POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

The Persona Effect: Analyzing How Personality Traits Amplify Gender Bias in Hindi and English Large Language Model Narratives

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Tanay Kumar, Shreya Gautam

Advisor: Prof. Francesco Pierri

Academic year: 2025-2026

1. Introduction

Large language models (LLMs) are increasingly deployed in persona-driven contexts such as creative writing assistants, educational tutors, customer service agents, and social simulation platforms, that condition models on combinations of demographic labels and psychological persona descriptions. A growing body of research confirms that demographic conditioning can systematically shift generated text toward culturally encoded stereotypes [6], with measurable behavioral consequences: in a controlled Prisoner's Dilemma experiment, users exploit female-labeled AI agents at higher rates and distrust male-labeled agents more than human counterparts carrying the same labels [6], thus establishing a direct pathway from stereotyped outputs to discriminatory user behavior.

A critical and underexamined question remains: does the *psychological* configuration of a persona, that is, its personality traits, independently modulate the expression of gender-stereotypical language, and if so, how does this interact with demographic labels, occupational context, language, and model architecture?

This thesis addresses that question through a large-scale multilingual experimental study. We systematically condition six state-of-the-art LMs on 9 personality traits from the HEXACO model and the Dark Triad framework, cross-factored with 3 persona genders, 50 occupational contexts, and 2 languages, English and Hindi, yielding a corpus of 23,400 generated narratives. Using a continuous, embedding-based bias metric grounded in cosine similarity to gender-stereotypical centroid representations, we measure story-level gender-stereotypical alignment and analyze the joint effects of personality conditioning, persona gender, occupational context, language, and model architecture.

The central finding is that personality conditioning is *not* a neutral stylistic parameter. Antagonistic Dark Triad traits act as bias amplifiers, prosocial HEXACO traits act as attenuators, and in several conditions the personality effect *exceeds* that of the explicit gender label. Amplification scales with trait intensity: high-score persona descriptions produce consistently larger distributional displacements than low-score descriptions, establishing personality as a continuous modulator. These effects persist across all six architectures, and Hindi exhibits a systematic compression of magnitude attributable to its obligatory morphological gender-marking system.

2. Background and Related Work

2.1. Gender Bias in LLMs

Prior work on gender bias in LLMs has primarily focused on how explicit demographic labels such as pronouns, gendered names, or role nouns alter the semantic content of model outputs [2, 6]. Embedding-based bias measurement originates with the Word Embedding Association Test (WEAT), which quantifies the association of target concept embeddings with attribute concept embeddings using cosine similarity differences [1]. Our approach extends this framework from static word embeddings to contextual sentence embeddings of multi-sentence narratives, enabling measurement of stereotype alignment at the discourse level.

2.2. Personality Frameworks

The HEXACO model decomposes personality into six personality dimensions: *Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness.* The Dark Triad comprises three constructs associated with antagonistic and self-enhancing behavior: *Machiavellianism, Narcissism, and Psychopathy.* Research in human social psychology has documented asymmetric relationships between these trait clusters and gender-stereotypical cognition: antagonistic traits are associated with dominance-coded behavior patterns culturally coded as masculine, while prosocial traits are more evenly distributed across gender categories [3]. Whether LLMs reproduce these regularities in generated text in a cross-linguistically generalizable way was unknown prior to this work.

2.3. Multilingual Considerations

All embedding computations use the L3Cube `indic-sentence-similarity-sbert` multilingual encoder [5]. Using a single encoder for both centroid construction and story embedding ensures that similarity values are computed within a shared geometric space. Hindi presents structurally distinct challenge: an obligatory morphological gender-marking system inflects adjectives and verb participles for grammatical gender, creating sentence-level gender signals absent from English's pronoun-centered system [4].

