How Humans and LLMs Read Gender into "Gender-Neutral" Physical Descriptions

Yingjia Wan[1*] **Lin L. Lin**[1*] **Elisa Kreiss**[1,2]

[1]Department of Communication
[2]Department of Computer Science
University of California, Los Angeles (UCLA)
{alisawan, llin, ekreiss}@ucla.edu

Abstract

When foundation models describe people, a growing body of work in AI fairness, accessibility, and ethics recommends avoiding inferred identity labels in favor of seemingly "objective" physical descriptions (e.g., "short hair," "a defined jawline"). Yet whether such descriptive language actually achieves gender-neutral communication remains an open empirical question. To study this, we introduce GAPA (Gender Associations of Physical Attributes), a dataset of 316 common physical attributes drawn from diverse sources and domains, paired with 14,706 gender-association ratings from 304 US-based annotators. Results of human annotations show that physical descriptions carry automatic, structured, and graded gender associations among human readers: 53% of attributes are significantly more associated with one gender than the others, with more consistent and distinctive associations for women and men than for non-binary identities. Next, we evaluate 16 large language models (LLMs) across model families, sizes, and post-training variants against human ratings. Models partially recover human associations but show systematic alignment biases, including compressed rating distributions, weaker alignment for associations with the *men* category, and asymmetric abstention disproportionately targeting non-binary categories which escalated among instruction-tuned and proprietary models. Finally, we release a proxy predictor model for predicting humans' gender associations of descriptive language at scale and demonstrate its utility through a sociolinguistic analysis of character descriptions in LitBank. Together, our findings challenge the assumption that replacing explicit gender labels with physical descriptions necessarily yields gender-neutral language, showing that such descriptions can retain systematic gender associations in human interpretation and introduce distinct challenges for model–human alignment. These findings provide first empirical evidence for the downstream challenges of providing an "objective" linguistic description of inherently subjective identity labels in human–AI interaction.[1]

1 Introduction

How should models describe a person? This question has become increasingly consequential as foundation models are deployed at scale in image-captioning systems, accessibility tools, content moderation pipelines, and assistive technologies. A range of research communities (e.g., fairness in machine learning, accessibility, critical AI studies, and human-computer interaction) have converged on a partial answer: when describing a person whose identity cannot be confirmed, models should avoid inferring categorical identity labels (race,

*Equal contribution.

[1]The dataset and code are available at https://github.com/Yingjia-Wan/GAPA, and the predictor at https://huggingface.co/alisa-yingjia-wan/gapa-predictor-olmo2-7b.

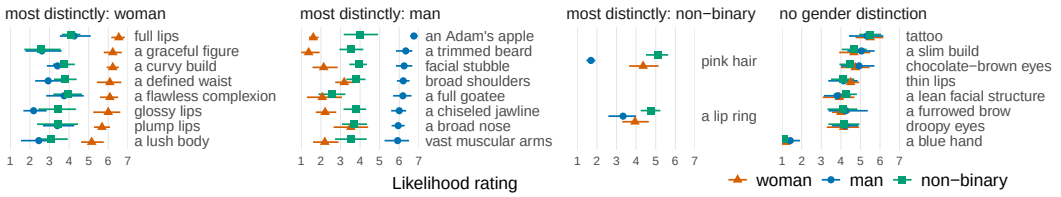


Figure 1: An excerpt of the most gender-distinctive attributes with significant gender differences ($p < 0.05$) in each ranking pattern, with their per-gender association ratings on the y-axis. See Figure 9 in the Appendix for the complete rating plots of all attributes with significant gender differences in the six ranking patterns.

gender, disability) and instead describe observable physical attributes (Bennett et al., 2021; Scheuerman et al., 2019; 2020; Hamidi et al., 2018). Such ideology is influential in shaping both human image-description practice and model behavior, and recent foundation models increasingly default to descriptive rather than identity-inferring language when describing people (Marin-Llobet et al., 2026; Hanley et al., 2021). The reasoning behind this looks epistemic and ethical: describing what is visually present, rather than what is inferred about the person, transfers the inference back to the reader and supposedly avoids imposing identity claims the system cannot verify.

However, such reasoning rests on an assumption that has not been empirically tested: that descriptive language is meaningfully gender-neutral. The twist here is: if "a chiseled jawline" or "a defined waist" carries gender associations as systematic as those carried by "a man" or "a woman," then the strategy may shift gendered communication into a less visible register rather than reducing it. This raises some critical, underexplored questions: do isolated physical descriptors, stripped of visual context, carry gender associations among human readers? If so, how strong, how consistent across people, and how distributed across types of attribute? Furthermore, do language models reproduce these associations, amplify them, or structurally distort them?

It remains a significant blind spot to what extent linguistic descriptions of physical characteristics themselves are gendered, with little evidence in prior work across disciplines. Sociolinguistic and psychological research has long established that humans infer gender from visual and acoustic cues within hundreds of milliseconds (Besson et al., 2017; Freeman & Johnson, 2016; Brown & Perrett, 1993; Freeman et al., 2012) and that language carries graded, context-sensitive gendered meanings (Eckert & McConnell-Ginet, 2003). In NLP, extensive work on gender bias has developed diagnostic benchmarks (Zhao et al., 2018; Bolukbasi et al., 2016; Caliskan et al., 2017) and mitigation methods (Dong et al., 2024; Zhao et al., 2019). These benchmarks typically evaluate whether foundational models associate gender with occupations, traits, pronouns, etc. However, they do not examine whether the descriptive vocabulary itself carries systematic gender associations. This gap is especially consequential because current practices that replace explicit identity labels with ostensibly neutral physical descriptions implicitly assume that such descriptions do not themselves transmit gendered information.

To address this gap, this study challenges the assumption that physical attribute descriptions function as semantically gender-neutral fallbacks, by presenting the first series of evidence from human and LLM experiments. We make the following contributions:

1. **Empirical evidence that physical attributes carry systematic gender associations.** We introduce GAPA (Gender Associations of Physical Attributes), a **human-annotated dataset** of 316 physical attribute descriptions sourced from three complementary channels (LLM generation, human elicitation, and contemporary fiction) annotated with 14,706 ratings from 304 U.S.-based participants on perceived associations with women, men, and non-binary gender categories (see Figure 1). While much prior work has focused on quantifying visual gender associations, GAPA allows us to do this in the linguistic modality, providing a crucial foundation for empirical studies on "gender-neutral" descriptions of people's appearances. Section 3 introduces the construction process and statistics of the dataset, while Section 4 reports the human experiments from

which we collect the human annotations of how physical attributes from GAPA are associated with each gender category.

2. **Systematic Evaluation of LLM Gender Representations.** After establishing that humans carry strong gender associations with physical attributes, it becomes imperative to understand to what extent models faithfully encode or distort them in meaningful ways, as this alignment marks the basis for human–AI interaction downstream. We evaluate 16 state-of-the-art LLMs on GAPA against human ratings, and find that models only partially recover the human pattern. This highlights that LLMs, too, carry gender associations with linguistic descriptors of physical attributes, yet exhibit systematic distortions on multiple levels. See Section 5 for detailed results and discussions.
3. **Studying Gender Associations at Scale.** Our findings also provide a new opportunity for studying linguistic gender representations at scale. To that end, we develop and release a proxy predictor of human gender associations for automatically quantifying how humans associate gender with descriptive language. Our selected predictor, fine-tuned on GAPA, achieves r=0.76 against held-out human ratings after extensive hyperparameter search and model ablations. We demonstrate its utility via a sociolinguistic analysis of character descriptions in LitBank (Bamman et al., 2019) (Section 6).

2 Related Work

Linguistic and social science research conceptualizes gender not as a fixed category but as a socially constructed, context-dependent phenomenon (West & Zimmerman, 1987; Eckert & McConnell-Ginet, 2003). From this perspective, gender is indexed through speech patterns and styles, making it a graded dimension rather than a binary one (Bucholtz & Hall, 2005). Empirical work demonstrates that humans infer gender from a wide array of cues—including voice, facial morphology, and motion (Johnson & Tassinary, 2005)—and that these associations are systematic and context-sensitive.

In NLP, extensive work has examined the explicit and implicit gender biases that language models encode, producing influential diagnostic benchmarks like WinoBias and developing mitigation strategies such as data balancing and fine-tuning (Bolukbasi et al., 2016; Zhao et al., 2018; Dong et al., 2024). More recent research challenges the predominantly binary treatment of gender, revealing that models struggle with gender-neutral pronouns, show lower performance on non-binary categories, and are inconsistent in generating gender-neutral language (Hossain et al., 2023; You et al., 2024; Savoldi et al., 2025).

