

Gender Representation in AI-Generated Language: A Conceptual Linguistic Analysis of Bias, Stereotypes, and Representation in Large Language Models

Sumaira Latif

Principal Lecturer, Faculty of Languages and Literature, University of Central Punjab, Lahore, Pakistan,
sumaira.latif@ucp.edu.pk

Muhammad Adeeb

Master of Arts in Digital English Studies, Lingnan University, Hong Kong
adeeb3389@gmail.com

Afaq Ahmad Khan

English Lecturer, University of South Asia, Lahore, Pakistan
afaq.ahmed@usa.edu.pk

Abstract

Large language models (LLMs) increasingly mediate writing, information retrieval, education, recruitment, and everyday communication. Their linguistic outputs therefore matter not only as technical products but also as forms of social representation. This conceptual paper examines gender representation in AI-generated language by synthesizing evidence from computational studies of bias with insights from critical discourse analysis and systemic functional linguistics. Rather than collecting new model outputs, the paper draws on peer-reviewed research and authoritative reports to identify recurrent patterns in how gender is represented, stereotyped, and evaluated in language-model behavior. The synthesis indicates that gender bias is often realized through occupational association, semantic prosody, emotional attribution, agency, evaluative wording, and recurring assumptions about social roles. The paper proposes a linguistic framework in which representational bias is analyzed across lexical-semantic, grammatical, interpersonal, and discursive levels. It argues that bias evaluation should move beyond counting explicitly stereotypical terms and examine how language choices distribute visibility, agency, authority, affect, and normality across genders. The framework offers a basis for future empirical studies of gendered AI discourse while acknowledging that model behavior remains prompt-, language-, task-, and version-sensitive.

Keywords: Large language models, gender representation, gender bias, stereotypes, AI-generated language, critical discourse analysis, systemic functional linguistics

1. Introduction

The advent of generative AI has altered the magnitude of machine-generated language dissemination within social and organizational contexts. Large language models have become popular for generating text, providing answers to queries, helping learning, recruitment, and offering guidance. The use of large amounts of human language in training these models has led to worries about the reproduction of stereotypes in society. Previous studies on word embeddings revealed that statistical language modeling can represent associations between women and family and men and career, which are evident from human language data (Bolukbasi et al., 2016; Caliskan et al., 2017). In subsequent research on language generation, it has been shown that output generated by models can have a different level of respect towards demographic groups and can reproduce stereotypes regarding occupations and society (Sheng et al., 2019).

The problem of gender bias in large language models (LLMs), nevertheless, cannot be explained simply as a list of offensive words and differences in frequency distributions. Language encodes social actors via choices of vocabulary,

transitivity, modality, evaluation, presuppositions, and associative relations. In that regard, an LLM may encode gender stereotypes even without using explicitly discriminatory language. AI-generated language may reproduce uniform and formulaic discourse patterns that can shape the representation of gender and other social identities (Rehman et al., 2026). Gender bias has been recently documented in emotion assignment, job decision making, multilingual stereotype evaluation, and general text generation (Plaza-del-Arco et al., 2024; Rowe et al., 2025; Soundararajan & Delany, 2024). UNESCO and the International Research Centre on Artificial Intelligence (IRCAI) have also reported consistent stereotypical associations of women with home and family spheres and of men with career and business spheres in chosen LLM texts (UNESCO & IRCAI, 2024).

This paper treats gender bias as a problem of linguistic representation. It does not generate a new corpus or claim new measurements. Instead, it offers a conceptual synthesis of published evidence and develops a linguistic framework for interpreting how gender stereotypes are realized in AI-generated discourse. The paper addresses three questions:

1. What patterns of gender representation and stereotyping have been documented in existing research on large language models?
2. How are gender biases linguistically realized in AI-generated language at lexical-semantic, grammatical, interpersonal, and discursive levels?
3. How can critical discourse analysis and systemic functional linguistics contribute to the conceptual analysis of gender representation in large language models?

2. Conceptualizing Gender Bias as Representation

The concept of bias is a contentious topic within natural language processing. According to Blodgett et al. (2020), the literature uses this term without clarifying what kind of harm, who was harmed, and what kind of norm is being considered in the context of research. In particular, this issue arises in connection with the study of gender because there are three related terms which researchers tend to use interchangeably: representation, stereotyping, and discrimination. Representation bias focuses on patterns of difference in terms of how genders are represented, described, connected, or framed. Stereotyping deals with associations between genders and traits, jobs, emotions, and roles. Discrimination or allocational harm have to do with consequences that affect decisions, resources, or opportunities.

This differentiation also avoids an excessive deterministic position of saying that models “hold” a certain quantity of sexism in themselves. The behavior of models is conditional and may change depending on prompt formulations, model structure, training instructions, languages used, safety measures, and tasks. The research of benchmark inconsistency provides evidence that various test datasets can detect various aspects of gender stereotypes and thus can provide inconsistent results (Zakizadeh & Pilehvar, 2025). Linguistic analysis, therefore, should look at how bias is embodied under particular discourse conditions.

