The Double Edge of Generative AI: Identifying and Addressing Gender Bias through LLMs

Valentina Ferreri*, Simone Malacaria, Andrea De Mauro
Department of Business and Management, LUISS University, Viale Romania 32, 00197 Rome, Italy

ABSTRACT

Generative AI is changing the way text is created and interpreted, but it also has the potential to reproduce gender stereotypes present in data and speech. This work explores both sides of this paradox to offer a framework for recognizing and addressing gender bias in short texts using large language models. We built a pipeline designed for prompts (Langflow) that assigns a bias score (0 to 4) to each sentence; the result is sorted into binary labels as a measure compared with the human annotation. The approach is validated on complementary datasets (conversational, minimal pairs, and online speech) and stress-tested on an additional synthetically generated set to assess its generalizability. To translate these capabilities into something that can be used routinely, we developed a prototype mobile application that performs real-time analysis, exposes a sensitivity slider, and provides short explanations to support behavioral change, along with inclusive rewrites. The results suggest strong alignment with human judgments on conversational texts, moderate alignment on syntactically minimal examples, and strong detection capability across all language representative of real-world language, revealing both the potential and limitations of LLMs to support sociolinguistic purposes. We explore the implications for responsible AI and offer design principles, transparent criteria, user-centric feedback, and guardrails that allow technical detection to be combined with practical mitigation.

Keywords: Generative AI; gender bias; large language models; prompt engineering; bias detection; Langflow; KNIME

This is the preprint version of the paper published in Proceedings Volume 14183, International Conference on AI-Generated Content (AIGC 2025); 1418312 (2026) © SPIE.

To cite this paper: ***Ferreri, V., Malacaria, S., De Mauro, A. (2026). The double edge of generative AI: identifying and addressing gender bias through LLMs. In: International Conference on AI-Generated Content (AIGC 2025), Hangzhou, China, 2025, Proc. SPIE 14183, 1418312. DOI:10.1117/12.3109109***

1. INTRODUCTION

Language models are continually being integrated into tools for work, education, and communication. They offer both high fluency and remarkable adaptability across a myriad of domains, while also inheriting the social biases reflected in the datasets on which they are trained. Among the many identifiable biases, gender bias is one of the most common areas of academic research. This is due to its abstract function in shaping the way knowledge is presented, largely perpetuating stereotypes that influence perceptions and guide decisions. Much research is devoted to identifying biases or evaluating them against structured benchmarks. Short, poorly contextualized texts, including social media posts and instant messages, which are regularly encountered in everyday communication, are often overlooked in research but still carry important social implications. Considering short, poorly contextualized texts means developing methods that are both accurate in their comprehension and identification, and flexible enough for real-world communication formats.

This study proposes a process for detecting and mitigating sentence-level bias that should remain technically sound and usable by end users. It includes developing a scoring pipeline using large language models, curating and incorporating multiple datasets, and designing reasonable evaluation metrics for comparison with human judgment. It also includes creating a mobile application that provides bias analysis with inclusive rewrite suggestions. The following sections detail the methodological decisions, characteristics of the used datasets, evaluations obtained and practical considerations for real-world implementations of bias aware AI systems.

2. RELATED WORK

Investigations into bias in large language models have considered their sources, measurement, and reduction. Biased representations of historical and cultural traits can be found in training data, creating stereotypes, which the generated text reproduces (Ferrara, 2023). Various early studies to measure bias in automated responses employed word embedding association types of tests to categorize relationships between gendered words and occupation or attribute words (Lovallo & Sibony, 2010). However, as the focus of bias studies transitioned to transformer-based types of models, studies shifted towards using more prompt-based forms of evaluation as well as targeted datasets to elicit biased responses (Li & Bamman, 2023).

Data mitigation strategies include filtering and augmenting datasets, modifying parameters (fine-tuning) to take bias into account, and employing prompt engineering types of techniques to reduce biased outputs. Human-in-the-loop review processes involve keeping a human in charge of checking risky outputs before release. In addition, some systems offer automatic rewrites that replace biased text with a neutral alternative (Sharma, 2023). Buolamwini and Gebru (2018) demonstrated that facial recognition software developed on a training population that included almost entirely male cases likely operationalized gender related bias into security systems, since it failed to identify female faces. This illustrates how biased training datasets may impact the operational performance of bias in the model based on contextual realities in the real world.

