Tag-First: Mitigating Distributional Bias in Synthetic User Profiles through Controlled Attribute Generation

Ismael Garrido-Muñoz[1], Fernando Martínez-Santiago[1], Arturo Montejo-Ráez[1]

[1]CEATIC, Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain

Correspondence: {igmunoz, dofer, amontejo}@ujaen.es

Abstract

Addressing the critical need for robust bias testing in AI systems, current methods often rely on overly simplistic or rigid persona templates, limiting the depth and realism of fairness evaluations. We introduce a novel framework and an associated tool designed to generate high-quality, diverse, and configurable personas specifically for nuanced bias assessment. Our core innovation lies in a two-stage process: first, generating structured persona `tags` based *solely* on user-defined configurations (specified manually or via an included agent tool), ensuring attribute distributions are controlled and crucially, are not skewed by an LLM's inherent biases regarding attribute correlations during the selection phase. Second, transforming these controlled `tags` into various realistic outputs—including natural language descriptions, CVs, or profiles—suitable for diverse bias testing scenarios. This tag-centric approach preserves ground-truth attributes for analyzing correlations and biases within the generated population and downstream AI applications. We demonstrate the system's efficacy by generating and validating 1,000 personas, analyzing both the adherence of natural language descriptions to the source tags and the potential biases introduced by the LLM during the transformation step. The provided dataset, including both generated personas and their source tags, enables detailed analysis. This work offers a significant step towards more reliable, controllable, and representative fairness testing in AI development.

1 Introduction

The imperative to ensure fairness and mitigate harmful biases in artificial intelligence (AI) systems is paramount (Garrido-Muñoz et al., 2021; Mehrabi et al., 2019), especially given their increasing deployment in high-stakes domains such as conversational agents, recommendation systems, and social modeling tasks. However, progress is frequently hindered by significant limitations in existing evaluation methodologies, particularly in how synthetic populations or personas are generated for bias testing.

Current persona generation approaches face significant hurdles for robust bias testing. Manual creation, while potentially rich, is hampered by scalability constraints, cost, and the risk of implicit creator bias (Jansen et al., 2021). Automated methods introduce their own set of challenges. Some rely on rigid templates that can produce stereotypical outputs (Li et al., 2025). More fundamentally, the evaluation benchmarks used to validate systems are often demographically skewed, which can hide critical performance gaps. The landmark "Gender Shades" study, for instance, audited commercial facial analysis systems and found substantial accuracy disparities across intersectional subgroups (Buolamwini and Gebru, 2018). The systems performed worst on darker-skinned females (with error rates up to 34.7%) compared to lighter-skinned males (with a max error rate of 0.8%), a disparity linked to the underrepresentation of darker-skinned women in the popular training and benchmark datasets (Buolamwini and Gebru, 2018). This highlights a critical flaw in AI evaluation: without balanced and representative test sets, harmful algorithmic biases can go undetected.

This problem of bias extends beyond evaluation data and is deeply embedded in the training corpora of generative models themselves. Foundational work by Bolukbasi et al. (2016) revealed this danger in word embeddings, showing that models trained on large text corpora absorb and reproduce stark gender stereotypes. This leads to harmful associations like "man is to computer programmer as woman is to homemaker" instead of neutral relationships (e.g., "doctor" being equally related to "man" and "woman") or biologically grounded ones (e.g., "man is to father as woman is to mother"). This issue of bias amplification

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 149–159

is even more pronounced in modern Large Language Models (LLMs). Directly using LLMs for end-to-end persona generation risks magnifying the societal biases present in their training data (Sheng et al., 2019; Bender et al., 2021) and provides little fine-grained control over attribute distributions (Raji et al., 2020; Liu et al., 2024). Furthermore, the immense scale and opacity of the datasets used to train these models create significant challenges for transparency and validation, a gap that has prompted calls for standardized documentation practices like Datasheets for Datasets (Gebru et al., 2021). These collective limitations underscore the need for a more flexible, controllable, and transparent methodology.

To address these limitations, we propose a novel framework centered around a **tag-first generation** methodology designed for creating flexible, realistic, and statistically controlled personas for rigorous bias testing. This framework tackles the core issue of uncontrolled attribute correlation bias inherent in direct LLM generation. The process involves two primary stages:

1. **Configurable Attribute Definition and Tag Generation:** First, desired persona characteristics (attributes) and their probability distributions are explicitly defined in a structured configuration (YAML). Based *only* on this configuration, structured attribute `tags` (key-value pairs) are probabilistically generated for each persona. This critical step ensures that the attribute distributions within the generated population strictly adhere to the user's specifications, preventing LLMs from skewing attribute selection based on their internal biases about real-world correlations (e.g., between occupation and gender).