This issue is supported by prior technical literature. Bender et al. (2021) stated that large language models face dangers that stem from the large-scale training data used in the development process, where these kinds of data become hard to analyze due to certain design choices in the process of creating them. What should be understood here is that the output of large language models does not necessarily reflect the social regularities in training data sets, unless they are changed somehow.

3. Theoretical Lens: Language as Social Representation

Two approaches from linguistics are especially effective in analyzing gendered language use in AI. First, Fairclough's (2015) critical discourse theory posits that language is socially powerful representation. Lexical selections, presuppositions, categories and agency can be used to reproduce some social relationships. This approach, when applied to LLMs, means that it is not important whether the output includes a forbidden word but whether the generated text positions men, women and gender roles in society.

Secondly, systemic functional linguistics considers grammar as a tool for constructing meanings. According to Halliday and Matthiessen (2014), there are ideational, interpersonal and textual functions of grammar. Ideational function is especially useful for analyzing transitivity, which determines who acts, who feels something, who receives something and who is positioned as an agent of action in case of gender representation. Interpersonal function can help to study modality and evaluation: how certain, authoritative, desirable, emotional, competent or socially valuable the gendered actor is constructed.

Together, critical discourse analysis and systemic functional linguistics support a multi-level model of representational bias. Gender can be constructed lexically through labels and descriptors, semantically through repeated associations,

grammatically through agency and process roles, interpersonally through evaluation and modality, and discursively through broader frames of normality, competence, care, leadership, or dependence.

4. Method

The current study adopts a conceptual review and secondary-evidence synthesis. No new prompts were submitted to LLMs, and no primary corpus was collected. The discussion section draws on peer-reviewed research in natural language processing, machine learning, and linguistics, together with one authoritative international report. Sources for secondary data collection were selected purposively for direct relevance to gender bias, representational stereotypes, language generation, or linguistic interpretation of representation.

The synthesis gives priority to studies that analyze generated language or gendered semantic associations and to recent work on LLMs, while retaining earlier foundational studies on computational gender bias. The reviewed evidence is interpreted through four linguistic levels: (1) lexical-semantic representation, (2) grammatical agency and transitivity, (3) interpersonal evaluation and stance, and (4) discourse-level framing. This procedure is conceptual rather than meta-analytic: reported findings are not pooled statistically, and differences in models, prompts, datasets, languages, and evaluation instruments are treated as limits on direct comparison.

5. Gender Representation in AI-Generated Language

5.1 Lexical and Semantic Association

Many computational studies have established that gender can be associated with occupations, societal fields, and personal traits. Bolukbasi et al. (2016) established gender stereotypes associated with professions in word embeddings. In study, Caliskan et al. (2017) revealed a relationship between female names and family and that between male names and careers. Such studies preceded the development of the current instruction fine-tuned LLMs; however, they show that distributional learning enables the embedding of semantic associations within socially structured data.

The existence of such associations in the language generated by generative systems can be evidenced from the research of Sheng et al. (2019). They noted demographic differences in the way that language models view people based on their attributes. Generative bias, in such a case, can be exhibited not by the use of slurs, but evaluative language, which means that there is a need to look at LLMs through descriptors, predicates, occupations, and roles.

From a linguistic perspective, lexical bias includes more than the presence of gendered nouns. It includes semantic prosody: the tendency for words referring to a gender to co-occur with positive, negative, domestic, professional, emotional, or authority-related meanings. A conceptual audit should therefore ask what kinds of predicates and attributes cluster around gendered actors and whether equivalent roles receive comparable lexical treatment.

5.2 Agency, Roles, and Transitivity

Gender stereotypes also come out in the ways in which participants in language are constructed using grammar. Output may present men as agents who decide, lead, build, manage, and solve, while women are constructed as receivers, supporters, caretakers, or experiencers. Even if each single sentence is grammatically neutral, it is the repeated distribution of roles that can cause representational imbalance.

The systemically functional transitivity theory provides a good terminology for this approach. Material processes emphasize the doing aspect, mental processes emphasize sensing and feeling, and relational processes attribute some attributes to the participants (Halliday & Matthiessen, 2014). Therefore, gendered imbalance can be analyzed from the point of view of which genders are presented as Actors, Goals, Sensors, Carriers, or beneficiaries. Theoretically, this analysis is in agreement with the research on occupational and role stereotypes of model behaviors that include such correlations as professional masculinity vs. femininity of caring and looks (Rowe et al., 2025; UNESCO & IRCAL, 2024).

5.3 Evaluation, Emotion, and Interpersonal Meaning

Representation of gender is also interpersonal. According to Plaza-del-Arco et al. (2024), there are gender-specific emotional attributions in LLMs that correlate with stereotypes associated with gender. This is important from a linguistic perspective because emotions are evaluative categories. Emotions are not neutral descriptions of one's feelings but representations of social actors according to socially accepted ideas concerning such things as anger, sadness, empathy, vulnerability, and control.

