# A Tale of Two Identities: An Ethical Audit of AI-Crafted Synthetic Personas

**Pranav Narayanan Venkit**[1,2], **Jiayi Li**[1], **Yingfan Zhou**[1],
**Sarah Rajtmajer**[1], **Shomir Wilson**[1]

[1]Pennsylvania State University, University Park, Pennsylvania, USA
[2]Salesforce AI Research, Palo Alto, USA
{pranav.venkit, jpl6207, yxz5975, smr48, shomir}@psu.edu

## Abstract

As LLMs (large language models) are increasingly used to generate synthetic personas, particularly in data-limited domains such as health, privacy, and HCI, it becomes necessary to understand how these narratives represent identity, especially that of minority communities. In this paper, we audit synthetic personas generated by 3 LLMs (GPT4o, Gemini 1.5 Pro, Deepseek v2.5) through the lens of representational harm, focusing specifically on racial identity. Using a mixed-methods approach combining close reading, lexical analysis, and a parameterized creativity framework, we compare 1,512 LLM-generated persona to human-authored responses. Our findings reveal that LLMs disproportionately foreground racial markers, overproduce culturally coded language, and construct personas that are syntactically elaborate yet narratively reductive. These patterns result in a range of sociotechnical harms, including stereotyping, exoticism, erasure, and benevolent bias, that are often obfuscated by superficially positive narrations. We formalize this phenomenon as algorithmic othering, where minoritized identities are rendered hypervisible but less authentic.

## Introduction

The expansion of LLMs as sociotechnical systems has transformed their usage in various spaces, including news and media, medicine and education (Sartori and Theodorou 2022; Narayanan Venkit 2023). Among their emerging uses is the automated generation of *personas*, fictional but realistic representations of user identities. Originally developed within human-centered design to foster empathy and represent end-users in system development, personas are now frequently produced by LLMs at scale to support applications such as chatbot character design (Kim et al. 2023; Dam et al. 2024), data augmentation (Whitehouse, Choudhury, and Aji 2023; Ding et al. 2024), and even healthcare simulation and social science research (BN et al. 2025; Biswas 2023; Yukhymenko et al. 2024; Manning, Zhu, and Horton 2024; Lehr et al. 2024).

Prior work has extensively documented biases in LLMs' classification (Ntoutsi et al. 2020; Gupta et al. 2024) and generation behaviors (Ferrara 2023; Ghosh and Caliskan 2023), yet their representational harms in identity performance remain understudied. This gap is particularly concerning given personas' dual role as design tools and increasingly social actors (Lehr et al. 2024; Biswas 2023; Barambones et al. 2024). Recent studies suggest LLM-generated text exhibit troubling patterns: overemphasizing trauma narratives for minority identities or flattening cultural complexity into reductive tropes (Cheng, Durmus, and Jurafsky 2023; Lazik et al. 2025; Haxvig 2024). These behaviors risk replicating historical harms in ethnographic research and sociotechnical cases, now automated at scale.

In this work, using the scope of representational harm identification and creativity analysis, we present a study of how LLMs construct identity through persona generation, with particular focus on representation within United States sociodemographic contexts. Through examination of 756 human-written self-descriptions (provided by 126 participants) alongside 1,512 LLM-generated personas from three leading models (GPT4o, Gemini, and DeepSeek), we address three research questions:

- How do LLM personas differ systematically from human self-descriptions in their identity representation patterns?
- What forms of representational bias emerge when LLMs simulate marginalized identities?
- What automated indicators can be used to identify the harms of synthetic personas from self-descriptions?

To answer these questions, we develop a multi-method framework rooted in computational sociolinguistics and HCI. First, we analyze the language of identity performance using TFIDF and log-odds ratio techniques to identify *markedness*, which depicts the linguistic and social differences between the unmarked default or majority group and marked groups that differ from the default (Wolfe and Caliskan 2022; Ghosh and Caliskan 2023). Building upon our linguistic analyses, we employ a creativity-based framework, where we parameterize persona expressiveness along four dimensions: semantic diversity, novelty, surprisal, and complexity, to quantify not only *what* is said but *how* narrations are told. We examine how LLM-generated personas differ not only from human counterparts but also how representations vary across racial groups when contrasted against the model's apparent 'default' persona outputs (typically reflecting dominant cultural norms). Drawing from prior work

on algorithmic stereotyping (Cheng, Durmus, and Jurafsky 2023; Kambhatla, Stewart, and Mihalcea 2022), we show how LLMs foreground demographic attributes like race even when other personal attributes are present.

Our findings show that LLM-generated personas consistently foreground racial identity through stylized but reductive narratives. Across all personas, LLMs disproportionately produce racially coded terms (e.g., *"heritage," "resilience," "abuela"*) that are absent from human-authored responses. Sentiment analysis reveals inflated positivity in LLM outputs (e.g., RoBERTa avg. 0.81 for Hispanic personas vs. 0.48 human-authored), masking stereotypical tropes under benevolent framing. Creativity metrics further show that LLMs generate text with higher syntactic complexity (1.717 vs. 0.656) but lower diversity (1.713 vs. 1.910) and surprisal (1.267 vs. 1.276). We term this pattern *algorithmic othering*, where minoritized identities are rendered hypervisible yet narratively 'hollow'.

