

What is my Age (and Gender) again? Profiling AI Systems Through the Lens of Gender Stereotype Detection

Davide Testa^{1,2,*}, Gloria Comandini³, Manuela Speranza¹, Sofia Brenna^{1,4},
Stefania Cavagnoli⁵ and Bernardo Magnini¹

¹Fondazione Bruno Kessler (FBK), Trento, Italy

²University of Rome La Sapienza, Rome, Italy

³Istituto Italiano di Studi Germanici (IISG), Rome, Italy

⁴Free University of Bozen-Bolzano, Bolzano, Italy

⁵University of Trento, Trento, Italy

Abstract

The identification of gender stereotypes is a deeply subjective and culturally grounded task, in which annotators' judgements are shaped by their background, lived experiences, and personal sensitivity. In this paper we investigate how the perception of gender stereotypes in Italian short texts varies across demographic groups, and we use group-level annotation profiles to characterise the behaviour of Large Language Models (LLMs) with respect to sociodemographic dimensions. Building on the GSI:detect dataset, we collected new annotations on a balanced subset of 124 texts from 281 non-expert participants, stratified by gender and age into four target groups (women and men, under and over 35). We then compared these annotations with the zero-shot outputs of seven LLMs. Our results show that (i) demographic background has a statistically significant effect on the perception of gender stereotypes, with differences concentrated at the intersection of gender and age; (ii) the considered LLMs align with different demographic profiles, with no uniform behaviour across systems; and (iii) the joint use of correlation- and distance-based metrics reveals how a model's sociodemographic profile can decompose into distinct but complementary dimensions of judgment, direction and intensity, suggesting that single-metric evaluations may oversimplify model behaviour in this kind of subjective tasks.

Code and data are publicly available on [GitHub](#).

Keywords

Gender stereotype detection, Subjectivity, Gender bias, Human-LLM alignment, Sociodemographic profiling

1. Introduction

Gender stereotypes (GSs) are a pervasive social phenomenon that both reflects and shapes collective expectations about gender. We define GSs as a set of generalized, simplistic and pre-constructed social beliefs about how a person should behave, what their role in society should be and what they should look like based on their gender [1, 2]. By prescribing socially accepted norms and behaviours, gender stereotypes contribute to reproducing and normalising the very expectations from which they originate. In Western societies GSs separate individuals into two groups, men and women, forcing them to conform to the cultural expectations of what it should mean to "be a man" or to "be a woman" [3]¹

However, the recognition and, therefore, the annotation of GSs (but also of stereotypes in general) is a very subjective task, shaped by cultural background, lived experience, and individual sensitivity, resulting in substantial disagreement among annotators. For instance, the annotation of racist stereotypes in [4] shows an inter-annotator agreement of 0.41 (Cohen's k). Although there are many reasons for such disagreement, two factors should be noteworthy in this context: first, belonging to a group

CLiC-it 2026: Twelfth Italian Conference on Computational Linguistics, September 14–16, 2026, Palermo, Italy

*Corresponding author.

✉ dtesta@fbk.eu (D. Testa); comandini@studigermanici.it (G. Comandini); manspera@fbk.eu (M. Speranza); sbrenna@fbk.eu (S. Brenna); magnini@fbk.eu (B. Magnini)

🆔 0009-0002-2489-5323 (D. Testa); 0000-0002-0740-5778 (B. Magnini)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Clearly, this polarized worldview excludes the people who do not recognize themselves in such a binary division of gender. In this study we focus only on the analysis of GSs regarding men and women but this is a purely methodological choice and does not assume in any way a binary view of gender.

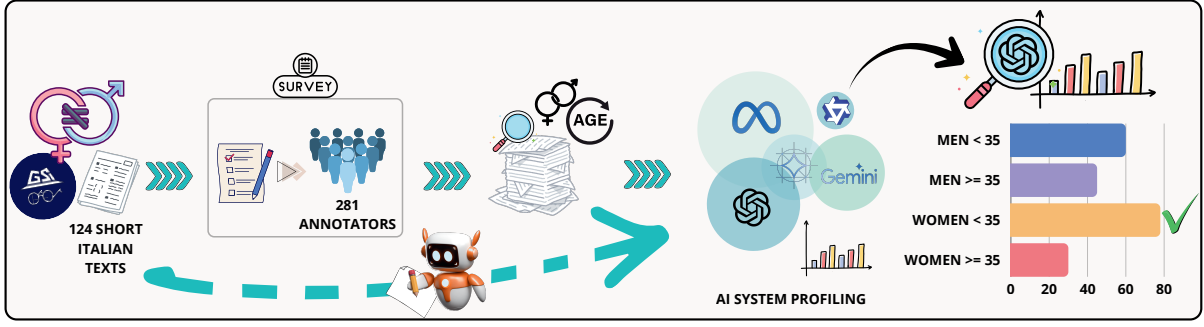


Figure 1: Workflow of the study. A subset of 124 short Italian texts from the GSI:detect dataset [10] was annotated by 281 participants, who assessed the presence of gender stereotypes using a continuous score in the interval $[0, 1]$. Participants were stratified by gender and age into four target groups: women under 35, women aged 35 or older, men under 35, and men aged 35 or older. Group-level annotation profiles were then compared with the outputs produced by LLMs. Through correlation-based analyses of human and model judgement patterns, a sociodemographic profiling of LLMs was attempted based on the human group with which their behaviour showed the strongest alignment.

targeted by the stereotype, as women and men tend to annotate sexist hate speech differently with women being more likely to classify less explicitly hateful comments (e.g., *pink quotas are useless*) as hate [5]; and second, internalised biases, which can subtly influence annotators’ judgements [6, 7]. Moreover, it should also be noted that these two reasons can intertwine, as some biases and stereotypes can also be interiorized by their own addressee, causing them to perpetuate such stereotypes [8].

The considerations above become especially relevant when the focus shifts from human annotators to LLMs. Since LLMs are trained on large collections of human-produced texts, their outputs may encode and reproduce not only the stereotypes found in those texts, but also the heterogeneous ways in which people perceive and interpret them.

In this context, the present study builds on the research conducted within the GSI:detect shared task organised at EVALITA 2026 [9], which resulted in the release of a specific dataset containing short Italian texts annotated by four trained annotators for the presence of GSs [10]. We extend the scope of the EVALITA work by investigating how GSs are perceived by a broader non-expert population. To this end, we selected a balanced subset of 124 texts from the original dataset and collected annotations from 281 participants, where 270 belong to the four target groups defined by the intersection of gender and age: women and men under and over 35. We then correlated the resulting group-level annotation profiles with the zero-shot outputs of seven LLMs evaluated on the same shared task. An overview of the overall workflow is shown in Figure 1. Our study addresses the following research questions:

RQ1: How do sociodemographic characteristics influence the perception of GSs in short Italian texts?

RQ2: To what extent can LLMs’ annotation patterns be associated with those of different sociodemographic groups?

In addressing these questions, our study provides a twofold contribution: (i) an analysis of how sociodemographic factors shape the perception of GSs; and (ii) a sociodemographic profiling of LLMs based on their alignment with distinct human groups.

2. Related Work

2.1. Subjectivity and Sociodemographic Variation in Annotation

The perception and annotation of socially grounded phenomena, such as hate speech, misogyny, and sexism, cannot always be reduced to a single objective ground truth. In highly subjective tasks, disagreements among annotators may reflect meaningful differences in their sociocultural backgrounds, lived experiences, and personal beliefs, rather than annotation noise alone. Aggregating individual

judgements into a single label may therefore obscure perspectives systematically expressed by specific individuals or social groups [11, 12, 13, 14]. Recent work has also investigated whether LLMs can approximate human judgement variation directly. For instance, Chen et al. [15] derive model judgement distributions from LLM probabilities and compare them with human label distributions in Natural Language Inference (NLI) tasks.