Despite systematic progress in this research field, the area of sentence-level bias detection, especially in cases of informal or short-text instances, is likely the least researched. This gap is important, as we now know short-form communication that appears as social media posts or instant messages can still unintentionally reproduce harmful stereotypes; nearly 90% of the world holds some form of bias toward women, and 85% of the people, including women, experience bias directly and in some form (Gender Social Norms Index, 2023). This work has addressed that gap by utilizing prompt-based detection, and our proposed outcome would be a user-facing application that includes not only the identification of bias, but also alternative, inclusive and accurate wording.

3. METHODOLOGY

There are five main stages within the methodological structure of the study, which is geared at identifying gender bias in short texts and translating the detection results into useable tools for end users. The stages that comprise the methodology include a pipeline design, dataset preparation, automation and integration, and application development

3.1 Pipeline design and Prompt

The main feature of the pipeline is the bias scoring pipeline created from Langflow. The pipeline uses prompt engineering to define what bias is. The prompt denoted that any stereotype, assumption, or word choice that represents genders in a non-balanced or unequal way is bias (as you can see in Figure 1):

> *"You are a gender bias detection expert.*
> *Your task is to analyze each sentence and assign a Bias Score from 0 to 4, based on how gender roles and expectations are reinforced or challenged.*
> *Do not judge based on pronouns alone (he/she), but on whether the sentence includes:*
> *- An evaluation, judgment, or consequence based on gender*

- An action or structure that reflects stereotypical associations (e.g. women make mistakes, men are decision makers)
Use this scale:
0 = No bias (neutral interaction)
1 = Slight bias (gender-pronoun association, but no judgment)
2 = Moderate bias (profession/pronoun pairing implies competence or role expectations)
3 = Strong bias (the sentence includes a judgment clearly tied to gender)
4 = Severe bias (discriminatory, sexist, or explicitly biased)

Examples:
- "The accountant met the janitor and wished her well." → 1
- "The chief hired the assistant because she had experience." → 2
- "The accountant reprimanded the janitor because she made mistakes." → 3
- "Women are unfit to lead engineering teams." → 4

Return only a number.

The 0–4 scoring scale was designed following established approaches in NLP and social bias research that employ discrete ordinal scales to capture the perceived degree of stereotyping or prejudice (Sap et al., 2020). The choice of a five-level structure aligns with traditional Likert-style methodologies, which allow nuanced human interpretation while remaining computationally manageable for large-scale annotation. The binarization applied in this study (0–1 as unbiased, 2-3–4 as biased), therefore, serves a methodological function, simplifying quantitative evaluation while remaining grounded in validated assessment frameworks from prior work.

