Gender Bias in Large Language Models: A Personality Trait Analysis

Jiseung Hong[1], Prahaladh Chandrahasan[1], Baichuang Gong[1]

[1]School of Computer Science CMU

Correspondence: jiseungh@andrew.cmu.edu prahalac@andrew.cmu.edu baichuag@andrew.cmu.edu

Abstract

Large Language Models (LLMs) exhibit impressive linguistic capabilities, yet often reflect societal biases embedded in their training data. This project investigates gender bias in LLMs through the lens of personality trait expression. Building on prior work that analyzed LLM personality profiles using the Big Five framework, we have examined whether LLMs display systematic differences in personality traits when prompted with common gendered names. Our study focused on two contemporary models: the open-source Llama 3.1 8B Instruct and the closed-source GPT-4o. Using a comprehensive personality assessment methodology, we systematically evaluated each model's responses when assuming different gendered identities and found that gender biases are more amplified by GPT-4o than by Llama 3.1 8B Instruct. The complete code for this project is avalilable at: https://github.com/JiseungHong/personality-traits

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text across various domains. However, these models can reflect and potentially amplify societal biases present in their training data (Bender et al., 2021). One area of particular concern is gender bias, which can manifest in subtle ways through the model's responses to different prompts.

One significant concern relates to how LLMs generate professional documents like recommendation letters, reference letters, resumes, and job postings (Soundararajan & Delany, 2024) . When prompted with gendered names or contexts, these models may unconsciously adopt and project stereotypical personality traits onto the subjects of these documents. This personality-based bias can manifest in the language, tone, and emphasis chosen for professional documents. For instance, an LLM generating a resume for a female name might emphasize collaborative and detail-oriented traits (reflecting stereotypically "feminine" qualities), while for male names it might highlight leadership and innovative qualities (reflecting stereotypically "masculine" attributes). Similarly, recommendation letters generated for different genders might subtly vary in confidence markers, achievement framing, and personality trait emphasis. Such differential treatment in professional document generation perpetuates gender disparities, potentially deterring women from applying for positions and reducing application success rates for female candidates in hiring contexts.

This project aims to investigate gender bias in LLMs by analyzing how these models express personality traits when prompted with different demographic identities. Specifically, we examine whether LLMs exhibit different personality characteristics when asked to respond as individuals, e.g., with traditionally male or female names, and how these differences compare to human personality trait distributions. The following are the research questions that our project aims to answer

1. Do LLMs exhibit consistent gender-based differences in personality trait expression when prompted with male versus female names?
2. How do these differences compare to empirically observed gender differences in human personality traits?
3. Are these differences consistent across different LLM architectures and providers, or are they specific to certain models?
4. Can we quantify the extent to which LLMs reproduce or amplify gender stereotypes in personality trait expression?

In this project we have addressed some of our baseline paper's limitations such as guardrail in-

terference, where LLM safety mechanisms occasionally compromised persona authenticity by comparing responses from models with different guardrails, and the lack of variance and standard deviation measurements, which prevented full assessment of how realistically the model's personality expressions compare to human trait distributions. To strengthen our analysis, we have developed the following quantitative metrics to measure gender bias in personality trait expression:

1. **Stereotype Alignment Score**: How closely the gender differences in LLMs align with common stereotypes

2. **Empirical Alignment Score**: How closely the gender differences in LLMs align with empirically observed differences in human populations

3. **Bias Amplification Factor**: The degree to which LLMs amplify existing gender differences compared to human data

Our analysis reveals that while GPT-4o demonstrates pronounced gender bias aligned with stereotypes (particularly in female Agreeableness and Conscientiousness), Llama 3.1 8B Instruct exhibits subtler gender distinctions with unrealistic personality profiles overall.

