Comparing human and Large Language Model judgments of gender stereotypes in Italian texts*

Elisa Famulari[1,†], Silvana Badaloni[1,†] and Antonio Rodà[1,*,†]

[1]Dept. of Information Engineering, University of Padua, Italy

Abstract

Large Language Models (LLMs) are increasingly employed as automatic annotators for subjective Natural Language Processing tasks, yet their ability to reproduce human judgments remains an open question, particularly in socially sensitive domains. This study investigates the suitability of LLMs as annotators by comparing their evaluations of gender stereotypes in Italian texts with judgments provided by human evaluators.

Using a corpus of 111 Italian textual passages, we compare the ratings produced by GPT-4.1-mini, Gemini-2.5-flash, and DeepSeek-chat against a human reference obtained through a shared five-point annotation protocol. Beyond conventional agreement measures, we adopt a multidimensional evaluation framework combining Spearman's rank correlation, Krippendorff's α, paired t-tests, and distributional metrics based on polarization and Shannon entropy to assess agreement, reliability, calibration, and judgment variability.

The results show that all models broadly reproduce the direction of human judgments, with DeepSeek and Gemini exhibiting the highest agreement. However, substantial differences emerge in calibration and score distribution. While the models identify similar stereotype patterns, they consistently produce more stable and conservative judgments than human annotators, particularly in ambiguous cases, a pattern consistent with the hypothesis that alignment procedures may contribute to reducing the variability naturally observed in human evaluations. These findings contribute to the growing literature on LLM-as-an-Annotator by showing that evaluating LLMs requires going beyond agreement alone and considering calibration, reliability, and the ability to reproduce the uncertainty inherent in human annotation.

1. Introduction

Large Language Models (LLMs) are increasingly being employed not only as text generators but also as evaluators of textual content. Recent research has shown that state-of-the-art models can perform annotation, classification, and comparative evaluation tasks with a level of consistency that often approaches human performance, opening the possibility of replacing or supporting human annotators in a wide range of Natural Language Processing (NLP) applications [1, 2]. This emerging research direction, commonly referred to as *LLM-as-a-Judge* or *LLM-as-an-Annotator*, investigates whether language models can reliably reproduce human judgments while offering substantial advantages in terms of scalability, cost, and reproducibility [3, 4]. Despite these promising results, the reliability of LLMs as evaluators remains an open question, particularly for tasks involving subjective social constructs rather than objective factual judgments. Human annotation of phenomena such as gender stereotypes, offensiveness, political orientation, or toxicity is intrinsically characterised by disagreement, contextual interpretation, and cultural variability. Consequently, assessing whether LLMs reproduce not only the average human judgment but also its variability and uncertainty has become an important research challenge [5].

Gender stereotypes constitute a particularly relevant testbed for this problem. Language both reflects and contributes to the construction of social representations, making stereotype perception a fundamentally interpretative task rather than a purely linguistic one. Because Large Language Models are trained on large-scale corpora that inevitably encode historical and societal biases, they may reproduce, attenuate, or amplify gender-related associations present in human language [6, 7].

BEWARE 2026 (AIxIA 2026), October 8, 2026, Perugia, Italy.

*Corresponding author.

†These authors contributed equally.

elisa.famulari@studenti.unipd.it (E. Famulari); silvana.badaloni@unipd.it (S. Badaloni); antonio.roda@unipd.it (A. Rodà)

0009-0008-6657-5911 (E. Famulari); 0000-0002-5287-0468 (S. Badaloni); 0000-0001-9921-0590 (A. Rodà)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-17

CEUR-WS.org/Vol-4332/paper2.pdf

Evaluating how closely LLM judgments align with human perception is therefore essential not only for understanding model behaviour but also for assessing their suitability as automatic annotators in socially sensitive domains.

Within this context, the study of gender stereotypes in Italian language, and in particular how gender stereotypes are represented in LLMs when they are prompted in Italian, remains comparatively underexplored. Building upon the dataset developed by Scagnet et al. [8] and a new corpus of Italian texts collected by Vismara [9], the present study investigates whether Large Language Models can reproduce human evaluations of gender stereotypes in Italian textual passages. Rather than introducing a new dataset, our objective is to evaluate the extent to which model-generated judgments resemble human annotations when the same evaluation protocol is applied to both. Specifically, we compare the evaluations produced by GPT-4.1-mini, Gemini-2.5-flash, and DeepSeek-chat with judgments collected from human annotators using an identical five-point gender-rating scale. While previous studies on LLM-as-a-Judge have primarily focused on benchmark accuracy or pairwise agreement, our analysis adopts a broader perspective by examining multiple complementary dimensions of annotation quality. In addition to rank-order agreement, we evaluate internal reliability, calibration with respect to human judgments, and the distributional properties of the assigned scores, combining Spearman's rank correlation, Krippendorff's α, paired t-tests, and measures of polarization and entropy. This approach allows us to assess not only whether the models assign similar ratings to humans, but also whether they reproduce the characteristic variability and uncertainty of human perception.

The present study is guided by the following research questions:

- **RQ1**: To what extent do the evaluations produced by Large Language Models reflect the judgments expressed by human annotators in terms of internal reliability and rank-order agreement?
- **RQ2**: Do the models exhibit a systematic deviation (bias) from the mean intensity assigned by human evaluators?
- **RQ3**: Do the models reproduce human variability and the human tendency to use the full range of the scale, or do they adopt a more conservative use of the available ratings?
- **RQ4**: (Exploratory) Which specific textual patterns – in terms of thematic domain, lexical salience, and structural ambiguity – generate the greatest interpretative divergences between humans and models?