## Related Work

### Sociotechnical Nature and Effects of AI

As AI systems become embedded in decision-making infrastructures within varying positions in society, their role as sociotechnical systems, systems shaped by both technological architecture and social context, has become increasingly visible and consequential (Dolata, Feuerriegel, and Schwabe 2022; Cooper and Foster 1971). The technical affordances and design choices of such systems do not operate in isolation; rather, they entangle with cultural norms, historical power dynamics, and social imaginaries (Bender et al. 2021; Gautam, Venkit, and Ghosh 2024; Venkit et al. 2025). This entanglement can result in disproportionate harms, particularly when AI systems are applied to socially situated tasks such as recidivism prediction, emotion detection, or language generation, where model behavior may reinforce structural inequities (Bender et al. 2021; Dev et al. 2022).

Recent work in HCI and critical algorithm studies has emphasized the need to move beyond performance metrics and interrogate the ways in which AI systems produce representational harms, harms that distort how individuals or groups are portrayed, perceived, or included in algorithmically mediated spaces (Blodgett et al. 2022; Ghosh et al. 2024; Qadri et al. 2023). These harms include stereotyping, erasure, and exoticism, often exacerbated by the decontextualized nature of large-scale training data and opaque generative processes. These issues become especially salient when models are tasked with simulating human identity: the production of synthetic personas by LLMs, an ostensibly neutral task, can carry embedded cultural priors and normative biases that shape how identity is expressed. Despite these growing concerns, there remains limited research investigating how such harms manifest in specific use cases like persona generation, particularly when synthetic outputs are used as stand-ins for real user data in sensitive domains such as healthcare, education, and policy.

### Ethical Issues in LLM-Generated Personas

Personas, representations of individuals based on personal, social, and contextual attributes (Jung et al. 2017; Cheng, Durmus, and Jurafsky 2023; Prpa et al. 2024), were traditionally crafted through manual qualitative methods, a process with significant limitations in scalability and diversity (Salminen, Jung, and Jansen 2019; Salminen et al. 2020). Recent advances in generative AI have positioned LLMs as a powerful tool for simulating complex identities across various domains (Amin et al. 2025; Park et al. 2023, 2024). Schuller et al. (2024) explored the use of LLMs to streamline persona creation for UX design, showing that LLM-generated personas are perceived as comparable in quality and acceptance to those written by human experts when crafted with strategic prompts. Shin et al. (2024) observed that while there are limitations in LLMs' ability to capture key user characteristics, the use of hybrid human-AI workflows for persona generation promotes more representative and empathy-evoking personas.

While bias in general GenAI applications has received considerable attention (Venkit et al. 2023; Wan et al. 2023; Gupta et al. 2023), limited research specifically interrogates the ethical issues embedded in LLM-generated personas. Sethi et al. (2025) found no significant linguistic bias in LLM-generated persona descriptions and suggests that LLMs can generate lexically diverse persona descriptions. In contrast, Salminen et al. (2024) investigated diversity and bias in LLM-generated personas and found evidence of bias towards age, occupation, and pain points, though these personas were perceived as informative, believable, and relatable by human evaluators. Lee, Montgomery, and Lai (2024) analyzed ChatGPT-generated portrayals of intersectional group identities and observed a tendency for LLMs to describe socially subordinate groups as more homogeneous than dominant groups, thereby reinforcing stereotypes.

These existing studies have primarily focused on specific aspects of ethical issues in LLM-generated personas and largely overlook how synthetic personas differ from human-authored self-descriptions in the ways they express identity, and what representational harms may emerge from those differences. In this work, we shift the focus from generation outputs to a systematic analysis of how LLM identity through persona generation by directly comparing human-written self-descriptions with matched LLM personas.

## Methodology

We examine how LLMs construct identity through synthetic personas, with particular attention to representational discriminations in *marginalized U.S. demographic groups*. We aimed to systematically compare LLM-generated personas to human-authored self-descriptions, assessing patterns across lexical, narrative, and identity-related dimensions. Through this work, we conceptualize personas as *dynamic, performative articulations of identity rather than static archetypes*. Drawing from prior works of persona construction (Prpa et al. 2024; Qin et al. 2024), we treat identity as emergent, constructed through narrative, shaped by social context, and reflective of intersecting positionalities.

## Self-Descriptive Data Collection

We established a ground truth dataset of authentic identity expressions through a survey of 141 participants recruited via Prolific, stratified to reflect U.S. demographic diversity in race and gender. Participants responded to six thematically designed open-ended questions that collectively captured variable aspects of persona construction, including personal values, daily experiences, and aspirational identity. They are as follows:

- Please describe yourself.
- What are your aspirations for your personal life?
- What are your most defining traits or qualities?
- Please describe your average day.
- What core values guide your decisions?
- What skills do you excel at, and how do you use them?

These questions were adapted from Kambhatla, Stewart, and Mihalcea (2022); Cheng, Durmus, and Jurafsky (2023) on stereotype analysis in self-presentation and refined through iterative discussion among authors to ensure comprehensive coverage of identity facets while avoiding priming effects. The survey introduction deliberately framed the study as exploring "how individuals express their identities through self-description", with no reference made to AI or persona generation, to elicit natural responses rather than performative or artificial narratives. We intended to capture how individuals naturally perform their identities when invited to reflect on their values, routines, and aspirations. Each participant was instructed to write at least 500 words per question, encouraging elaboration in their self-descriptive narratives. We collected sociodemographic metadata alongside textual responses to enable persona replication. The survey took up to 30 minutes to complete, and participants were compensated $5 for their participation. The study and the human collection procedure received *Exempt status by Institutional Review Board (IRB) review*. The complete survey and the questions are present in Appendix.