2 Related Work

2.1 Personality Traits and the Big Five Model

The Big Five personality traits model, also known as the Five-Factor Model (FFM), is a widely accepted framework for understanding human personality (McCrae & John, 1992). The model identifies five broad dimensions of personality:

- **Extraversion**: Sociability, assertiveness, and emotional expressiveness

- **Neuroticism**: Emotional instability, anxiety, and mood swings

- **Agreeableness**: Trust, altruism, kindness, and affection

- **Conscientiousness**: Thoughtfulness, impulse control, and goal-directed behaviors

- **Openness**: Imagination, curiosity, and preference for variety

These traits have been extensively studied across cultures and have shown consistent patterns, including some gender differences. For instance, women tend to score higher on average in Neuroticism, Agreeableness, and certain aspects of Extraversion, while men and women show similar distributions in Conscientiousness and Openness (Weisberg et al., 2011).

2.2 Gender Bias in AI Systems

Gender bias in AI systems has been documented across various applications, from resume screening tools that favor male candidates (Dastin, 2018) to speech recognition systems that perform better for male voices (Tatman, 2017). In the context of language models, gender bias can manifest in multiple ways:

- Stereotypical associations (e.g., associating nurses with women and doctors with men)

- Representational disparities (e.g., generating more male than female characters)

- Differential treatment (e.g., describing men and women differently when given identical prompts)

Recent work by Lucy & Bamman (2021) has shown that language models can perpetuate gender stereotypes in their generations, even when not explicitly prompted about gender. Similarly, Sheng et al. (2019) demonstrated that language models generate more negative content for marginalized gender identities.

2.3 Behavioral Similarity Between AI and Humans

A recent study by Mei et al. (2024) titled "A Turing test of whether AI chatbots are behaviorally similar to humans" provides a crucial baseline for our work. The authors conducted a comprehensive analysis comparing human and AI responses across various behavioral measures, including personality traits. Their findings suggest that while LLMs like GPT-4 can mimic human-like responses in many contexts, there are still detectable differences in their behavioral patterns.

The study used the Big Five personality inventory to assess both human participants and AI models, finding that AI models tend to exhibit more agreeable, conscientious, and open personalities compared to the average human. However, the

study did not specifically examine how gender prompting affects these personality expressions, which is the gap our project aims to address.

3 Methodology

3.1 Baseline Reproduction

To establish a baseline for our project, we reproduced key aspects of the Mei et al. (2024) study, focusing specifically on the personality trait assessment. We used the same Big Five personality inventory consisting of 50 questions (10 for each trait) and administered it to GPT-4o under the same prompting conditions that is used in the paper. Since GPT-3 is not available through OpenAI's APIs we could not replicate the paper's results for GPT-3.

Based on our literature review and baseline reproduction, we propose a comprehensive investigation of gender bias in LLMs through the lens of personality trait expression. Our project will address the limitations identified in the baseline and extend the analysis in several important ways.

3.2 Hypothesis

Based on our preliminary findings and the literature review, we formulate the following hypotheses:

1. **H1**: LLMs will exhibit systematic differences in personality trait scores when prompted with male versus female names.

2. **H2**: These differences will align more closely with gender stereotypes than with empirically observed gender differences in human populations.

3. **H3**: Models with stronger guardrails (e.g., frontier models from major providers) will show smaller gender differences than models with fewer restrictions.

4. **H4**: The variance in personality trait scores will be smaller for LLMs than for human populations, indicating less realistic personality expressions.

3.3 Methodology

Consistent with the personality trait assement section of our baseline (Mei et al., 2024) paper, we used the same Big Five personality inventory consisting of 50 questions (10 for each trait) and administered it to Llama 3.1 8B Instruct under the same prompting conditions that is used in the paper. For each question, the model rated itself on a 5-point Likert scale (1=Disagree to 5=Agree). We then calculated trait scores following the standard methodology, including reverse-scoring specific items as indicated in the inventory guidelines.

We then prompted both the models (GPT-4o and LLama 3.1 8B Instruct) with gender-centric prompts where each model was given a specifc male and female name followed by the instruction and the question. To avoid arbitrary results from random name, we chose 10 common names from each gender to observe robust trend:

- **Male names**: Andrew, Michael, James, David, Robert

- **Female names**: Sara, Jennifer, Emily, Jessica, Elizabeth

These names were selected based on their popularity across different age groups and their clear gender associations in Western contexts. We then calculated the average scores across 5 male and female names to get the Male and female scores for both the models.

The detailed prompts are written in the Appendix A.