2. **Controlled Transformation:** Second, these generated structured `tags` serve as a controlled input foundation. A Large Language Model (LLM) then transforms these tags into richer, realistic outputs (e.g., natural language descriptions) suitable for specific testing scenarios, while maintaining the link to the source tags.

This tag-centric approach offers significant advantages beyond the controlled attribute assignment achieved in Stage 1. It provides transparency regarding the exact attributes assigned to each persona, and the persistent `tags` serve as ground truth. This enables systematic analysis of how generated content correlates with specific attributes and how downstream AI systems respond to these controlled variations.

To facilitate the potentially complex task of creating the initial configuration (Stage 1), we have developed an interactive tool featuring a conversational agent. This tool guides users, including non-experts, through the process of defining persona attributes and distributions using natural language dialogue. It assists in creating the necessary structured configuration file, incorporating configurable attribute randomization and offering suggestions informed by the user's specified testing context. Manual creation or modification of the configuration file remains possible for expert users.

2 Related Work

Our work intersects with several research areas: persona generation methodologies, the study and mitigation of bias in Large Language Models (LLMs), the use of personas for evaluating AI systems, and the inherent challenges of bias in manual processes.

2.1 Approaches to Persona Generation

Personas, as archetypal representations of users, are widely employed in Human-Computer Interaction (HCI), software design, and increasingly, AI evaluation and training (Cooper, 1999; Nielsen, 2019). Traditionally, personas were meticulously crafted by researchers based on qualitative user data. While these manual personas can be rich and context-grounded, their creation is resource-intensive, does not scale well, and, critically, can inadvertently embed the creators' own conscious or unconscious biases and stereotypes (Jansen et al., 2020; Chapman and Milham, 2006). This underscores the challenge of **human bias in manual creation**, where designers might unintentionally oversimplify or stereotype user groups.

To address scalability and potentially reduce individual bias, various automated and semi-automated persona generation techniques have emerged (Şengün et al., 2018). Early approaches often relied on rule-based systems or templates populated from data analytics (Jansen et al., 2021). While scalable, these methods could lack nuance or enforce overly rigid structures. Other techniques utilize clustering algorithms on user data to identify common behavioral patterns and derive per-

sona archetypes (An et al., 2018). However, such data-driven methods risk directly inheriting and potentially amplifying biases present in the source data (e.g., reflecting historical inequities or sampling biases) (Jansen et al., 2020).

More recently, the advent of powerful LLMs has spurred interest in leveraging them for persona generation (Jiang et al., 2024; Park et al., 2022). LLMs can produce fluent and seemingly detailed persona descriptions from relatively simple prompts. However, achieving fine-grained control over specific attributes and ensuring representative diversity often relies heavily on complex and brittle prompt engineering (Raji et al., 2020). Furthermore, systematically validating the generated personas for internal consistency and adherence to desired attributes remains a significant challenge (Zhao et al., 2023). Our approach contrasts with purely LLM-driven generation by employing a **structured YAML configuration** to explicitly define attribute possibilities and their probability distributions before generation. This affords explicit control over the persona population's characteristics. The subsequent LLM-based transformation step (e.g., generating natural language) then builds upon this controlled, tag-based foundation, separating attribute selection from narrative generation.

2.2 Bias Testing in Large Language Models

The potential for LLMs to perpetuate and even amplify societal biases encoded in their vast training data is well-documented (Bender et al., 2021; Weidinger et al., 2021). Research has extensively investigated biases related to **gender, race, ethnicity, religion, age, disability, socioeconomic status, and other demographic factors** within LLMs (Bolukbasi et al., 2016; Caliskan et al., 2017; Blodgett et al., 2021). These biases can manifest as stereotypical associations (e.g., linking genders to specific occupations (Sheng et al., 2019)), disparate performance across demographic groups for downstream tasks, or the generation of harmful, offensive, or denigrating content (Garrido-Muñoz et al., 2021; Mehrabi et al., 2019).

