GBEval: A SHAP-based Interpretable Gender Bias Assessment Framework for LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) are increasingly being used in fairness-critical tasks, making it essential to evaluate gender bias. We propose GBEval, a systematic approach to the identification and explanation of gender stereotypes through probabilistic assessment in six domains of behavior. Our corpus contains 17 subcategories of domestic work, professional work, technical ability, emotional work, physical work, and cognitive work. We tested six leading LLMs with 20 runs per question type, revealing long-standing domain-specific biases: female associations prefer domestic and emotional work, and males prefer technical skills and physical labor. Bias scores ranging from 0.664 (Gemma2-9B) to 0.767 (GPT-3.5-turbo) are reported. SHAP analysis identifies bias-causing tokens like "cooking," "cleaning," and "coding" as primary gender markers, offering interpretable reasons behind systematic patterns of stereotyping.

1 Introduction

The extensive utilization of large language models (LLMs) in various applications from screening job candidates and helping students to supporting doctors with medical treatment suggestions has spurred algorithmic fairness as the primary research problem in AI [1, 2]. These models can have a drastic influence on people's opportunities and well-living, and ensuring their lack of prejudice is therefore highly important [3]. One of the major concerns is gender bias, where an LLM may repeatedly show preference for one gender over another in its responses [4, 5, 6]. This bias can show up subtly, such as linking technical jobs more with men and caregiving roles more with women. [7] Studies on AI educational writing assistance reveal how bias transfers through AI writing support pipelines and impacts human writing [8]. Recent research, such as [9], confirms the persistence of gender stereotypes in open-source LLMs also. If not addressed, these patterns can strengthen outdated social norms and worsen existing inequalities in areas like hiring, education, and healthcare [10]. Establishing trustworthy AI requires researchers to quantify the degree of gender bias present in LLM outputs, and understand which words or phrases contribute to this bias, so that effective mitigation strategies can be developed.

Recent studies have demonstrated that LLMs, despite impressive in capability, are likely to reflect systematic forms of bias that reflect historical patterns of discrimination present in their training data [11, 12, 13, 14]. Even highly advanced LLMs reflect persistent stereotyping across domains and languages [15]. The forms of bias are varied and span from occupational stereotyping and role assignment to differential performance on gender-related tasks [16]. Though current studies have gone far in identifying the presence of gender bias in language models, there are some important gaps in understanding how these biases operate across different behavioral domains and how they can be systematically detected and explained [17, 18].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Current bias assessment methods exhibit significant limitations: most rely on single-instance evaluations or limited domain coverage, failing to capture the probabilistic nature of model responses and complex stereotypical associations. Existing methodologies lack interpretable mechanisms for identifying specific linguistic features responsible for biased outputs. This work addresses these gaps by introducing GBEval, a systematic probabilistic framework combining comprehensive domain coverage with interpretability analysis. Our contributions are: (1) quantifying gender bias using probabilistic sampling across model response variability, (2) identifying domain-specific bias patterns revealing stereotype manifestations, and (3) providing token-level interpretable explanations of bias mechanisms through SHAP analysis.

2 Methodology

2.1 GBEval Framework Design

We used a quantitative experimental approach with controlled prompts to measure probabilistic bias across multiple state-of-the-art LLMs as shown in Figure 1.

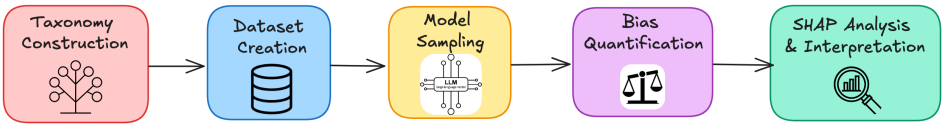

Figure 1: **The five-phase methodology of the GBEval framework.** *Starts with taxonomy construction, followed by dataset creation across 6 domains and 17 subdomains, model selection, bias quantification through repeated iterations, and concludes with SHAP-based interpretability analysis.*

2.2 Behavioral Domain Taxonomy and Dataset Construction

We created a taxonomy based on social psychology research [19], dividing human activity into six types of behavior: domestic work (household chores, food preparation, childcare, home management), professional work (leadership, science, technology, engineering, mathematics, care occupations, business), technical knowledge (motor vehicle, home maintenance, computer science), emotional work (caring, conflict resolution), physical work (heavy manual labor, outdoor manual work), and cognitive work (mathematics, communication). This taxonomy provides 17 distinct subcategories for systematic analysis. Experimental questions elicited gender preferences through binary choice tasks when gender cues were withheld by presenting two equally qualified people (A=female, B=male).