This issue, which has received increasing attention in NLP, is relevant to the detection of gender-related phenomena, including the automatic identification of misogynistic content [16]. Previous studies have shown that annotators’ characteristics and attitudes may influence their judgements in tasks involving sexist or harmful language [17, 18, 19, 20]. More specifically, Abercrombie et al. [21] investigated the relationship between annotator identities, underlying attitudes, and responses to a sexism-labelling task.

2.2. Biases and Stereotypes in LLMs

A further line of research has examined social biases in LLMs, particularly those related to race and gender [22, 23, 24]. Gender biases have been documented both in machine translation systems [25, 26] and in generative LLMs. For example, Muti et al. [27] showed that GPT-2 and GPT-3 reproduce gender-stereotypical associations in their representations of feminists in Italian. Likewise, Valderrey and Bloem [28] found that stereotypes such as “Women are beautiful” are consistently encoded by models including LLaMA-2-Chat and LLaMA-2-7B. Finally, Garrido-Munoz et al. [29] showed that adjectives related to physical appearance are more frequently associated with female-subject templates, suggesting that LLMs tend to characterise women through physical attributes.

2.3. Profiling LLM Behaviour Through Human Traits and Perspectives

Several studies in recent years have investigated whether LLM behaviour can be described through human-related traits and profiles. One line of research focuses on *personality induction*, using persona-based prompting or role-playing techniques to elicit outputs consistent with specific personality traits or identities. Early studies explored how predefined personas can shape dialogue generation and user interactions [30, 31], while more recent work has examined the use of persona prompting for applications such as data synthesis, personalisation, and role-playing with LLMs [32, 33, 34, 35, 36]. Beyond predefined personas, Sorensen et al. [37] introduce natural-language value profiles derived from individuals’ previous annotations and use them to predict their held-out judgements. Their approach provides an interpretable representation of human perspectives that can capture behavioural variation beyond demographic attributes alone. However, behaviours induced through persona prompting do not always remain stable across contexts, raising questions about the consistency of personality simulation in LLMs [38]. Gendered personas have also been investigated in relation to commercial voice assistants and chatbots, which are often assigned feminine-coded identities [39]. Similarly, Fadaei et al. [40] show that LLM-based agents may exhibit fluctuating patterns of gender-coded language over time. A complementary line of research examines the behavioural patterns that emerge when LLMs are prompted without explicit persona instructions. Psychometric studies suggest that models may exhibit distinctive and relatively consistent tendencies, which can vary across systems and may be affected by alignment procedures [41, 42]. These findings have motivated the use of human-related traits and profiles as interpretative tools for analysing model behaviour.

Building on the theoretical and experimental background discussed in the previous sections, we approach the analysis of LLM behaviour in GS detection by explicitly accounting for sociodemographic variation in human annotations. Rather than inducing predefined personas or attributing demographic identities to the models, we compare their annotation patterns with those of different human groups.

3. Methodology

This section outlines the methodology followed in this study, which is designed around our two complementary objectives: a sociodemographic analysis of how different groups of people perceive the presence of, or reference to, GSs in short Italian texts, and the sociodemographic profiling of LLMs, aimed at investigating whether, and to what degree, the behaviour of LLMs in a GS detection task aligns with the judgements of human demographic groups (see Figure 1 for an overview of the workflow).

To pursue our objectives, we designed a survey in which native Italian-speaking non-expert participants annotated a subset of 124 texts taken from the GSI:detect dataset by assigning a continuous score reflecting the perceived presence of gender stereotypes (see Section 4.1 for details about survey design and data collection). Participants were recruited through two channels – voluntary participation and the Prolific crowdsourcing platform – and were stratified into four demographic groups defined by the intersection of gender and age.

The resulting annotations were analysed to characterise the sociodemographic composition of the sample and to assess whether different groups exhibit systematically different judgement patterns. As for models, we reused the zero-shot outputs previously submitted to the GSI:detect shared task rather than running the models specifically for the present study.² The group-level annotation profiles were used as reference points to measure how closely the judgement patterns captured by these outputs align with those of each human group.

GSI:detect Sociodemographic Analysis of Human Annotations. The data collected through the survey are analysed along two dimensions. First, we provide a descriptive characterisation of the full participant sample across all the sociodemographic variables we collected, including gender, age, education, geographic provenance, political orientation, and sexual orientation. Second, we focus on the four main demographic groups and examine whether they exhibit systematically different judgement patterns. To this end, each annotator is represented by their personal mean score across all 124 items, and the resulting distributions are compared across groups through inferential statistical tests. This analysis allows us to assess whether demographic background is a significant predictor of how GSs are perceived in text.

Sociodemographic Profiling of LLMs. We employ the zero-shot outputs produced by the selected systems for the same items within the GSI:detect shared task [9]. No additional model inference is performed in the present study. Human groups and models are represented by their scores on the same items, with human scores averaged within each group. The alignment between a model and each human group is quantified through a set of complementary metrics, which capture both the directionality and the absolute proximity of their respective judgements (see below). The goal is to identify the human group that each LLM’s behaviour most closely resembles, while also taking into account the level of confidence associated with this comparison, in order to describe each LLM in sociodemographic terms.

Metrics. We quantify the LLM–human group alignment through three complementary metrics:

i. **Pearson correlation (r)** measures the linear association between the two vectors and serves as the primary criterion for profile assignment, following prior work that uses Pearson correlation to quantify the alignment between model behaviour and human or dataset-specific profiles [43].

ii. **Spearman rank correlation (ρ)** measures agreement in score ranking, making it more robust to distributional differences and discrete LLM outputs. It is used as a confirmatory criterion, with disagreement between Pearson and Spearman indicating ambiguous cases.

iii. **Mean Absolute Error (MAE)** measures the average absolute distance between scores, capturing magnitude differences to contextualise calibration differences and interpret ambiguous assignments.

²For the prompt templates and model-specific implementation details, we refer the reader to the shared task system reports describing the different models, as explained in more detail in Section 5.

4. Data

The present study builds on GSI:detect [10], an Italian linguistic resource specifically designed for the detection and classification of GSs in written texts. The dataset consists of 1,010 short Italian texts manually collected from social media platforms and informative websites, in order to capture naturally occurring language across a variety of communicative contexts (family, gender, politics, etc.). Unlike many existing resources, it focuses on GSs as they spontaneously emerge in everyday written communication, going beyond overtly sexist or hateful discourse and including the ambiguous cases in which stereotypes are produced unconsciously or even with apparently positive intentions.

Each text in GSI:detect is annotated along two dimensions: a GS value, a numerical score in the interval $[0, 1]$ indicating the degree to which the text reflects or refers to a GS, and a GS category, indicating the semantic dimension along which the stereotype is articulated (*role*, *personality*, *competence*, *physical*, *sexual*, or *relational*). For the purposes of the present study, we focus exclusively on the GS value dimension (leaving the categorical annotation aside) and on a subset of 124 selected texts. As an illustrative example, consider, for instance, the following text from the selected subset:

«Quando le donne sono stressate, diventano emotive e si sfasciano, mentre quando gli uomini sono stressati, rimangono calmi e lucidi.»³

4.1. Survey Design and Data Collection

Building on the resource described above, we designed a survey to collect annotations on the GS value dimension from a broader, non-expert population. While the original GSI:detect dataset reflects the perspectives of four trained annotators following pre-defined criteria, our goal here is to investigate how the perception of gender stereotypes varies across the general public, collecting annotations from people whose judgments about the phenomenon solely reflect their experience and sensitivity.