Numerous benchmarks and techniques exist for detecting and measuring such biases. These range from analyzing geometric properties of word embeddings (Caliskan et al., 2017) and probing model outputs with carefully crafted templates (Nadeem et al., 2020) to evaluating performance disparities on downstream tasks across different demographic contexts (Blodgett et al., 2021; Mehrabi et al., 2019). Understanding these biases is critical for our work for two primary reasons: first, our framework utilizes LLMs (within the optional agent tool, for the controlled transformation step, and potentially for validation), making awareness and mitigation of their inherent biases crucial; second, the diverse and controlled personas generated by our framework are intended precisely for use in evaluating biases within AI systems. Our adjective-based bias check (§4) represents a preliminary step towards monitoring potential biases introduced specifically during the LLM-based transformation phase of our pipeline.

2.3 Using Personas for Bias Evaluation

Recognizing the limitations of purely quantitative metrics or evaluations based on aggregate data, researchers have increasingly turned to **using personas to conduct more qualitative or contextualized evaluations of AI systems**, particularly regarding fairness, bias, and safety (Ghai, 2023). Personas allow for testing system responses across a spectrum of intersecting user characteristics and backgrounds, offering potentially richer insights than abstract benchmarks. For instance, personas representing different demographics can interact with chatbots to assess response quality, identify potential harms, and evaluate safety guardrails (akin to structured red teaming approaches, e.g., (Perez et al., 2022)), or they can be used as simulated users to evaluate recommendation systems for fairness in exposure or disparate outcomes across groups (Misztal-Radecka and Indurkhya, 2020).

However, the effectiveness of this evaluation paradigm hinges critically on the quality, diversity, and representativeness of the personas employed. If the personas themselves are biased, lack diversity along relevant axes, or are not well-validated, the resulting evaluation may produce misleading or incomplete conclusions (Salminen et al., 2018). Our work aims to contribute directly to this area by providing a methodology for generating diverse, validated personas with explicitly controlled attribute distributions. By enabling the systematic creation of persona sets tailored to specific fairness concerns (facilitated by the structured configuration and optional agent), our framework provides more reliable and reproducible artifacts for downstream bias testing compared to ad-hoc, manually created, or unvalidated LLM-generated persona sets (Ghai, 2023).

3 Methodology

Our persona generation framework operationalizes the tag-first methodology introduced in Section 1 (illustrated in Figure 1). The process is orchestrated through several key components designed for flexibility and control over persona attributes. Central to the framework is a structured YAML configuration file that defines the desired attributes and their distributions. An optional agent tool assists users in creating this configuration. Based solely on the YAML specifications, the system first generates structured persona `tags`, which then serve as controlled input for subsequent transformation into richer outputs like natural language descriptions. This section details these components, starting with the configuration structure.

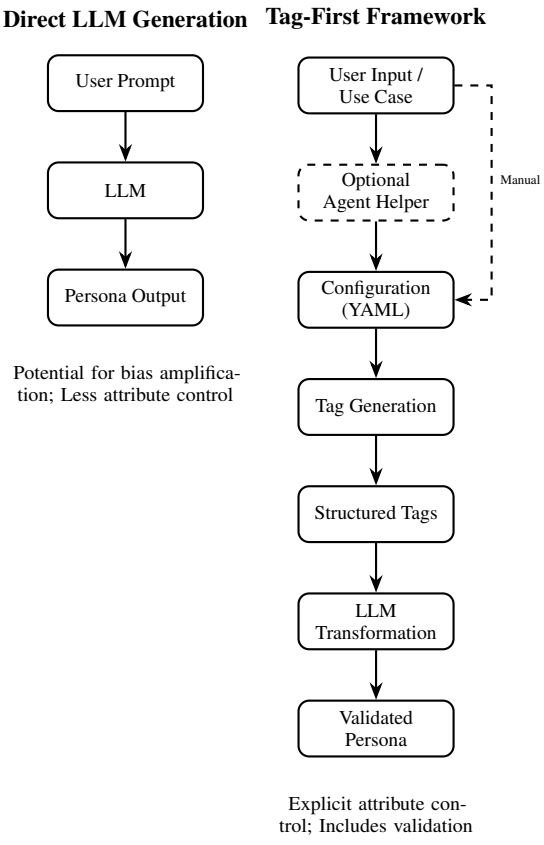

Figure 1: Direct LLM persona generation vs our proposed tag-first approach

3.1 Structured Persona Configuration (YAML)

Our persona framework leverages structured YAML configurations to specify diverse attributes comprehensively. Users define attributes such as gender, race, religion, socioeconomic status, geography, political affiliation, disability status, age, sexual orientation, working experience, hobbies, and education. Each attribute is defined using detailed YAML sections containing parameters such as quantity (how many values to generate for each feature), potential values with associated probabilities, desired levels of detail for the values and dynamic property names to help the LLM.