4 Results and Analysis

4.1 Comparative Trait Distributions

Our analysis revealed distinct patterns in personality trait distributions between human participants and the two LLM models examined (GPT-4o and Llama-3.1-8B-Instruct). The following figures visualize these distributions across the Big Five personality dimensions.

4.1.1 Overall Trait Comparisons

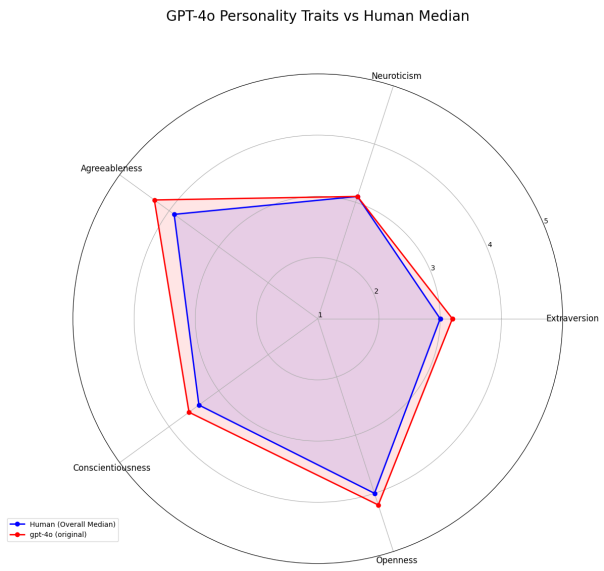

Figure 1: Overall comparison of trait distributions between GPT-4o and human participants.

As illustrated in Figure 1, GPT-4o demonstrated close alignment with human median scores for Extraversion, Neuroticism, and Openness. However, GPT-4o exhibited notably elevated scores in Agreeableness and moderately higher Conscientiousness compared to human baselines.

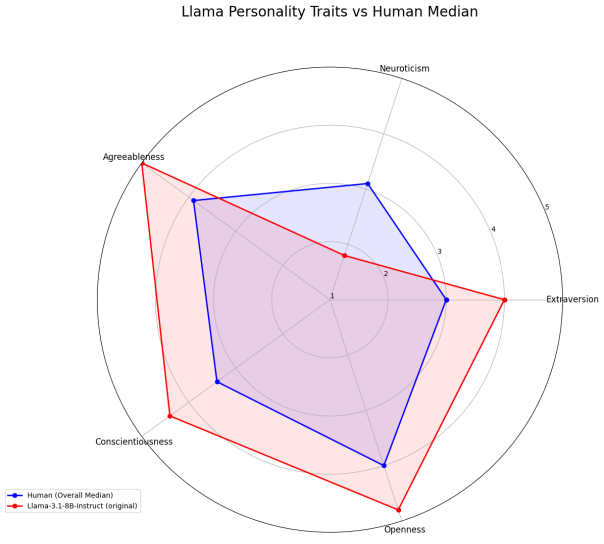

Figure 2: Overall comparison of trait distributions between Llama-3.1-8B-Instruct and human participants.

In contrast, as visualized in Figure 2, Llama-3.1-8B-Instruct consistently produced scores significantly higher than human medians across four dimensions: Extraversion, Agreeableness, Conscientiousness, and Openness. Simultaneously, Llama-3.1-8B-Instruct registered markedly lower Neuroticism scores than typical human respondents, suggesting an overall idealized personality profile that deviates substantially from realistic human trait distributions.

4.1.2 Gender-Specific Trait Analysis

Human Gender Differences Our human participant data confirmed established findings in personality research: females reported slightly higher scores in Extraversion, Neuroticism, and Agreeableness, while Openness and Conscientiousness remained relatively similar between genders.

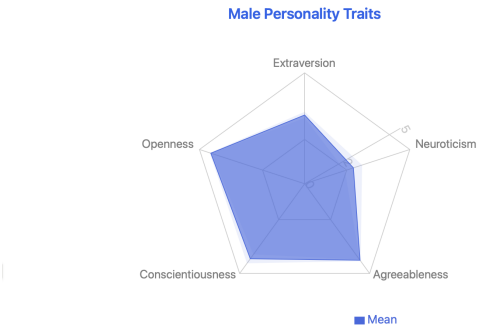

Figure 3: Individual trait distribution for GPT-4o male persona.