However, this body of work has almost exclusively focused on biases associated with labels (e.g., occupations, pronouns) rather than on the descriptive vocabulary itself. It remains underexplored whether seemingly objective physical descriptions—which are increasingly recommended as a neutral alternative to identity labels (**?**)—carry systematic gender associations. Our work addresses this gap by investigating the gendered nature of physical description language and how it is represented by both humans and language models.

3 GAPA: A Collection of Physical Attributes

In order to systematically investigate how humans and language models associate gender with text descriptions of physical attributes, we built GAPA, the largest existing collection of linguistic descriptions of physical attributes. Building a lexicon of physical attributes serves two purposes: (1) to provide a structured resource for analyzing variation in appearance-based language across sources, and (2) to enable evaluation of whether physical attributes are systematically perceived as gendered.

GAPA Set	Attributes	Participants	Questions	Total Ratings
LLM-generated	170	178	510	8,406
Novel-extracted	96	80	288	2,300
Human-written	50	46	150	4,000
Overall	315	304	945	14,706

Table 1: Summary statistics of GAPA.

To facilitate diversity of physical attributes, we collected them from distinctive sources based on LLM-generated attributes (providing broad categorical coverage; n=170), human-

written attributes (ensuring high-frequency terms; n=50), and attributes extracted from six novels (covering more unusual and stylized references; n=96), resulting in a comprehensive collection of 316 physical attributes. We provide all procedural specifications on the selection process in Appendix A.

4 Physical Attributes Carry Systematic Gender Associations for Humans

After compiling the physical attribute lexicon, we conducted a human-subject experiment to measure how strongly participants associated each attribute with different gender categories: *woman*, *man*, and *non-binary person*[2]. Specifically, we investigate: (1) whether there is an overall systematic difference in gender association ratings in GAPA, averaged across attributes and participants; (2) whether there exists a consistent shared association of gender for each attribute, alongside heterogeneity in how gender differences manifest at the attribute level in terms of direction and magnitude.

4.1 Experiment Details

The annotated dataset of GAPA includes 304 participants (152 female; ages 18–77, $M = 41.36$) who passed at least four of five attention checks. Analyses were conducted controlling for participant demographics (e.g., gender, age) and attribute source. Full details of human participant demographics and the recruitment process are provided in Appendix B.1.

On each trial, participants used a 7-point Likert scale to answer: *"How likely is it for someone to say that a woman/man/non-binary person has a given physical attribute?"* Asking what *someone* would say captures culturally available associations rather than participants' explicit endorsement of those associations, as individuals may possess stereotype knowledge that is different from personal belief (Devine, 1989; Kunda & Spencer, 2003).

Using a cross-classified design, participants rated randomly sampled attributes. To minimize carryover effects and direct cross-gender comparisons, each participant rated a given attribute for only one target gender. Counterbalancing ensured an even distribution of target gender and trial order. Each participant completed 55 trials, yielding an average of 17 ratings per attribute, per gender category.

4.2 Metrics

We define two complementary metrics to quantify the gender association of physical attributes. **Gender Association** is the mean rating for each attribute–gender pair, capturing how strongly participants associate a physical attribute with a given gender category. **Gender Distinctiveness** is the difference between the highest and second-highest mean ratings across the three gender categories for a given attribute, capturing how clearly that attribute differentiates one category from the others.

Together, these measures distinguish attributes that are strongly associated with a gender category from those that are also clearly differentiated across categories. This approach is consistent with prior work showing that category membership and social judgments are represented in graded rather than binary terms (Rosch, 1975; Rosch & Mervis, 1975; Barsalou, 1985), and that person perception depends not only on category activation but also on how diagnostic cues are for distinguishing among categories (Freeman & Ambady, 2011; Oosterhof & Todorov, 2008).

4.3 Results

Dataset-Level Gender Effects Across the dataset, attributes were associated more with women than men ($\beta = 0.26$, $SE = 0.04$, $z = 6.77$, $p < .001$), and least with non-binary people ($\beta = 0.43$, $SE = 0.04$, $z = 11.68$, $p < .001$), providing first evidence that physical attributes

[2]We include *woman*, *man*, and *non-binary person* because they reflect commonly recognized gender categories in contemporary U.S. English. These categories are not intended to represent the full range of gender identities, nor to imply that gender is naturally discrete. Instead, they serve as reference points for examining how physical attributes are associated with gender in everyday language *beyond a simple woman–man contrast*.

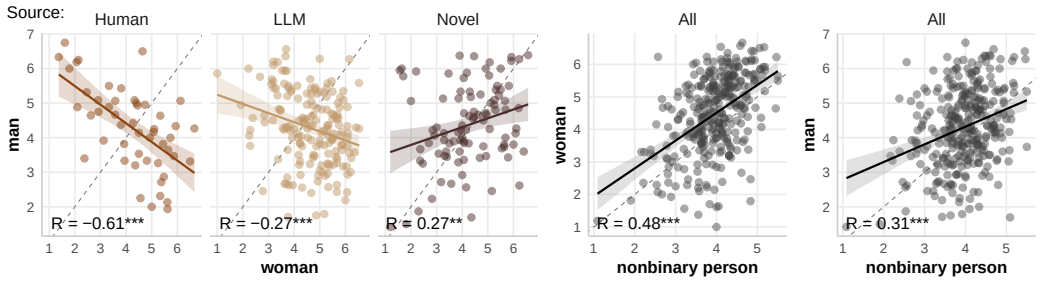


Figure 3: Correlations among mean gender-association ratings across physical attributes. Each point represents an attribute ($p < 0.05$). Man–Woman correlations are further split by source domain of the physical attributes, showing significant variation across attributes.

show structured gender profiles and are not perceived as generally gender-neutral. There are two potential explanations for the significant distinctions between genders: the dataset may overrepresent attributes that are culturally coded as feminine, or the pattern may reflect broader biases in language use, where physical descriptors are unevenly distributed across gender categories. Ultimately, the data clearly reveals that people have nuanced gender associations with physical attributes that we now aim to unpack.

Attribute-Level Gender Effects Since different attributes may be associated with gender categories to varying extent, we also conduct a significance test for each attribute in GAPA, to examine heterogeneity in gender effects across attributes.

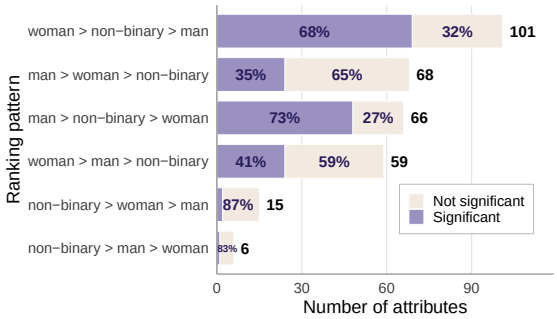


Figure 2: Attribute counts for gender-ranking patterns and their significance percentages ($p < 0.05$).

Over half of the attributes (53%) are significantly more associated with one particular gender than other genders. Importantly, the direction and strength of attribute-level gender differences vary systematically across ranking patterns (Figure 2). Attributes that align with conventional binary gender distinctions (e.g., *woman* > *non-binary* > *man* and *man* > *non-binary* > *woman*) are both more frequent and substantially more likely to yield significant effects (68%–73%). In contrast, attributes where non-binary individuals are rated highest are the rarest and exhibit much lower rates of significance. Figure 1 provides the qualitative case studies for the attributes in each ranking pattern of gender ratings.

Overall, the *attribute-level heterogeneity* suggests that gender associations carried in physical attributes are structured around a binary axis, with non-binary categories occupying a less strongly encoded region of the semantic space. In this sense, ranking patterns can be interpreted as approximations of culturally shared notions of "femininity" and "masculinity," where attributes strongly aligned with these poles produce clearer and more detectable statistical effects. This intuition is supported by a significant negative correlation of man vs. woman attribute associations ($r = -.16$, $p = .004$).

In contrast, ratings for non-binary gender associations with both women and men were significantly positively correlated (see Figure 3), highlighting how even when attributes are judged as common for a gender, it doesn't mean they are gender-diagnostic. Interestingly, splitting up the man–woman correlations by how we sourced attributes shows significant variation. While the correlation is highly negative for the human-sourced examples ($r = -0.61$), this reverses for novel-extracted physical attributes ($r = 0.27$). This variation highlights the importance of employing diverse sampling techniques and is the reason why replications across source domains are especially challenging but meaningful.