This structured approach enables fine-grained control over the persona population. Examples of detailed configurations include:

Race Configuration Example: Users can enable mixed-race profiles by specifying probabilities for generating one or two race tags, potentially using different property names for each case.

```
race:
  type: categorical
  quantity:
    1: 80
    2: 20
  quantity_properties:
    1: race
    2: [father_race, mother_race]
  level_of_detail_values:
    low: [white, black, hispanic, asian,
    native_american, pacific_islander]
```

Political Affiliation Example: Users can specify varying granularity (e.g., general orientation vs. specific party), mixing broad labels with detailed, weighted options.

```
political:
  type: categorical
  quantity: 1
  level_of_detail_values:
    low: [left, center, right]
    detailed: [
        Party A: 30,
        Party B: 25,
        Party C: 20,
        Party D: 15,
        Party E: 10
        ]
  level_of_detail_properties:
    low: political_orientation
    detailed: political_party
```

Geography: Configuring geographical detail from broad to specific.

```
geography:
  type: categorical
  quantity:
```

```
  1: 60
  2: 40
quantity_properties:
  1: country
  2: [born_country, current_city]
level_of_detail_values:
 countries: [USA, Spain, Germany, Italy]
 cities: [New York, Madrid, Berlin, Rome]
```

Using this configurations, values are generated based on predefined probability distributions specified within the YAML file. This flexibility ensures realistic and diverse personas closely aligned with user-defined requirements.

3.2 Agent-Assisted Configuration

To facilitate the creation of a potentially complex YAML configuration file, especially for users less familiar with YAML syntax or the nuances of persona attribute design for bias testing, we developed an interactive agent. This agent guides the user through the configuration process using natural language interaction, leveraging Large Language Models (LLMs) for specific tasks such as understanding context, suggesting adaptations, explaining YAML, and processing updates based on user feedback. The agent's workflow is implemented as a state machine using the LangGraph framework (LangChain, 2024), managing the conversation state and orchestrating the different steps involved.

1. **Use Case Definition:** The agent begins by prompting the user to define the specific context or system they intend to test (e.g., "CV screening system for software engineers in Germany" "loan application evaluation").

2. **Feature Prioritization (LLM-driven):** Based on the defined use case and a predefined list of potential persona attributes (features), an LLM categorizes these features into groups: those expected to be directly relevant to the system's function, those expected *not* to be relevant but crucial for bias testing (e.g., demographics), and those deemed irrelevant to the use case. This step helps focus the configuration effort on attributes pertinent to bias evaluation.

3. **Insight Generation (LLM-driven):** For features identified as important for bias testing, the agent uses an LLM to generate brief, potentially non-obvious insights about how these features might relate to bias within the specified use case, aiming to inform the subsequent configuration choices.

4. **Iterative Feature Configuration:** The agent then enters an iterative loop, processing each feature one by one. For each feature:

   - *Adaptation (LLM-driven):* An LLM mutates and proposes an initial YAML configuration for the feature, attempting to tailor value distributions, levels of detail, or ranges based on the use case and any generated insights.
   - *Explanation (LLM-driven):* The agent presents the proposed YAML snippet and uses an LLM to generate a plain-language explanation of what the configuration implies (e.g., "female and male each have a 40% chance of being chosen and non-binary has a 10%").
   - *User Feedback & Refinement (LLM-driven):* The user can then accept the configuration or provide natural language feedback to request modifications (e.g., "Tweak the 'non-binary' probability up to 15%", "Add the of 'Hispanic' ethnicity" or "let's go with the top 3 religions in Spain with their respective probabilities"). If the user request a change, an LLM processes the feedback and attempts to update the YAML snippet accordingly. This sub-loop allows for interactive refinement until the user is satisfied or chooses to proceed.

5. **Finalization:** Once all prioritized features have been configured, the agent saves the complete YAML and let the user download the configuration to a file for later use in the persona generator. Optionally, the agent can then generate a sample persona immediately using this final configuration.

The detailed workflow of this agent is illustrated in Figure 2.