This is true for adjectives and modal verbs. An artificial intelligence system can portray a male manager as 'decisive' and a female manager as 'caring', or add different nuances of certainty, admiration, criticism, and obligation to the same social actor. Even if the difference is small, it will define authority and social value of those actors. A linguistic study of AI bias must, therefore, take into account evaluative adjectives; affect categories, judgment, appreciation, modality, and intensification.

5.4 Discursive Framing and Normalization

Discourse-level bias relates to what is considered normal, unusual, natural, or worthy of explanation. The theory of Fairclough's (2015) can be applied in this situation since representation is linked to wider social assumptions. An LLM can perpetuate stereotypes due to presupposition, narrative role allocation, back grounding, and framing. For instance, the recurrent creation of women in family contexts and men in leader roles can create naturalized separation of social space regardless of any explicit statements made by sentences.

This fact is emphasized by UNESCO and IRCAI (2024) that discovered gender stereotypes in tested LLMs, and by Rowe et al. (2025) who found the presence of stereotyped reasoning in multilingual testing of models. Multilingual results are particularly crucial since gender representation is dependent on grammatical features, cultural traditions, and translating practices. Thus, generalizations made in one language cannot be applied to all languages.

6. A Conceptual Linguistic Framework for Gender Representation

The synthesis suggests that there is a need to use a four-layer approach to analyzing representational bias in AI-generated language.

The first layer includes such issues as naming, descriptors, occupations, collocations, semantic prosody, and the existence of gender-trait combinations. The second layer involves the analysis of transitivity, agency, passivization, participants' roles, and the action-experience ratio. The third layer deals with the issues of modality, affect, judgment, appreciation, certainty, politeness, and intensification. Finally, the fourth layer includes the analysis of framing, presupposition, foregrounding, omission, narration roles, and social expectation normalization.

These factors cannot be considered separately. A stereotype is linguistically potent once many of these resources are combined in the text. A female character might get an occupation which is related to care, adjectives referring to her looks, position as an Experienced and not as an Actor, and she is involved in a family-based story. On the other hand, a male character will be connected with vocabulary about leadership, physical action, high certainty modality, and institutional power.

This approach also makes it possible to see the connection between the computational and linguistic methods. Scores from benchmarks can show whether stereotypes are present in the text. Linguistic approach is able to prove how these preferences are linguistically expressed.

7. Implications, Limitations, and Conclusion

In evaluating gender bias in LLMs, there is no point in asking whether the models are 'biased.' Instead, one should ask what representations of genders occur, under what circumstances, using what language, and with what possible social consequences. This statement aligns well with the argument of Blodgett et al. (2020), who state that the research on bias should clearly define behavior, harm, and context of the evaluation.

There are several limitations to the current analysis because of the conceptual nature of the paper. First, there are no experiments testing specific models, nor any estimations of prevalence of gender bias. Moreover, the studies cited above differ in terms of prompts, languages used; model families, benchmarks, definitions and their results should not be generalized into some universal measure. Finally, most of the gender bias research still focuses on the binary notion of gender, thus providing no data on non-binary and intersectional representation.

However, the literature provides clear grounds for viewing gender bias in LLMs not only as a technical challenge to fairness but also as a linguistic issue of representation. Existing findings indicate the presence of gender bias in terms of semantic associations, occupations, emotions, role assignment, and evaluations. The combination of the critical discourse and systemic functional approach will make these patterns visible in terms of representation of certain actors, their attributes, the actions they perform, evaluations and normalization of their assumptions. Future empirical research can apply the four-level framework to the corpus of outputs of LLMs and compare different models, languages, prompts, and even different versions of the same model without reducing the gender representation to a single bias score.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of 'bias' in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454-5476). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (Vol. 29, pp. 4349-4357).
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. <https://doi.org/10.1126/science.aal4230>
- Fairclough, N. (2015). *Language and power* (3rd ed.). Routledge.
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.
- Plaza-del-Arco, F. M., Cercas Curry, A., Curry, A., Abercrombie, G., & Hovy, D. (2024). Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume I: Long Papers) (pp. 7682-7696). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.415>
- Rowe, J., Klimaszewski, M., Guillou, L., Vallor, S., & Birch, A. (2025). EuroGEST: Investigating gender stereotypes in multilingual language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 32074-32096). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1632>.
- Rehman, Z., Shamim, M., Sajid, K., & Mahmud, A. (2026). Stance and authorial voice in human and AI-generated academic texts: A corpus-based discourse analysis. *The Critical Review of Social Sciences Studies*, 4(1), 3241–3249. <https://doi.org/10.59075/s292j004>.
- Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3407-3412). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1339>
- Soundararajan, S., & Delany, S. J. (2024). Investigating gender bias in large language models through text generation. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024) (pp. 410-424). Association for Computational Linguistics. <https://aclanthology.org/2024.icnlsp-1.42/>.
- UNESCO, & International Research Centre on Artificial Intelligence. (2024). Challenging systematic prejudices: An investigation into bias against women and girls in large language models. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000388971>.
- Zakizadeh, M., & Pilehvar, M. T. (2025). Blind men and the elephant: Diverse perspectives on gender stereotypes in benchmark datasets. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 22827-22840). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1162>,