Given our study's goal of comparing human-authored and LLM-generated personas, it was important to ensure that all participant responses were authentically written by humans. To mitigate the risk of AI-generated or copy-pasted responses, we employed a data quality assurance protocol that included preventative deterrents, based on best practices from prior work (Zhang, Xu, and Alvero 2024; Christoforou, Demartini, and Otterbacher 2024), and a final automatic validation step, as detailed in the Appendix. After removing flagged responses from an initial collection of 846 responses, our final corpus comprised **126 verified human participants** yielding **756 authentic self-descriptions** (6 responses × 126 participants).

An essential component of our data collection strategy was to ensure diverse and balanced representation across race and gender. Drawing on demographic categories defined by the U.S. Census[1], we intentionally recruited participants to achieve equitable distribution across racial and gender groups. Our recruitment protocol targeted seven racial/ethnic groups[2]. The final sample comprised of intentional oversampling of typically underrepresented groups (e.g., Native populations) to ensure adequate subgroup analysis power.

The final sample's racial/ethnic distribution was: African American/Black ($n = 28$), American Indian/Alaskan Native ($n = 22$), Asian ($n = 26$), Hispanic/Latino ($n = 21$), White ($n = 23$), and Multiracial/Other ($n = 6$). With respect to gender, we obtained a balanced representation across binary categories while including diverse gender identities *(Female:* $n = 63$ *(50.4%), Male:* $n = 61$ *(48.8%), Non-Binary:* $n = 2$ *(0.1%))*.

## Generation of AI Personas

To examine how LLMs construct identity representations in synthetic personas, we designed a comparative study using matched human-authored and model-generated texts.

We then prompted three LLMs—**GPT-4o**, **Gemini 1.5 Pro**, and **DeepSeek v2.5**—to generate synthetic counterparts for these human-authored personas. These models were selected due to their broad public availability, strong performance on general-purpose tasks, and increasing deployment in sociotechnical applications such as conversational agents and data augmentation systems (Sartori and Theodorou 2022; Sun, Zhan, and Such 2024).

To mirror the human study setup, we issued API calls to each model using the same six self-description questions. Prompts were framed in the second person (e.g., "You are a 32-year-old Black woman...") to induce the model to write as a fictional persona. We tested *four prompting conditions* that varied the demographic information provided to the model: **Race only; Race and age; Race, gender and age; full sociodemographic profile** *(race, age, gender, occupation, nationality, and relationship status)* We chose these settings to evaluate how different levels of demographic detail shape identity construction in LLM-generated personas. Prior work shows that even minimal cues can activate stereotypical outputs (Cheng, Durmus, and Jurafsky 2023; Gupta et al. 2023). Since our analysis revealed that race remains the dominant narrative anchor even with full prompts, we center race in this study. These settings allow us to audit how racial othering persists across varied input conditions while also reflecting real-world scenarios where user data may be partial. The sociodemographic factors were selected mirroring works that demonstrate usage of synthetic personas (Staab et al. 2024; Yukhymenko et al. 2024).

The prompt design and strategy were inspired by (Staab et al. 2024) and (Cheng, Durmus, and Jurafsky 2023) which had used similar method to generate synthetic personas and responses. All prompts used a default temperature of **1.0**, which according to model documentation, balances determinism and creativity (Peeperkorn et al. 2024). The latest API calls were made on *May 2025* to obtain the latest responses from 3 chosen LLM models.

[1] https://www.census.gov/topics/population/race/about.html

[2] African American/Black, American Indian/Alaskan Native, White, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, and multiracial/other identities

| Race | LLM Marked Words | Human Marked Words |
|---|---|---|
| **African American / Black** | **african**, *community*, life, american, **atlanta**, resilience, old, **black**, personal, *vibrant*, family, work, woman, young, stories | people, like, work, life, want, time, try, make, day, person, love, good, family, use, things |
| **American Indian / Alaska Native** | *community*, *heritage*, life, *generation*, **indian**, *ancestors*, *traditions*, *cultural*, stories, **american**, **values**, day, *wisdom*, family, *respect* | people, like, time, try, work, make, want, help, day, good, life, just, things, love, able |
| **Asian** | **asian**, life, *cultural*, *heritage*, personal, family, *diverse*, living, *values*, *community*, work, old, year, **american**, *vibrant* | work, like, life, good, want, day, person, people, time, make, new, family, love, home, try |
| **Hispanic / Latino** | family, **hispanic**, *community*, life, **latino**, *heritage*, *vibrant*, *cultural*, **abuela**, day, living, work, **latina**, *roots*, *values*, old | people, like, work, day, want, life, just, things, love, time, say, good, try, new, family |
| **Multiracial / Other** | life, personal, *diverse*, *community*, *tapestry*, *identity*, **aloha**, **native**, *heritage*, empathy, old, living, experience, *vibrant*, **different** | work, kids, day, time, life, family, good, day, job, doing, things, make, children, home, like |
| **White** | life, *community*, day, work, personal, family, old, year, *local*, living, **values**, skills, new, time, love | like, work, life, people, want, make, time, day, things, home, good, think, try |

Table 1: Top 15 significant words in LLM-generated and human-authored personas identified using TF-IDF analysis, aggregated across all prompt settings. Words in **bold** denote racially coded terms suggesting elevated attention to racial features; *italicized* words (if any) reflect culturally coded associations; red words indicate adversity-oriented narration patterns.