4 Validation and Analysis

Using the finalized YAML configuration (created manually or via the agent), we generated a dataset of 1,000 personas following a systematic, multi-step approach designed to ensure both adherence to the configuration and internal consistency.

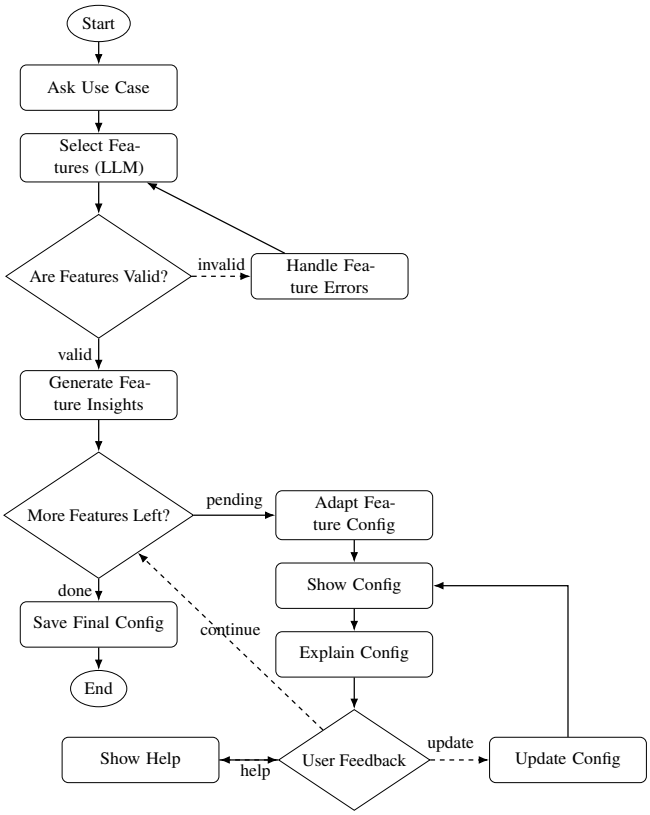

Figure 2: Workflow diagram of the interactive agent for YAML configuration generation. Diamonds represent decision points based on state or user input.

Persona Tags Generation: First, for each persona, the system generates a set of structured `tags` by sampling values for each attribute according to the probabilities, ranges, and constraints defined in the YAML configuration. This critical step ensures the resulting tag distributions align strictly with the user's specifications before any LLM generation occurs. An example tag set for a single persona might be:

```
gender: female
race: hispanic
past_religion: agnostic
current_religion: none
socio_economic_status: high
born_location_country: Spain
current_location_world_region: Africa
political_orientation: conservative
disability: none
age: 22
sexual_orientation: heterosexual
job_title: research assistant
first_hobby: sailing
studied: Psychology
```

Validation and Transformation Pipeline: The generated tags then proceed through a validation and transformation pipeline:

1. **Tag Validation:** The initial set of tags for each persona is validated to identify potential logical contradictions or highly improbable combinations (e.g., conflicting age and occupation). This is done by asking an LLM to spot inconsistencies; if any issue is found, the persona is discarded.

2. **Controlled Transformation:** Validated tag sets serve as input to an LLM, which is prompted to synthesize a coherent natural language description based *only* on the provided tags, aiming to weave them into a realistic narrative without introducing unstated information. The tags can be applied to any other use case, e.g., generating CVs, creating tweets, or even answering questions from the perspective of the persona based on its tags.

3. **Tag Adherence Validation:** After generating the natural language description, an automated validation step assesses how well the text reflects the original source tags. We use an LLM to check each source tag against the generated text, classifying its presence as 'explicitly mentioned', 'implied by context', or 'absent'. Personas failing to meet a predefined adherence threshold (in our case, at least 90% of tags classified as explicitly mentioned or clearly implied) are discarded. This step aims to ensure the final personas remain faithful to the controlled, structured attributes.

This systematic process is designed to yield personas whose underlying attributes are known and controlled.

Adjective Extraction for Bias Analysis: Finally, for the validated personas that passed the adherence checks, adjectives were automatically extracted from their natural language descriptions. This step provides structured data that can be used for subsequent quantitative analysis, particularly for preliminary checks on potential biases or stereotypical language patterns introduced by the LLM during the transformation (natural language generation) stage.