2. Background

2.1. Gender stereotypes and language

In the perspective of developing a trustworthy AI able to build a new gendered innovation [10] it is important to understand how we can re-design the scientific theories taking into account the gender dimension: we have to re-formulate the scientific questions that can produce gendered innovations in every field of knowledge [11] [12]. Stereotypes can be defined as shared cognitive representations through which generalised characteristics are attributed to members of a specific social group. Although such representations serve an important simplifying function in the interpretation of reality, they may also contribute to the formation and persistence of expectations and prejudices toward social groups [13]. Gender stereotypes are particularly significant because gender constitutes one of the primary dimensions of social identity. The World Economic Forum's *Global Gender Gap Report* highlights the persistence of substantial inequalities between men and women in economic participation, access to opportunities, and political representation, demonstrating how such representations continue to influence contemporary society [14].

The formation of stereotypes is generally associated with processes of social categorisation and with the distribution of roles within society. From this perspective, individuals tend to associate specific characteristics, behaviours, or competencies with particular social groups, thereby consolidating different expectations of men and women over time [13, 15]. These representations are also expressed through language. Within the framework of *Critical Discourse Analysis*, Fairclough [16] argues that

language does not merely describe reality, but actively contributes to the construction and reproduction of social representations and power relations. Consequently, associations between specific semantic domains and particular gender identities may emerge implicitly in texts, making linguistic material a privileged context for analysing how stereotypes are perceived and interpreted by both human evaluators and artificial intelligence systems.

2.2. Bias and alignment in Large Language Models

Because training data inevitably reflect the societies from which they originate, Large Language Models (LLMs) may learn and reproduce stereotypical associations contained in the corpora [6]. Numerous studies have shown that such associations emerge in linguistic representations and may influence model behaviour in tasks involving occupations, social roles, or characteristics attributed to different genders [15, 13]. Observable bias in LLMs may therefore be interpreted as a consequence of the statistical distributions learned during pre-training and of the optimisation mechanisms employed to model language [17].

In contemporary models, however, final behaviour does not depend exclusively on pre-training. Following this stage, models are generally subjected to *alignment* procedures, such as *Supervised Fine-Tuning* and *Reinforcement Learning from Human Feedback* (RLHF), with the aim of making their responses more consistent with safety criteria, quality standards, and normative expectations [18, 19]. These techniques do not necessarily eliminate associations learned during training, but they may alter how such associations are manifested in model outputs, for example by encouraging more cautious responses or reducing the tendency to express strongly polarised evaluations.

Alignment strategies, together with differences in training procedures and the policies adopted by individual developers, may therefore contribute to behavioural differences across model families [20, 21]. Accordingly, any differences observed among GPT, Gemini, and DeepSeek cannot be attributed to a single factor, but should instead be interpreted as the combined result of training data, optimisation procedures, and the alignment strategies adopted.

In light of these considerations, systematically comparing model-generated evaluations with those expressed by human evaluators [8] makes it possible to analyse not only the presence of potential stereotypes, but also the manner in which such associations are expressed, calibrated, and distributed across different systems.

It is therefore particularly relevant to assess the extent to which the attributions produced by Large Language Models are consistent with human perception when these models are used to analyse gender stereotypes in real-world texts [9]. The present study addresses this issue by comparing the evaluations produced by three language models with those obtained through human annotation on an Italian dataset, analysing not only the level of agreement, but also the stability, calibration, and distribution of the judgments.

3. Methodology

3.1. Dataset construction

The dataset used in this study builds upon the pilot corpus introduced by Vismara (2024)[9]. In the original study, textual passages were collected from three different types of Italian-language sources, selected according to their intended audience: magazines targeting women, hereafter referred to as *Feminine* sources (F); magazines targeting men, referred to as *Masculine* sources (M); and institutional university websites, considered *Neutral* sources (N). Accordingly, each text was assigned *a priori* to one of these three categories on the basis of its source, rather than through perceptual annotation.

In the present work, we extend Vismara's corpus following the same data collection methodology and source-selection criteria. Additional textual passages were gathered from the same three categories of sources. This extension enables a statistically more robust comparison between human and LLM-based annotations without altering the underlying experimental design. The final corpus comprises a total

of 180 Italian-language textual passages, evenly distributed across the three categories, with 60 texts per category. The new excerpts were selected using the same criteria adopted for the pilot corpus in terms of length (30–80 words), content type, and linguistic register, thereby ensuring overall thematic consistency. Class balance was maintained to prevent distortions in subsequent statistical analyses and to enable controlled comparisons.

3.2. Human evaluation of the dataset

To obtain a human annotation reference for the subsequent comparison with the language models, we adopted the evaluation protocol originally proposed by Scagnet (2023) [8]. Participants were asked to assess the perceived target audience of each textual passage by answering the following question: *"A quale genere pensa che il testo sia rivolto?"*[1]

Responses were collected using a five-point ordinal bipolar scale designed to capture both the direction and the intensity of the perceived gender orientation of each text. For statistical analyses, the response categories were encoded as follows:

- $-2 \rightarrow$ Completely feminine
- $-1 \rightarrow$ More feminine than masculine
- $\ \ 0 \rightarrow$ Neutral
- $+1 \rightarrow$ More masculine than feminine
- $+2 \rightarrow$ Completely masculine

This scale, which is symmetric around zero, makes it possible to quantify both the direction and the intensity of the perceived target audience along a gender continuum. A non-response option was also provided, and missing data were consequently excluded. The survey was administered through the PsyToolkit platform [22],[23], with each participant receiving a randomised subset of the corpus in order to minimise order effects and fatigue. Following a filtering procedure that excluded incomplete submissions, failed attention checks, and anomalous response times, the final analysed sample consisted of 75 participants. The gender distribution of participants was 56% female, 41.3% male, 1.3% non-binary, and 1.3% who preferred not to specify. The sample was linguistically homogeneous (82.7% native Italian speakers, 98.7% Italian nationality), with most participants falling within the 18–30 age group. Finally, to ensure the statistical robustness of the mean *gender scores*, textual passages with fewer than five independent evaluations were discarded. All subsequent comparative analyses were therefore conducted exclusively on the resulting 111 valid sections. All participants provided informed consent before taking part in the study, in accordance with ethical guidelines for research involving human participants.