In total, each for six responses across the four prompting conditions and three LLMs, we generated **1512 synthetic personas**. This yielded **9072 model-generated texts**, which we then analyzed alongside the 756 human-authored self-descriptions. The full prompt templates used for generation are included in the **Appendix**.

## Analysis of Identity Representation in Personas

Our analysis first concerns a critical question: How do LLM-generated personas systematically differ from human-authored self-descriptions in their identity representation? Our analysis reveals key divergences in how human and LLMs construct identity, particularly for individuals from marginalized groups.

### Exploration of Thematic and Lexical Divergence

Close reading of our corpora revealed a key divergence: LLM personas tend to overemphasize demographic markers, especially racial identity, often foregrounding stereotypical or reductive narratives, while human-authored responses tend to emphasize personal values, social relationships, or life experiences (**Appendix** for examples).

To systematically validate this observation, we conducted a term frequency–inverse document frequency (TF-IDF) analysis over the entire corpus, stratified by racial identity and generation source (LLM vs. human). This TF-IDF analysis surfaces the most distinctive lexical features associated with specific identity groups, helping us identify which terms disproportionately define LLM-generated versus human-authored personas.

Across all three models, and prompting settings combined, we find that the top-weighted terms in synthetic responses are disproportionately anchored in racially-coded features. For instance, terms such as *diverse, vibrant, heritage, roots* and similar racial identifiers (like atlanta for Black and African American) surface with high TF-IDF weights, regardless of whether race was the only demographic provided or part of a fuller profile.

By contrast, human-authored responses showed a markedly different lexical profile. Common TF-IDF terms included *people, work, life* and *like*, highlighting relational and experiential facets of identity rather than categorical labels. Importantly, these terms appeared *across racial groups, indicating a more universally grounded and less racialized framing of identity*. This divergence shows the model's tendency to amplify perceived difference, often at the expense of authenticity or nuance.

### Analysis of *Algorithmic Othering* in Minority Narratives

Building on our lexical analyses using TF-IDF, we apply the framework of *markedness*, drawing from sociolinguistic theory (Eckert 2011; Wolfe and Caliskan 2022) to investigate representational bias emerge when LLMs simulate marginalized identities. In this view, language used to describe minoritized groups often contains distinctive lexical signals, *marked words*, that index deviation from an assumed normative baseline (e.g., white, Western, able-bodied). These markers are not merely descriptive but ideologically loaded: they signify *otherness*.We adopt a computational approach to identify such markers using log-odds ratio analysis with informative Dirichlet priors (Monroe, Colaresi, and Quinn 2008; Cheng, Durmus, and Jurafsky 2023), which quantifies how significantly certain words distinguish a target group from a reference group, while controlling for background word frequency. Formally, let $w$ denote a word, $c_1(w)$ and $c_2(w)$ be its count in the target and reference corpora respectively, and $p(w)$ the word's count in the full corpus. We calculate the smoothed log-odds ratio $\delta_w$ as:

$$\delta_w = \frac{\log\left(\frac{c_1(w)+p(w)}{N_1-c_1(w)+P-p(w)}\right) - \log\left(\frac{c_2(w)+p(w)}{N_2-c_2(w)+P-p(w)}\right)}{\sqrt{\frac{1}{c_1(w)+p(w)} + \frac{1}{c_2(w)+p(w)}}}$$

where $N_1$ and $N_2$ are the total word counts in the target and reference corpora, and $P$ is the total count in the prior. Words with $|\delta_w| > 1.96$ are considered statistically distinctive at a 95% confidence threshold.

We applied this method across five racial groups, *African American or Black*, *Asian*, *American Indian or Alaska Native*, *Hispanic or Latino*, using *White* personas as the default group. As shown in in Table 2, LLM-generated personas for

minoritized groups frequently contain racially salient or culturally coded, even when full sociodemographic profiles are provided. This suggests a model-level overreliance on racial identity as the primary narrative.

In contrast, White-coded personas feature fewer marked terms, relying on broad, neutral descriptors like *town, work, hiking*, aligning with unmarked social norms. Human-authored responses across all races, meanwhile, converge on vocabulary rooted in everyday life and relational experience, frequent terms include *kids, husband, work*, all of which are notably downplayed in LLM outputs for marginalized groups. The full breakdown of marked words for each model and results from four different prompt settings are provided in **Appendix**. We interpret these findings as evidence of *algorithmic othering*. As defined by (Ehlebracht 2019), othering occurs when individuals are mentally classified as "not one of us," reducing their complexity to a simplified, often dehumanized identity.

### Obfuscation through Positive Narratives

Our analysis reveals how LLMs obscure racial markedness through affectively positive but narratively reductive language. This rhetorical framing, which uses terms like *resilient, vibrant,* and *diverse*, makes underlying stereotypes difficult to detect and aligns with prior findings on positive yet stereotypical portrayals of minority communities Cheng, Durmus, and Jurafsky (2023). To quantify this, we performed sentiment analysis on both LLM-generated and human-authored persona responses using two models: VADER (Hutto and Gilbert 2014), a rule-based sentiment classifier, and RoBERTa (Liu et al. 2019), a transformer-based neural model. Each response was scored for sentiment polarity (-1 to +1), with group-level averages computed across racial identities.