Together, these correlations indicate that gender associations in physical description are structured rather than random. This pattern suggests that physical attributes are not simply

distributed along a single binary opposition, but instead occupy partially shared and partially distinct regions of gendered meaning. In particular, non-binary associations appear to overlap more with attributes linked to women than with those strongly opposed to them.

Human Uncertainty Interpreting the gender association of physical attributes is inherently subjective and diverse among participants. To characterize this variability, we quantify *human uncertainty* using two complementary reliability statistics: a leave-one-rater-out (LOO) correlation and the intraclass correlation coefficient (ICC). The formal calculation and scientific interpretation of the two measures for human uncertainty and LLM correlation baselines are elaborated on in Appendix B.3.

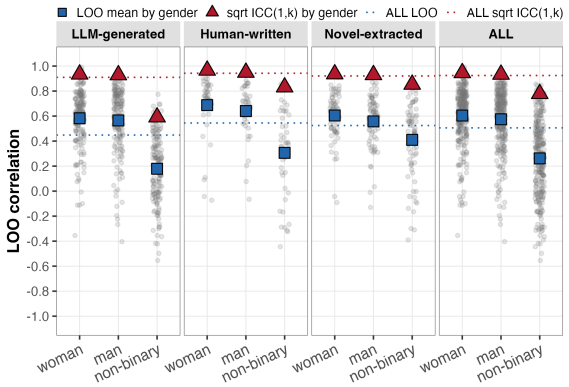


Figure 4: Human rater reliability in GAPA across three attribute sources, measured by r^{LOO} and $\sqrt{\text{ICC}(1,k)}$. The scatterplot shows r^{LOO} per rater.

In short, LOO measures how well a typical individual human rater agrees with the consensus by correlating each rater's ratings with the mean ratings of all remaining raters across items, and then averaging across raters. This reflects the correlation performance of a *typical* human annotator relative to the group. In contrast, $\text{ICC}(1,k)$ quantifies the reliability of the *aggregated* human ratings by decomposing variance into signal (between-item variability) and noise (within-item, across-rater variability), estimating how much noise remains after averaging across raters.

Figure 4 shows the noise level of human ratings across attribute sources. Human rater reliability (measured by LOO and ICC) is consistently the highest for women, intermediate for men, and lowest for non-binary targets, indicating that ratings for non-binary attributes are less consistent across raters.

5 Physical Attributes Carry Systematic Gender Associations for LLMs

After establishing that humans have reliable gender associations with text descriptions of physical attributes, we now turn to compare these to LLM gender associations. We conduct a zero-shot evaluation similar to the human experiment, across a diverse set of 16 state-of-the-art language models, varying across open-weight vs. proprietary and instruct vs. base versions.

5.1 Method

For each physical attribute in our test set, we prompt the LLMs the same question as in the human subject experiment, with minor modifications to instruct the model to directly respond with a single integer rating on a 1–7 Likert scale (see full prompt in Appendix D.1).

Two complementary strategies to elicit ratings were explored: (1) direct generation: we directly extract the verbalized ratings in the model-generated response; (2) top-logit token extraction: we examine the model's output distribution at the first generated token position. Two methods showed consistency in results, as illustrated in subsection D.2 in detail. We report the results from method (1).

5.2 Evaluation Metric & Baselines

To assess how well a model's predictions align with the averaged human ratings across attributes, we adopt the Pearson (r) correlation coefficient between *LLM-predicted ratings* and the *averaged human ratings* across items as the primary evaluation metric, and root mean squared error (RMSE) as the auxiliary metric. Higher Pearson correlation and lower RMSE suggest higher alignment with human interpretations of gender associations.

For a clearer interpretation of LLMs' correlation performance in consideration of inherent subjectivity in the human ground truth ratings, we also adopted two human performance baselines by using the uncertainty measures from section 4.3: (i) the leave-one-rater-out (LOO) correlation corresponds to the agreement level of a typical human rater, while (ii) the intraclass correlation coefficient $\sqrt{\text{ICC}(1,k)}$ approximates the maximal attainable correlation with the aggregated human judgments (i.e., an ideal model that perfectly recovers the latent true rating from the human population, so that the error only comes from measurement noise). The two human baselines suggest the correlation ceilings for the model evaluation due to rater uncertainty and measurement noise.

5.3 Results

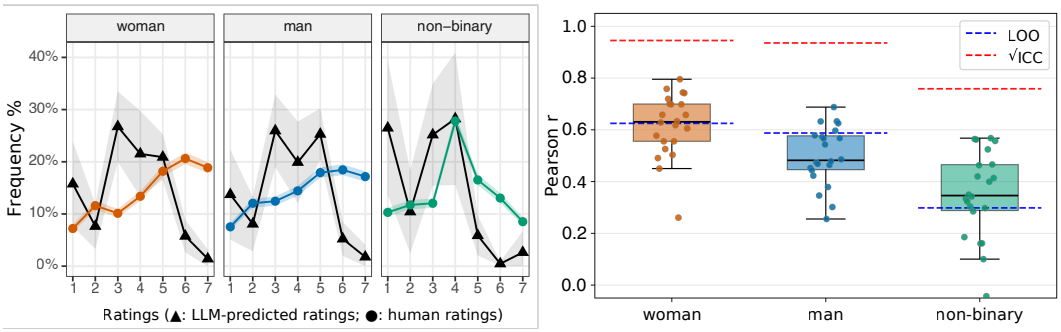
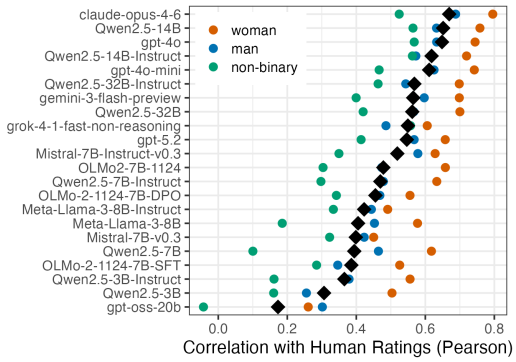


Figure 5: *Left*: the frequency% distribution of predicted rating scores (1-7) per gender category of all LLMs, compared to human ratings. *Right*: Pearson r (correlation between LLM and averaged human ratings) distribution of all models per gender category; each point represents an evaluated LLM.

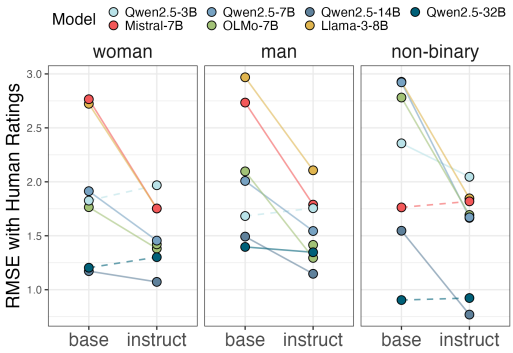
Overview of Model Alignment We first examine how well state-of-the-art (SOTA) language models align with human judgments about how gender is associated with physical attributes in GAPA at the collective level. Figure 5 (right) presents model–human correlations alongside the two reliability-based reference baselines. Overall, SOTA language models show moderate alignment with human gender associations across attributes. However, alignment varies systematically across gender categories. In particular, models tend to achieve higher correlations for *woman and non-binary person* than for *man*, centering around or exceeding the LOO (typical-rater) baseline. Interpreting these results relative to the reliability ceilings helps account for measurement noise, suggesting that the observed differences reflect genuine gender variation in model–human alignment in SOTA LLMs: SOTA LLMs are better aligned with humans in interpreting the semantic association of physical attributes with woman and non-binary person than the male category.

Notably, model rating distributions are more concentrated in the mid-range (3–5) and less frequent at the extremes (6–7) compared to human ratings, which are more evenly distributed across the scale (Figure 5, left). This pattern suggests that models tend to avoid extreme judgments, producing more conservative or less differentiated ratings than human annotators, and thus only partially capture the strength of gender associations present in human interpretations.