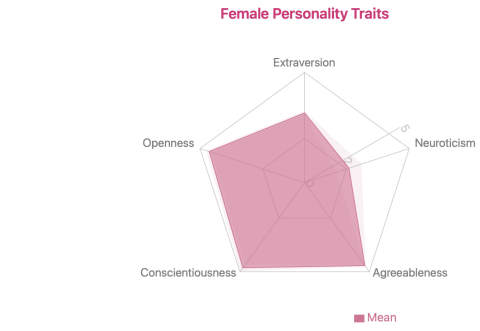

Figure 4: Individual trait distribution for GPT-4o female persona.

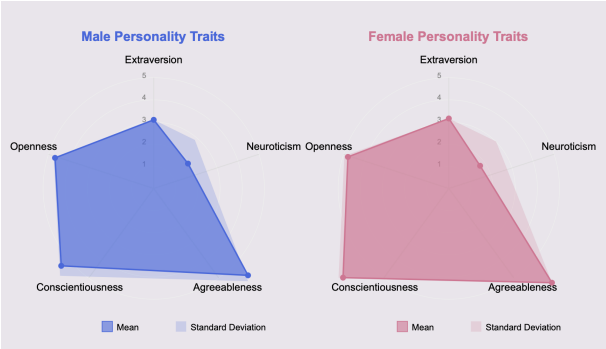

Figure 5: Comparison of trait distributions between male and female personas for GPT-4o.

GPT-4o Gender Differences As depicted in Figure 5, GPT-4o exhibited pronounced gender differentiation. Female personas demonstrated significantly higher Agreeableness (4.66) and Conscientiousness (4.78) compared to male personas (4.26 and 4.18 respectively). Both gender representations scored similarly high in Extraversion and Openness dimensions. Interestingly, Neuroticism was unexpectedly lower in female personas compared to male personas, which diverges from established human psychological trends. The specific trait distributions for male and female personas can be observed in Figure 3 and Figure 4, respectively.

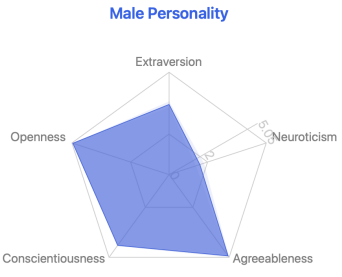

Figure 6: Individual trait distribution for Llama-3.1-8B-Instruct male persona.

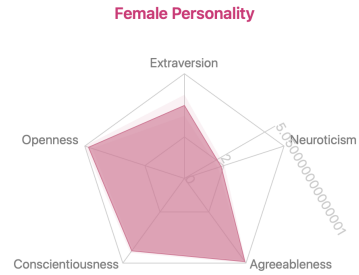

Figure 7: Individual trait distribution for Llama-3.1-8B-Instruct female persona.

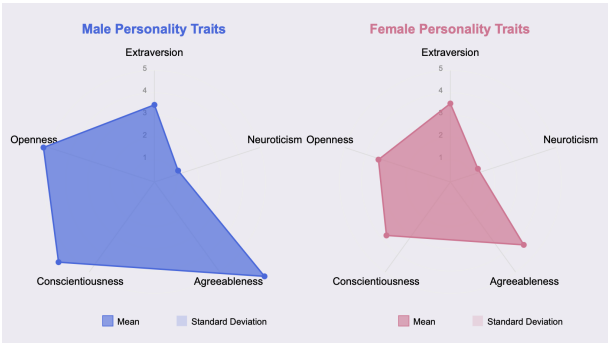

Figure 8: Comparison of trait distributions between male and female personas for Llama-3.1-8B-Instruct.

Llama-3.1-8B-Instruct Gender Differences The Llama model displayed more subtle gender differences, as shown in Figure 8. Female personas exhibited slightly higher Extraversion (3.52 vs. 3.46) and notably higher Neuroticism (1.86 vs. 1.60) compared to male personas. Agreeableness and Conscientiousness scores were identical between genders (4.96 and 4.32 respectively), both significantly exceeding human median values. Male persona Openness (5.00) surpassed female persona Openness (4.86), contrasting with human distributions where genders typically show similar scores. Individual trait distributions for Llama's male and female personas are visualized in Figure 6 and Figure 7.