Demographic stratification. Several sociodemographic variables are collected from each participant (described in detail below). Among these, we identified gender and age as the two primary discriminants in this study, on the basis of their well-documented relevance in shaping perceptions of gender-related phenomena [5, 8]. Accordingly, participants are stratified into four target categories defined by the intersection of these two variables: women < 35 , women ≥ 35 , men < 35 , and men ≥ 35 .⁴ The four groups constitute the primary units of analysis throughout the rest of the paper.⁵

Power analysis. To determine the minimum sample size required for reliable group-level estimates and statistically meaningful comparisons, we conducted a Monte Carlo power analysis prior to data collection (for further details, see Appendix A). The simulation generated synthetic annotations in which each judgement was influenced by three factors: the demographic group’s average tendency (base rate), the stereotypicality of the text (fixed per item), and the annotator’s individual sensitivity (fixed per participant). We evaluated designs with $[50, 100]$ annotators per group and $[70, 150]$ texts per annotator, running 200 simulations for each configuration with group-level base rates ranging from 0.50 to 0.65. Each design was assessed against two criteria: *precision*, requiring the 95% confidence interval on each group mean to have a half-width below 0.05; and *power*, requiring pairwise group differences to reach statistical significance ($p < 0.05$, with Bonferroni correction) in at least 80% of the simulations. The minimum suitable design comprised approximately 60–80 annotators per group and 130 texts per annotator. Accordingly, we set up a dataset of 130 texts composed of 124 texts selected

³Translation: «When women are stressed, they become emotional and fall apart, whereas when men are stressed, they remain calm and clear-headed.»

⁴We adopted 35 years as an operational cut-off to distinguish younger and older participants. The choice is consistent with the Italian youth-policy framework, as documented by the institutional Youth Wiki platform, according to which some national initiatives target individuals aged up to 35 [44].

⁵Although our study focuses on women and men, this choice is purely methodological and does not assume a binary view of gender. We acknowledge that we were not able to investigate the perception of gender stereotypes among non-binary individuals, due to the difficulty of recruiting a sample of sufficient size to support robust statistical analysis within the budget and timeframe. The analysis of GS perception among non-binary individuals may be the object of future work.

from the GSI: detect test set (balanced by the original GS values assigned by the trained annotators) and 6 quality-control items to assess participants' attention and annotation reliability.⁶

Survey design. The survey was implemented as a custom web application built on Google Apps Script, with a Google Sheets back-end serving as the database. Each participant was automatically assigned an anonymous incremental identifier (e.g., GS1, GS2, ...) upon opening the form, with no link to personal data; progress was continuously autosaved both locally and server-side, allowing participants to interrupt and resume the task on the same device. Before starting the annotation task, participants were asked to read the task introduction – including an operational definition of *Gender Stereotype*⁷, a content warning, the annotation instructions and quality-control rules, and a privacy statement followed by a mandatory consent checkbox. They were then asked to provide sociodemographic information used in the descriptive and inferential analyses presented in the following sections: GENDER, AGE, LEVEL OF EDUCATION, GEOGRAPHIC PROVENANCE (categorized as Northern, Central, or Southern Italy, or Other), POLITICAL ORIENTATION, and SEXUAL ORIENTATION (specifically, whether the participant identifies as belonging to the *LGBTQIA+* community or not).

In the annotation task, participants are presented one text at a time (e.g., *"Se le donne in pensione ricevono meno, significa che hanno versato meno contributi. Basta con queste continue lamentele!"*⁸), together with an instructional prompt asking how strongly the text expresses or refers to a GS. The judgment is provided by moving a slider on a continuous scale from 0 to 1 (Appendix A.1, Figure 3). To support intuitive interpretation, qualitative anchors are displayed below the slider at evenly spaced positions: *Per nulla* (Not at all), *Poco* (Slightly), *Moderatamente* (Moderately), *Molto* (Strongly), *Estremamente* (Extremely). Prior to public deployment, the survey underwent a pilot phase aimed at refining the user experience of the tool and at estimating the average completion time. During this phase, the form was iteratively administered to a group of five external participants not involved in the project.

Recruitment. The survey was then deployed in two consecutive phases. In the first phase, the form was distributed to the general public on a voluntary, unpaid basis through institutional mailing lists and word-of-mouth dissemination. In the second phase, the survey was administered on the Prolific crowdsourcing platform⁹ until the target sample size was reached, compatibly with the available budget.¹⁰ Besides gender and age, participants in both channels provided the same set of sociodemographic information described above, which is leveraged in the descriptive analysis presented in Section 4.2.

4.2. Data Analysis

In this section we present the analyses carried out on the data collected through the survey. The first part of the analysis is descriptive and inferential in nature: we characterise the composition of the sample along the relevant sociodemographic dimensions and we test whether the four target groups differ in a statistically significant way in their perception of GSs. The second part is concerned with the internal consistency of the collected annotations and with the patterns that emerge at the level of individual items, with the aim of assessing the overall reliability of the data and their suitability to serve as a reference for the LLM profiling experiments presented in the following section.

Sample composition. A total of 281 participants (98 recruited on a voluntary basis and 183 through the Prolific platform) completed the annotation task, with 270 belonging to one of the four target groups defined by the intersection of gender and age. The remaining 11 participants self-identified as *Other* on the gender question and are therefore not included in any of the group-level analyses based on the four target categories, nor in the LLM profiling experiments presented in the following sections; their annotations are however retained for descriptive analyses that do not rely on the man/woman binary distinction (e.g., the distributions in Figure 2 and the inter-annotator agreement in Table 3). Figure 2

⁶The quality-control items were removed from all subsequent analyses and are not part of the distributed dataset.

⁷Defined in the form as *"the belief that men and women have, or should have, certain behaviours or qualities solely because they belong to a given gender"*, accompanied by an illustrative example.

⁸Translation: "If retired women receive less, it means they paid fewer contributions. Enough with all this constant complaining!"

⁹<https://www.prolific.com>

¹⁰On Prolific, recruitment was filtered by the four target demographic categories; 183 participants were recruited for a task taking approximately 35 minutes, for a total cost (including platform service fees) of approximately 1,400£.

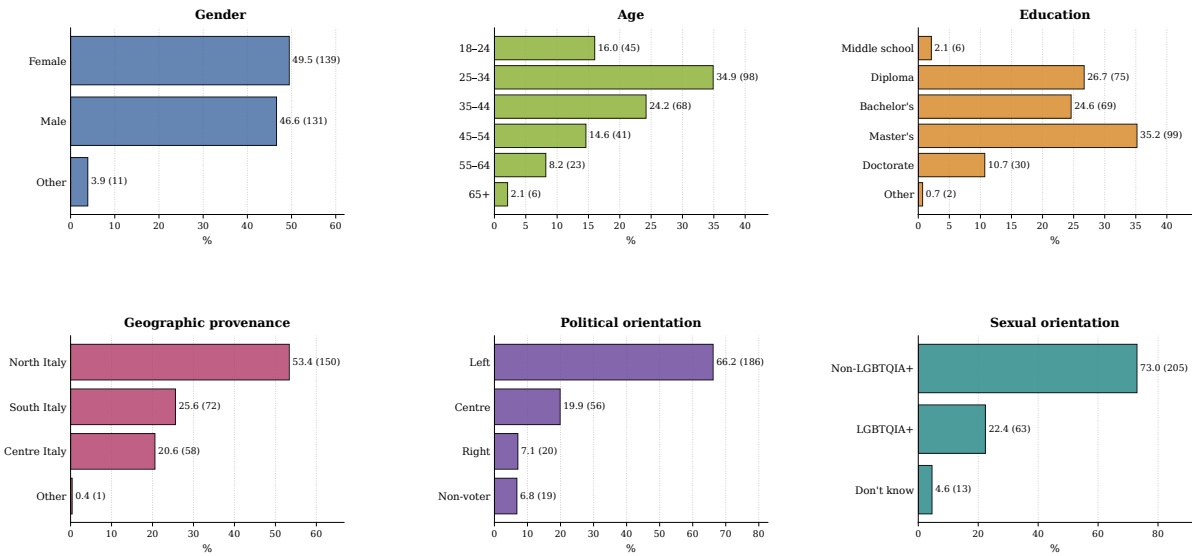


Figure 2: Distribution of the 281 participants along all sociodemographic variables collected through the survey.

provides a general picture of the sample, reporting the distribution of all sociodemographic variables collected through the survey.

Table 1 focuses on the four target groups, reporting group size together with the central tendency of the annotation scores produced by each group. The highest mean score is observed for women ≥ 35 (0.610), followed by women < 35 (0.584), men < 35 (0.558), and men ≥ 35 (0.546). Two descriptive patterns emerge. First, women assign higher GS values than men on average, suggesting a greater sensitivity to the presence of gender stereotypes. Second, age appears to modulate this tendency differently across genders: the mean score increases with age among women but decreases among men. The largest difference is therefore observed between women ≥ 35 and men ≥ 35 , while the two younger groups fall in between.