3. Methodology

3.1. Dataset Construction

The experimental corpus consists of 23,400 short narratives generated by prompting six LLMs with fully crossed combinations of:

- **Persona gender**: male, female, neutral;
- **Personality condition**: 9 traits at high and low intensity, and a no-personality baseline;
- **Occupation**: 50 roles balanced between male and female stereotyped professional categories;
- **Language**: English and Hindi;
- **Model**: 6 architecturally diverse models that are *LLMs* (GPT-5 nano, LLama-3.3-70B), *SLM* (Gemma-3-1B) *MoE* (Mistral-8x7B), *RLM* (DeepSeek-R1), *SSM* (Falcon-Mamba-7B)

For each occupation–gender–language–model cell, three baseline stories are generated without personality specification to establish occupational and demographic priors. Generation stability analysis confirms a mean intra-configuration cosine distance of approximately 0.01, justifying three baseline observations per cell.

3.2. Personality Prompt Design

A central methodological contribution of this work is the construction of explicit high-score and low-score persona descriptions for all nine personality traits. Each trait is operationalized at two intensity levels: a **high-score** description articulates the trait in its most salient and behaviorally extreme form, while a **low-score** description characterizes the opposing pole of the same dimension. For Dark Triad traits, high-score descriptions foreground manipulation, grandiosity, and callousness; low-score descriptions foreground cooperativeness, modesty, and empathy. Descriptions deliberately avoid occupation-specific or gender-specific vocabulary, isolating the personality signal from potential confounds. The descriptions are reviewed by the two authors and validated using SD3 test [].

3.3. Bias Metric

Each story S is segmented into sentences $\{s_1, \ldots, s_n\}$, each embedded into a 768-dimensional contextual representation $\mathbf{s}_i$. Male

and female stereotype centroids $\mathbf{c}_{\text{male}}$ and $\mathbf{c}_{\text{female}}$ are constructed as mean embeddings of curated gender-stereotypical attribute word sets. The sentence-level bias score is:

$$\text{bias}(s_i) = \cos(\mathbf{s}_i, \mathbf{c}_{\text{male}}) - \cos(\mathbf{s}_i, \mathbf{c}_{\text{female}}), \quad (1)$$

where positive values indicate net male-stereotypical alignment. The story-level score aggregates via max-absolute selection:

$$Y(S) = \text{bias}(s_k), \quad k = \arg\max_i |\text{bias}(s_i)|, \quad (2)$$

motivated by the psychological salience of the most evaluatively extreme sentence in a narrative. To isolate personality effects from occupational priors, the baseline-adjusted score is:

$$\Delta Y_{m,g,p,o,\ell} = Y_{m,g,p,o,\ell} - \bar{Y}^{\text{base}}_{m,g,o,\ell}, \quad (3)$$

where $\bar{Y}^{\text{base}}_{m,g,o,\ell}$ is the mean bias of the three baseline stories for model m, gender g, occupation o, and language ℓ. Positive ΔY indicates amplification; negative ΔY indicates attenuation.

3.4. Statistical Framework

Inferential analyses use an ordinary least squares (OLS) regression model implemented in `statsmodels` [7] with ΔY as the dependent variable.

$$\begin{aligned}\Delta Y_{m,g,p,o,\ell} = \beta_0 &+ \beta_g \,\text{Gender}_g + \beta_t \,\text{Trait}_p \\ &+ \beta_{\text{lang}} \,\text{Lang}_\ell + \beta_{\text{oc}} \,\text{OccType}_o \\ &+ \beta_{\text{mod}} \,\text{Model}_m + \varepsilon,\end{aligned} \quad (4)$$

with reference categories: neutral gender, baseline personality, GPT-5 nano, and English. Stratified models are additionally fitted separately for male/female personas and English/Hindi subsets to isolate subgroup-specific personality effects.

4. Results

4.1. Baseline Gender Bias

Prior to any personality conditioning, baseline narratives exhibit a consistent positive bias score across all six models and both languages.As illustrated for GPT-5 nano in Table 1, 52% of English baseline generations lean toward the male centroid (median +0.014), and 57.33% of Hindi baseline generations lean male (median +0.020). This asymmetry is statistically significant (Wilcoxon signed-rank test, $p < 0.001$ for both languages) and persists even when persona gender is explicitly specified as female. Female persona prompts reduce but do not eliminate male alignment.

Table 1: Proportion of generations leaning toward male vs. female stereotype centroids, with and without personality conditioning.