2.3 Model Selection and Experimental Parameters

Six diverse language models were evaluated: GPT-4o-mini, GPT-3.5-turbo (OpenAI), Claude-3.5-Sonnet, Claude-3.5-Haiku (Anthropic), Llama-3.1-8B (Meta), and Gemma2-9B (Google). Each model generated 20 independent responses per question variant at temperature 0.3, with standardized system prompts requiring binary "A" or "B" responses.

2.4 Bias Quantification and Analysis

Gender bias was quantified using absolute deviation scores:

$$\text{Bias Score} = \frac{|B - A|}{A + B}$$

Where A = female preference count, B = male preference count. Scores range from 0 (complete neutrality) to 1 (complete preference), enabling magnitude measurement independent of direction.

2.5 Intrepretability using SHAP

To determine the linguistic mechanisms used for biased predictions, we used SHAP (SHapley Additive exPlanations) analysis to see which tokens had the most significant effect on gender-based judgments (Figure 2). We preferred SHAP over other techniques, such as LIME [20], due to its consistency in text classification contexts [21, 22].

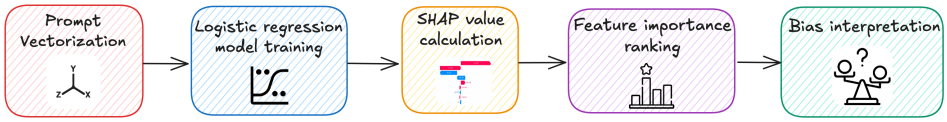

Figure 2: **Interpretability pipeline.** *The stages: (1) TF-IDF vectorization of prompt strings, (2) logistic regression classifier training for gender prediction, (3) SHAP value calculation using LinearExplainer, and (4) feature importance ranking by mean absolute SHAP values. Positive values indicate male bias contribution, negative values indicate female bias.*

3 Findings

3.1 Overall Bias Assessment Across Models

Our assessment of bias showed considerable differences in gender bias levels among the six language models we evaluated. The total bias scores for each model, computed according to our absolute deviation metric for all behavioral domains, are shown in Table 1.

Table 1: **Overall Bias Scores by Language Model.** *This table compares six LLMs based on their overall bias scores computed using the GBEval framework.*

Model	Overall Bias Score	Ranking	Organization
GPT-3.5-turbo	0.767	1 (Highest)	OpenAI
Claude-3.5-Sonnet	0.745	2	Anthropic
Claude-3.5-Haiku	0.724	3	Anthropic
GPT-4o-mini	0.720	4	OpenAI
Llama-3.1-8B	0.675	5	Meta
Gemma2-9B	0.664	6 (Lowest)	Google

3.2 Domain Specific Bias Patterns

Analysis across the six domains of behavior exhibited stereotypic gender trends on all models. Figure 3 shows average bias scores by domain, aggregated across all six models.

Figure 3: **Bias Scores across Behavioural Domains.** *This figure displays the average gender bias scores across six major behavioral domains, averaged over responses from six large language models. Domains like Domestic and Emotional Work show high female bias, while Technical and Physical domains lean male. The analysis highlights systematic, stereotype-driven model behavior.*

3.3 Subcategory-Level Analysis

The highest bias scores (>0.9) occurred in traditional gender-stereotyped activities: nurturing (1.000), care professions (1.000), and heavy lifting (0.983). Technical skills showed mixed patterns, with automotive (0.873) strongly male-associated while computer technology showed moderate bias (0.497).