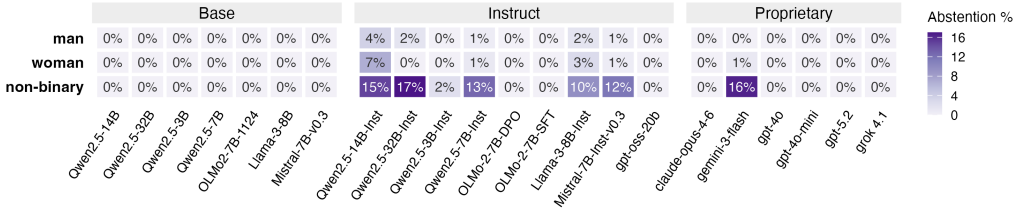
Model Comparisons Figure 6a demonstrates the results of model-human alignment in gender associations of physical attributes evaluated on the complete set of 16 models. Consistent with Figure 5 (right), gender differences remain systematically salient within each model's performance. Two other results stand out. First, Claude-Opus-4.6 achieves the strongest alignment with human gender associations, with an average Pearson correlation of r = 0.669. In contrast, GPT-OSS-20B has surprisingly the weakest overall alignment (r = 0.173) and even shows a negative correlation for the non-binary category (r = -0.043). Second, compared with prior gender-bias benchmarking work (Yang et al., 2025; Liang et al., 2022; Felkner et al., 2023), proprietary models do *NOT* show a clearly dominant separation from open-source models on GAPA. At a collective level, proprietary models show only a modest advantage, with substantial overlap in practical performance rather than a distinct lead.



(a) Correlation results of all evaluated models on GAPA. ◆ means the overall average.



(b) The comparison of model-human alignment (RMSE) between base and instruct models.



(c) Model heatmap of abstention responses on GAPA. Instruct models show significant abstention rates among models; abstention is significantly more likely to occur for the non-binary gender category.

Figure 6: Zero-shot inference experiment results of the state-of-the-art LLMs.

We also directly compare base and instruct variants of the same model backbone, as illustrated in Figure 6b. Instruction-tuned variants of open-source models generally improve over their base counterparts in RMSE. This pattern suggests that post-training alignment reduces absolute prediction divergence from the human ground-truth. However, no comparable gains are consistently observed in Pearson r, indicating that it does not substantially improve the rank-order alignment of attribute–gender associations.

Abstention Analysis We also investigate to what extent LLMs abstain from answering the question. We count a model response to be an abstention if it specifically expresses an abstention, such as "physical attributes are not related to gender identity", and "non-binary people can have any eye shape, just like anyone else." In such abstention cases, models either choose the lowest 1 (not likely at all) as the rating to express refusal and disagreement on the question assumption, or simply end the response without a rating.

We conduct a more systematic quantitative analysis on abstention counts by analyzing the model responses, which reveal two salient patterns (Figure 6c). Firstly, we observed highly asymmetric abstention in the model responses against the non-binary gender category. This indicates that the burden of refusal or deflection is not uniform across gender groups, but is disproportionately concentrated on the non-binary category. This finding crucially warns of practical accessibility bias caused by over-triggering of sensitivity and safety flags, and the behavior of AI refusal disproportionately constraining more vulnerable groups (Luo et al., 2024; Abramovich & Ma, 2024). Secondly, abstention is not a broadly shared behavior across all systems. Rather, it is concentrated almost entirely in the proprietary model Gemini-3-flash and instruct-tuned open-source models, while the corresponding base models show little to no comparable tendency. This distribution suggests that abstention is not simply a property of model scale or architecture, but is closely associated with post-training and alignment procedures. Taken together, these patterns are consistent with the interpretation that post-training can introduce asymmetric over-sensitivity, causing some models to over-trigger safety or sensitivity heuristics specifically in response to non-binary-targeted prompts.

6 Scalable Socio-Linguistic Analyses Using a Trained Proxy Predictor

6.1 Training a Proxy Model for Predicting Humans' Gender Associations

To extend the contribution of this study beyond human experiments and model evaluation analyses, we further developed a proxy predictor model on the full GAPA's data by selecting and training from the open-source models evaluated in section 5.

To seek the best model candidate for training the proxy predictor, we first conduct supervised fine-tuning across an extensive search grid of LLMs and hyperparameters on GAPA (65/15/20 split for train/validation/test). To match the prediction goal, each model is augmented with a linear regression head to directly map aggregated human ratings for gender categories $c \in \{woman, man, non\text{-}binary\ person\}$. Given an input sentence describing an attribute for a target gender (e.g., "a woman has broad shoulders"), the final-token representation is projected to a scalar score, framing the task as continuous regression to directly assess model–human alignment. The training optimizes mean squared error (MSE) on averaged human ratings, encouraging recovery of shared annotator signals while reducing individual noise.

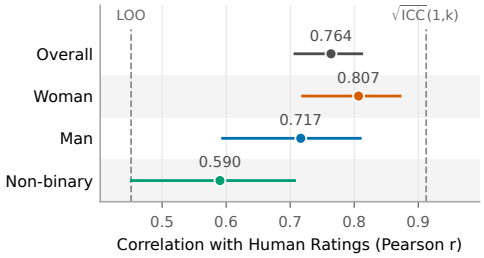


Figure 7: Training results of the predictor model, evaluated on GAPA (test).

For a fair comparison, we use a standardized hyperparameter search with model-specific ranges (e.g., learning rate, batch size), and select the best-performing configuration for each model. All models share the same tuning budget and are evaluated under consistent cross-validation and multi-seed settings. Full details including design choices and hyperparameter specifications are provided in Appendix E.

The final proxy model takes the backbone of `olmo2_7b_base`, and achieves the highest test-set correlation of $r = 0.764$ and lowest RMSE $= 0.633$. This predictor can serve as a simulator for a scalable collection of human-like gender-association interpretations on unseen items, supporting larger-scale sociolinguistic analysis.

6.2 Scaled-Up Analyses on LitBank (Bamman et al., 2019)

As a preliminary demonstration of what the predictor model can do at scale, we utilize the extraction pipeline previously used in section 3 to extract physical attributes from LitBank (Bamman et al., 2019; Sims et al., 2019), a literature corpus of 100 novels from Project Gutenberg, which consists of fiction (both novels and short stories) in a mix of genres and literary styles before 1923. After extracting the attributes, we employed the proxy model to estimate the gender association ratings of each attribute with men, women, and non-binary people, yielding a three-dimensional gender-association profile for every attribute. This enabled us to conduct exploratory analyses of their relationships with two types of metadata: author gender and the gender of the character each physical attribute describes in the original novel. We primarily explore the hypothesis[3]:

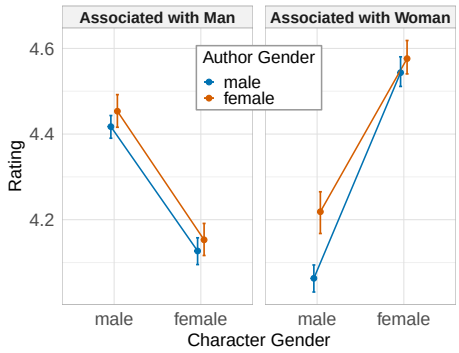


Figure 8: Predicted gender association ratings of physical attributes describing LitBank characters, by character gender and author gender.

[3]The analyses reflect findings of contemporary reader perceptions in the US due to the participant demographics of GAPA. See Appendix F for a more detailed interpretation.

the physical descriptions used to describe a lit-
erary character are systematically gendered on their own; their gender associations can be predicted by who the writer is (author gender), and about whom the description is written (character gender).

Character Gender. As shown in Figure 8, physical descriptions differentiate male and female characters through the direction rather than the magnitude of their gender associations. Attributes describing male characters are rated as more associated with man than with woman, and those describing female characters show the reverse ordering, but the two groups are comparably distinctive in absolute terms. vs. 0.71, $p = .20$).

Author Gender. Author gender modulates this structure along only one of the two binary dimensions. Ratings of association with man are largely invariant to the author gender , and female characters elicit similar descriptions from both groups of authors. Male characters do not: male authors describe men using attributes with markedly weaker feminine associations than female authors do (4.06 vs. 4.22; $\beta = -0.12$, $p = .03$). Authorial variation is thus confined to the degree of feminine association admitted into the description of a man, while masculine association follows a convention shared across author groups. The results carry the central claim of this work from controlled annotation into naturalistic text at scale: descriptions that seem independent of gender assignment remain systematically gendered by whom they describe and by who describes them.

7 Conclusion

This study challenges a widely held assumption in AI fairness, accessibility, and ethics research: that *replacing categorical identity labels with "objective" physical descriptions yields gender-neutral communication*. Introducing GAPA, a large-scale human dataset of gender associations for 316 physical attributes, we show that **this assumption does not hold**: physical descriptions carry automatic, structured, and graded gender associations among human readers, with over half of attributes significantly more associated with one gender than others. Evaluating 16 LLMs against these human ratings, we find that models partially recover human patterns yet with systematic misalignment: model ratings are compressed toward the midpoint, they align less well with human judgments of the man category, and instruction-tuned and proprietary models show asymmetric abstention that disproportionately targets the non-binary category, raising distinct equity concerns. Finally, we developed a proxy predictor to analyze gendered descriptions beyond controlled annotation settings, but at scale in naturalistic text. Together, these findings provide the first empirical evidence that "objective" physical descriptions do not neutralize gender, but instead shift gendered communication into a less visible register with downstream consequences for how models should be evaluated and designed for human–AI interaction.