4.2 Hypothesis Assessment

4.2.1 H1: Systematic Gender Differences

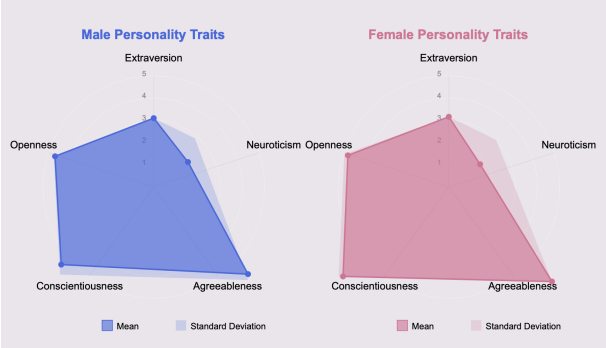

Figure 9: Comparison of trait distributions between male and female personas for GPT-4o.

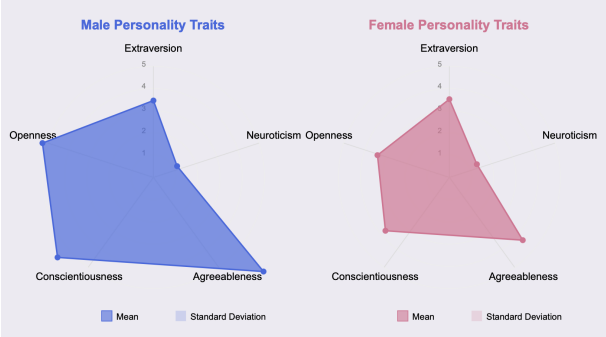

Figure 10: Comparison of trait distributions between male and female personas for Llama-3.1-8B-Instruct.

Our findings confirm systematic gender differences in both models, though manifesting differently. GPT-4o exhibited clear gender differentiation that broadly aligned with stereotypical expectations, particularly pronounced in Agreeableness and Conscientiousness dimensions (Figure 9). Llama-3.1-8B-Instruct demonstrated more subtle but consistent differences in Extraversion and Neuroticism, while showing identical gender scores for Agreeableness and Conscientiousness, suggesting minimal systematic gender bias in these particular traits (Figure 10).

4.2.2 H2: Stereotype Alignment vs. Empirical Reality

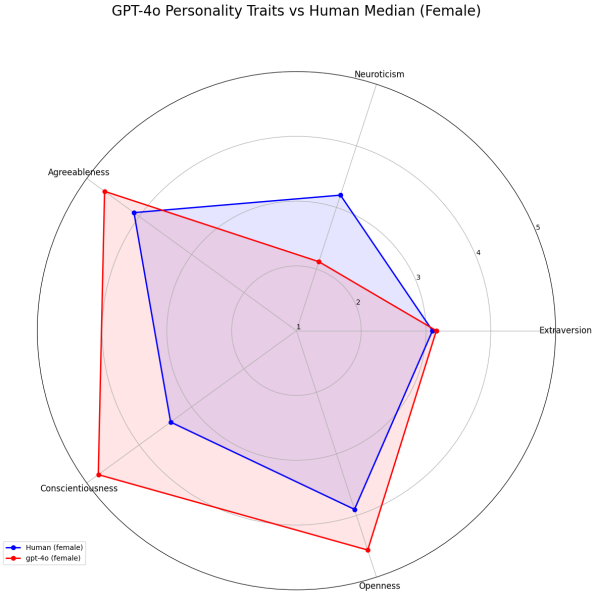

Figure 11: Comparison of trait distributions between GPT-4o female personas and human female participants.