3.3. Language-model evaluation protocol

The complete corpus of 180 textual passages was submitted to three generative language models: GPT-4.1-mini, Gemini-2.5-flash, and DeepSeek-chat. The choice of *lightweight* versions (mini" or flash") was motivated by the discrete and narrowly defined nature of the classification *task*, for which very high-capacity models would have entailed greater computational costs without proportionate benefits. This choice also enables a homogeneous comparison among systems belonging to a similar application tier but developed according to different training and *governance* paradigms: alongside the established proprietary US-based models GPT and Gemini, DeepSeek was included to extend the comparison to a different geographical and strategic context.

To ensure maximum comparability and reproducibility, all queries were processed through a single automated Python *pipeline*, made available on GitHub, which interfaced with the official API *endpoints*. The *prompt* submitted to the models was identical to the question used in the human questionnaire (*"A quale genere pensa che il testo sia rivolto?"*) and explicitly required the output of a single numerical value

[1] "Which gender do you think this text is addressed to?"

constrained to the set $-2, -1, 0, +1, +2$. Restricting the output to a strictly numerical format prevented discursive or ambiguous generations and ensured perfect alignment with the human rating scale. To assess the internal stability of the models with respect to generative stochasticity, five independent executions (*runs*) were performed for each *prompt*, varying the temperature parameter: one *run* at $T = 0$ (deterministic setting), two at $T = 0.3$, and two at $T = 0.5$ (moderately stochastic settings). Higher temperatures were not explored in order to avoid introducing excessive and unjustified randomness, which would have compromised comparative stability in a purely classificatory *task*. The maximum output token limit was set to 64 tokens for all models, a value determined during preliminary testing to be sufficient for GPT-4.1-mini and DeepSeek-chat given the strictly numerical nature of the expected output; this configuration was kept identical across models for consistency.

Finally, the scores generated across the five executions were aggregated by computing the arithmetic mean and standard deviation for each textual passage. Consistently with the filtering applied to the human sample, subsequent comparative analyses were conducted exclusively on the 111 valid sections.

3.4. Statistical analysis and evaluation metrics

The following metrics were selected to investigate alignment between human perception and computational-model responses along four complementary dimensions: internal reliability, rank-order agreement, the presence of systematic bias, and the distributional properties of the judgments.

Internal reliability and stability (Krippendorff's Alpha) To quantify inter-annotator agreement among human evaluators and the internal stability of model responses across different temperature settings, Krippendorff's Alpha (α) with an ordinal distance metric was employed. This choice is methodologically preferable to alternative coefficients for two main reasons: it natively and statistically robustly handles missing data, which are an inherent feature of the human dataset because each participant evaluates only a subset of the texts, and it accounts for both the ordered nature of the scale and the metric distance between its categories. The index ranges from -1 to $+1$, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate systematic disagreement.

Overall agreement (Spearman's correlation) To measure convergence between the mean scores assigned by human annotators and the averages obtained across the Large Language Model executions, Spearman's rank correlation coefficient (ρ) was used. This non-parametric statistic replaces the Pearson linear correlation employed during the initial stages of the pilot study. The change is motivated by the fact that the *gender scores* originate from a five-level discrete ordinal scale and exhibit strongly asymmetric and non-normal distributions, conditions under which Pearson's correlation would be distorted and insufficiently robust to extreme values or rank ties.

Detection of systematic bias (Paired t-test) Because high rank-order agreement indicates alignment in the direction of interpretative trends but does not rule out a constant deviation in intensity—for example, a model could correlate strongly with the human mean while systematically assigning shifted or attenuated scores—a paired t-test was employed. This analysis is computed on the paired differences for each textual section ($D = \text{human} - \text{model}$) in order to formally assess whether the mean attribution error is statistically significant and differs from the ideal value of zero. It therefore quantifies the systematic tendency of the models to overestimate or underestimate stereotype intensity relative to human evaluators.

Distributional properties (Polarization and Shannon entropy) Finally, to examine how alignment strategies and *safety policies* influence the models' effective use of the interpretative spectrum relative to humans, two distributional descriptors were calculated directly on the unaggregated raw responses:

- **Polarization:** defined as the mean absolute value of the ratings (|score|), this measure captures the overall tendency of a source to move away from the neutral point (0) and toward either gender pole.
- **Shannon entropy (H):** calculated from the relative-frequency distribution p_i of the five available categories according to the formula:

$$H(X) = -\sum_{i=-2}^{+2} p_i \log_2(p_i) \tag{1}$$

Entropy quantifies the diversity, informational richness, and heterogeneity of the judgments. High entropy indicates uniform and diversified use of the full scale, whereas a lower value reflects distributional compression, namely repetitive behaviour concentrated on a small number of classes, a phenomenon commonly observed in models subjected to strict bias-mitigation procedures.

4. Results

Table 1
Descriptive statistics of gender scores by source and textual category (111 valid sections).