Limitations

Several limitations qualify the human norming results. First, the task measures socially shared associations between physical attributes and gender categories, rather than whether attributes are inherently gendered. Ratings may therefore reflect stereotype knowledge, perceived prevalence, and expectations about language use in addition to semantic association.

Second, the attribute set is heterogeneous, spanning bodily morphology, grooming, accessories, age related cues, and figurative descriptions that may signal gender through different mechanisms. Attributes were also evaluated in isolation, without the discourse context that can shape gendered interpretation. Finally, *non-binary person* may have less culturally shared appearance-based prototypes than *woman* or *man*, so lower agreement may reflect greater conceptual heterogeneity. The findings should therefore be interpreted as evidence of structured *socially shared gender associations in physical description*, rather than a complete account of gendered meaning in language.

Acknowledgments

We thank the members of the UCLA Coalas Lab for their valuable comments throughout this project. Special thanks goes to Hillary Nguyen, Ella Han, Troy Tian, Jonathan Pak, and Ryan Zheng for their assistance with manual verification of the novel-extracted attributes in GAPA to ensure the quality of the automated pipeline of physical attribute extraction. We further thank Rohan Jain, Maureen Widjaja and Karin Yamaoka for their feedback early on in the project. Finally, we're grateful to Google's GiG and GCP Credit Program for their support of this research.

Ethics Statement

This study examines socially shared gender associations in linguistic descriptions of physical attributes. We emphasize that our work does not define or infer individuals' gender from appearance, but instead analyzes patterns in language and perception. Because such analyses risk reinforcing stereotypes if misused, we frame all findings descriptively and caution against using the dataset or models for gender inference, profiling, or decision-making about real individuals.

Human-subject data were collected with informed consent, fair compensation, and no personally identifiable information beyond basic demographics. All analyses are conducted at the aggregate level. The dataset and models are released for research purposes, such as auditing implicit bias and improving AI transparency, and should be used with appropriate safeguards and contextualization.

References

Shira Abramovich and Anna J Ma. Evaluating refusal. In *Workshop on Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI at NeurIPS*, 2024.

David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2138–2144, 2019.

Lawrence W Barsalou. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4):629–654, 1985.

Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. "it's complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pp. 1–19, 2021.

Gabriel Besson, Gladys Barragan-Jason, Simon J Thorpe, Michèle Fabre-Thorpe, Sébastien Puma, Mathieu Ceccaldi, and Emmanuel J Barbeau. From face processing to face recognition: Comparing three different processing levels. *Cognition*, 158:33–43, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in neural information processing systems*, 29, 2016.

Elizabeth Brown and David I Perrett. What gives a face its gender? *Perception*, 22(7):829–840, 1993.

Mary Bucholtz and Kira Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614, 2005.

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

Lee Joseph Cronbach. The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, pp. 1–33, 1972.

Patricia G Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5–18, 1989.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*, 2024. URL https://arxiv.org/abs/2402.11190.

Penelope Eckert and Sally McConnell-Ginet. *Language and Gender*. 2003.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9126–9140, 2023.

Jonathan B Freeman and Nalini Ambady. A dynamic interactive theory of person construal. *Psychological review*, 118(2):247–279, 2011.

Jonathan B Freeman and Kerri L Johnson. More than meets the eye: Split-second social perception. *Trends in cognitive sciences*, 20(5):362–374, 2016.

Jonathan B Freeman, Kerri L Johnson, Reginald B Adams Jr, and Nalini Ambady. The social-sensory interface: category interactions in person perception. *Frontiers in integrative neuroscience*, 6:81, 2012.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–13, 2018.

Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 543–554, 2021.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. Misgendered: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5352–5367, 2023.

Kerri L. Johnson and Louis G. Tassinary. Perceiving sex directly and indirectly: Meaning in motion and morphology. *Psychological Science*, 16(11):890–897, 2005. doi: 10.1111/j.1467-9280.2005.01633.x.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Ziva Kunda and Steven J Spencer. When do stereotypes come to mind and when do they color judgment? a goal-based theoretical framework for stereotype activation and application. *Psychological bulletin*, 129(4):522–544, 2003.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Sha Luo, Sang Jung Kim, Zening Duan, and Kaiping Chen. Refusal as silence: Gendered disparities in vision-language model responses. *arXiv preprint arXiv:2406.08222*, 2024.

Arnau Marin-Llobet, Simon Henniger, and Mahzarin R Banaji. Vision-language models suppress female representations under ambiguous input. *arXiv preprint arXiv:2605.31556*, 2026.

Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

Nikolaas N Oosterhof and Alexander Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008.

Eleanor Rosch. Cognitive reference points. *Cognitive psychology*, 7(4):532–547, 1975.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

Beatrice Savoldi, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mgente. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13709–13731, 2025.

Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1):1–35, 2020.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Matthew Sims, Jong Ho Park, and David Bamman. Literary event detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3623–3634, 2019.

Charles Spearman. The proof and measurement of association between two things. 1961.

Candace West and Don H Zimmerman. Doing gender. *Gender & society*, 1(2):125–151, 1987.

Jinrui Yang, Xudong Han, and Timothy Baldwin. Demographics and democracy: Benchmarking llms' gender bias and political leaning in european parliament. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pp. 416–439, 2025.

Zhiwen You, HaeJin Lee, Shubhanshu Mishra, Sullam Jeoung, Apratim Mishra, Jinseok Kim, and Jana Diesner. Beyond binary gender labels: Revealing gender bias in llms through gender-neutral name predictions. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 255–268, 2024.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 629–634, 2019.

A Supplementary Information of Dataset Collection for GAPA

Physical attributes were collected from three distinct source domain to facilitate both large coverage, cognitively prominent concepts, and rare and stylized descriptors.

LLM-generated attributes were collected by prompting GPT-4o and Gemini 2.0 to list physical attributes that follow the pattern of "{ADJ} + {NOUN}", where {ADJ} describes the properties of a body part that is referred to by {NOUN} (e.g., "an hourglass figure", "a chiseled jawline").

Human-written attributes were collected by presenting participants with a diverse set of images depicting individuals whose facial appearance contained both stereotypically gender-typical and gender-atypical features: $N = 26$ images generated by GPT 4o and $N = 22$ naturalistic images selected from the Flickr-Faces-HQ dataset (Karras et al., 2019). During the study, each participant saw a total of 10 images. Participants viewed one image at a time and were asked, "How strongly do you associate this person's appearance with that of a woman/man/non-binary person in your culture?" using a 7-point scale. After providing their ratings, participants were prompted to list three features of the person's appearance that led them to associate the individual with the gender category they rated highest. They were also given the option to identify at least one feature of the person's appearance that seemed unusual or atypical for someone they would associate with the gender category they rated highest.

Attributes extracted from novels were collected using a specially designed extraction pipeline for sourcing physical attributes from any long-context documents or literature corpora such as novels. The design and implementation details of the extraction pipeline is elaborated in Appendix C. To construct the novel-extracted set of the current dataset, we chose six popular contemporary novels of similar size, genre, and a balanced woman-man author ratio: *Game of Thrones* (Vol. 1), *Harry Potter* (Vol. 1), *Twilight* (Vol. 1), *Hunger Games* (Vol. 1), *Lord of the Rings* (Vol. 1), and *Maze Runner*. From each book source, we randomly sampled 16 attributes from the extracted physical attributes.

B Supplementary Information for Human Annotations on GAPA

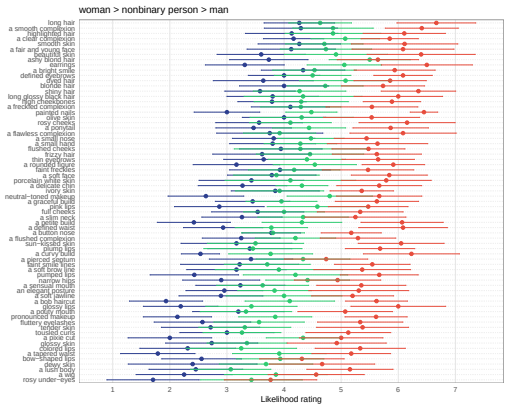
B.1 Participant Demographics

The study was certified as exempt from IRB review by the UCLA Office of the Human Research Protection Program (IRB-24-5529; 45 CFR 46.104). All participants provided informed consent prior to participation. We recruited 320 US participants via Prolific ($12/hour). Recruitment criteria required participants to be native English speakers, residing in the United States, with a Prolific approval rate of at least 95% and passing an attention check. The final sample included 75 female. The participants ranged in age from 19 to 83 years (M = 42.70).