While the recruitment goal of approximately 80 participants per category was not fully met in the time available for data collection because of practical recruitment difficulties, the final per-group sample sizes remain within the range of configurations identified as valid by the power analysis (Section 4.1).

Group	N	Mean	Median	St.Dev
Women < 35	71	0.584	0.587	0.100
Women ≥ 35	68	0.610	0.613	0.097
Men < 35	66	0.558	0.567	0.111
Men ≥ 35	65	0.546	0.549	0.127
Overall	281	0.578	0.585	0.111

Table 1

Size of the four target demographic groups and descriptive statistics of GS value annotations. Each participant is summarised by their personal mean across the 124 items, and the reported mean, median and standard deviation are computed over these per-annotator values. The overall sample mean (across all 281 participants) is 0.578.

Statistical significance. To assess whether the differences in mean GS value across the four target groups are statistically significant, we represented each annotator by their personal mean score across the 124 items and then compared the resulting per-annotator distributions across groups. A Shapiro-Wilk test confirmed that the distribution of personal mean scores does not deviate significantly from normality in any of the four groups (all $p > 0.05$, see Table 2, top block), allowing us to apply parametric tests. A one-way ANOVA yielded a statistically significant difference between the four groups ($F = 4.52$, $p = 0.0041$), with an effect size of $\eta^2 = 0.049$ that falls in the small-effect range. We then performed pairwise comparisons via Tukey's HSD post-hoc test, which controls the family-wise error rate across

all six pairwise comparisons (Table 2, bottom block). Two comparisons reach statistical significance at $\alpha = 0.05$: women ≥ 35 differ significantly from both men ≥ 35 ($p = 0.0047$) and men < 35 ($p = 0.0327$). All other comparisons – including women < 35 vs. men of either age group – do not reach significance. This suggests that the observed effect cannot be straightforwardly reduced to gender or age alone, but may instead reflect their interplay, with significant differences concentrated in comparisons involving women aged 35 or older, who show a stronger perception of gender stereotypes than both male groups.

<i>(a) Shapiro-Wilk normality test</i>		
Group	W	p-value
Women < 35	0.9936	0.9772
Women ≥ 35	0.9854	0.6114
Men < 35	0.9720	0.1406
Men ≥ 35	0.9741	0.1904
<i>(b) One-way ANOVA</i>		
$F = 4.52, p = 0.0041, \eta^2 = 0.049$		
<i>(c) Tukey HSD pairwise comparisons</i>		
Comparison	Mean diff.	p-adj
Women ≥ 35 vs. Men ≥ 35	+0.064	0.0047
Women ≥ 35 vs. Men < 35	+0.052	0.0327
Women ≥ 35 vs. Women < 35	+0.026	0.4895
Women < 35 vs. Men ≥ 35	+0.038	0.1865
Women < 35 vs. Men < 35	+0.026	0.5218
Men < 35 vs. Men ≥ 35	+0.012	0.9194

Table 2: Statistical tests on per-annotator mean GS value across four groups. Each annotator contributes one value (the mean of their 124 annotations). Significant pairwise comparisons ($p < 0.05$) in bold.

Variable	Group	N	Alpha
–	All participants	281	0.42
Gender $\times$ Age	Women < 35	71	0.48
	Women ≥ 35	68	0.39
	Men < 35	66	0.47
	Men ≥ 35	65	0.37
Political orientation	Left	186	0.47
	Centre	56	0.40
	Right	20	0.34
	Non-voter	19	0.28
Sexual orientation	LGBTQIA+	63	0.48
	Non-LGBTQIA+	205	0.41
Geographic provenance	North Italy	150	0.40
	Centre Italy	58	0.45
	South Italy	72	0.45

Table 3: Within-group inter-annotator agreement (Krippendorff’s Alpha, interval-level). Higher values indicate greater internal cohesion in the perception of GSs.

Inter-annotator agreement. We computed inter-annotator agreement using Krippendorff’s Alpha, taking into account the numerical distance between annotation scores, both for the full sample and separately for each demographic group (Table 3). The overall agreement is moderate, with $\alpha = 0.42$: while annotators broadly converge in their perception of gender stereotypes, substantial individual variation remains. A consistent pattern emerges within the four target groups: younger participants exhibit higher agreement than older ones in both genders (women < 35 : $\alpha = 0.48$ vs. women ≥ 35 : $\alpha = 0.39$; men < 35 : $\alpha = 0.47$ vs. men ≥ 35 : $\alpha = 0.37$). In this sample, age is more clearly associated with within-group cohesion than gender. However, this pattern may partly reflect the composition of the age groups rather than a general effect of age: the younger groups cover a relatively narrow age range (18–34), whereas the older groups include participants from 35 to 65 years and above. The latter may therefore encompass more heterogeneous life stages, experiences, and perspectives, which could contribute to their lower internal agreement. Consequently, this finding should be interpreted as sample-specific and requires confirmation in larger samples with more finely balanced age ranges. Moreover, two further patterns deserve attention. First, participants who identify as belonging to the LGBTQIA+ community exhibit higher internal agreement than non-LGBTQIA+ participants ($\alpha = 0.48$ vs. 0.41); this is consistent with the hypothesis that belonging to a group potentially targeted by gender stereotypes correlates with a more shared and articulated sensitivity to the phenomenon. Second, descriptive differences emerge across political-orientation groups, with higher agreement among left-wing participants ($\alpha = 0.47$) than among centrist ($\alpha = 0.40$), right-wing ($\alpha = 0.34$), and non-voting participants ($\alpha = 0.28$). Given the strong imbalance in group sizes ($N = 186, 56, 20$, and 19, respectively), these estimates, particularly those of the smaller groups, should be regarded as exploratory and may partly reflect the composition of the sample. By contrast, geographic provenance does not show any meaningful effect: agreement levels are comparable across the three Italian areas (North: $\alpha = 0.40$; Centre: $\alpha = 0.45$; South: $\alpha = 0.45$), with the North showing slightly lower internal

cohesion. We also computed mean scores and saw that they are nearly identical (0.58, 0.56, 0.59). This suggests that regional origin is not a confounding factor in our analysis.

5. Experiments

This section presents the selected LLMs and the experiments conducted to characterise their behaviour in sociodemographic terms. We investigate whether each model’s pattern of judgements in the GS detection task aligns more closely with one of the four demographic groups. Rather than evaluating absolute model performance against a gold standard, we treat the four annotation profiles as alternative reference points and measure the degree of alignment between each model and each group.

Experimental setup. As described in Section 3, the analysis compares human and model judgements on the same 124 texts. Both the four human demographic groups and each LLM are represented as 124-dimensional vectors of GS values, one entry per item in the subset described in Section 4. For the human groups, the vector entries are obtained by averaging, item by item, the scores assigned by all annotators belonging to that group; for the LLMs, the vector entries are simply the GS values returned by the model when prompted on each of the 124 items. The alignment between an LLM vector and a human group vector is then quantified through three different metrics (Pearson correlation, Spearman rank correlation, and MAE), each computed over the 124 aligned pairs of scores. This procedure is repeated for every model against every human group, yielding for each LLM a profile of alignment scores across the four target groups. Following the criterion defined in Section 3, the group with the highest Pearson correlation is reported as the model’s highest-alignment profile; Spearman is used as a confirmatory criterion, and cases in which Pearson and Spearman disagree are flagged and discussed in light of the corresponding MAE values. Both the human annotations and the evaluated LLM outputs are expressed as scores in the $[0, 1]$ interval.