Source	Overall (M, SD)	F (M)	M (M)	N (M)
Humans	-0.012, 0.639	-0.395	0.446	-0.081
GPT	-0.135, 0.642	-0.380	0.179	-0.206
Gemini	0.070, 0.955	-0.390	0.790	-0.226
DeepSeek	0.059, 0.488	-0.190	0.421	-0.065

Table 1 reports the basic descriptive statistics—mean and standard deviation—of the gender scores assigned by each source, both overall and by textual category. All sources exhibit an overall mean close to zero, with GPT showing the strongest tendency toward the feminine pole ($M = -0.135$). At the category level, the distinction between Feminine and Masculine texts is consistent across all sources, although its intensity varies: Gemini exhibits the strongest polarization in the Masculine category ($M = 0.790$), whereas DeepSeek produces more moderate scores ($M = 0.421$). These aggregate values provide the descriptive basis for the reliability, agreement, and distributional analyses presented in the following sections.

4.1. RQ1: Internal reliability and agreement across judgments

The first research question (*Research Question 1*) examines the extent to which the evaluations produced by Large Language Models (LLMs) reflect the judgments expressed by human annotators when attributing gender stereotypes to texts. To address this question, the analysis considers two complementary dimensions: the internal *reliability* of the different annotation sources and the degree of relational agreement among the assigned scores.

Table 2
Krippendorff's Alpha (ordinal metric): overall reliability and reliability by textual category.

Source	**Type of agreement**	**Overall**	**F**	**M**	**N**
Humans	Inter-rater	0.328	0.226	0.313	0.162
GPT	Inter-run	0.938	0.869	0.961	0.999
Gemini	Inter-run	0.970	0.942	0.954	1.000
DeepSeek	Inter-run	0.992	0.968	0.995	1.000

The reliability analysis, measured using ordinal Krippendorff's Alpha (Table 2), reveals a pronounced and structural difference between human evaluators and language models. Human participants exhibit a moderate overall level of agreement ($\alpha = 0.328$), a result that is fully consistent with the natural heterogeneity and subjectivity characterising the social perception of gender-related nuances in a text. At the level of individual textual categories, human agreement decreases substantially for texts classified a priori as Neutral ($\alpha = 0.162$) compared with Masculine ($\alpha = 0.313$) and Feminine texts ($\alpha = 0.226$), confirming greater interpretative uncertainty and fragmentation of judgments in the absence of explicit markers.

By contrast, all models display extremely high Alpha values (GPT: $\alpha = 0.938$; Gemini: $\alpha = 0.970$; DeepSeek: $\alpha = 0.992$), indicating that the generated evaluations remain highly stable and consistent despite the stochastic variation introduced by the five executions at different temperatures. A striking inverse pattern emerges specifically for the Neutral category (N): whereas humans exhibit their greatest interpretative divergence, the computational models converge toward almost complete determinism, with Gemini and DeepSeek achieving perfect agreement ($\alpha = 1.000$) and GPT reaching $\alpha = 0.999$. The LLMs therefore behave as rigid classifiers whose internal stability appears largely independent of the pragmatic complexity or ambiguity of the textual stimulus.

To determine whether the high internal consistency of the individual models also translates into mutual agreement, Krippendorff's Alpha was recalculated by treating each source as an independent evaluator of the same set of texts, rather than separately aggregating the five runs of each model (Table 3). In the first configuration, GPT, Gemini, and DeepSeek were treated as three distinct evaluators, yielding an agreement value of $\alpha = 0.631$, which is substantially lower than the individual consistency of each model. In the second configuration, the mean human judgment was added as a fourth evaluator, causing Alpha to decrease further to $\alpha = 0.417$. This declining pattern confirms that the high inter-run reliability observed within individual models is not conceptually equivalent to genuine convergence among different sources of judgment. Rather, it reflects the stability of each system relative to itself, which weakens as soon as the model is compared with other independent evaluators, whether human or artificial.

Table 3
Krippendorff's Alpha calculated by treating each source as an independent evaluator.

Configuration	No. of evaluators	α
GPT + Gemini + DeepSeek	3	0.631
GPT + Gemini + DeepSeek + Humans	4	0.417

Table 4
Overall Spearman correlation matrix (ρ) across all sources.

	Human	GPT	Gemini	DeepSeek
Human	1.000	0.584	0.672	0.694
GPT	0.584	1.000	0.733	0.632
Gemini	0.672	0.733	1.000	0.758
DeepSeek	0.694	0.632	0.758	1.000

To quantify interpretative convergence across sources, Table 4 reports Spearman's rank correlation coefficients (ρ). The overall results show that all models exhibit a strong and statistically significant positive correlation with the mean human judgments. Consistently with the initial hypothesis, DeepSeek displays the highest degree of alignment ($\rho = 0.694$), closely followed by Gemini ($\rho = 0.672$), whereas GPT exhibits the lowest agreement, although its correlation remains clearly significant ($\rho = 0.584$). Strong internal affinity is also observed among the models themselves, particularly between Gemini and DeepSeek ($\rho = 0.758$), suggesting that distinct computational architectures tend to share similar categorisation trends.

Table 5
Spearman correlation by textual category (Human vs. LLM).

Category	Comparison	ρ	p-value	Sig.
F	Human – GPT	0.578	7.67e-05	***
	Human – Gemini	0.461	2.43e-03	**
	Human – DeepSeek	0.606	2.67e-05	***
M	Human – GPT	0.532	4.84e-04	***
	Human – Gemini	0.667	3.60e-06	***
	Human – DeepSeek	0.679	2.00e-06	***
N	Human – GPT	0.453	1.06e-02	*
	Human – Gemini	0.446	1.18e-02	*
	Human – DeepSeek	0.427	1.65e-02	*

However, when the analysis is disaggregated by textual category (Table 5), alignment between human perception and model-generated judgments reveals substantial contextual asymmetries. For strongly gendered texts (M and F), the models maintain a strong association with human judgments, with DeepSeek reaching $\rho = 0.679$ for Masculine texts and $\rho = 0.606$ for Feminine texts. By contrast, for Neutral texts (N), agreement systematically deteriorates across all analysed systems, with ρ coefficients falling to values between 0.427 and 0.453, together with a marked reduction in statistical significance ($p > 0.01$).