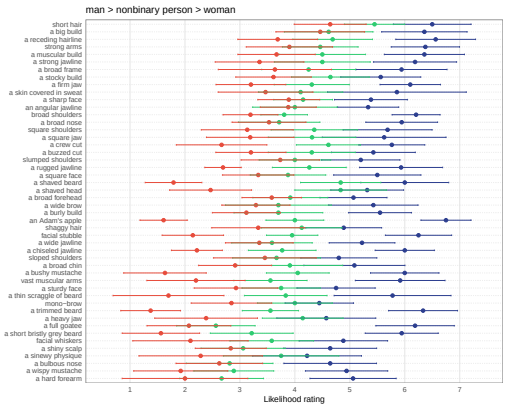
B.2 Descriptive Results

Table 2 reports descriptive statistics of human-rated gender likelihood for attributes across three sources (human-written, LLM-generated, and novel-derived) for each target gender. A consistent pattern emerges across sources: ratings are highest for women, followed by men, and lowest for non-binary targets. This ordering holds within most individual sources and in the overall aggregates. This suggests a systematic difference in how our collected attributes are overall associated with gender categories in human judgments.

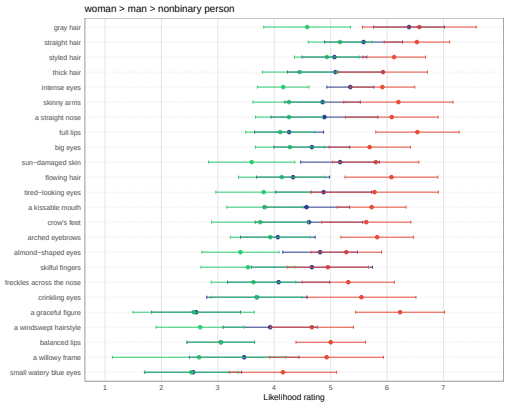
Figure 9 illustrates the per-gender ratings of all attributes with significant gender differences ($p < 0.05$) in the complete six ranking patterns.



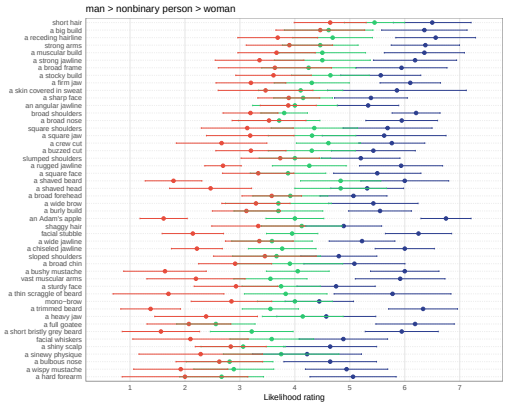
(a) W > NB > M



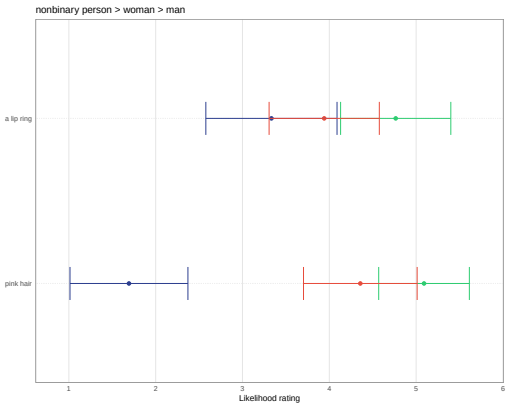
(b) M > NB > W



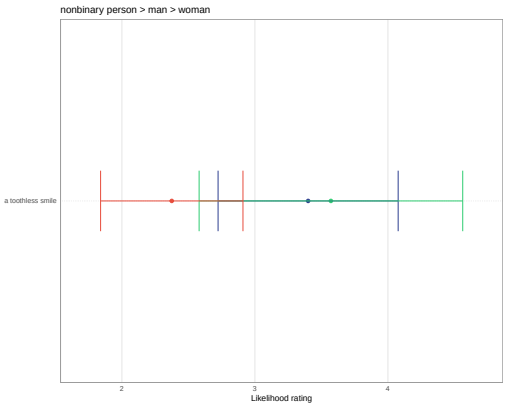
(c) W > M > NB



(d) M > NB > W



(e) NB > W > M



(f) NB > M > W

Figure 9: Per-gender ratings of all attributes with significant gender differences ($p < 0.05$) in the six ranking patterns.

Table 2: Descriptive statistics by gender and attribute source

Gender	Source	n	Mean	SD	CI Low	CI High
Man	Human	744	4.27	1.83	4.14	4.40
	LLM	2760	4.25	1.85	4.19	4.32
	Novel	1364	4.38	1.93	4.28	4.48
	Overall	4868	4.29	1.87	4.24	4.34
non-binary	Human	762	4.05	1.55	3.94	4.16
	LLM	2766	4.07	1.66	4.01	4.14
	Novel	1354	3.76	1.85	3.66	3.86
	Overall	4882	3.98	1.71	3.93	4.03
Woman	Human	794	4.20	1.93	4.07	4.33
	LLM	2880	4.71	1.80	4.64	4.77
	Novel	1282	4.24	1.97	4.13	4.35
	Overall	4956	4.50	1.88	4.45	4.56

B.3 Correlation Ceilings from Human Uncertainty

As the target labels $\bar{y}_i$ are derived from the average of noisy human judgments that are highly subjective and diverse (see subsection 4.3), **correlation performance is fundamentally bounded by inter-rater variability** (Cronbach, 1972). We therefore estimate two complementary noise ceilings: a *leave-one-rater-out* (LOO) ceiling and an *intraclass correlation* (ICC) ceiling, illustrated in Figure 4. They later serve as interpreting references for models of different strengths Figure 5 (right).

Leave-One-Out Reliability Ceiling. The LOO ceiling estimates the agreement between a typical individual rater and the consensus of other raters. For each rater j, we compute r_j^{LOO}, the Pearson correlation between that rater's ratings and the mean ratings of all remaining raters across items:

$$r_j^{\mathrm{LOO}} = \mathrm{corr}\big(\{y_{ij}\}, \{\bar{y}_{i,(-j)}\}\big),$$

$$\text{where} \quad \bar{y}_{i,(-j)} = \frac{1}{k-1} \sum_{j' \neq j} y_{ij'}.$$

The LOO ceiling r^{LOO} is then obtained by averaging r_j^{LOO} across raters. This ceiling value reflects *human-level performance*: a model (with parameter $\theta_{typical}$) achieving LOO-level correlation performs at the same level as any *typical* individual annotator. Consequently, a model with a comparable performance to a typical human rater with idiosyncratic rater noise is expected to satisfy

$$r(\theta_{typical}) \;\sim\; r^{\mathrm{LOO}}.$$

Intraclass Correlation (ICC) Ceiling. In contrast, the intraclass correlation coefficient $\mathrm{ICC}(1,k)$ estimates the reliability of the *mean* rating across k raters, describing the distribution of the collected human rating data. ICC is a variance-partitioning reliability coefficient derived from a random-effects model and is widely used to quantify inter-rater reliability in behavioral measurement (Shrout & Fleiss, 1979; McGraw & Wong, 1996). Under a standard random-effects model,

$$y_{ij} = \tau_i + \varepsilon_{ij},$$

where τ_i denotes the *latent true attribute* value for item i and ε_{ij} represents rater-specific noise assumed to be independent with mean zero, the population-level ICC for the mean of k raters is:

$$\mathrm{ICC}(1,k) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2/k}.$$

$\mathrm{ICC}(1,k)$ represents the proportion of variance in the averaged human ratings attributable to the latent signal rather than noise, which has been adopted as the reliability of the averaged human judgment across psychometrics literature (Shrout & Fleiss, 1979; Cronbach, 1972).