As shown in (**Appendix**), LLM-generated personas consistently receive higher positive sentiment scores than human-authored ones across all racial groups, a disparity most pronounced for minoritized identities such as Hispanic/Latino. These findings reveal a form of representational inflation, where LLMs produce uniformly positive portrayals that mask underlying stereotyping, which we term *benevolent bias*. While flattering, these portrayals rely on tropes like strength through struggle or cultural pride, reducing individual experiences to narratives of racial resilience. This aligns with positive stereotyping (Czopp, Kay, and Cheryan 2015; Kay et al. 2013), where admiration still confines identity to pre-scripted roles, reinforcing stereotypes.

## Parameterization of Creativity in Synthetic and Human Persona

### Quantifying the Creativity Framework

Our prior section primarily evaluate content at the lexical themes or emotional valence level. To understand the synthetic identity construction we extend our analysis to the *structural and stylistic qualities* of narrative expression. Specifically, we ask: beyond *what* is conveyed, *how* do LLMs differ from humans in the way they tell stories?

To address this, we draw on recent advances in computational creativity evaluation (Chakrabarty et al. 2024; Ismayilzada, Stevenson, and van der Plas 2024), which provide scalable frameworks for assessing narrative quality. Traditional creativity assessments have depended on human or expert raters evaluating factors such as originality, coherence, or emotional resonance (Chakrabarty et al. 2024). However, emerging parameterized approaches allow for more systematic, replicable benchmarking of large corpora. Creativity in these frameworks largely follows Torrance's dimensions of *fluency, flexibility, originality,* and *elaboration* to define creativity used in Torrance Tests of Creative Thinking (TTCT) (Torrance 1966). Building on this foundation, we adopt a multifaceted view of creativity, operationalizing it along four core computational axes. Our framework, though inspired by Ismayilzada, Stevenson, and van der Plas (2024), takes into account group based creativity analysis, focusing on race as a primary factor. We analyze how each group differs from others and hence to the human written persona responses. The quantification is as follows:

- **Semantic Diversity**: Stemming from the *flexibility* component of TTCT, we assess the breadth of thematic variation within a group by computing the average pairwise semantic distance between all responses. Each narrative is embedded using a pretrained Sentence-BERT model, and diversity is quantified as:

$$\textbf{Diversity} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (1 - \cos(\mathbf{e}_i, \mathbf{e}_j))$$

  where $\cos(\mathbf{e}_i, \mathbf{e}_j)$ is the cosine similarity between sentence embeddings. Higher diversity indicates richer expression. Diversity measures the average semantic distance among all pairs of responses *within a group*.
- **Semantic Novelty**: Using the *originality* facet of TTCT, we measure how distinct a group's narratives are from the corpus norm, we calculate the deviation of a group's internal semantic distance from that of the overall corpus to measure semantic novelty:

$$\textbf{Novelty} = 2 \times |d_{\text{group}} - d_{\text{corpus}}|$$

  where $d_{\text{group}}$ and $d_{\text{corpus}}$ are the average pairwise semantic distances *within the group and corpus*, respectively. Higher novelty reflects greater deviation from learned or expected storytelling patterns.
- **Semantic Complexity**: We capture narrative sophistication, *fluency* of TTCT, through a composite measure combining lexical rarity and semantic spread. We first compute the average TF-IDF score to quantify term distinctiveness and then calculate the average pairwise cosine distance among word embeddings within each narrative using a Word2Vec model trained on the corpus. The combined semantic complexity for a story is given by:

$$\textbf{Complexity} = 0.5 \times \frac{\text{Cmplxity}_{\text{TFIDF}}(s)}{\max(\text{Cmplxity}_{\text{TFIDF}})} + 0.5 \times \frac{\text{Cmplxity}_{\text{W2V}}(s)}{\max(\text{Cmplxity}_{\text{W2V}})}$$

  where higher values indicate more intricate narratives.

| Race | LLM Marked Words | Human Marked Words |
|---|---|---|
| **African American / Black** | **black**, **african**, *atlanta*, community, resilience, *music*, young, justice, uplift, *jazz* | skills, home, before, large, analytical, generally, friends, career, tech, friendly |
| **American Indian / Alaska Native** | *ancestor*, *land*, *generations*, *traditions*, **indian**, *stories*, *passed*, *heritage*, *elders*, wisdom | learn, helping, business, world, festival, past, people, *native*, morals |
| **Asian** | **asian**, *cultural*, parents, *heritage*, **american**, *immigrant*, *blend*, *expectations*, *sanfrancisco*, *kimchi* | learning, goal, *math*, patient, believe, language, *anime*, student, games |
| **Hispanic / Latino** | **latino**, **hispanic**, *abuela*, *spanish*, **latina**, *salsa*, cultural, *roots*, *arroz*, *heritage* | technical, science, lord, god, start, design, hope, playing, learning, freedom |
| **White** | **white**, pretty, *midwest*, hiking, kids, honesty, town, husband, work, straightforward | we, husband, home, keep, big, southern, house, son, children, number |

Table 2: Top 10 marked words in LLM-generated personas for each racial group, identified using log-odds ratio analysis with White personas as the reference group. Words in **bold** directly denote racial or ethnic identity, *italicized* terms reflect culturally coded or identity-relevant associations, and words in red indicate adversity-oriented language associated with *trauma scripting*.