Models. The models analysed in this study were drawn from the pool of systems participating in the GS Detection task within the EVALITA 2026 GSI: detect shared task, including both submissions from the participating teams [45, 46, 47, 48] and the baseline models that we tested in our own contribution as task organisers [9]. From this pool, we selected 7 systems with the aim of representing a diverse set of LLM families and sizes, including both proprietary and open-weight models, without the intention of isolating differences across model families at comparable scale, nor differences across model sizes within the same family. All systems are considered in their zero-shot configuration, since our objective is to profile the socio-demographic tendencies that emerge from their post-training abilities, without any form of in-context learning, fine-tuning, or other adaptations that could alter such tendencies.¹¹

The final selection of systems comprises Gemini 2.5-Flash [49], whose experimental setup is described in Festa [47]; Llama 3.3-70B [50], described in Greco et al. [46]; Gemma 3-12B [51], described in Labruna [48]; GPT-5 [52], described in Sorokoletova et al. [45]¹²; and GPT-4o [53], GPT-5-nano [52], and Qwen3-14B [54], described in Comandini et al. [9]; see the papers describing each submission for information about prompt templates and inference details for each zero-shot setting run.

6. Results and Discussion

This section presents the outcome of the sociodemographic profiling experiments described in Section 5. We report here the full set of alignment scores between each LLM and each human demographic group

¹¹Some of the models that participated in the shared task are not included in this study, either because the submitting team did not provide any zero-shot run, or because no system description was made available.

¹²Due to the experimental design adopted by the shared task participants, GPT-5 was prompted to output only discrete values (0, 0.25, 0.50, 0.75, and 1), unlike the other models considered in this study. For the present analysis, these values were treated uniformly as numerical scores within the $[0, 1]$ continuum.

(Table 4), discussing the patterns that emerge, with particular attention to the cases in which the three metrics agree on the assigned profile, the cases in which they diverge, and the broader trends that connect these results to the sociodemographic findings of Section 4.2.

We assessed the **statistical robustness of the observed profile alignments** using a hierarchical bootstrap procedure (described in Appendix B).¹³ Although all model–group correlations were significantly different from zero, none of the models showed a highest Pearson correlation that was significantly greater than those of all alternative profiles.¹⁴ Accordingly, the profile labels in Table 4 should be interpreted as indicating relative alignment rather than statistically unique demographic classifications.

Model	Metric	W <35	W ≥35	M <35	M ≥35	Highest-alignment profile
Gemini 2.5-Flash	Pearson ↑	0.766	0.769	0.761	0.761	W/M ≥35*
	Spearman ↑	0.768	0.765	0.769	0.784	
	MAE ↓	0.215	0.215	0.228	0.251	
Gemma 3-12B	Pearson ↑	0.625	0.625	0.634	0.619	M <35
	Spearman ↑	0.577	0.570	0.577	0.560	
	MAE ↓	0.233	0.237	0.235	0.253	
GPT-4o	Pearson ↑	0.777	0.765	0.780	0.762	M <35*
	Spearman ↑	0.744	0.707	0.740	0.718	
	MAE ↓	0.126	0.121	0.138	0.144	
GPT-5	Pearson ↑	0.797	0.791	0.818	0.787	M <35
	Spearman ↑	0.786	0.772	0.808	0.786	
	MAE ↓	0.197	0.208	0.195	0.222	
GPT-5-nano	Pearson ↑	0.594	0.598	0.602	0.577	M <35*
	Spearman ↑	0.579	0.585	0.579	0.562	
	MAE ↓	0.188	0.183	0.189	0.200	
Llama 3.3-70B	Pearson ↑	0.734	0.729	0.696	0.711	W <35
	Spearman ↑	0.686	0.685	0.648	0.682	
	MAE ↓	0.144	0.139	0.156	0.155	
Qwen3-14B	Pearson ↑	0.588	0.602	0.574	0.571	W ≥35
	Spearman ↑	0.534	0.555	0.517	0.525	
	MAE ↓	0.188	0.178	0.199	0.202	

Table 4

Alignment scores between each LLM and the four human demographic groups (W = Women, M = Men). The best-aligned group per metric is in bold. The highest-alignment profile is identified primarily through Pearson correlation, with Spearman used as a confirmatory measure; cases marked with * indicate disagreement between the two correlation metrics. Since none of the highest observed Pearson correlations was significantly greater than all alternative correlations (see Appendix B), profile labels should therefore be interpreted as descriptive closest matches.

Consistent descriptive alignments. For four out of the seven systems – GPT-5, Llama 3.3-70B, Gemma 3-12B, and Qwen3-14B – Pearson and Spearman identify the same highest-alignment group. GPT-5 stands out as the system with the highest overall alignment with human judgements, with Pearson and Spearman above 0.78 across all four groups and the strongest alignment with men <35 profile (Pearson = 0.818, Spearman = 0.808), which is also the group selected most frequently in the bootstrap samples: approximately 74% under Pearson and 71% under Spearman (see Appendix B). This makes GPT-5 the model with the most internally consistent descriptive tendency in the pool, although this should be interpreted as the most stable tendency rather than as evidence of a uniquely

¹³Since Pearson correlation serves as our primary criterion for identifying the highest-alignment profile, this inferential analysis focuses on whether the highest observed Pearson correlation is significantly greater than the correlations with the alternative demographic groups. The same bootstrap replicates are also used to descriptively assess profile stability under Spearman correlation and MAE, which serve as complementary criteria.

¹⁴This may partly reflect the limited number of evaluation items ($N = 124$), since a larger item set would yield more precise estimates and narrower confidence intervals. However, the observed differences between the group correlations are also generally small, suggesting that limited statistical power is unlikely to be the only explanation and that the human demographic profiles themselves may share a substantial common structure.

differentiated demographic alignment. Qwen3-14B displays a similarly coherent, although weaker, tendency toward women ≥ 35 . Pearson, Spearman, and MAE all identify this group as the closest profile, and women ≥ 35 are selected in approximately 59% of the Pearson bootstrap samples, 66% of the Spearman samples, and 87% of the MAE samples. For Llama 3.3-70B, Pearson and Spearman both identify women <35 , whereas MAE identifies women ≥ 35 . The model therefore shows a tendency toward the female profiles, but the distinction between the two age groups is not stable across metrics. In particular, the Pearson difference between women <35 and women ≥ 35 is only 0.005. Gemma 3-12B shows its highest Pearson and Spearman correlations with men <35 , but with lower correlation values than GPT-5 (Pearson = 0.634, Spearman = 0.577): the demographic signal is consistent across metrics, but the overall alignment with the human data is weaker, suggesting that the model captures the general direction of human judgements only moderately well, independently of the specific group.

Ambiguous alignments. For the remaining three systems – Gemini, GPT-4o, and GPT-5-nano – Pearson and Spearman point to different highest-alignment groups, while MAE provides an additional and sometimes divergent indication. Gemini represents the clearest case of limited profile differentiation, whose Pearson values are nearly flat across all four groups (range 0.761–0.769), and none of the pairwise differences approaches statistical significance. Pearson and MAE identify women ≥ 35 , while Spearman identifies men ≥ 35 . Rather than supporting a specific demographic profile, these results indicate that Gemini does not exhibit a clearly distinguishable demographic alignment within the present sample. The cases of GPT-4o and GPT-5-nano are structurally similar and methodologically informative. In both systems, Pearson converges on men <35 , while the smallest MAE is observed with women ≥ 35 . This dissociation between correlation-based and distance-based metrics suggests that both models share the same *directional* pattern as men <35 (i.e., they identify the same items as more or less stereotyped) but produce GS values whose *absolute magnitude* is closer to that of women ≥ 35 . It is worth noting that GPT-4o exhibits the lowest MAE in the entire pool (MAE = 0.121 with women ≥ 35), making it the best-calibrated model in absolute terms with the lowest absolute deviation from the human group scores, despite not achieving the highest correlations. This case illustrates the value of the three-metric design proposed in Section 3: relying on a single metric would have masked this dissociation between the direction of judgement and the intensity of the response.

