







← Go to **NAACL 2025 Workshop QAI** homepage (/group?id=aclweb.org/NAACL/2025/Workshop/QAI)

# Some Myths About Bias: A Queer Studies Reading of Bias Evaluation and Mitigation Techniques in NLP

 (/pdf?id=GE6X5tLjcK)

Filipa Calado (/profile?id=~Filipa\_Calado1) 

 30 Jan 2025 (modified: 31 Jan 2025)  NAACL 2025 Workshop QAI Submission  QAI, Reviewers, Authors  
 Revisions (/revisions?id=GE6X5tLjcK)  CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/)

**Keywords:** Queer Studies, gender bias, mitigation and evaluation methods, word embeddings


**TL;DR:** This paper takes a Queer Studies critique of binary forms to read and analyze popular gender bias evaluation and mitigation techniques in NLP.


## Abstract:

This paper critically examines gender bias in large language models (LLMs) by integrating concepts from Queer Studies, particularly the theory of gender performativity and the critique of binary forms. It argues that many existing bias detection and mitigation techniques in Natural Language Processing (NLP), such as the Word Embedding Association Test (WEAT) and gender swapping methods, rely on outdated conceptualizations of gender, which take for granted the gender binary as a symmetrical and stable form. Drawing from Queer Studies, the paper highlights three "myths" about gender bias: that bias can be excised, that it is categorical, and that it can be leveled. Due to their operationalizing of the gender binary, each of these myths effectively reduce and flatten bias into a measure that fails to represent real-world workings of semantics, discrimination, and prejudice. The paper concludes by suggesting that bias mitigation in NLP should focus on amplifying diverse gender expressions and incorporating non-binary perspectives, rather than attempting to neutralize or equalize them. By reworking that which is outside the binary form, against which the binary defines itself, one may fashion more inclusive and intersectional approaches to mitigating bias in language systems.

**Archival Submission:** Archival




**Submission Number:** 3




Filter by reply type... 


Filter by author... 


Search keywords...

Sort: Newest First











 Everyone Program Chairs Submission3 Reviewers Submission3 Authors

4 / 4 replies shown

Submission3... Submission3... Submission3... Submission3... 

Add: **Withdrawal**

## Paper Decision

Decision by Program Chairs  07 Mar 2025, 14:56 (modified: 07 Mar 2025, 17:49)  
 Program Chairs, Reviewers, Authors  Revisions (/revisions?id=6mSOiXmGhF)

**Decision:** Accept (Non-archival)

**Comment:**

This paper presents a queer studies perspective on NLP bias evaluation and mitigation techniques. Reviewers agree that the submission draws on the relevant queer studies literature and serves to introduce NLP practitioners to this area; however, reviewers also expressed that the paper could be strengthened by adding further analysis and interpretation of the literature it discusses. Consider these recommended changes from reviewers: (1) additional engagement with gender studies literature and feminist analyses of gender bias, including specific references from Reviewer aD53; (2) add additional discussion to tie the cited works to the analysis of NLP; (3) add additional analysis of quotes, and possibly move some quotations to paraphrases to make space for this analysis. Also please adjust to Judith Butler's preferred pronouns (thank you Reviewer kSyt for noting this!). Furthermore, the work is of nice quality but it's accepted as non-archival because the required formatting guidelines for archival were not followed.

===== IMPORTANT INFO BELOW =====

**Note:** If you plan to present your work (either in-person or virtually), you must be registered for NAACL. That is, if you would like to present, at minimum, you need to be registered for the workshop-only option (either virtual or in-person is acceptable). **Please email us by April 1 with the name of the author who will be presenting, or let us know if you do not plan to present.**

Please note the costs of registration here (<https://2025.naacl.org/registration/registration/>). We STRONGLY recommend that authors who need registration support apply for the following calls by **March 10**:

- D&I subsidies ([https://2025.naacl.org/calls/dei\\_subsidies/](https://2025.naacl.org/calls/dei_subsidies/))
- Virtual registration subsidies ([https://2025.naacl.org/calls/virtual\\_dei\\_subsidies/](https://2025.naacl.org/calls/virtual_dei_subsidies/))
- Volunteering (<https://2025.naacl.org/calls/volunteers/>)

The NAACL D&I chairs are meeting on March 12 at noon Pacific Time to make decisions about the subsidies, so please submit by then at latest.

**If you are concerned that the registration cost will prohibit you from attending, please let us know ASAP.** Queer in AI is unable to subsidize registrations for authors due to much fewer renewals of corporate sponsorship this year. We are actively working with NAACL to procure more free registrations for Queer in AI authors, but we do not currently have any guarantees on the amount of support we will receive. Please do not hesitate to email us if you have any additional comments or concerns ([queer-in-nlp@googlegroups.com](mailto:queer-in-nlp@googlegroups.com) (<mailto:queer-in-nlp@googlegroups.com>)).

## Official Review of Submission3 by Reviewer kSyt

Official Review by Reviewer kSyt 📅 04 Mar 2025, 13:44 (modified: 07 Mar 2025, 18:51)

👁 Program Chairs, Reviewers Submitted, Reviewer kSyt, Authors 📁 Revisions (/revisions?id=KlXhii4h1u)

### Summary:

The paper primarily discusses some myths about bias. First it briefly mentions schemas of gender bias in NLP and then dives into Queer Studies and talks about gender (using gender performativity) and harms of binary conceptualisation of gender. Then using some works in NLP bias research, it examines three myths about bias (that it can be removed, it is categorical and it can be leveled). In conclusion, it highlights directions for mitigating bias in NLP inspired by queer studies and some works in NLP on amplifying gender expression and including non-binary perspectives.