Model	Condition	% male	% female	med.	std
GPT-eng	w/o personality	52.00%	48.00%	0.014	0.034
	w/ personality	76.56%	23.44%	0.028	0.029
GPT-hindi	w/o personality	57.33%	42.67%	0.020	0.032
	w/ personality	78.00%	22.00%	0.029	0.029

Occupational context reinforces this tendency: male-stereotyped occupations exhibit stronger baseline male alignment than female-stereotyped occupations independent of persona gender, thus occupational stereotype structure is a strong structural predictor of bias magnitude.

4.2. Personality Effects and Trait Intensity (RQ1)

Introducing personality conditioning produces a substantial aggregate shift. The proportion of male-leaning generations increases English (52% to 76.56%) and in Hindi (57.33% to 78%). The median bias score approximately doubles in both languages while standard deviation decreases, indicating that personality conditioning not only shifts the center of the distribution but concentrates it, reducing the proportion of neutral or counter-stereotypical outputs.
Figures 1 plots per-trait bias score across all six models under high-score for male personas in English.

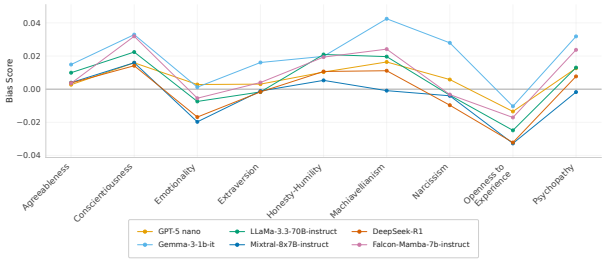

Figure 1: Gender-bias scores for nine personality traits under **high-score** conditioning for male personas in English.

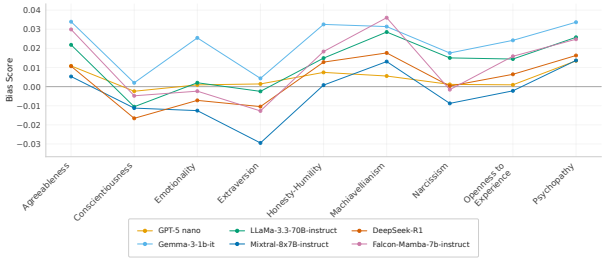

Figure 2: Gender-bias scores for nine personality traits under **low-score** conditioning for male personas in English.

Three structural properties emerge: (i) the *direction* of each trait's effect is preserved across intensity levels; (ii) *magnitude* scales with intensity across all Dark Triad traits, establishing a dose-response relationship rather than a threshold effect; and (iii) *inter-model variance* is larger under high-score conditions.

The trait-level results conform to a clear *antagonism–prosociality gradient*. Machiavellianism and Psychopathy are the strongest amplifiers, with consistently large positive scores under high-score conditioning that are substantially reduced but directionally preserved under low-score conditioning (2). Narcissism shows greater architectural heterogeneity, with most models producing positive values. Conscientiousness unexpectedly produces large positive values under high-score conditioning, likely activating competence vocabulary with male-coded statistical associations, though this effect collapses to near zero at low intensity. Openness to Experience and Emotionality are the primary attenuators, with Openness showing the most consistent attenuation across both intensity levels, while Emotionality deflects negatively only under high-score conditioning.

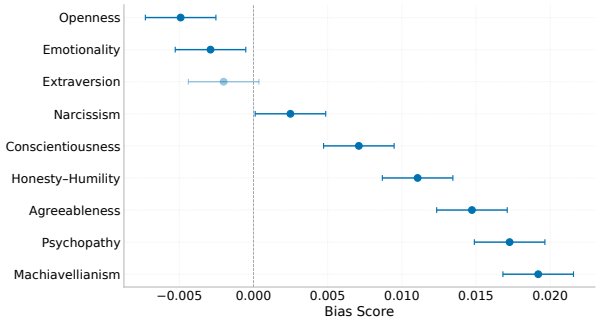

Figure 3: Personality trait regression coefficients (β_t) pooled across all conditions (reference = neutral baseline).