Connecting LLM profiles to the human sociodemographic findings. Three broader patterns emerge when the LLM profiles are read together with the findings of Section 4.2. First, Llama 3.3-70B is the only model whose highest Pearson and Spearman correlations are observed for women <35 , despite this being the most internally cohesive human group in our sample, with the highest within-group Krippendorff’s Alpha ($\alpha = 0.48$, Table 3). The results therefore do not show that the most internally cohesive human profile is uniquely reproduced by the models, which instead tend to align with groups that exhibit lower internal cohesion (e.g., men under and over 35). Second, we observe that all three GPT-based systems (GPT-5, GPT-4o, GPT-5-nano) obtain their highest Pearson correlation with men < 35 , while the other models distribute differently. This family-level regularity is noteworthy from a descriptive point of view, particularly because GPT-5 also shows substantial bootstrap stability in this direction (Appendix B). However, the pattern is not supported by statistically significant pairwise differences, and GPT-4o and GPT-5-nano show disagreement across Pearson, Spearman, and MAE. It should therefore be treated as a preliminary observation rather than as evidence of a systematic effect of model family, training data, or alignment procedures.¹⁵ Third, the ambiguous cases of GPT-4o and GPT-5-nano show that the direction and the intensity of a model’s judgements can dissociate: a model can agree with one demographic group on which items are stereotyped while expressing this judgement at an intensity closer to that of a different group. This dissociation, made visible by the joint use of correlation- and distance-based metrics, suggests that single-metric evaluations may oversimplify the way models reproduce (or fail to reproduce) specific human perspectives.

¹⁵The small number of systems per family and the absence of statistically differentiated demographic profiles prevents us from drawing conclusions about model-family effects.

7. Conclusion

In this work we addressed two complementary objectives: characterising how the perception and identification of gender stereotypes varies across human demographic groups, and using the resulting group-level annotation profiles as a reference to investigate the sociodemographic alignment of LLMs in the same task. Building on the GSI:detect dataset, we collected a new set of annotations from non-expert participants stratified by gender and age, and we compared the resulting human profiles with the zero-shot outputs of a set of selected LLMs.

Our findings can be summarised as follows: (i) demographic background has a statistically significant effect on the perception of gender stereotypes, with the divergence concentrated at the intersection of gender and age: women over 35 differ significantly from men above and below this age, while no other pairwise comparison reaches significance; (ii) age emerges as a stronger predictor of internal annotator cohesion than gender, with younger participants of both genders consistently showing higher within-group agreement than older ones; (iii) LLMs exhibit alignment patterns that can be meaningfully characterised along sociodemographic dimensions, with substantial variation across model families and configurations rather than a uniform behaviour; (iv) the most internally cohesive human group (women under 35) is also the least replicated by the models, with only one system out of seven aligning with it, while the three GPT-based systems investigated tend to align most closely with the men under 35 profile in terms of judgement direction; and (v) the joint use of correlation- and distance-based metrics reveals that a model can identify the same items as stereotyped as a given demographic group, but score them with a different intensity, often closer to that of another group. This shows that a model’s sociodemographic profile is not a single, monolithic label, but can decompose into independent dimensions of perception and calibration, a distinction that single-metric evaluations would have masked, and that we believe should be taken into account in future profiling studies of subjective tasks.

8. Limitations

The limitation of this work concerns the size of the annotator and evaluation samples. Although the per-group sample sizes fall within the range identified as valid by the power analysis (Section 4.1), due to time and budget constraints we did not reach the original recruitment target of approximately 80 participants per demographic group. Larger samples would strengthen the robustness and generalisability of the reported patterns, while also increasing the precision and statistical power of the comparisons between demographic profiles. This could help determine whether the small observed differences reflect genuinely overlapping group profiles or insufficient data to detect more subtle alignment effects. A further limitation concerns the use of scalar scores: human scale calibration may vary, and direct LLM score elicitation may be unreliable. However, this study reuses the outputs already available from the original task on which the profiling analysis is based (GSI:detect shared task).

Acknowledgments

We would like to thank the survey participants who were recruited voluntarily and unpaid for their valuable contributions. Davide Testa conducted the experimental work presented in Sections 3, 5, and 6; Gloria Comandini contributed Sections 1 and 2; Sofia Brenna and Manuela Speranza contributed Section 4. This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with FBK.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 4.7 in order to: perform grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] L. Arcuri, M. Cadinu, *Gli Stereotipi. Dinamiche psicologiche e contesto delle relazioni sociali*, Il Mulino, Bologna, 1998.
- [2] S. Cavagnoli, F. Dragotto, *Sessismo*, Mondadori, Milano, 2021.
- [3] G. Peruzzi, V. Bernardini, R. Lombardi, Le questioni di genere dei giovani. un'indagine sulle percezioni e le esperienze di studenti e studentesse universitari., in: G. Peruzzi, V. Bernardini, R. Lombardi, C. Rinaldi, M. Bacio, L. Bainotti, G. Viggiani (Eds.), *Il bias del gender*, Durango, Andria, 2019, pp. 13–50.
- [4] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, S. M., An italian twitter corpus of hate speech against immigrants, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Paris, 2018, pp. 2798–2805.
- [5] M. Wojatzki, T. Horsmann, D. Gold, T. Zesch, Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgement, in: A. Barbaresi, H. Biber, F. Neubarth, R. Osswald (Eds.), *Proceedings of the 14th conference on Natural Language Processing (KONVENS 2018)*, 2018, pp. 110–120.
- [6] V. Basile, F. Cabitza, A. Campagner, Toward a perspectivist turn in ground truthing for predictive computing., in: *Toward a Perspectivist Turn in Ground Truthing for Predictive Computing*, Association for the Advancement of Artificial Intelligence, Washington DC, 2023, pp. 6860–6868.
- [7] B. Muscato, C. Sree Mala, M. Marchiori Manerba, G. Gezici, F. Giannotti, An overview of recent approaches to enable diversity in large language models through aligning with human perspectives, in: G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, S. Tonelli (Eds.), *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, ELRA and ICCL, 2024, p. 49–55.
- [8] D. Sue, *Microaggressions in everyday life. Race, gender and sexual orientation*, John Wiley & Sons, Hoboken, 2010.
- [9] G. Comandini, M. Speranza, S. Brenna, D. Testa, S. Cavagnoli, B. Magnini, GSI:detect at EVALITA 2026: Overview of the task on detecting gender stereotypes in Italian, in: *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, CEUR.org, Bari, Italy, 2026.
- [10] D. Testa, S. Brenna, M. Speranza, G. Comandini, S. Cavagnoli, B. Magnini, Gsi:detect - a perspectivist approach to gender stereotypes identification in italian, in: *Proceedings of the Perspectivist Approaches to Disagreement in NLP Workshop*, ELRA Language Resources Association (ELRA), 2026, pp. 21–32.
- [11] V. Basile, It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks, in: G. Vizzari, M. Palmonari, A. Orlandini (Eds.), *Proceedings of the AIXIA 2020 Discussion Papers Workshop*, CEUR Workshop Proceedings, 2020, pp. 31–40.
- [12] V. Prabhakaran, A. Mostafazadeh Davani, M. Diaz, On releasing annotator-level labels and information in datasets, in: C. Bonial, N. Xue (Eds.), *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 133–138. URL: <https://aclanthology.org/2021.law-1.14/>. doi:10.18653/v1/2021.law-1.14.
- [13] A. Mostafazadeh Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, *Transactions of the Association for Computational Linguistics* 10 (2022) 92–110. URL: <https://aclanthology.org/2022.tacl-1.6/>. doi:10.1162/tacl_a_00449.
- [14] E. Fleisig, S. L. Blodgett, D. Klein, Z. Talat, The perspectivist paradigm shift: Assumptions and challenges of capturing human labels, in: K. Duh, H. Gomez, S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, 2024, pp. 104–118.