4.1 Analysis and Mitigation Potential for Linguistic Bias in LLM Transformation

Our tag-first generation framework is designed primarily to ensure that core persona attributes (like gender, race, etc.) adhere strictly to user-defined distributions, mitigating bias in attribute *selection*. However, the subsequent step of transforming these controlled tags into natural language using an LLM can still introduce subtler linguistic biases, reflecting patterns learned from the LLM's training data. We investigated this by analyzing the adjectives generated within the descriptions of our 1,000 validated personas, comparing frequencies based on the 'male' vs. 'female' gender tags.

The results, summarized in Table 1, confirm the presence of such residual linguistic bias. Despite the balanced input distribution for the gender tag itself, noticeable differences emerged in the adjectives the LLM used. For instance, descriptions for male personas in our sample more frequently included adjectives like *diverse* (+1.10% weight difference), *financial* (+0.55%), and *physical* (+0.41%), while descriptions for female personas were more likely to contain *dynamic* (-2.06%), *vibrant* (-1.17%), *resilient* (-0.95%), and *strong* (-0.67%). These deviations range from -2,06% and 1,10%, which is a marginal bias difference.

This finding highlights that LLMs carry inherent linguistic associations (Bolukbasi et al., 2016; Bender et al., 2021) which can manifest even when provided with controlled, structured input like our tags. However, a key advantage of our tag-centric framework is that it provides potential avenues to actively mitigate this linguistic bias, which are unavailable in direct end-to-end LLM generation. Because we control the precise set of tags fed into the LLM transformation step, we can strategically modify the tag generation process itself:

- **Enriching Tag Sets:** The YAML configuration could be extended beyond core attributes to include specific 'style', 'tone', or 'personality' tags. Generating these alongside demographic tags could provide explicit guidance to the LLM during transformation, potentially overriding default linguistic tendencies. For example, explicitly adding a tag like 'personality: analytical' might encourage the LLM to use related adjectives more evenly across genders.
- **Counter-Stereotypical Tag Combinations:** The configuration could be designed to intentionally generate combinations of tags that challenge stereotypical associations. For instance, frequently pairing the 'female' tag with tags related to typically male-associated fields (e.g., 'job_sector: finance', 'hobby: coding') might nudge the LLM to adjust its descriptive language during transformation.
- **Feedback-Driven Configuration Refinement:** The type of adjective analysis presented here (Table 1) can serve as direct feedback. These results could inform iterative adjustments to the YAML configuration probabilities or the inclusion of specific guiding tags in future generation runs, aiming to systematically reduce observed linguistic disparities.

Therefore, while the existence of residual linguistic bias necessitates careful validation and awareness, our framework's explicit control over the intermediate tag representation offers concrete pathways for addressing it. This contrasts sharply with direct generation approaches where influencing the nuanced linguistic choices of the LLM is far more opaque and difficult.

The implications remain significant: validation beyond tag adherence is crucial, users should be aware of potential linguistic nuances, and further research is needed. However, this research can now explore leveraging the configurable tag-generation process itself as a primary tool for linguistic bias mitigation, in addition to developing better LLM prompting or fine-tuning strategies for the transformation step.

In conclusion, our analysis confirms that linguistic bias can persist even with controlled input attributes. Critically, however, the proposed tag-first methodology provides tangible mechanisms—through richer configuration, strategic tag combination, and feedback loops—to actively steer the LLM's linguistic output and work towards generating persona descriptions that are not only demographically representative but also linguistically equitable.

4.2 Potential Use Cases for AI System Evaluation

The primary strength of our flexible persona generation system lies in its ability to create controlled, diverse, and validated user representations for the