- **Surprisal (Semantic Entropy)**: We measure the unpredictability of narrative progression, to evaluate *elaboration* of TTCT, by calculating the semantic distance between consecutive sentences *within each story*.

$$\textbf{Surprisal} = \frac{2}{n-1}\sum_{i=2}^{n}(1-\cos(\mathbf{e}_{i-1},\mathbf{e}_i))$$

where $\cos(\mathbf{e}_{i-1},\mathbf{e}_i)$ denotes the semantic similarity between adjacent sentences. Higher surprisal suggests greater topic shifts and less formulaic sentence chaining.

## Analysis of Creativity in Narration

Table 3 shows the averaged scores across four LLM prompting conditions and compares to human-authored responses. Across all dimensions, we observe statistical differences between synthetic and human-authored narratives conducted via two-sample Welch's t-tests ($p < 0.001$).

**Surprisal:** Surprisal scores show modest but consistent differences. Human-authored responses exhibit the highest average surprisal (1.276), indicating greater unpredictability in narrative progression. LLM-generated personas, *while approaching human levels under full-profile prompting* (1.267), *remain marginally lower overall*. A t-test revealed that the difference was statistically significant ($t = -3.57, p < 0.001$), suggesting that LLM narratives, although fluent, remain slightly more formulaic.

**Semantic Diversity.** Diversity analysis reveals a larger gap. Human-authored narratives display substantially higher semantic diversity (1.910) compared to LLM outputs across all prompting conditions (range: 1.621–1.713), ($t = -6.24, p < 0.001$), indicating that humans draw on a wider thematic space when constructing self-narratives.

**Semantic Novelty.** In contrast, novelty scores were consistently higher for LLM narratives (0.453 to 0.314) than for human-authored ones (0.105), ($t = 7.89, p < 0.001$). However, closer qualitative analysis (Appendix) suggests that this apparent novelty arises from over-emphasis on racialized tropes rather than genuinely individualized storytelling. As novelty measures the difference in themes between race groups, we see that *LLMs tend to create narrations that are very dissimilar between groups*. However human experience are not intended to be that different just based on an individuals race, as seen by the low novelty scores.

**Semantic Complexity.** Semantic complexity shows the most pronounced divergence. LLM personas exhibit significantly greater complexity (1.690–1.720) than human-authored narratives (0.656), ($t = 9.12, p < 0.001$). While complexity is often seen as a marker of creativity, in this case it *reflects elaborative rather than authentic narrative depth*. This marker shows that LLMs tend to use unnecessarily complex words, in contrast to the simpler vocabulary found in human-written text (as shown in Table 1).

## Race-Based Creativity Patterns

To examine how different racial groups are differentially represented, we conducted a group-level analysis under the prompting condition, where only race information was provided. As shown in Table 4, LLM-generated personas exhibit distinct creativity patterns across racial groups, as compared to participant responses:

**Surprisal** varied moderately, with Hispanic/Latino (1.227) personas showing slightly higher surprisal than African American/Black (1.158), Asian (1.160), or Native Hawaiian or Other Pacific Islander (1.141) personas. While surprisal differences suggest greater narrative fragmentation for some minoritized identities, overall values remain close across groups. In comparison, human-authored narratives displayed consistently higher surprisal.

**Semantic diversity** was highest for White personas (1.684), while groups such as Asian (1.110) and American Indian or Alaska Native (1.101) exhibited substantially lower diversity.

**Novelty** scores were sharply racialized. Minoritized groups such as American Indian or Alaska Native (0.334) exhibited much higher novelty compared to White personas (0.060) By contrast, human-authored responses showed consistently low novelty across races.

**Semantic complexity** in LLM personas was consistently high across groups, with minoritized personas such as Hispanic/Latino (0.885) and Asian (0.873) slightly exceeding White (0.837). Human-authored narratives exhibited lower complexity across all races.

While LLM-generated personas produce linguistically sophisticated and seemingly novel narratives, these outputs remain less thematically diverse and authentic, and more stereotypically for minoritized groups.

| Metric | LLM(Race) | LLM(Race-Age) | LLM(Race-Age-Sex) | LLM(All) | Human |
|---|---|---|---|---|---|
| Surprisal | 1.165 | 1.237 | 1.242 | 1.267 | 1.276 |
| Diversity | 1.621 | 1.648 | 1.691 | 1.713 | 1.910 |
| Novelty | 0.453 | 0.388 | 0.337 | 0.314 | 0.105 |
| Complexity | 1.720 | 1.692 | 1.690 | 1.717 | 0.656 |

Table 3: Average creativity scores across all LLM-generated personas combined for all prompt conditions compared to human-authored personas.

| | Surprisal | | Diversity | | Novelty | | Complexity | |
|---|---|---|---|---|---|---|---|---|
| | LLM | Human | LLM | Human | LLM | Human | LLM | Human |
| **African American/Black** | 1.158 | 1.202 | **1.388** | 1.878 | 0.172 | 0.015 | 0.853 | 0.417 |
| **American Indian/Alaska Native** | 1.105 | 1.267 | 1.101 | 1.905 | **0.334** | 0.015 | 0.858 | 0.432 |
| **Asian** | 1.160 | 1.216 | 1.110 | 1.857 | **0.290** | 0.036 | 0.873 | 0.401 |
| **Hispanin/Latino** | **1.227** | 1.246 | 1.190 | 1.974 | 0.246 | 0.051 | 0.885 | 0.427 |
| **White** | 1.185 | 1.253 | **1.684** | 1.917 | **0.060** | 0.007 | 0.837 | 0.409 |

Table 4: Comparison of creativity metrics *[Surprisal, Diversity, Novelty, and Complexity]* across racial groups for all **LLM-generated** and **human-authored** personas under race-only prompting.