Based on classical measurement theory, correlation between a noisy measurement and any predictor is bounded by the square root of that measurement's reliability (i.e., known as attenuation due to measurement error) (Spearman, 1961; McGraw & Wong, 1996). Consequently, $\mathrm{ICC}(1,k)$ provides a *natural upper bound* on the attainable correlation of an ideal model (with parameter θ_{ideal}) that does not contain any rater noise in any rating prediction with the averaged human ratings. In other words, $\mathrm{ICC}(1,k)$ describes the correlation of an ideal model that perfectly recovers the latent true rating from the human population (by always predicting τ_i i.e., $\hat{y}_i(\theta_{ideal}) = \tau_i$)), offering a stricter theoretical noise ceiling than r^{LOO} for correlation-based evaluation:

$$r(\theta_{ideal}) \sim \mathrm{corr}(\tau_i, \bar{y}_i) = \sqrt{\mathrm{ICC}(1,k)}.$$

While r^{LOO} and $\sqrt{\mathrm{ICC}(1,k)}$ provide the theoretical ceiling references for interpreting model correlation values, it is important to note that the empirical results of model performances may occasionally exceed them, due to the sampling noise in the finite set of collected data for evaluation.

C Extraction Pipeline

We describe the data source, the extraction pipeline implemented, and the schema and features of the extracted attributes used in our experiments.

C.1 Extraction Pipeline Process

1. Lexicon-based sentence filter. Sentences are obtained with NLTK's `sent_tokenize`. We then restrict LLM calls to sentences that are likely to contain physical descriptions. Two hand-defined lexicons are used: (a) **physical-attribute nouns**: body-part terms (e.g., head, face, hair, eyes, nose, jaw, hands, skin, build, posture) and related nouns (e.g., silhouette, frame); (b) **physical-attribute adjectives**: descriptors for build/size (e.g., thin, broad, tall), musculature (e.g., muscular, toned), face and eyes (e.g., sharp, round, deep-set), hair (e.g., curly, blond), skin (e.g., freckled, pale-skinned), and evaluative terms (e.g., beautiful, handsome). Tokens are normalized (lowercased, hyphens removed) and matched after NLTK lemmatization.

2. LLM-based extraction. Each filtered sentence is sent to gpt-4o-mini with a fixed prompt that instructs the model to: (i) extract descriptions of a person's physical attributes from the given text by outputing each description as a noun phrase that removes gender pronouns (e.g., "his narrow jaw" $\rightarrow$ "a narrow jaw"); this requires both a body part and a characterizing modifier (e.g., "dark brown hair," "piercing green eyes") and to skip bare body-part mentions (e.g., "a beard") or purely evaluative phrases without a body part (e.g., "a slender girl"); (ii) for each extracted attribute, assign a **gender** label (`male`, `female`, `non-binary`, or `unknown`) and provide reasoning.

3. Postprocessing and filtering. The pipeline postprocesses the then normalizes gender pronouns in the attribute strings: possessive "his"/"her" are replaced by "a"/"an"; pronouns such as "he"/"she"/"him"/"herself"/"himself"/"hers" are removed. Specifically, we also classify the extracted attributes from novels in two categories in terms of whether their formats are consistent with the LLM-generated or human-written physical attributes: **(a) in-domain-format attributes** are attributes that follow exactly the same pattern {[(a/an)] + ADJ + NOUN} (e.g., "a narrow jaw," "black hair," "a long beard," "a wide smile"); **(b) out-of-domain-format attributes** are physical attributes that do not exactly match the pattern, including: more than two content words (e.g. "bright green eyes," "vast muscular arms"),

different word order or POS (e.g. noun before adjective, or no clear single adjective + single noun), longer or more complex noun phrases.

C.2 Prompt for LLM-Based Extraction

```
You are an expert at physical attribute extraction from text. Your task is to
extract physical attributes and gender of a person from a book.
You will be given unstructured text from a book that potentially describes physical
 features of a person. Ignore descriptions of clothes.
You should first extract all the descriptions of a person's physical attributes and
 convert them into a **noun phrase structure that removes gender pronouns like 'his
' or 'her'**. Descriptions of a person's physical attributes are defined as
adjectives or noun phrases describing the characteristics of a body part. You
should extract complete noun phrases including both the body part and its
description, such as "a narrow jaw", "dark brown hair", "piercing green eyes", and
skip extraction of a body part without its description (e.g., "a beard"). A word or
 phrase merely referring to a body part without any chacracteristic description in
the text, such as "a beard" or "his hair", should not be extracted; a mere
description without a body part in the text also should not be extracted (e.g., "a
slender girl"). If there is no description of physical attributes that meet the
extraction criteria, simply return an empty list.
Then, for each physical attribute, you should extract or infer the gender of the
person described by the attribute, by labeling it as "male", "female", or "non-
binary" if there is solid evidence in the text supporting one of them, or "unknown"
 if there is no evidence for inference.
Provide your reasoning and justification for each attribute's gender label.

For example:
<text>
Harry Potter was a young man with a thin, angular face, sharp features, and a
narrow jaw.
</text>
[{"attribute":"a thin face", "reasoning":"The attribute is used to describe 'a
young man', explicitly suggesting the person's gender; 'Harry' is also a common
male name.", "gender":"male"}, {"attribute":"an angular face", "reasoning":"The
attribute is used to describe 'a young man', explicitly suggesting the person's
gender; 'Harry' is also a common male name.", "gender":"male",}, {"attribute":"
sharp features", "reasoning":"The attribute is used to describe 'a young man',
explicitly suggesting the person's gender; 'Harry' is also a common male name.", "
gender":"male", }, {"attribute":"a narrow jaw", "reasoning":"The attribute is used
to describe 'a young man', explicitly suggesting the person's gender; 'Harry' is
also a common male name.", "gender":"male"}]

<text>
Her hair was dark brown and curly, and her eyes were a piercing green, staring at
the man intently.
</text>
[{"attribute":"dark brown hair", "reasoning":"The text uses the pronoun 'her' to
refer to the person with the current attribute ('dark brown hair'), explicitly
suggesting the person's gender as female.", "gender":"female"}, {"attribute":"curly
 hair", "reasoning":"The text uses the pronoun 'her' to describe the person whose
current attribute ('currly hair') is being described, explicitly suggesting the
person's gender as female.", "gender":"female"}, {"attribute":"piercing green eyes",
 "reasoning":"The text uses the pronoun 'her' to describe the person whose eyes are
 being described, explicitly suggesting the person's gender as female.", "gender":"
female"}]

<text>
A thick lump grew in his throat.
</text>
[]

<text>
The nurse's eyes were dark blue.
```

```
</text>
[{"attribute":"dark blue eyes", "reasoning":"The text describes the nurse's eyes as
 'dark blue' without any clues about the person's gender.", "gender":"unknown"}]

<text>
His  foot  pushed  mine  off  the  gas  pedal.
</text>
[]

<text>
He took her small pink hand in his own frail spotted one and gave it a gentle
squeeze.
</text>
[{"attribute":"a small pink hand", "reasoning":"The text uses the pronoun 'her' to
refer to the person with the current attribute ('a small pink hand'), explicitly
suggesting the person's gender as female.", "gender":"female"}, {"attribute":"a
frail spotted hand", "reasoning":"The text uses the pronoun 'his' to refer to the
person with the current attribute ('a frail spotted hand'), explicitly suggesting
the person's gender as male.", "gender":"male"}]

<text>
Hers was not a pretty face, alas.
</text>
[{"attribute":"a pretty face", "reasoning":"The text uses the pronoun 'hers' to
refer to the person with the current attribute ('a pretty face'), explicitly
suggesting the person's gender as female.", "gender":"female"}]
```

D Zero-Shot LLM Inference Experiments

D.1 Prompt Template

```
How likely is it for someone to say a {person_term} has {attribute}? Your answer
should only be a single number between 1 (Not at all) and 7 (Extremely likely).
Answer:
```

D.2 Method Details

Two complementary strategies were adopted to elicit ratings based on varied constraints from different LLMs: **(1) direct generation**: we directly extract the verbalized ratings in the model-generated response; **(2) top-logit token extraction**: we examine the model's output distribution at the first generated token position: for open-weight models, we rank top k tokens with the highest logit values, select the token that contains any number in 1–7 can be extracted, and use the number as model rating. This is a more tolerant approach than simply selecting the highest-logit token among the seven tokens 1-7, as it considers common cases of zero-shot output that contain blanks adjacent to the actual answering, causing unexpected tokenization results; for API-based models, we similarly retrieve log-probabilities for the top-ranked tokens (if available) and identify the highest-probability token containing a valid rating. Incorporating both strategies ensures the collection of model rating preferences even without demanding their instruction-following abilities, and provides a robust and thorough investigation of models' encoding of gender in physical descriptions from not only the model's explicit textual response but also its underlying distributional preferences.