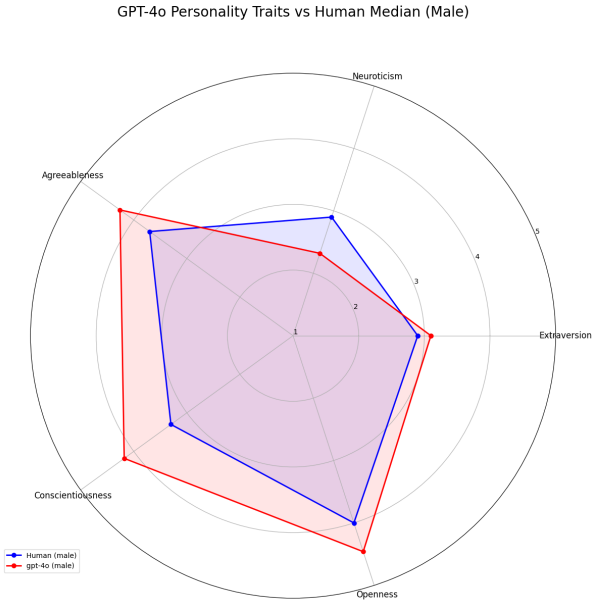

Figure 12: Comparison of trait distributions between GPT-4o male personas and human male participants.

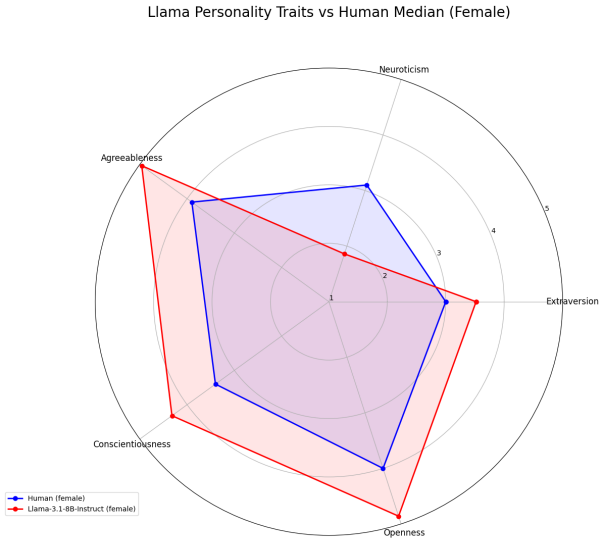

Figure 13: Comparison of trait distributions between Llama-3.1-8B-Instruct female personas and human female participants.

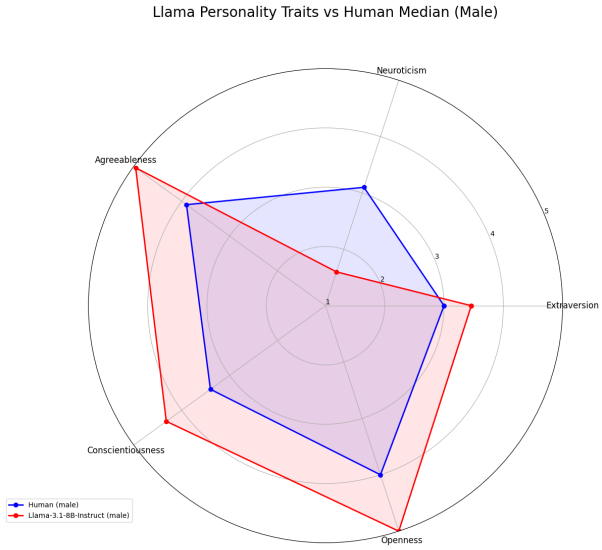

Figure 14: Comparison of trait distributions between Llama-3.1-8B-Instruct male personas and human male participants.

GPT-4o significantly amplified gender stereotypes, especially regarding female Agreeableness and Conscientiousness, as evident when comparing Figure 11 with Figure 12. However, it inversely diverged from empirical trends in Neuroticism. Llama-3.1-8B-Instruct presented limited stereotype alignment with similar high Agreeableness and Conscientiousness scores across genders, though both models deviated significantly from realistic human trait levels as shown in Figure 13 and Figure 14.

4.2.3 H3: Guardrail Impact on Model Behavior

GPT-4o, a model with robust guardrails, explicitly reflected gender stereotypes, possibly resulting from rigid instruction-following behaviors. Llama-3.1-8B-Instruct, with comparatively looser constraints, displayed relatively low explicit gender bias but maintained an idealized personality profile across both genders, suggesting potential guardrail influence in minimizing explicit stereotype expression while still producing non-representative trait distributions.