Adjective	Male Count	Female Count	Male Weight	Female Weight	Weight Difference	Count Difference
diverse	486	457	8.18%	7.09%	**+1.10%**	+29
personal	315	298	5.30%	4.62%	**+0.68%**	+17
rich	322	307	5.42%	4.76%	**+0.66%**	+15
hispanic	102	75	1.72%	1.16%	**+0.55%**	+27
financial	75	46	1.26%	0.71%	**+0.55%**	+29
unique	355	357	5.98%	5.53%	+0.44%	-2
physical	109	92	1.84%	1.43%	+0.41%	+17
spiritual	108	93	1.82%	1.44%	+0.38%	+15
fascinating	75	57	1.26%	0.88%	+0.38%	+18
asian	66	53	1.11%	0.82%	+0.29%	+13
moderate	72	61	1.21%	0.95%	+0.27%	+11
middle-class	93	84	1.57%	1.30%	+0.26%	+9
progressive	96	89	1.62%	1.38%	+0.24%	+7
conservative	109	106	1.84%	1.64%	+0.19%	+3
traditional	108	106	1.82%	1.64%	+0.18%	+2
modern	78	75	1.31%	1.16%	+0.15%	+3
different	75	73	1.26%	1.13%	+0.13%	+2
comfortable	65	62	1.09%	0.96%	+0.13%	+3
profound	87	87	1.46%	1.35%	+0.12%	+0
deep	142	148	2.39%	2.29%	+0.10%	-6
analytical	122	126	2.05%	1.95%	+0.10%	-4
balanced	83	84	1.40%	1.30%	+0.10%	-1
intriguing	73	73	1.23%	1.13%	+0.10%	+0
complex	92	94	1.55%	1.46%	+0.09%	-2
innovative	62	64	1.04%	0.99%	+0.05%	-2
keen	76	80	1.28%	1.24%	+0.04%	-4
multicultural	101	108	1.70%	1.67%	+0.03%	-7
young	68	72	1.14%	1.12%	+0.03%	-4
christian	60	63	1.01%	0.98%	+0.03%	-3
political	88	94	1.48%	1.46%	+0.02%	-6
academic	102	110	1.72%	1.71%	+0.01%	-8
liberal	87	97	1.46%	1.50%	-0.04%	-10
open	62	70	1.04%	1.09%	-0.04%	-8
new	70	79	1.18%	1.22%	-0.05%	-9
socio-economic	109	123	1.84%	1.91%	-0.07%	-14
cultural	209	236	3.52%	3.66%	-0.14%	-27
compassionate	58	72	0.98%	1.12%	-0.14%	-14
global	85	102	1.43%	1.58%	-0.15%	-17
intellectual	61	77	1.03%	1.19%	-0.17%	-16
professional	193	225	3.25%	3.49%	-0.24%	-32
social	73	95	1.23%	1.47%	-0.24%	-22
adventurous	63	84	1.06%	1.30%	-0.24%	-21
creative	162	193	2.73%	2.99%	-0.26%	-31
bustling	84	111	1.41%	1.72%	-0.31%	-27
multifaceted	109	143	1.84%	2.22%	-0.38%	-34
demanding	44	79	0.74%	1.22%	-0.48%	-35
strong	114	167	1.92%	2.59%	**-0.67%**	-53
resilient	45	110	0.76%	1.71%	**-0.95%**	-65
vibrant	295	396	4.97%	6.14%	**-1.17%**	-101
dynamic	151	297	2.54%	4.60%	**-2.06%**	-146

Table 1: Top 50 adjectives compared between male and female

rigorous evaluation of AI systems, particularly concerning fairness, robustness, and safety. Key evaluation scenarios include:

- **Auditing Conversational AI for Bias:** Systematically testing chatbots and virtual assistants with personas representing diverse demographic backgrounds (gender, race, age, disability), socioeconomic statuses, and communication styles. This allows for detecting differential treatment, biased responses (e.g., variations in politeness, helpfulness, or accuracy), or safety failures triggered by specific user profiles.
- **Evaluating Fairness in Recommendation Systems:** Generating sets of personas with controlled preference distributions and demographic attributes (Misztal-Radecka and Indurkhya, 2020) to audit recommendation engines (e.g., for job listings, news, products, financial services) for fairness issues like exposure disparities, filter bubbles, or inequitable outcomes across different user groups.
- **Assessing Automated Content Moderation Tools:** Simulating user interactions and content submissions (text, potentially images/video concepts linked to persona tags in future work) from personas with varying political affiliations, cultural backgrounds, or sensitivities. This helps identify biases in moderation decisions, such as disproportionate flagging or removal of content associated with certain groups.
- **Probing Personalization Algorithms:** Using personas to evaluate how personalization algorithms (e.g., in search engines, social media feeds) tailor content and whether this leads to undesirable outcomes like information cocoons, biased information exposure, or discriminatory targeting based on inferred persona characteristics.
- **Structured Red Teaming for Bias Discovery:** Employing personas (Perez et al., 2022) specifically designed to represent vulnerable groups, edge cases, or adversarial inputs to proactively uncover hidden biases, stereotypes, or failure modes in AI systems before deployment.
- **Generating Controlled Synthetic Data for Bias Testing:** Creating balanced or specifically skewed datasets of synthetic user interactions based on personas when real-world data is unavailable, sensitive, or lacks sufficient representation of minority groups. This enables controlled experiments to isolate and measure algorithmic bias.
- **Standardized Fairness Auditing Benchmarks:** Leveraging the system to create shareable, reproducible benchmark suites of diverse personas, allowing for standardized testing and comparison of fairness properties across different AI models or platforms (Felt et al., 2023).