Method 1 and 2 showed a close-to-perfect agreement rate, serving as sanity check, except for the Mistral-7b models. Considering the higher accessibility of method 1, we report the ratings of the directly generated responses from models as the primary results.

D.3 Complete Results

Figure 6a in the main body text presents the complete ratings in comparison with human ratings across genders for all models. Below, Table 3 shows the per-gender RMSE & Pearson r statistics of all evaluated models in the experiment of gender association on GAPA, with significance values included in the brackets.

Table 3: Per-gender RMSE & Pearson r statistics of all evaluated models in the experiment of gender association on GAPA.

Model	Dataset	woman		man		non-binary	
		RMSE	Pearson r (p)	RMSE	Pearson r (p)	RMSE	Pearson r (p)
Meta-Llama-3-8B	LLM-generated	2.67	0.60 (0.000)	2.76	0.50 (0.001)	2.95	0.16 (0.308)
	Novel-extracted	2.82	0.56 (0.000)	3.09	0.49 (0.000)	2.85	0.19 (0.060)
	Human-written	2.57	0.59 (0.000)	2.90	0.39 (0.005)	3.06	0.23 (0.102)
	All	2.72	0.58 (0.000)	2.97	0.45 (0.000)	2.93	0.19 (0.011)
Llama-3-8B-Inst	LLM-generated	1.70	0.36 (0.019)	2.00	0.47 (0.002)	1.96	0.05 (0.775)
	Novel-extracted	1.90	0.48 (0.000)	2.21	0.48 (0.000)	1.84	0.32 (0.002)
	Human-written	1.48	0.59 (0.000)	2.00	0.38 (0.007)	1.77	0.49 (0.000)
	All	1.75	0.49 (0.000)	2.11	0.44 (0.000)	1.85	0.33 (0.000)
Mistral-7B-v0.3	LLM-generated	3.13	0.42 (0.006)	2.90	0.59 (0.000)	2.11	0.19 (0.227)
	Novel-extracted	2.75	0.36 (0.000)	2.67	0.38 (0.000)	1.76	0.34 (0.001)
	Human-written	2.46	0.63 (0.000)	2.72	0.42 (0.002)	1.40	0.42 (0.003)
	All	2.77	0.45 (0.000)	2.73	0.42 (0.000)	1.76	0.32 (0.000)
Mistral-7B-Inst	LLM-generated	1.43	0.58 (0.000)	1.79	0.52 (0.000)	1.85	0.22 (0.160)
	Novel-extracted	1.89	0.61 (0.000)	1.86	0.57 (0.000)	1.87	0.32 (0.002)
	Human-written	1.72	0.69 (0.000)	1.65	0.65 (0.000)	1.69	0.46 (0.001)
	All	1.75	0.63 (0.000)	1.79	0.58 (0.000)	1.82	0.35 (0.000)
OLMo2-7B-1124	LLM-generated	1.49	0.72 (0.000)	2.05	0.20 (0.199)	2.92	0.25 (0.117)
	Novel-extracted	1.90	0.62 (0.000)	2.20	0.53 (0.000)	2.70	0.35 (0.001)
	Human-written	1.71	0.68 (0.000)	1.93	0.53 (0.000)	2.82	0.29 (0.040)
	All	1.77	0.66 (0.000)	2.10	0.47 (0.000)	2.78	0.30 (0.000)
OLMo-2-7B-DPO	LLM-generated	1.37	0.36 (0.018)	1.34	0.24 (0.128)	1.66	0.25 (0.110)
	Novel-extracted	1.41	0.56 (0.000)	1.38	0.48 (0.000)	1.71	0.36 (0.000)
	Human-written	1.34	0.64 (0.000)	1.08	0.61 (0.000)	1.69	0.29 (0.042)
	All	1.38	0.56 (0.000)	1.30	0.47 (0.000)	1.69	0.34 (0.000)
Qwen2.5-14B	LLM-generated	1.04	0.72 (0.000)	1.44	0.60 (0.000)	1.52	0.41 (0.007)
	Novel-extracted	1.22	0.75 (0.000)	1.60	0.59 (0.000)	1.64	0.60 (0.000)
	Human-written	1.19	0.82 (0.000)	1.32	0.80 (0.000)	1.38	0.51 (0.000)
	All	1.17	0.76 (0.000)	1.49	0.63 (0.000)	1.55	0.56 (0.000)
Qwen2.5-14B-Inst	LLM-generated	0.99	0.68 (0.000)	1.03	0.53 (0.000)	0.64	0.26 (0.097)
	Novel-extracted	1.17	0.72 (0.000)	1.31	0.58 (0.000)	0.87	0.61 (0.000)
	Human-written	0.92	0.78 (0.000)	0.89	0.70 (0.000)	0.67	0.40 (0.004)
	All	1.07	0.72 (0.000)	1.15	0.57 (0.000)	0.77	0.56 (0.000)
gpt-4o	LLM-generated	0.98	0.76 (0.000)	1.08	0.59 (0.000)	0.61	0.20 (0.211)
	Novel-extracted	1.23	0.72 (0.000)	1.36	0.69 (0.000)	0.84	0.62 (0.000)
	Human-written	0.92	0.80 (0.000)	1.01	0.65 (0.000)	0.72	0.43 (0.002)
	All	1.10	0.74 (0.000)	1.22	0.63 (0.000)	0.76	0.57 (0.000)

E Training Experiments for the Proxy Predictor Model

E.1 Training Details

We use the LLM-generated subset of GAPA for training, validation, and in-domain evaluation (split ratio = 60/15/25), and reserve the novel-derived and human-written attribute sets as out-of-domain evaluation benchmarks. Complementing the zero-shot generation approach in section 5, we train each model with an added linear regression head to predict aggregated human ratings of physical attributes for different gender categories $c \in \{woman, man, non\text{-}binary\ person\}$. For each training instance, the model takes as input a sentence describing a physical attribute with respect to a target gender (e.g., "a woman has broad shoulders"). The final-token hidden representation is passed through a one-dimensional linear projection to produce a scalar prediction corresponding to the human likelihood rating for that attribute–gender pair. This formulation treats the task as

continuous regression, avoiding reliance on explicit numerical generation and providing a more direct probe of model–human alignment.

The model is optimized using a mean squared error (MSE) loss between predicted scores and averaged human ratings. Let y_{ij} denote the rating provided by rater j for item i, and $\bar{y}_i = \frac{1}{k}\sum_{j=1}^{k} y_{ij}$ the mean across k raters. Training minimizes $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2$, where $\hat{y}_i$ is the model prediction and N is the number of samples. Training on aggregated ratings encourages the model to capture the shared signal across annotators while attenuating idiosyncratic noise.

To ensure fair comparison across models with different architectures and scales, we adopt a standardized hyperparameter search template, with model-specific ranges for sensitive parameters such as learning rate and effective batch size. Models are compared based on their best performance across the search. Each model is allocated the same tuning budget and evaluated under consistent cross-validation and multi-seed settings to improve robustness. Full training details and configurations are provided in Appendix E.

E.2 Full Configurations for Hyperparameter Search

Table 4 lists the hyperparameter configuration for hyperparameter search in the training ablation experiments, aiming for a relatively controlled comparison across model family, scale, and instruction-tuning variants.

Models	Llama 3-8B (base and instruct), Mistral-7B (base and instruct), Qwen2.5-3B(base and instruct), Qwen2.5-7B(base and instruct), Qwen2.5-14B(base and instruct), Qwen2.5-32B(base and instruct), OLMo-2-7B (base, SFT, and DPO), and GPT-OSS-20B.
Learning rate	$\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$
Batch size	$\{2, 4\}$
LoRA rank (r)	$\{8, 16\}$
LoRA scaling (α)	$\{16, 32\}$
Early stopping	Enabled (patience = 6)
Cross-validation	K-fold ($K = 5$)
Seeds/Replicates	Aggregated from 3 independent runs per fold (seeds: 42, 123, 456)

Table 4: Training hyperparameter configuration, aiming for a relatively controlled comparison across model family, scale, and instruction-tuning variants.

F How to Interpret the Exploratory LitBank Analyses

It is important to note that the analyses in subsection 6.2 reflect predicted gender associations as they are held by the US participants we recruited in recent years. Since gender associations change over time, the following results should not be interpreted as a clear window into the authors' absolute biases decades ago. Instead, it's a window into the associations a common reader may take away today and comparative analyses between groups of authors can inform hypotheses about potentially persistent gendered perceptions.