4.3. Personality–Gender Interactions (RQ2)

The interaction between personality conditioning and persona gender is among the most theoretically significant findings of this work. Female personas conditioned on high-score Dark Triad traits produce story embeddings aligned with the male-stereotypical centroid at levels approaching those of male personas under the same conditions, a finding that directly motivates the two-intensity prompt design.

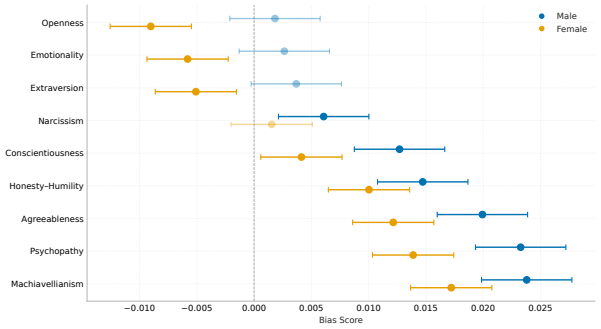

Figure 4: Regression coefficients (β_t) by persona gender (reference = neutral).

Figure 4 confirms that female Dark Triad regression coefficients ($\approx +0.014$ to $+0.018$) are three to four times larger in magnitude than the female gender label coefficient in Figure 5 itself (≈ -0.005), constituting direct empirical evidence that personality conditioning overrides demographic specification under high-intensity antagonistic conditioning.

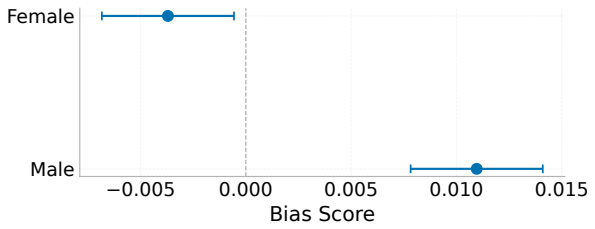

Figure 5: Regression coefficients for gender (reference = neutral).

A female persona described with high-score Machiavellianism does not produce a feminized version of Machiavellian language; instead, it absorbs the male-stereotypical semantic signature of dominance-coded vocabulary, partially overriding the female persona label in embedding space.

4.4. Cross-Linguistic and Occupational Moderation (RQ3)

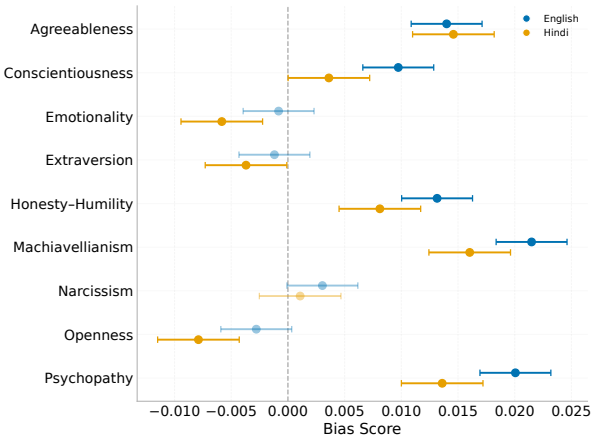

Figure 6: Personality regression coefficients by language.

Language. Figure 6 reveals a consistent compression of personality-induced shifts in Hindi relative to English. Machiavellianism and Psychopathy retain the strongest cross-linguistic consistency ($\approx +0.022$ vs. $+0.018$ and $+0.021$ vs. $+0.020$ respectively), while Conscientiousness shows the largest single compression gap, dropping from $\approx +0.010$ in English to $\approx +0.001$ in Hindi. Hindi's obligatory morphological gender-marking anchors narrative representations more firmly to the occupational baseline [4], imposing a ceiling on personality-induced displacement.

Occupationally, Dark Triad conditioning further amplifies male-aligned baselines in male-stereotyped roles, while amplification is attenuated in female-stereotyped roles. Model-level coefficients are uniformly smaller in magnitude than any Dark Triad trait coefficient, confirming that personality trait and intensity level is more predictive of a story's bias score than the choice of model architecture and that no architecture eliminates the amplification pattern.