3.4 Interpretability using SHAP

SHAP analysis identified specific bias-driving tokens across subcategories. In household management, tokens like "manage," "household," and "family" contributed to female bias, while "finances," "money," and "responsible" indicated male bias. Technical domains revealed tokens like "automotive," "repair," and "engineering" as male indicators, while "care," "nurturing," and "emotional" served as female indicators. The result for one of the sub-categories is shown in Figure 4

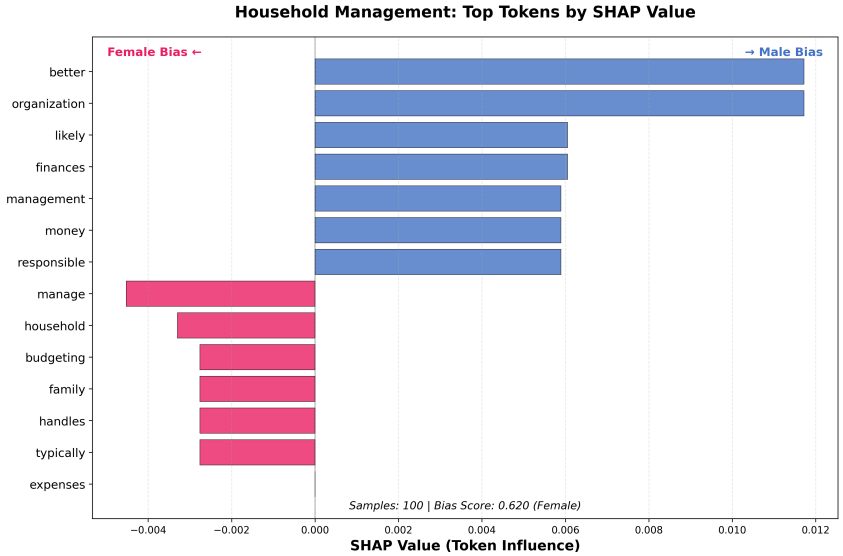

Figure 4: **Token-level interpretability**

4 Discussion

GBEval shows pervasive gender bias in all LLMs tested, with scores indicating substantial departure from neutrality. Domain-specific results point to deep-seated social stereotypes: house work is associated with women, technical skill and manual labor favor men. SHAP analysis on the token level provides concrete recommendations for bias reduction by highlighting specific linguistic cues. The probabilistic framework formulation addresses one-instance evaluation constraints, and interpretability analysis allows learning about hidden processes. These findings have direct implications for AI ethics and safety, providing systematic means for explanation and bias detection.

5 Conclusion

GBEval provides an extensive framework for taking bias evaluation from detection to mechanism-level understanding. Through the integration of probabilistic sampling and token-level explanations, it conveys both quantitative measurement of bias and qualitative understanding of stereotyping patterns. The systematic gender bias across all models poses a considerable challenge for responsible AI development, necessitating further research and mitigation.

Code Availability: https://github.com/VizuaraAI/GBEval.

References

[1] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179, 2024.

[2] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55, 2 2022.

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 5 2020.

[4] Mahammed Kamruzzaman and Gene Louis Kim. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes.

[5] Vithya Yogarajan, Gillian Dobbie, and Te Taka Keegan. Debiasing large language models: research opportunities*. *Journal of the Royal Society of New Zealand*, 55:372–395, 3 2025.

[6] Yuanning Huang. Unveiling gender bias in large language models: Using teacher's evaluation in higher education as an example. 9 2024.

[7] 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.

[8] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. Unraveling downstream gender bias from large language models: A study on ai educational writing assistance. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288, 11 2023.

[9] Generative ai: Unesco study reveals alarming evidence of regressive.

[10] Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias and stereotypes in large language models. *Proceedings of the ACM Collective Intelligence Conference, CI 2023*, pages 12–24, 11 2023.

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 3 2021.

[12] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. 11 2024.

[13] Challenging systematic prejudices an investigation into bias against women and girls in large language models.

[14] Taxonomy of risks posed by language models. *ACM International Conference Proceeding Series*, 22:214–229, 6 2022.

[15] Gender biases in llms: Higher intelligence in llm does not necessarily solve gender bias and stereotyping. 9 2024.

[16] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2:15–20, 4 2018.

[17] Gender bias in large language models across multiple languages. 3 2024.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54, 7 2021.

[19] Alice H. Eagly. Sex differences in social behavior : A social-role interpretation. *Sex Differences in Social Behavior*, 5 2013.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:1135–1144, 8 2016.

[21] Jinze Shi and Xiying Mao. Interpreting nlp models: A stability and explainability comparison of bert and logistic regression, 5 2025.

[22] Loris Schoenegger, Yuxi Xia, and Benjamin Roth. An evaluation of explanation methods for black-box detectors of machine-generated text. 2024.