This distributional contrast outlines a clear pattern: LLMs are effective at tracking the broad direction of a stereotype when it is supported by explicit lexical or thematic cues. However, when confronted with neutrality or structural ambiguity, the semantic space of the models diverges from that of human evaluators. Whereas human annotators disperse their judgments to reflect uncertainty, artificial systems select a rigid and stereotyped response and reproduce it consistently across runs, while achieving lower agreement with actual social perception. This phenomenon suggests that convergence in the direction of a judgment does not necessarily correspond to genuine calibration of its intensity, an issue explored in the following sections.

4.2. RQ2: Analysis of systematic deviation (bias)

The second research question (*Research Question 2*) investigates the presence of systematic deviations *bias* in the attribution of gender stereotypes. Despite the high overall relational agreement demonstrated in RQ1, a model may still consistently overestimate or underestimate gender intensity relative to humans. To formally test this hypothesis, a *paired t-test* was conducted on the paired mean differences for each textual section ($D = \text{Human} - \text{Model}$).

Table 6
Results of the paired t-test ($D = \text{Human} - \text{Model}$).

Model	N	Mean D	95% CI lower	95% CI upper	t	p-value
DeepSeek	111	-0.0714	-0.1472	0.0045	-1.864	0.06498
Gemini	111	-0.0822	-0.2039	0.0396	-1.338	0.18380
GPT	111	0.1232	0.0247	0.2217	2.479	0.01470

The results (Table 6) reveal two distinct behavioural profiles. For Gemini and DeepSeek, the statistical test does not identify a mean difference that is significantly different from zero ($p = 0.1838$ and $p = 0.0649$, respectively). This indicates that, in aggregate, the evaluations produced by these systems remain close to the human mean and do not exhibit a systematic deviation toward either gender pole.

By contrast, GPT exhibits a statistically significant difference ($p = 0.0147$). The positive mean difference ($M_D = 0.1232$) suggests that, when deviations occur, the mean human scores tend to be

systematically higher, and therefore shifted further toward the masculine pole of the scale, than the scores assigned by GPT. This finding points to a slight baseline systematic deviation (bias) in the OpenAI model for this specific task.

Overall, as anticipated in the theoretical framework, most systems (Gemini and DeepSeek) exhibit no systematic deviation (bias) in their overall mean values. However, a null mean difference does not imply that the artificial and human judgment systems are equivalent in terms of intensity and dispersion. To determine whether the models use the scale as conservatively or as heterogeneously as humans, it is necessary to shift the analysis from measures of central tendency to the distributional properties of the judgments.

4.3. RQ3: Distributional properties and conservativeness of judgments

The absence of pronounced systematic deviations in the mean values (RQ2) does not imply that artificial intelligence systems and humans use the interpretative spectrum in the same way. The third research question investigates the distributional properties of the judgments in order to determine whether the models reproduce human variability and the human tendency to use the extreme values of the scale. To quantify this aspect, mean Polarization and Shannon entropy (H) were calculated directly from the raw ratings.

Table 7
Polarization and entropy for each evaluation source.

Source	N	Polarization	Entropy (bit)
Humans	1238	0.6575	1.9472
DeepSeek	555	0.2360	1.0292
Gemini	547	0.5137	1.4297
GPT	555	0.3261	1.2749

Note. The slightly lower number of observations for Gemini ($N = 547$ vs. $N = 555$ for GPT and DeepSeek) is due to a small number of missing responses under the shared 64-token output limit (see Section 5 for details).

The results (Table 7) reveal a profound behavioural asymmetry. Human evaluators exhibit the highest values for both Polarization (0.6575) and Entropy (1.9472 bits). This indicates that humans tend to move confidently away from neutrality and make use of the full richness of the available scale, distributing their evaluations heterogeneously in order to reflect the perceived intensity of the stereotype.

By contrast, the indicators confirm that artificial intelligence systems exhibit much more concentrated and less heterogeneous distributions. The LLMs adopt a markedly conservative" approach. DeepSeek, in particular, emerges as the most centrist" model (Polarization 0.2360) and the least diversified one (Entropy 1.0292), displaying repetitive behaviour and a reluctance to use the extreme values of the scale. Gemini is the computational system that most closely approximates human behaviour (Polarization 0.5137; Entropy 1.4297), although its dispersion remains substantially lower than that of the *ground truth*.

To facilitate a visual comparison of these distributional dynamics, Figure 1 presents a *violin plot* summarising the dispersion of individual ratings. The visualisation unequivocally corroborates the quantitative metrics: whereas human judgments form broad distributions that extend smoothly to the extremes (± 2), the central bodies of the violins associated with the LLMs appear enlarged and compressed around the central values (0 and ± 1), indicating a collapse in the frequency of extreme judgments.