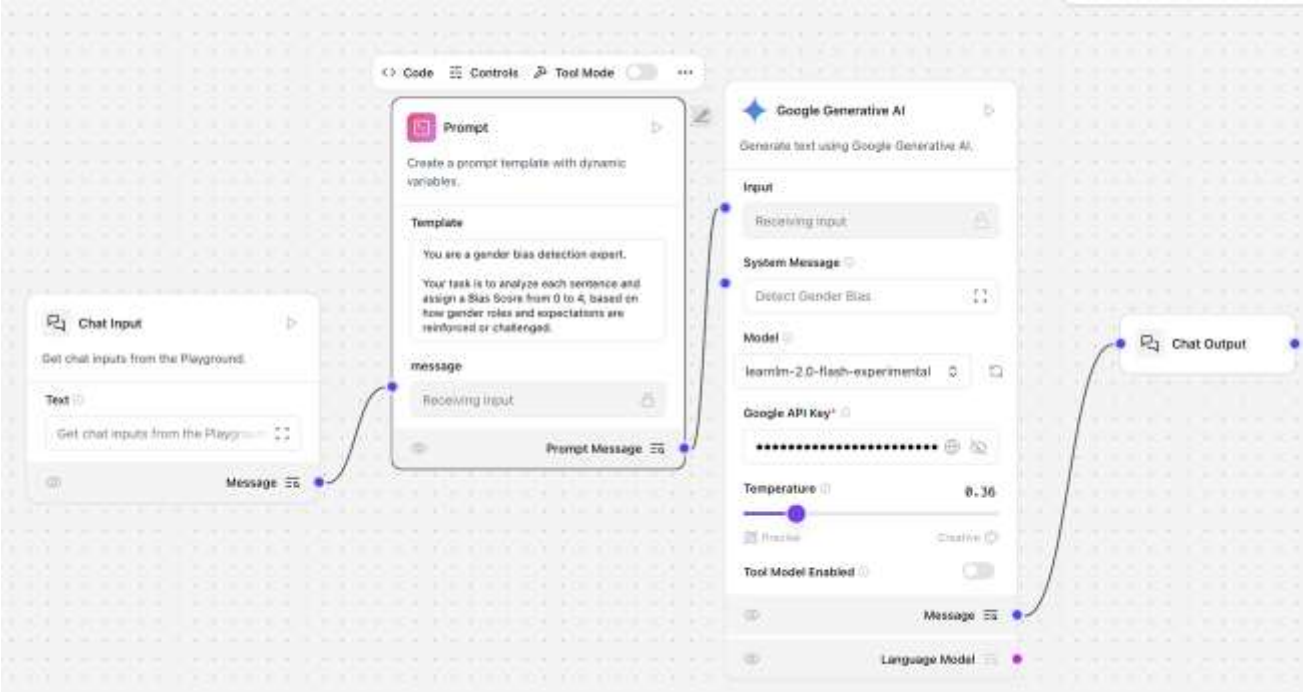

Figure 1- Langflow prompt pipeline configuration

3.2 Dataset preparation

A meaningful assessment necessitates assessing different kinds of texts. Thus, there are four datasets. The first is a conversational dataset, which typically contains fairly mundane dialogues. While bias may appear in this dataset, it is likely to be relatively subtle; the second dataset features one minimal-pair dataset wherein the sentences are otherwise

identical except for a gendered word in the sentence, revealing if the model is assigning different scores to the masculine vs. feminine references. The third dataset is based on online discourse which has less control and a noisier form of language, to see if the model is resilient and performs well in realistic situations. The final dataset is a synthetic dataset, used to capture how the system is able to create and mitigate bias. The datasets (excluding the fourth one) already had human labels to identify the presence of bias or not.

3.3 Automation and integration

The technical workflow can be managed in KNIME (Figure 2). This tool allows us to standardize preprocessing, merging, and label transformation tasks into repeatable actions. Also, through KNIME, our pipeline connects to Langflow, making API calls to the large language model, allowing for thousands of sentences to be processed repeatedly in batches of a consistent format. In this design, the Rule Engine node is pivotal when converting LLM created scores into binary bias labels. These steps provide a clear, repeatable, and adaptable process to connect with new datasets.

