Self-Description	10.333***	5.041***	4.604***	4.448***	16.398***
Prompt Phrasing v2	1.089	0.441	0.974	1.478***	1.840
Intercept	15.548***	5.037***	2.513*	2.531***	14.448***

Table 9: **Regression on the marked word count.** We conduct OLS regression analyses per LLM using the number of marked words (↓) as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), gender (reference: male), race (reference: White), task (reference: Bio), and prompt phrasing (reference: v1). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Llama-3.3-70B	Llama-3.1-8B	OLMo-2-32B	OLMo-2-7B	Gemma-3-27b
Name	0.068***	0.021***	0.034***	0.036***	0.068***
Structured	−0.006	−0.009	0.002	0.007*	−0.001
Interview	0.009	0.041***	0.030***	0.018***	0.006
Third Person	0.004	−0.009*	−0.006*	−0.011**	−0.002
Female	−0.006	−0.020***	−0.012***	−0.011**	−0.015**
Nonbinary	−0.006	−0.047***	−0.023***	−0.036***	−0.017**
Asian	0.004	−0.000	−0.011***	−0.023***	0.006
Black	0.004	−0.005	−0.017***	−0.020***	0.006
Hispanic	−0.026**	−0.027***	−0.025***	−0.018***	−0.016*
Middle-Eastern	−0.004	−0.016**	−0.024***	−0.032***	−0.013*
Self-Description	−0.057***	−0.101***	−0.128***	−0.121***	−0.052***
Prompt Phrasing v2	0.006	0.010**	0.003	0.007*	0.010*
Intercept	0.300***	0.459***	0.486***	0.506***	0.354***

Table 10: **Regression on semantic diversity.** We conduct OLS regression analyses per LLM using semantic diversity (↑) as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), gender (reference: male), race (reference: White), task (reference: Bio), and prompt phrasing (reference: v1). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Llama-3.3-70B	Llama-3.1-8B	OLMo-2-32B	OLMo-2-7B	Gemma-3-27b
Name	−0.048***	−0.027***	−0.082***	−0.035***	−0.040***
Structured	−0.040***	−0.018***	−0.081***	−0.017*	−0.024*
Interview	−0.013	0.000	−0.058***	−0.033***	−0.027**
Third Person	0.023*	−0.001	−0.015	−0.009	0.004
Female	−0.026*	−0.003	−0.006	0.003	−0.007
Nonbinary	−0.020	−0.012*	−0.015	−0.013	−0.011
Asian	−0.001	0.001	0.003	−0.000	0.001
Black	−0.000	−0.000	0.006	0.002	−0.000
Hispanic	0.097***	0.053***	0.144***	0.100***	0.094***
Middle-Eastern	−0.001	−0.002	0.021	−0.001	0.001
Self-Description	−0.002	−0.012**	0.030**	−0.035***	0.024**
Prompt Phrasing v2	0.011	0.004	0.023*	0.012	−0.000
Intercept	0.038*	0.028***	0.061**	0.053***	0.023

Table 11: **Regression on the share of non-English responses.** We conduct OLS regression analyses per LLM, using the share of non-English responses (↓) as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), gender (reference: male), race (reference: White), task (reference: Bio), and prompt phrasing (reference: v1). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Llama-3.3-70B	Llama-3.1-8B	OLMo-2-32B	OLMo-2-7B	Gemma-3-27b
Name	−0.040***	−0.036***	−0.017***	−0.013***	−0.042***
Structured	−0.001	−0.007***	−0.011***	−0.009***	−0.002
Interview	−0.017***	−0.032***	−0.024***	−0.019***	−0.022***
Third Person	−0.001	0.002	−0.001	−0.001	−0.001
Female	−0.003*	−0.004*	−0.004*	−0.003	−0.005*
Nonbinary	−0.008***	−0.010***	−0.003	−0.004*	−0.010***
Asian	−0.009***	−0.010***	−0.017***	−0.017***	−0.008**
Black	−0.005***	−0.012***	−0.008***	−0.010***	−0.010***
Hispanic	−0.001	−0.006*	−0.003	−0.011***	−0.005
Middle-Eastern	−0.007***	−0.012***	−0.015***	−0.011***	−0.007**
Self-Description	0.001	0.011***	0.023***	0.015***	0.002
Prompt Phrasing v2	0.002	−0.003	0.006***	0.003*	0.003*
Intercept	1.005***	0.982***	0.942***	0.945***	1.005***

Table 12: **Regression on accuracy.** We conduct OLS regression analyses per LLM, using classification accuracy (↓) as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), gender (reference: male), race (reference: White), task (reference: Bio), and prompt phrasing (reference: v1). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Llama-3.3-70B	Llama-3.1-8B	OLMo-2-32B	OLMo-2-7B	Gemma-3-27b
Name	0.000	0.005***	−0.010***	0.003	0.001
Structured	0.001	−0.004***	−0.002	0.001	−0.001
Interview	−0.002	−0.019***	−0.013***	−0.009***	−0.006***
Third Person	−0.001	−0.005***	0.001	0.007***	−0.001
Female	0.001	0.004***	0.002	0.001	0.000
Asian	−0.031***	−0.020***	−0.017***	−0.016***	−0.029***
Black	−0.041***	−0.035***	−0.038***	−0.030***	−0.040***
Hispanic	−0.038***	−0.026***	−0.028***	−0.025***	−0.037***
Prompt Phrasing v2	0.000	0.001	−0.006***	−0.003**	−0.001
Intercept	0.273***	0.196***	0.209***	0.151***	0.276***

Table 13: **Regression on opinion distance.** We conduct OLS regression analyses per LLM, using opinion distance (as measured by Wasserstein distance) (↓) as a dependent variable and report the regression coefficients. The independent variables include: demographic priming (reference: explicit), role adoption (reference: direct), gender (reference: male), race (reference: White), and prompt phrasing (reference: v1). We note that the opinion distance of Llama-3.3-70B and Gemma-3-27b was worse than that of a random baseline. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.