- tics, Mexico City, Mexico, 2024, pp. 2279–2292. URL: <https://aclanthology.org/2024.naacl-long.126/>. doi:10.18653/v1/2024.naacl-long.126.
- [15] B. Chen, X. Wang, S. Peng, R. Litschko, A. Korhonen, B. Plank, “seeing the big through the small”: Can LLMs approximate human judgment distributions on NLI from a few explanations?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 14396–14419. URL: <https://aclanthology.org/2024.findings-emnlp.842/>. doi:10.18653/v1/2024.findings-emnlp.842.
 - [16] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), EVALITA Evaluation of NLP and Speech Tools for Italian. Proceedings of the Final Workshop, Accademia University Press, Torino, 2018, pp. 59–66.
 - [17] F. Poletto, V. Basile, C. Bosco, V. Patti, M. Stranisci, Annotating hate speech: Three schemes at comparison, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), CEUR Workshop Proceedings, Bari, Italy, 2019, pp. 304–311. URL: <https://aclanthology.org/2019.clicit-1.46/>.
 - [18] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, N. A. Smith, Annotators with attitudes: How annotator beliefs and identities bias toxic language detection, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5884–5906. URL: <https://aclanthology.org/2022.naacl-main.431/>. doi:10.18653/v1/2022.naacl-main.431.
 - [19] N. Tahaei, S. Bergler, Analysis of annotator demographics in sexism detection, in: A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, D. Nozza (Eds.), Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 376–383. URL: <https://aclanthology.org/2024.gebnlp-1.24/>. doi:10.18653/v1/2024.gebnlp-1.24.
 - [20] A. Jiang, N. Vitsakis, T. Dinkar, G. Abercrombie, I. Konstas, Re-examining sexism and misogyny classification with annotator attitudes, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15103–15125. URL: <https://aclanthology.org/2024.findings-emnlp.887/>. doi:10.18653/v1/2024.findings-emnlp.887.
 - [21] G. Abercrombie, N. Vitsakis, A. Jiang, I. Konstas, Revisiting annotation of online gender-based violence, in: G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, S. Tonelli (Eds.), Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 31–41. URL: <https://aclanthology.org/2024.nlperspectives-1.3/>.
 - [22] X. Dong, Y. Wang, P. S. Yu, J. Caverlee, Disclosure and mitigation of gender bias in llms, arXiv preprint arXiv:2402.11190 (2024).
 - [23] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of The ACM Collective Intelligence Conference (CI ’23), Association for Computing Machinery, 2023, pp. 12–24.
 - [24] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 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), ACM, 2023, pp. 610–623.
 - [25] B. Savoldi, J. Bastings, L. Bentivogli, E. Vanmassenhove, A decade of gender bias in machine translation, *Patterns* 6 (2025).
 - [26] O. M. Mastromichalakis, G. Filandrianos, M. Symeonaki, G. Stamou, Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 32233–32249.
 - [27] A. Muti, E. Bassignana, E. Moscato, D. Nozza, Probing feminist representations: A study of bias in

- LLMs and word embeddings, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 815–825. URL: <https://aclanthology.org/2025.clicit-1.77/>.
- [28] A. Valderrey, J. Bloem, Investigating how llms propagate female stereotypes: Comparing what models say via prompts with what they represent in their embeddings, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), 2026, pp. 77–92.
- [29] I. Garrido-Munoz, F. Martínez-Santiago, A. Montejo-Raez, From body to mind: Analyzing gender representation in spanish generative language models, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), 2026, pp. 3861–3874.
- [30] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, B. Dolan, A persona-based neural conversation model, arXiv preprint arXiv:1603.06155 (2016).
- [31] S. Zhang, Y. Du, G. Liu, Z. Yan, Y. Cao, G4: grounding- guided goal-oriented dialogues generation with multiple documents, in: Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, 2022, pp. 108–114.
- [32] P. Hämäläinen, M. Tavast, A. Kunnari, Evaluating large language models in generating synthetic hci research data: a case study, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–19.
- [33] S. Mao, N. Zhang, X. Wang, M. gru Wang, Y. Yao, Y. Jiang, P. Xie, F. Huang, H. Chen, Editing personality for llms, arXiv preprint arXiv:2310.02168 (2023).
- [34] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models, arXiv preprint arXiv:2307.00184 (2023).
- [35] M. Shanahan, K. McDonell, L. Reynolds, Role play with large language models, *Nature* 623 (2023) 493–498.
- [36] C. Zaghi, Team prisma at gsi:detect: Comparing pb&j persona-based and few-shot approaches to gender stereotype detection at evalita 2026, in: F. Cutugno, A. Miaschi, A. Aproso, G. Rambelli, L. Siciliani, M. S. (eds.) (Eds.), Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, 2026.
- [37] T. Sorensen, P. Mishra, R. Patel, M. H. Tessler, M. A. Bakker, G. Evans, I. Gabriel, N. Goodman, V. Rieser, Value profiles for encoding human variation, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 2047–2095. URL: <https://aclanthology.org/2025.emnlp-main.106/>. doi:10.18653/v1/2025.emnlp-main.106.
- [38] P. Rajput, Y. Song, I. E. Olatunji, J. Klein, T. F. Bissyandé, Evaluation drift in llm personality induction: Are we moving the goalpost?, in: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), 2026, pp. 11272–11285.
- [39] J. C. Moran, Artificial intelligibility: the role of gender in assigning humanness to natural language processing systems, *Journal of Gender Studies* 34 (2025) 300–310.
- [40] F. Fadaei, J. C. Moran, T. Yasseri, Gender dynamics and homophily in a social network of llm agents, arXiv preprint arXiv:2602.02606v2 (2026).
- [41] S. Lee, S. Lim, S. Han, G. Oh, H. Chae, J. Chung, M. Kim, B. woo Kwak, Y. Lee, D. Lee, J. Yeo, Y. Yu, Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics, arXiv preprint arXiv:2406.14703v2 (2024).
- [42] T. F. Heston, J. Gillette, Large language models demonstrate distinct personality profiles, *Cureus* 17 (2025) e84706.
- [43] S. Santy, J. Liang, R. Le Bras, K. Reinecke, M. Sap, NLPositionality: Characterizing design biases of datasets and models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),

- Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9080–9102. URL: <https://aclanthology.org/2023.acl-long.505/>. doi:10.18653/v1/2023.acl-long.505.
- [44] Youth Wiki, Italy: 1.1 target population of youth policy, National Policies Platform, European Union, 2026. URL: <https://national-policies.eacea.ec.europa.eu/youthwiki/chapters/italy/11-target-population-of-youth-policy>, last updated: 30 March 2026. Accessed: 6 June 2026.
 - [45] O. E. Sorokoletova, E. Musumeci, D. Nardi, Diag-sapienza at gsi:detect: A technical report, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
 - [46] S. Greco, M. La Quatra, M. Marchiori Manerba, R. Muñoz Sánchez, A. T. Cignarella, Stereobusters at gsi:detect: Llm-based detection and human qualitative analysis of gender stereotypes in italian short texts, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
 - [47] D. Festa, Festa at gsi:detect: In-context learning vs. fine-tuning for gender stereotype detection, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
 - [48] T. Labruna, Tiz at gsi:detect: Modeling gender stereotypes detection as multi-category gender stereotype scoring, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
 - [49] Google, Gemini 2.5 Flash – model card, <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf>, 2025. Last updated: December 2025. Accessed: 2026-06-06.
 - [50] A. G. et al., The llama 3 herd of models, 2024. URL: <https://arxiv.org/abs/2407.21783>. arXiv:2407.21783.
 - [51] G. Team, A. K. et al., Gemma 3 technical report, 2025. URL: <https://arxiv.org/abs/2503.19786>. arXiv:2503.19786.
 - [52] OpenAI, Gpt-5 system card, <https://cdn.openai.com/gpt-5-system-card.pdf>, 2025. Accessed: October 2025.
 - [53] OpenAI, Gpt-4o system card, 2024. URL: <https://arxiv.org/abs/2410.21276>. arXiv:2410.21276.
 - [54] Qwen-Team, Qwen3 technical report, 2025. URL: <https://arxiv.org/abs/2505.09388>. arXiv:2505.09388.

A. Defining Survey Design

A.1. Additional Materials

We present an example prompt fed to LLMs in the context of the GSI:detect shared task for predicting the GS value:¹⁶

«Voglio sapere se la seguente frase contiene uno stereotipo di genere o un riferimento ad uno stereotipo di genere. Non rispondere con sì o no ma dammi un valore numerico compreso tra 1 (frase molto stereotipata) e 0 (frase poco stereotipata) usando fino a due cifre decimali.»