The agent-driven configuration and explicit control over attribute probabilities are crucial for designing targeted evaluation studies that systematically explore how AI systems respond to the diversity inherent in real-world user populations.

Limitations

While our framework provides enhanced control over persona attribute distributions, several limitations should be acknowledged. First, despite mitigating attribute selection bias by design, the reliance on Large Language Models (LLMs) for the transformation stage (generating natural language descriptions, etc.) means that linguistic biases inherent in the LLM can still manifest in the output, as discussed in Section 4.1. Continuous monitoring and the proposed mitigation strategies are important. Second, the quality and representativeness of the generated personas are fundamentally dependent on the comprehensiveness and accuracy of the initial YAML configuration. Crafting highly nuanced configurations may still require significant domain expertise, even with the aid of the agent tool. Third, the overall effectiveness of the framework, including the agent's utility and the realism of the generated outputs, is tied to the capabilities and potential failure modes of the chosen LLMs. Finally, the current implementation focuses on attributes explicitly defined within the configuration schema, primarily emphasizing mainstream demographic categories, and generates text-based outputs. This focus may overlook the complex overlap between social categories and diverse communication styles across different cultures. Extending the attribute ontology to be more inclusive or supporting diverse output modalities represents important avenues for future work.

Availability

The source code for our framework, the conversational agent, and the generated persona dataset are publicly available on GitHub at: https://github.com/IsGarrido/Gender_Agent_Frozen.

Bias Statement

In this work, we define bias as the tendency of a generative model to produce synthetic user profiles with stereotypical correlations between demographic attributes (e.g., gender, race) and personal characteristics (e.g., occupation). This behavior is harmful because it creates a **representational harm** by reinforcing damaging societal stereotypes about different social groups. Consequently, when these biased profiles are used to evaluate downstream AI systems (e.g., for hiring), this can lead to **allocational harm**, where systems validated on stereotypical data may unfairly discriminate against real individuals from underrepresented groups.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. pages 2450–2461.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Christopher N. Chapman and Russell P. Milham. 2006. The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(5):634–636.

Alan Cooper. 1999. *The Inmates are Running the Asylum*, pages 17–17. Vieweg+Teubner Verlag, Wiesbaden.

Gillian Felt, Paula Cho, and Meredith Ringel Morris. 2023. Approaches for measuring and reducing gendered biases with personas. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–16. ACM.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago , and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Bhavya Ghai. 2023. *Towards Fair and Explainable AI using a Human-Centered AI Approach*. Ph.D. thesis, Stony Brook University.

Bernard J. Jansen, Joni O. Salminen, and Soon-Gyo Jung. 2020. Data-driven personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management*, 4(1):1–17.

Jim Jansen, Joni Salminen, Soon-Gyo Jung, and Kathleen Guan. 2021. Data-driven personas. *Synthesis Lectures on Human-Centered Informatics*, 14:i–317.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

LangChain. 2024. LangGraph. `https://github.com/langchain-ai/langgraph`. Accessed: 2025-04-01.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. *arXiv e-prints*, arXiv:2503.16527.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. 2019. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35.

Joanna Misztal-Radecka and Bipin Indurkhya. 2020. Persona prototypes for improving the qualitative evaluation of recommendation systems. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, page 206–212, New York, NY, USA. Association for Computing Machinery.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *Preprint*, arXiv:2004.09456.

Lene Nielsen. 2019. *Personas - User Focused Design*, 2nd edition. Springer Publishing Company, Incorporated.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Joni Salminen, Jim Jansen, Jisun An, Haewoon Kwak, and Soon-Gyo Jung. 2018. Are personas done? evaluating their usefulness in the age of digital analytics. *Persona Studies*, 4:47.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Hendricks, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Sercan Şengün, Joni Salminen, Haewoon Kwak, Jim Jansen, Jisun An, Soon-Gyo Jung, Sarah Vieweg, and D. Harrell. 2018. From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday*, 23.