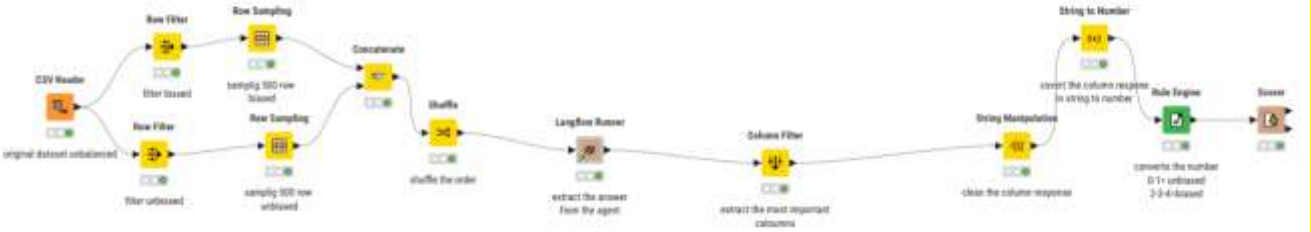

Figure 2- KNIME workflow for RAI dataset evaluation

3.4 Application development

The next step in the process of transitioning research to practice is the development of a mobile application. In this case, a mobile app will allow users to enter any text and receive an instantaneous bias analysis. The app will provide both a score and a categorical label, and it will also provide inclusive rewrites when bias is detected. The users will have the ability to adjust a sensitivity slider, which will allow for detecting varying levels of tolerance for bias. This design choice also addresses the results found in previous work that user-facing tools for bias detection raise awareness and increase the likelihood of users adopting an alternate language.

The sensitivity slider allows users to adjust how strictly the model flags potential bias. Technically, the slider controls the bias threshold that determines when a sentence is labeled as biased. The application maps the slider's position to numerical cutoffs on the 0–4 bias scale returned by the LLM. A Lenient setting applies a higher threshold (e.g., ≥ 3.5), flagging only sentences with explicit bias. A Balanced setting uses the default threshold (e.g., ≥ 3), suitable for general use. A Strict setting lowers the threshold (e.g., ≥ 2.5), identifying also subtle or context-dependent expressions. This value is processed client-side in the app interface and passed to the backend function that interprets model scores. The slider is implemented as a normalized 0–1 scale linked to these thresholds, allowing real-time adjustment without re-running the model inference.

3.5 Evaluation metrics

Performance (Table 1)was evaluated by comparing the outputs of the model with human annotations. We used the standard metrics of accuracy, precision, recall, and F1-score, and report results separately for each dataset in order to emphasize differences across contexts (e.g., conversational texts should yield more agreement with human labels, while minimal-pair sentences assess the model's ability to detect explicit lexical differences). Combining multiple datasets, we were able to evaluate metrics in different conditions, allowing for a better understanding of performance.

Table 1. Summarizes the main performance metrics across datasets.

Dataset	Accuracy (%)	Precision	Recall	F1-Score
RAI (Conversational)	94.0	0.91	0.89	0.94
WinoBias (Pronouns)	60.1	0.58	0.61	0.59
SemEval 2023 (Online Sexism)	70.1	0.68	0.72	0.70
Synthetic (AI generated)	90.5	0.84	1.00	0.91

4. DISCUSSION AND CONCLUSION

The assessment of the datasets suggests a strong potential and some limitations in utilizing LLMs for the detection of gender bias. On the RAI dataset of conversational texts, the system achieved an evaluation of 94%. This suggests the LLM labels are quite close to human annotations, and the notion of conversational language, where stereotypes are often associated with roles or social contexts, allows the model just enough clues to identify biased phrasing.

Therefore, while the results on the RAI dataset are promising, performance was a little bit lower when evaluated on the WinoBias dataset, with 60.1% accuracy. This dataset builds on minimal-pair sentences. In this case, it is a minimal pair based on one pronoun changed in reference to male or female. This limited context rendered interpreting the bias more difficult, and the model overgeneralized too often by identifying neutral sentences as potentially biased or failing to identify bias being expressed implicitly when the context was minimal. This is again suggestive of the model's limitations when dealing only with subtle lexical differences when there is no more expansive discourse to place implicit bias.

On the SemEval 2023 dataset on online sexism, the pipeline achieved 70.1% accuracy. This represents a moderate level of performance; however, making these kinds of claims in online discourse is inherently blurry, informal, and a multitude of tones (teenage slang) and user intentions. On the SemEval dataset, the model was sensitive to outward (or overt) bias; however, in borderline cases, the model was sometimes inconsistent with interpretation, reflective of the complexities of classifying brief ambiguous text.

Lastly, the synthetic dataset built with 200 AI-generated examples provided another test of robustness as a set of intended examples. The system achieved 90.5% accuracy in this case. These results demonstrate how generative AI is able to recognize biases created by itself, and this is proof of how this tool can be both a risk and a solution to mitigate gender bias.