## Discussion

As LLMs are increasingly used to generate synthetic personas across domains such as healthcare, civic tech, and user research (BN et al. 2025; Whitehouse, Choudhury, and Aji 2023), our study reveals that these personas often foreground racial identity, even when provided with full sociodemographic profiles. Drawing on prior work on language model harms (Blodgett et al. 2020; Dev et al. 2022; Ghosh et al. 2024), we analyze these failures across sociotechnical harm.

**Stereotyping:** LLM personas feature racially marked and culturally coded terms (e.g., *heritage, salsa, resilience*) regardless of prompt richness. This results in *stereotyping through presence* (Noble 2018), where identity is reduced to racial symbols. Diversity metrics corroborate this flattening: within-group variation for minoritized personas is markedly lower than in human-authored texts.

**Disparagement:** Although explicit negativity is rare, LLMs encode adversity-centered narratives framed through positive affect; a form of *benevolent bias*. The frequent scripting of resilience, struggle, or cultural pride as defining characteristics renders minoritized identities primarily as sites of hardship and admiration.

**Dehumanization:** LLM outputs often omit mundane, relational, or contradictory aspects of lived experience, resulting in a narrative flattening. Reduced surprisal and semantic diversity point to an underlying omission of narrative depth, especially for marginalized groups, where personas become archetypes (e.g., the resilient survivor).

**Erasure:** Creativity analysis shows that LLMs produce narrower thematic ranges for minoritized groups. Compared to White personas or human-authored responses, synthetic outputs lack intra-group variability, signaling a form of representational erasure. These personas become not only stereotyped but interchangeable, undermining identity.

**Exoticism:** Marked word analyses reveal symbolic exoticism, where cultural markers (e.g., *aloha, abuela, jazz*) are stylized and aestheticized. These identities are narrated as colorful and distinct, but not ordinary. Novelty scores are inflated for minoritized groups, suggesting that the models treat difference as novelty rather than lived norm, amplifying a spectacle of otherness.

**Quality of Service:** Finally, narrative-level creativity metrics (e.g., complexity, surprisal, diversity) show that LLMs generate less expressive, less faithful personas for marginalized identities. This creates potential disparities in contexts where synthetic personas may influence system design, user experience, or data-driven policy.

As LLM-generated personas gain traction in high-stakes domains, we argue for stricter auditing, narrative-aware evaluation, and community-involved validation to mitigate harms, which is provided as a framework in the Appendix.

## Conclusion

As LLMs become embedded in sociotechnical systems, synthetic personas increasingly stand in for human narratives. Yet, our study reveals key representational gaps: LLM-generated personas consistently foreground racial identity, flatten lived experience, and follow formulaic storytelling. Analyzing 1,512 synthetic and 126 human-authored personas, we find that minoritized identities are disproportionately scripted through tropes of diversity, and adversity, a dynamic we call *algorithmic othering*. Our findings and accompanying recommendations therefore (**Appendix**)[3] show the risks of deploying synthetic personas. Our findings therefore call for a fundamental reframing of persona generation in AI systems: not as a task of realism or plausibility alone, but as a site of ethical design and cultural accountability. As generative models grow more capable, so too must our standards for how they simulate—and serve—the people they claim to represent.

## Acknowledgements

This work was partially supported by the National Science Foundation award #2247723.

[3]The complete Appendix of this work can be found in this repository[4]

## References


Amin, D.; Salminen, J.; Ahmed, F.; Tervola, S. M.; Sethi, S.; and Jansen, B. J. 2025. How Is Generative AI Used for Persona Development?: A Systematic Review of 52 Research Articles. *arXiv preprint arXiv:2504.04927*.

Barambones, J.; Moral, C.; de Antonio, A.; Imbert, R.; Martínez-Normand, L.; and Villalba-Mora, E. 2024. ChatGPT for learning HCI techniques: A case study on interviews for personas. *IEEE Transactions on Learning Technologies*, 17: 1460–1475.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Biswas, S. S. 2023. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5): 868–869.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Blodgett, S. L.; Liao, Q. V.; Olteanu, A.; Mihalcea, R.; Muller, M.; Scheuerman, M. K.; Tan, C.; and Yang, Q. 2022. Responsible language technologies: Foreseeing and mitigating harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–3.

BN, S.; Mattioli, D.; Abdullah, S.; Arriaga, R. I.; Wiese, C. W.; and Sherrill, A. M. 2025. Thousand Voices of Trauma: A Large-Scale Synthetic Dataset for Modeling Prolonged Exposure Therapy Conversations. *arXiv preprint arXiv:2504.13955*.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–34.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1504–1532.

Christoforou, E.; Demartini, G.; and Otterbacher, J. 2024. Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 2097–2103.

Cooper, R.; and Foster, M. 1971. Sociotechnical systems. *American Psychologist*, 26(5): 467.