Overall, the combined analysis of polarization and entropy provides a plausible interpretative framework: the observed compression in judgment intensity is consistent with the effects generally attributed to fine-tuning and normative alignment procedures. In other words, the results suggest that the models

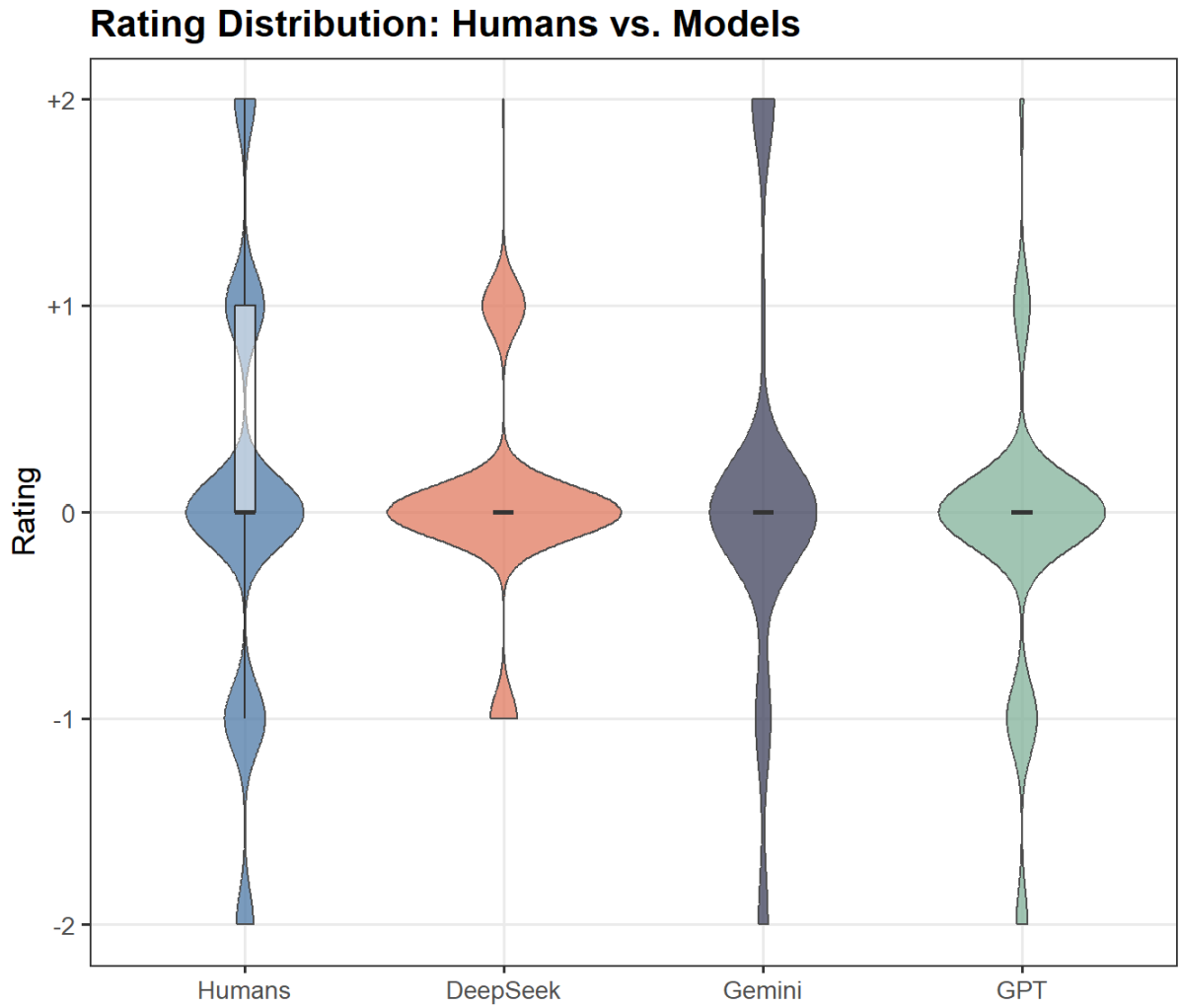


Figure 1: Distribution of individual ratings by source. The width of each violin is proportional to the data density. The horizontal lines inside each violin indicate the first quartile (Q1), the median, and the third quartile (Q3).

agree with humans in identifying *which* texts are polarised, as demonstrated by the Spearman correlations, but systematically diverge in the *intensity* with which that polarization is expressed, preferring to remain within statistically safer regions of the scale.

4.4. RQ4: (Exploratory) Qualitative analysis and patterns of misalignment

The quantitative analysis showed that LLMs share the direction of gender stereotypes with humans, but differ substantially in the intensity and stability of their judgments. The fourth research question adopts an exploratory, qualitative approach to illustrate — rather than systematically catalogue — the textual contexts in which the most substantial interpretative divergences emerge. The examples discussed below were selected as the cases showing the largest absolute divergence between human and model scores, and are intended as illustrative rather than exhaustive.

The examination of misalignment cases (Table 8) shows that Neutral texts (e.g., N3_1 and N15_2) containing narratives related to gender-based violence or vulnerability trigger a "shared social bias": both the models and the human annotators converge strongly toward the feminine pole.

The fashion-related excerpt in the third row of Table 8 is particularly emblematic: although classified a priori as Masculine (M3_3), it is "feminised" by all the artificial systems. This suggests that, in the absence of strong semantic cues or in the presence of structural ambiguity, the models rely on latent statistical stereotypes, such as the association between the fashion domain and the feminine gender, often amplifying the intensity of the judgment relative to human annotators.

Table 8
Examples of the most misaligned texts between human and model judgments.