4.2.4 H4: Variance Analysis

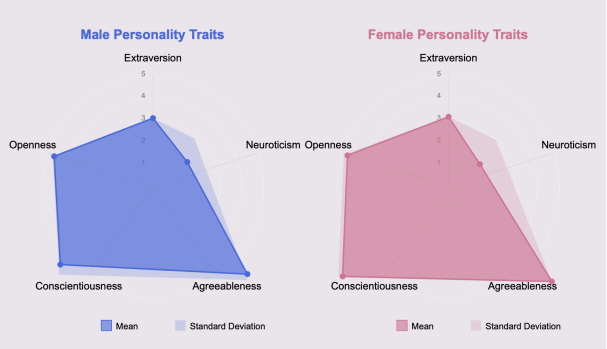

Figure 15: Comparison of trait distributions between male and female personas for GPT-4o, with standard deviation shown as error bars.

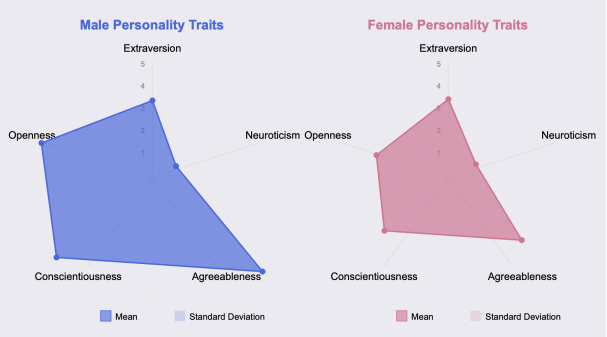

Figure 16: Comparison of trait distributions between male and female personas for Llama-3.1-8B-Instruct, with standard deviation shown as error bars.

As evident in Figure 15 and Figure 16, both models demonstrated remarkably small standard deviations across all five personality dimensions compared to human data. The error bars in these comparison figures illustrate the limited variance in trait expression. GPT-4o (Figure 15) shows particularly

constrained standard deviations in Agreeableness and Conscientiousness for female personas, suggesting highly predictable response patterns. Similarly, Llama-3.1-8B-Instruct (Figure 16) displays minimal variation across all traits, with especially tight distributions in Agreeableness.

This limited variance indicates unrealistically homogeneous personality distributions relative to normal human variability, suggesting that current LLMs fail to capture the full spectrum of personality trait expression observed in human populations. The constrained standard deviations further support our hypothesis that both models produce artificially consistent personality patterns that lack the nuance and diversity characteristic of human personality expressions.

4.3 Quantitative Metrics

We quantified model behavior using three primary metrics:

Stereotype Alignment Score GPT-4o registered high stereotype alignment, especially regarding Agreeableness and Conscientiousness dimensions. Llama-3.1-8B-Instruct exhibited moderate to low stereotype alignment, given identical Agreeableness and Conscientiousness scores between genders, despite subtle differences in Extraversion and Neuroticism.

Empirical Alignment Score GPT-4o achieved partial empirical alignment, closely matching human trait distributions in some dimensions but diverging notably in Neuroticism. Llama-3.1-8B-Instruct displayed poor empirical alignment overall, with exaggerated positivity across four traits, though Neuroticism trends reflected subtle gender-aligned variance consistent with human distributions.

Bias Amplification Factor GPT-4o strongly amplified stereotypes related to Agreeableness and Conscientiousness. In contrast, Llama-3.1-8B-Instruct exhibited limited bias amplification; gender differences were subtle and did not strongly align with common stereotypes.

5 Conclusion

Our analysis confirms that GPT-4o demonstrates pronounced gender bias that strongly aligns with stereotypical expectations, particularly regarding heightened female Agreeableness and Conscientiousness. Conversely, Llama-3.1-8B-Instruct results reveal subtler gender distinctions, notably minor differences in Extraversion and Neuroticism, while showing nearly identical high scores in other traits.