Czopp, A. M.; Kay, A. C.; and Cheryan, S. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10(4): 451–463.

Dam, S. K.; Hong, C. S.; Qiao, Y.; and Zhang, C. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; et al. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 246–267.

Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Tuan, L. A.; and Joty, S. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, 1679–1705.

Dolata, M.; Feuerriegel, S.; and Schwabe, G. 2022. A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4): 754–818.

Eckert, P. 2011. Language and power in the preadolescent heterosexual market. *American speech*, 86(1): 85–97.

Ehlebracht, M. 2019. Social media and othering: Philosophy, algorithms, and the essence of being human. *Consensus*, 40(1): 3.

Ferrara, E. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1): 3.

Gautam, S.; Venkit, P. N.; and Ghosh, S. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776*.

Ghosh, S.; and Caliskan, A. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 901–912.

Ghosh, S.; Venkit, P. N.; Gautam, S.; Wilson, S.; and Caliskan, A. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 476–489.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Gupta, V.; Venkit, P. N.; Wilson, S.; and Passonneau, R. J. 2024. Sociodemographic Bias in Language Models: A Survey and Forward Path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 295–322.

Haxvig, H. A. 2024. Concerns on Bias in Large Language Models when Creating Synthetic Personae. *arXiv preprint arXiv:2405.05080*.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2024. Evaluating Creative Short Story Generation in Humans and Large Language Models. *arXiv preprint arXiv:2411.02316*.

Jung, S.-G.; An, J.; Kwak, H.; Ahmad, M.; Nielsen, L.; and Jansen, B. J. 2017. Persona generation from aggregated social media data. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 1748–1755.

Kambhatla, G.; Stewart, I.; and Mihalcea, R. 2022. Surfacing racial stereotypes through identity portrayal. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1604–1615.

Kay, A. C.; Day, M. V.; Zanna, M. P.; and Nussbaum, A. D. 2013. The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology*, 49(2): 287–291.

Kim, J. K.; Chua, M.; Rickard, M.; and Lorenzo, A. 2023. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5): 598–604.

Lazik, C.; Katins, C.; Kauter, C.; Jakob, J.; Jay, C.; Grunske, L.; and Kosch, T. 2025. The Impostor is Among Us: Can Large Language Models Capture the Complexity of Human Personas? *arXiv preprint arXiv:2501.04543*.

Lee, M. H.; Montgomery, J. M.; and Lai, C. K. 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*, 1321–1340.

Lehr, S. A.; Caliskan, A.; Liyanage, S.; and Banaji, M. R. 2024. ChatGPT as research scientist: probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35): e2404328121.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manning, B. S.; Zhu, K.; and Horton, J. J. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.

Narayanan Venkit, P. 2023. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 1004–1005.

Noble, S. U. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

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

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is Temperature the Creativity Parameter of Large Language Models? *CoRR*.

Prpa, M.; Troiano, G.; Yao, B.; Li, T. J.-J.; Wang, D.; and Gu, H. 2024. Challenges and Opportunities of LLM-Based Synthetic Personae and Data in HCI. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 716–719.

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.

Qin, H. X.; Jin, S.; Gao, Z.; Fan, M.; and Hui, P. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Salminen, J.; Guan, K.; Jung, S.-g.; Chowdhury, S. A.; and Jansen, B. J. 2020. A literature review of quantitative persona creation. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Salminen, J.; Jung, S.-g.; and Jansen, B. J. 2019. The Future of Data-driven Personas: A Marriage of Online Analytics Numbers and Human Attributes. In *ICEIS (1)*, 608–615.

Salminen, J.; Liu, C.; Pian, W.; Chi, J.; Häyhänen, E.; and Jansen, B. J. 2024. Deus ex machina and personas from large language models: investigating the composition of AI-generated persona descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20.

Sartori, L.; and Theodorou, A. 2022. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1): 4.

Schuller, A.; Janssen, D.; Blumenröther, J.; Probst, T. M.; Schmidt, M.; and Kumar, C. 2024. Generating personas using LLMs and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

Sethi, S.; Salminen, J.; Amin, D.; and Jansen, B. J. 2025. " When AI Writes Personas": Analyzing Lexical Diversity in LLM-Generated Persona Descriptions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8.

Shin, J.; Hedderich, M. A.; Rey, B. J.; Lucero, A.; and Oulasvirta, A. 2024. Understanding human-AI workflows for generating personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 757–781.

Staab, R.; Vero, M.; Balunovic, M.; and Vechev, M. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Sun, G.; Zhan, X.; and Such, J. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 1–6.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Venkit, P. N.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122.

Venkit, P. N.; Laban, P.; Zhou, Y.; Huang, K.-H.; Mao, Y.; and Wu, C.-S. 2025. DeepTRACE: Auditing Deep Research AI Systems for Tracking Reliability Across Citations and Evidence. *arXiv preprint arXiv:2509.04499*.

Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748.

Whitehouse, C.; Choudhury, M.; and Aji, A. 2023. LLM-powered Data Augmentation for Enhanced Cross-lingual Performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 671–686.

Wolfe, R.; and Caliskan, A. 2022. Markedness in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1269–1279.

Yukhymenko, H.; Staab, R.; Vero, M.; and Vechev, M. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37: 120735–120779.

Zhang, S.; Xu, J.; and Alvero, A. 2024. Generative ai meets open-ended survey responses: Participant use of ai and homogenization.