Textual excerpt	Human	GPT	Gemini	DeepSeek
(N3_1) *"Il primo dato che balza all'occhio è quello delle vittime di sesso femminile. Dal 1 gennaio 2023 al 31 luglio 2024 sono state 175. Conosciamo bene purtroppo i fatti di cronaca dietro queste storie, così come sappiamo che il 31,5% delle donne in Italia ha subito nel corso della propria vita una qualche forma di violenza fisica o sessuale."*	−1.28	−2.00	−2.00	−1.00
(N15_2) *"La donna ha trovato il coraggio di denunciare l'aggressione solo la mattina seguente. I carabinieri, dopo aver ricevuto il racconto, hanno allertato il 118 e l'hanno trasferita all'ospedale di Galliera. Qui sono stati eseguiti i primi esami e test, e la donna ha parlato con una psicologa."*	−1.00	−2.00	−2.00	−1.00
(M3_3) *"Invece di risultare un po' piatto (come può accadere a volte con la monocromia), il mix di colori è un ottimo modo per rendere un outfit interessante, aggiungendo l'elemento cruciale e apparentemente spontaneo del 'l'ho appena indossato'."*	−1.00	−1.00	−2.00	0.00

Conversely, the models display considerable resistance to assigning Masculine-pole scores to technically ambiguous texts for which humans—guided by pragmatic inferences or lexical style—tend toward a masculine interpretation. In these scenarios, the models instead return a judgment of complete neutrality (value 0), a behaviour plausibly attributable to bias-mitigation mechanisms such as RLHF, although this cannot be directly verified without access to non-aligned model variants. This behaviour demonstrates that the "conservativeness" measured in RQ3 is ultimately the result of a specific form of normative alignment that tends to neutralise stereotypical inferences whenever they are not supported by unambiguous lexical evidence.

5. Discussion

The results of this study outline a coherent picture: the analysed Large Language Models (LLMs)—GPT, Gemini, and DeepSeek—do not operate within a semantic space that is alien to human perception, but instead reflect the same structures of gender association encoded in language [16]. The strong Spearman correlations confirm that these systems capture the direction of the stereotype, supporting the claim that gender, understood as a discursive construction, is intrinsically embedded in pre-training data [24, 17].

A central finding of this study is that Large Language Models reproduce the overall direction of human judgments while differing substantially in the way those judgments are calibrated and expressed. The discrepancy between high directional agreement, as measured by Spearman's correlation, and pronounced distributional compression, as measured by Entropy and Polarization, is consistent with the hypothesis that mitigation policies and RLHF operate as a form of "methodical neutralisation" [18, 19], although alternative explanations — such as inherent differences in architecture or pre-training corpora — cannot be ruled out on the basis of the present data. While the models correctly identify the polarity of a text, they tend to minimise the expressive intensity of their judgments, displaying behaviour that is systematically more conservative than that of human annotators [25].

In ambiguous contexts, particularly Neutral texts, this divergence becomes more pronounced. Whereas humans exhibit natural interpretative variability, which may be understood as a form of complex social categorisation [13, 15], the models respond with almost artificial inter-run stability and avoid polarised judgments. This suggests that, for the analysed systems, neutrality is not merely the absence of gender, but also a precautionary output strategy [26]. The comparative analysis reveals that, although the models share a common semantic basis, their "thresholds" for departing from neutrality

differ: DeepSeek favours stability at the expense of interpretative richness, whereas Gemini displays greater flexibility in reflecting natural social dynamics, although it does not reach the interpretative boldness of human annotators validated according to the methodologies of [8].

It should also be noted that the high inter-run reliability observed in the models is not directly comparable, at the conceptual level, to the inter-rater reliability observed among humans. In the human case, a moderate α reflects genuine interpretative variability shared among different individuals. In the case of the models, an α value close to 1 may instead result from substantially deterministic behaviour that is insensitive to textual nuances, particularly under conditions of structural ambiguity. In other words, high inter-run stability does not necessarily imply superior discriminatory ability. On the contrary, it may indicate a system that returns a content-insensitive default response rather than providing a substantively more reliable judgment.

Although the present framework offers a rigorous perspective on model calibration, several intrinsic limitations should be acknowledged. First, the sample size of 111 textual sections, while enabling an in-depth and granular analysis, represents a domain-specific subset of the broad heterogeneity of everyday language. Second, the analysis is restricted to Italian; model behaviour, influenced by predominantly English-language pre-training architectures and corpora, may differ in multilingual contexts. Finally, the use of smaller models ("mini/flash" versions) requires caution when extending these conclusions to higher-capacity models, which may exhibit even more pronounced alignment dynamics. Moreover, this study did not include a comparison with non-aligned or base ('vanilla') versions of these models, which would be necessary to more directly isolate the specific contribution of alignment procedures to the observed conservativeness of judgments. Relatedly, the 64-token output limit was calibrated during preliminary testing on GPT and DeepSeek, for which it proved sufficient given the purely numerical nature of the task, and was then kept constant across models to preserve a homogeneous experimental protocol. Because Gemini was evaluated last and exhibited slower response times, its tendency to consume part of the allocated budget for internal reasoning only became apparent at a later stage. Rather than increasing the token limit selectively for Gemini — which would have introduced an asymmetry across models — the original configuration was retained, resulting in a small number of missing responses (see Table 7). This represents a deliberate trade-off between protocol homogeneity and completeness, and points to a possible refinement for future replications, for instance by calibrating token budgets independently for each model prior to the main data collection. These limitations do not undermine the validity of the findings, but provide a foundation for future investigations using larger-scale corpora.

Beyond their methodological implications, these findings also have broader educational and societal relevance. The observed differences between human and LLM judgments, particularly the models' tendency to produce more conservative and less variable evaluations, highlight that the interpretation of socially sensitive phenomena such as gender stereotypes is not merely a technical task. Rather, it reflects design choices, training data, and governance strategies that shape how AI systems represent and reproduce social knowledge. In this sense, the results reinforce the view that these technologies are not neutral, but are influenced by the values and assumptions embedded throughout their development.