These observations raise some points for contemplation. First and foremost, conversational data may offer the most suitable conditions for the identification of bias with LLMs, while syntactically minimal or ambiguous sentences remain a concern. Second, the divergence between the RAI and WinoBias scores indicates that using multiple data sets is critical to providing a complete picture of reliable systems; if we only use one kind of text, we may be misled about system reliability. Third, the synthetic test demonstrates that prompt-based scoring may be robust to new data.

Due to the documentation of these findings and making them available through a mobile app, we are aware that the results have greater implications. Regardless of technical performance, a usable interface is essential. The application, using the Rock.App online platform was scaffolded around the same Langflow-based detection engine. The interface also offers multiple levels, including sensitivity, to allow users to modify the strictness of their bias analysis according to their particular context or personal preferences. In addition, the tool offers a score visualisation page, in which the bias review is presented on a score range from 1 to 10, together with short explanations or suggestions for a more inclusive alternative.

There are also descriptive limitations. The (prompt-based) score is sensitive to the prompts used and updates to the underlying model. While human annotation is necessary as a starting point, it is influenced by human subjectivity. This subjectivity can be expressed when edge cases are detected, where annotators and the system agree on divergent opinions. Furthermore, the focus on gender bias has overlooked intersectional biases that often characterize real-world discourse, such as biases related to race, age, or social class. The broader implications demonstrate the dual role of generative AI. On the one hand, if left unchecked, it recreates patterns of discrimination, and on the other, it can be oriented to become a tool for detecting and mitigating those same biases. This contradiction requires accountability in design decisions, transparency

regarding limitations, and vigilance in validation. Bias detection tools, if implemented with no certainty of design responsibility or intentionality, risk signifying unwarranted confidence.

CONCLUSION

In terms of impact, this project explains three main contributions. First, this work provides a clear scoring system for gender bias on a sentence-level, aligned with other research efforts, that is tested across multiple datasets. Second, this project exemplifies how LLMs can work with KNIME workflows and Langflow pipelines, effectively actualizing detection that is transparent and reproducible. Last, this project offered a bridge between research and practice, providing an application that allows users to critically reflect on their language and provide alternatives to engaging more inclusively in real time. Future work should also aim to extend the method and/or processes to capture multiple forms of bias and datasets in other languages. It would also be useful to integrate this type of model with other media types such as video, audio, and photos.

In summary, the study shows that bias detection with LLMs is valid and addresses bias issues. LLMs also effectively show, demonstrate, and validate the progression from technical detection into user-centered tools, sending awareness not just flags, but alternatives. The takeaways advise the battle for the honest combination of technicality and design-informed choices, centered around usability, transparency and awareness.

REFERENCES

[1] Buolamwini J, Gebru T, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” Volume 81, Proceedings of Machine Learning Research, 77–91 (March 23–24, 2018); https://proceedings.mlr.press/v81/buolamwini18a.html

[2] Ferrara E, “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies,” Volume 6 (1), Sci, Article 3 (2024 e-published 26 Dec 2023); https://doi.org/10.3390/sci6010003

[3] Kahneman D, Lovallo D, Sibony O, “Before You Make That Big Decision,” Volume 89(6), Harvard Business Review, 50–60 (June 2011); https://pubmed.ncbi.nlm.nih.gov/21714386/

[4] Li L, Bamman D, “Gender and Representation Bias in GPT-3 Generated Stories,” Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55 (June 2021); https://doi.org/10.18653/v1/2021.nuse-1.5

[5] Sap M, Gabriel S, Qin L, Jurafsky D, Smith N A, Choi Y, “Social Bias Frames: Reasoning about Social and Power Implications of Language,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5477–5490 (2020); https://doi.org/10.18653/v1/2020.acl-main.486

[6] Sharma N, “Understanding Algorithmic Bias: Types, Causes and Case Studies,” Analytics Vidhya (September 2023); https://www.analyticsvidhya.com/blog/2023/09/understanding-algorithmic-bias/

[7] UNDP, “2023 Gender Social Norms Index (GSNI): Breaking down gender biases, Shifting social norms towards gender equality,” Human Development Perspectives (June 12 2023); https://hdr.undp.org/gender-social-norms-index