- English Translation:

«I would like to know whether the following sentence contains or refers to a gender stereotype. Do not answer yes or no; instead, provide a numerical value between 0 (the sentence is not or only weakly stereotypical) and 1 (the sentence is highly stereotypical), using up to two decimal places.»

Below an example of how we proposed the annotation survey to the annotators.

¹⁶Each model-specific prompt template, as designed by the GSI:detect shared task participants, are provided in the systems' reports referenced in Section 5.

Figure 3: Example of item annotation from the Survey.

A.2. Power analysis

To determine the minimum sample size required to support both reliable group-level estimates and statistically meaningful comparisons between the four groups, we conducted a Monte Carlo power analysis prior to data collection. The simulation generates synthetic annotations that mimic how real participants would behave: each judgment is influenced by the average tendency of the annotator’s demographic group (its base rate), by how stereotypical the text actually is (a fixed value drawn for each text), and by the personal sensitivity of the individual annotator (a fixed value drawn for each participant). We tested many possible study designs by varying the number of annotators per group and the number of texts each annotator would label. For each design, we ran 200 simulations using group-level base rates between 0.50 and 0.65, chosen to reflect plausible variation across the four demographic groups. Each design was evaluated against two criteria:

- **Precision** – the 95% confidence interval on the mean GS value of each group must have a half-width below 0.05, ensuring sufficiently narrow group-level estimates.
- **Power** – pairwise comparisons between groups (e.g., men ≥ 35 vs. men < 35 ; women vs. men) must reach statistical significance ($p < 0.05$ with Bonferroni correction for multiple testing) in at least 80% of the simulations.

The grid explored configurations with [50, 100] annotators per group and [70, 150] texts per annotator. The analysis identified configurations in the neighborhood of approximately 60–80 annotators per group and 130 texts per annotator as the minimum design satisfying both criteria.

Based on the power analysis, we selected a subset of 124 texts from the GSI:detect test set, balanced with respect to the GS values assigned by the original trained annotators. To these we added 6 additional quality-control items used to verify participants’ attention and annotation reliability, bringing the total annotation load to 130 items per participant – in line with the design identified by the power analysis.¹⁷ We then aimed at recruiting approximately 80 annotators per group, slightly above the minimum threshold identified by the simulation to account for incomplete or invalid submissions.

B. Bootstrap Analysis of Demographic Profile Comparisons

The significance tests associated with the Pearson and Spearman coefficients reported in Table 4 evaluate whether each individual model–group correlation differs from zero. They do not establish whether the demographic group with the highest observed correlation is significantly better aligned with the model than the alternative groups. We therefore performed a hierarchical bootstrap analysis to quantify the uncertainty of the profile comparisons.

¹⁷The quality-control items were subsequently removed from all analyses and are not part of the dataset reported in the following sections.

For each of 10,000 bootstrap replicates, we resampled annotators with replacement within each demographic group and resampled the 124 evaluation items with replacement. The resampled items were used for all four groups, preserving the paired structure of the comparisons. Group profiles were reconstructed in each replicate, and Pearson correlations between each LLM and each demographic profile were recomputed. We then calculated bootstrap confidence intervals for the pairwise differences between correlations. Bootstrap-based two-sided p -values were also computed for all six pairwise comparisons between demographic groups and adjusted within each model using Holm’s correction. The procedure accounts for uncertainty arising from both the sampled annotators and the sampled evaluation items.

Table 5 summarises the stability of the demographic profile with the highest Pearson correlation for each model. The *selection frequency* is the proportion of the 10,000 bootstrap replicates in which the reported group obtained the highest Pearson correlation among the four profiles. It therefore indicates how consistently the same group remains the highest-alignment profile when annotators and items are resampled, but should not be interpreted as a probability that the assignment is correct or as a test of statistical significance.

The *runner-up* is defined as the demographic group with the second-highest Pearson correlation in the original sample. We report Δr as the difference between the Pearson correlation of the highest-alignment profile and that of the runner-up. This represents the most conservative pairwise comparison, because the runner-up is the closest observed alternative to the highest-alignment profile. The 95% bootstrap confidence interval quantifies the uncertainty of this difference: an interval including zero indicates that the two profiles cannot be statistically distinguished.

Model	Highest profile	Selection freq.	Δr vs. runner-up [95% CI]
Gemini 2.5-Flash	$W \geq 35$	33.6%	0.003 [−0.047, 0.048]
Gemma 3-12B	$M < 35$	38.8%	0.009 [−0.046, 0.066]
GPT-4o	$M < 35$	43.0%	0.003 [−0.038, 0.046]
GPT-5	$M < 35$	74.0%	0.021 [−0.021, 0.070]
GPT-5-nano	$M < 35$	39.6%	0.004 [−0.060, 0.072]
Llama 3.3-70B	$W < 35$	55.2%	0.005 [−0.044, 0.057]
Qwen3-14B	$W \geq 35$	59.3%	0.014 [−0.044, 0.068]

Table 5

Hierarchical bootstrap analysis of the demographic profiles with the highest observed Pearson correlation. Selection frequency reports the proportion of the 10,000 bootstrap replicates in which the reported group obtained the highest Pearson correlation. Δr is the observed difference between the highest-alignment profile and the runner-up. All 95% confidence intervals include zero.

The results show different degrees of stability across models. GPT-5 exhibits the clearest tendency: Men < 35 remain the highest-correlated profile in 74.0% of the bootstrap replicates. Qwen3-14B and Llama 3.3-70B show moderately stable tendencies toward Women ≥ 35 and Women < 35 , respectively. By contrast, the selection frequencies for Gemini, Gemma 3-12B, GPT-4o, and GPT-5-nano are more evenly distributed across demographic groups, indicating uncertainty in their descriptive profile matches.

Nevertheless, the 95% confidence interval for the difference between the highest and second-highest correlations includes zero for every model. The set of pairwise comparisons, including the Holm-adjusted tests, leads to the same conclusion: no model’s highest Pearson correlation is significantly greater than its correlations with all three alternative demographic profiles. Accordingly, the demographic profiles discussed in the main text represent relative alignment tendencies rather than statistically unique or categorical assignments.

Complementary stability under Spearman correlation and MAE. Although Pearson correlation serves as the primary criterion for descriptive profile identification, we also use the same bootstrap replicates to assess the stability of the profile matches under Spearman correlation and MAE. For each replicate, we record the demographic group obtaining the highest Spearman correlation and the group obtaining the lowest MAE, respectively. The resulting selection frequencies are reported in Table 6.

These frequencies provide a descriptive indication of how consistently each complementary metric identifies the same closest-matching group under resampling. Unlike the Pearson-based analysis above, we do not use these results to perform pairwise inferential comparisons between profiles. They should therefore not be interpreted as tests of statistical significance or as probabilities that a demographic assignment is correct.

Model	Spearman		MAE	
	Highest profile	Selection freq.	Closest profile	Selection freq.
Gemini 2.5-Flash	$M \geq 35$	51.9%	$W \geq 35$	48.3%
Gemma 3-12B	$M < 35$	34.9%	$W < 35$	43.1%
GPT-4o	$M < 35$	46.0%	$W \geq 35$	66.0%
GPT-5	$M < 35$	71.1%	$M < 35$	56.7%
GPT-5-nano	$W \geq 35$	39.5%	$W \geq 35$	59.1%
Llama 3.3-70B	$W < 35$	34.8%	$W \geq 35$	65.6%
Qwen3-14B	$W \geq 35$	66.3%	$W \geq 35$	86.9%

Table 6

Complementary bootstrap stability under Spearman correlation and MAE. For Spearman, selection frequency is the proportion of bootstrap replicates in which the reported group obtains the highest correlation. For MAE, it is the proportion of replicates in which the reported group obtains the lowest error. These frequencies are descriptive and are not tests of statistical significance.