For this reason, fostering awareness of the relationship between gender, ethics, and intelligent technologies has become an essential component of engineering education. Since 2021, the School of Engineering at the University of Padua has been pioneering an innovative course entitled *Gender Knowledge and Ethics in Artificial Intelligence*, representing the first explicit introduction of gender-related topics within an engineering curriculum [27]. The findings presented in this work provide a concrete case study for the course, illustrating how gender-related language is interpreted differently by humans and AI systems and highlighting the importance of critically assessing human–machine interaction in socially sensitive contexts.

6. Conclusions

This study investigated the extent to which Large Language Models can reproduce human judgments when evaluating gender stereotypes in Italian texts. By comparing GPT-4.1-mini, Gemini-2.5-flash,

and DeepSeek-chat against human annotations through a multidimensional evaluation framework, we showed that agreement between humans and LLMs cannot be adequately characterised by correlation alone. Overall, all three models successfully captured the direction of human judgments, confirming that they identify similar gender-related semantic patterns. DeepSeek and Gemini achieved the highest agreement with human evaluations, whereas GPT exhibited a small but statistically significant systematic deviation (bias). However, the analysis also revealed substantial differences in the way models express their judgments. Compared with humans, all LLMs produced markedly more stable and less variable annotations, assigning fewer extreme ratings and concentrating their responses around the centre of the evaluation scale. These findings are consistent with the view that current alignment strategies may enable LLMs to reproduce the general structure of human perceptions while simultaneously reducing the uncertainty and variability that naturally characterize subjective human annotation. Consequently, although LLMs appear well suited to supporting large-scale annotation tasks, they should not yet be considered direct replacements for human evaluators in socially sensitive domains, where disagreement and interpretative diversity constitute meaningful aspects of the phenomenon being measured.

More broadly, our results demonstrate the importance of evaluating LLMs along multiple dimensions, including agreement, reliability, calibration, and distributional behaviour. Considering these complementary perspectives provides a more complete understanding of the strengths and limitations of LLMs as annotators than relying on agreement measures alone.

Data and code availability

The expanded dataset, the evaluation pipeline implemented in Python, the statistical analysis scripts, and the complete results of the experimental runs are publicly available in the project's GitHub repository. The repository also includes the documentation necessary to replicate the entire experimental process, including the parameter configurations used for each model: https://github.com/elisafamulari-coder/GenderSte-

Ethics Statement

The annotation task involved adult participants who were asked to read sentences extracted from magazine and website articles and provide judgments on a Likert scale. The task did not involve sensitive or offensive material. All annotators were informed about the nature and purpose of the task and provided informed consent prior to participation. Participation was voluntary, and the collected annotations were used exclusively for research purposes.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools (ChatGPT, Gemini) to support grammar and spelling checks, sentence rephrasing, and improvements in textual clarity and readability. All AI-assisted content was critically reviewed and edited by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript. Additionally, as detailed in the paper, several LLMs were employed for experimental purposes as part of the research methodology.

References

[1] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120.

[2] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, T. Zhao, An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 5880–5895.

[3] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 15607–15631.

[4] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.
[5] P. Törnberg, Large language models outperform expert coders and supervised classifiers at annotating political social media messages, Social Science Computer Review 43 (2025) 1181–1195.
[6] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).
[7] 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, 2021, pp. 610–623.
[8] M. Scagnet, Sviluppo e validazione di un dataset in italiano per l'analisi di stereotipi di genere nei documenti testuali, 2023.
[9] F. Vismara, Un dataset italiano per l'analisi degli stereotipi di genere nei documenti testuali, 2024.
[10] L. Schiebinger, Gendered innovations, http://genderedinnovations.stanford.edu/index.html, 2026. Accessed: 2026-07-15.
[11] S. Badaloni, F. A. Lisi, Towards a gendered innovation in ai, in: Proceedings of the AIxIA 2020 Discussion Papers Workshop, 2020. URL: http://ceur-ws.org/Vol-2776/paper-2.pdf.
[12] S. Badaloni, A. Rodà, Gender knowledge and artificial intelligence, in: Proceedings of the BEWARE Workshop, AIxIA 2022, 2022. URL: https://ceur-ws.org/Vol-3319/paper12.pdf.
[13] H. Tajfel, Social identity and intergroup behaviour. (2010).
[14] R. Hausmann, L. Tyson, S. Zahidi, The global gender gap report: World economic forum, 2009.
[15] A. H. Eagly, W. Wood, Social role theory, Handbook of theories of social psychology 2 (2012) 458–476.
[16] N. Fairclough, Language and power, 1989, Harlow: Longman (2001).
[17] D. Biasion, A. Fabris, G. Silvello, G. A. Susto, Gender bias in italian word embeddings, in: Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020), 2020, pp. 32–39.
[18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[19] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, S. McCandlish, et al., Constitutional ai: Harmlessness from ai feedback, arXiv preprint arXiv:2212.08073 (2022).
[20] OpenAI, Gpt-4 system card, https://cdn.openai.com/papers/gpt-4-system-card.pdf, 2023. Accessed: 2026-02-24.
[21] Google, Google ai principles, https://ai.google/principles/, 2018. Accessed: 2026-02-24.
[22] G. Stoet, Psytoolkit: A software package for programming psychological experiments using linux, Behavior research methods 42 (2010) 1096–1104.
[23] G. Stoet, Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments, Teaching of Psychology 44 (2017) 24–31.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[25] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, A. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359 (2021).
[26] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 2019, pp. 220–229.
[27] S. Badaloni, C. Ferrari, A. Rodà, Teaching gender knowledge and ethics in ai to stem students, in: Proceedings of the 2nd Workshop on Education for Artificial Intelligence (ECAI 2025), 2025. URL: https://ceur-ws.org/Vol-4114/11_paper.pdf.