5. Discussion

The convergence of distributional, regression, and cross-linguistic evidence establishes personality conditioning as a consequential and underexamined pathway through which gender stereotypes enter LLM outputs. The two-intensity prompt design makes this claim precise: by showing that amplification scales monotonically from low-score to high-score conditions (Figures 1–2), we establish that personality functions as a continuous modulator rather than an incidental categorical cue. The underlying mechanism is corpus-linguistic: dominance-related vocabulary is statistically proximate to male-stereotypical centroid regions in multilingual embedding space, inherited from pretraining corpora in which male-coded professional roles are the unmarked narrative baseline [8]. When a high-score Dark Triad prompt activates this vocabulary region, it simultaneously pulls generated text toward male-stereotypical embedding space, even when the persona is explicitly female, and even across architectures as distinct as a 1B-parameter dense model and a 70B reasoning-focused system.

The practical implications are significant. Personality descriptors are routinely specified in deployed AI systems as stylistic choices unrelated to demographic representation. Our findings demonstrate this assumption is unwarranted: personality conditioning systematically tilts the semantic register of generated narratives toward culturally encoded gender stereotypes, and this effect grows stronger as trait intensity increases. Current fairness benchmarks that probe demographic labels in isolation miss this pathway entirely. Evaluation protocols must treat persona configuration, including psychological trait specification at varying intensity levels, as a first-class variable alongside demographic labels [2].

6. Conclusions

This thesis demonstrates that gender bias in LLM-generated narratives is not a static property of demographic labeling alone, but a context-dependent phenomenon shaped by psychological persona conditioning. Across 23,400 narratives generated by six model architectures in English and Hindi, four principal conclusions emerge:

1. **Personality traits modulate bias systematically.** Dark Triad traits amplify male-stereotypical alignment; prosocial HEXACO traits (Agreeableness, Honesty-Humility) attenuate it; Emotionality and Openness additionally deflect negatively owing to their female-stereotypical cultural associations The aggregate shift in English and Hindi is substantively large and statistically robust across all conditions.
2. **Amplification scales with trait intensity.** High-score descriptions produce consistently larger distributional displacements than low-score descriptions across all Dark Triad traits, both languages, and all six architectures, establishing personality as a continuous dose-response modulator rather than a binary cue.
3. **Personality can dominate gender labeling.** Under high-score Dark Triad conditioning, female personas undergo a mode reversal in embedding space, shifting from female-leaning baselines to male- stereotypical alignment. Female Dark Triad coefficients ($\approx +0.014$ to $+0.018$) exceed the female gender label coefficient (≈ -0.005) by a factor of three to four, constituting direct evidence that psychological conditioning overrides demographic specification.
4. **Effects are cross-architecturally and cross-linguistically robust.** Directionality is identical in English and Hindi at both intensity levels, with a systematic magnitude compression in Hindi attributable to obligatory morphological gender-marking. No architecture eliminates the amplification pattern; model choice modulates bias magnitude marginally but cannot override personality directionality.

These findings call for fairness evaluation frameworks that treat persona configuration—including personality trait intensity—as a first-class variable. As persona-driven AI systems become embedded in education, customer service, and social interaction, understanding how psychological conditioning modulates stereotype expression is essential for designing systems that avoid unintended amplification of harmful social representations.

References

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

[2] Yixin Chen et al. More women, same stereotype: Gender bias in large language models. *arXiv preprint*, 2025.

[3] Tobias Greitemeyer. The dark triad and prosocial behavior revisited. *Current Psychology*, 2023.

[4] Kira Hall. Unnatural gender in Hindi. *Gender Across Languages*, 2002.

[5] Mayur Joshi, Aakanksha Goel, and Raviraj Joshi. L3cube-hindsbert and hindroberta: Hindi sentence embeddings for clustering and classification. *arXiv preprint arXiv:2211.11018*, 2022.

[6] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, 2023.

[7] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010.

[8] Yida Wang et al. Exploring the impact of toxicity on bias in large language models. *arXiv preprint*, 2025.