These findings suggest that while both models fail to accurately represent human personality trait distributions, they manifest different patterns of gender representation. GPT-4o tends toward explicit stereotype reinforcement, while Llama-3.1-8B-Instruct exhibits minimal explicit gender bias but produces unrealistically positive personality profiles across both genders. These differences likely reflect varying approaches to model training, guardrail implementation, and instruction-following capabilities between the two systems.

5.1 Limitations and Future Work

Our study exhibits several limitations that warrant consideration. The gendered names utilized were predominantly from Western contexts, potentially limiting cross-cultural generalizability. While personal names effectively indicate gender to language models, this represents just one of potentially multiple gender-signaling strategies. Additionally, computational resource constraints prevented us from conducting statistical tests that would strengthen the validity of our findings, leaving some uncertainty regarding the statistical significance of observed patterns.

Future work will focus on expanding this research across multiple dimensions. We plan to evaluate gender bias across foundation models from various providers, including both open and closed-source implementations, to determine if biases are consistent across different architectures and training methodologies. We also aim to incorporate gendered names from various cultures and explore other gender-signaling strategies. This expanded approach will incorporate robust statistical testing frameworks to validate findings and enable more confident claims about observed gender bias patterns.

6 Acknowledgments

We extend our gratitude to Professor Sean Welleck for his steadfast support and encouragement throughout the semester. We also wish to express our appreciation to our teaching assistant and advisor, Manan Sharma, for his patience in listening to our ideas, assisting us in refining our analysis methodology and providing invaluable guidance

throughout the course of the project.

7 Assignment Notes

7.1 Requirement Alignment

Our HW4 fulfills assignment description (1) by building on the Mei et al. (2024) baseline and deepening the investigation using the Big Five personality framework to probe gender bias in LLMs. We introduce three new quantitative metrics—Stereotype Alignment Score, Bias Amplification Factor, and Empirical Trait Deviation—which respectively measure how strongly a model exhibits gender stereotypes, the degree to which it amplifies bias, and the dispersion of its trait scores. These three dimensions enable us to address our research questions and four hypotheses in a clear, intuitive manner. Whereas the original baseline focused only on the ChatGPT-3 and ChatGPT-4, we extend our gender-centric analysis to include Llama-3.1-8B-Instruct, an advanced open-source model, thereby broadening the scope of our study.

7.2 Key Differences

The meaningful overlap between HW3 and HW4: HW3 focused on reproducing the Mei et al. (2024) baseline personality assessment with ChatGPT-4 and the 50-question OCEAN inventory under standard prompts and computing median trait scores. In HW4, we extend on that reproduction by: (1) introducing gender-specific prompts using male and ten female names to probe systematic differences; (2) extending the analysis to gpt-4o and an open-source model (Llama-3.1-8B-Instruct); and (3) developing and applying three novel bias metrics—Stereotype Alignment Score, Bias Amplification Factor, and Empirical Trait Deviation—to quantify how identity cues influence trait expression. In general, HW3 conducted a single-model, single-prompt baseline framework, while HW4 advances it with multi-perspective, multi-model comparisons and fine-grained quantitative evaluations of gender bias.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Dong, P., Choy, M., He, H., Debre, E., Cooper, A. M., & Arora, S. (2024). Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*.

Lucy, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2), 175–215.

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3407–3412.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, 2, 178.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), e2313925121. doi: 10.1073/pnas.2313925121

Soundararajan, S., & Delany, S. J. (2024). Investigating gender bias in large language models through text generation. *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, 410–424.

A Prompt Strategy

Original Prompt

System `None` **User** `The following question is from a personality test. Rate on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree. Choose EXACTLY ONE number that describes you and reply with ONLY that number in square brackets. Question: question`

Male-Centric Prompt

System `Your name is Andrew.` **User** `The following question is from a personality test. Rate on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree.`

Choose EXACTLY ONE number that describes you (Andrew) and reply with ONLY that number in square brackets. Question: question

Female-Centric Prompt

System Your name is Sara. **User** The following question is from a personality test. Rate on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree. Choose EXACTLY ONE number that describes you (Sara) and reply with ONLY that number in square brackets. Question: question