### Quality And Clarity:

I find this paper to be quite clear in its arguments and word usage. For example, using the words 'excise' and 'mitigate', essentially one point the paper presents is that bias can't be excised directly but can be mitigated (made less severe). Section 2.2 on Queer Studies uses the theory of gender performativity (also preferred in this (<https://dl.acm.org/doi/10.1145/3531146.3534627>) work) to define gender and further based on relevant works discuss the harms of using binary categories such as male/female.

The main section (Section 3) of the paper is on the myths about bias, which primarily talks about some popular works in NLP gender bias research, specifically: word-based data filtering (for 'bias is excisable' myth), WEAT (for 'bias is categorical' myth) and Hard DeBias/Counterfactuals (for 'bias can be levelled' myth). So while this is not as exhaustive as a survey, it still suffices to use them to illustrate these myths.

Section 3.2 can be made more clear. WEAT seems to dominate most of the discussion, however there are several other examples to illustrate how bias is categorically conceptualised in research. WEAT measurement involves measuring associations with concepts of different variety for example, measuring gender associations with maths vs arts, career vs family, etc. So it is not always (infact, rarely) good vs bad conceptualisation. Although since WEAT uses binary gender representations in doing so, binary category argument holds which can be highlighted more clearly in the discussion of this myth (Section 3.2). I do believe that WEAT is often unreliable and unstable, but I think the resulting score is a measure of strength of association with these concepts which is more interpretative/suggestive rather than plainly good/bad.

I believe this paper presents an original perspective by conceptualizing these myths, summarizing critical considerations for NLP bias research. While these myths build on existing critiques of bias evaluation and mitigation, their grounding in Queer Studies makes them especially relevant for NLP bias researchers and practitioners.

### Impact And Relevance:

The paper is relevant for QIA workshop as it explores how queer studies can be accounted for in NLP bias research and points out myths about gender bias conceptualisation. We often forget to understand bias and what it means for us before starting to measure and mitigate it.

By highlighting these myths, this paper provides a necessary direction on how we should carry NLP bias research. For example, being cautious about the use of data filtering techniques which can lead to erasure, treating bias as categorical or even quantifiable can undermine the importance of measuring its actual effect on diverse users, and strategies that seem to level it could actually just mask these biases which can be manifested in other forms like downstream applications.

### Changes And Additional Comments:

**For formal reasons, since this paper does not follow standard ACL style formatting, if selected, camera ready version should be in official ACL style as required by the workshop.**

### Suggestions/Comments:

The paper illustrates myths using specific methods from bias evaluation and mitigation literature, such as WEAT, HardDebias, which may not represent the current state-of-art, albeit very popular. So I would suggest also mentioning more works to illustrate the varieties in which these myths continue to exist in research. This paper references this work (<https://arxiv.org/pdf/2205.02526>), which has a detailed survey (till 2022) including how bias is theorised? (See Table 4), where many works will be relevant to support the points made about the myths.

Since we are moving towards generative models and evaluation is moving away from intrinsic embedding-based metrics or extrinsic template-based methods to purely generation setting (such as measuring gender bias in LLM generated stories, biographies, reference letters (<https://aclanthology.org/2023.findings-emnlp.243.pdf>) etc.), you can connect these ideas with this era of bias evaluation and mitigation (whether they are based on prompting, fine-tuning, alignment etc.). For example, according to this (<https://aclanthology.org/2024.findings-eacl.121/>) paper about bias suppression: "We show that, using CrowsPairs dataset, our textual preambles covering counterfactual statements can suppress gender biases in English LLMs such as LLaMA2. Moreover, we find that gender-neutral descriptions of gender-biased objects can also suppress their gender biases." Some LLM oriented papers which can be useful for your analysis are: 1 (<https://arxiv.org/pdf/2401.15585>), 2 (<https://arxiv.org/pdf/2402.11190>), 3 (<https://aclanthology.org/2024.icnlp-1.42.pdf>), 4 (<https://aclanthology.org/2023.acl-short.30.pdf>), 5 (<https://aclanthology.org/2024.emnlp-main.13.pdf>), 6 (<https://aclanthology.org/2025.coling-main.450.pdf>), 7 (<https://aclanthology.org/2023.findings-emnlp.689.pdf>) etc. Multilinguality of bias is another ignored aspect, since many methods are based on resources (template set, specific word banks) which might not contextualise well for other languages.

Also, the paper only briefly mentions the possible directions to address these myths, that too in the conclusion. A separate section such as Discussion would be ideal. I would suggest referring back to aspects from queer studies

and how they can be used to improve NLP bias research while discussing these myths to draw more connections. I know the main focus of this paper is the discussion of Myths about bias, but some original actionable insights from the author would give a better directive to the readers on how this gap can be filled.

#### Minor suggestions:

You can use "they/them" pronouns to refer to Judith Butler which they prefer (<https://www.tagesspiegel.de/gesellschaft/queerspiegel/das-pronomen-ist-frei-vom-korper-aber-es-ist-nicht-frei-vom-geschlecht-4149826.html>).

Abstract and Introduction mention that this paper examined bias in LLMs, although based on its contents, using a more general word such as NLP will be more suitable. Subtitle can use gender bias instead of just bias.

On page 5, Para 3 uses AIT instead of possibly IAT.

**Rating:** 3: Accept

**Confidence:** 4: The reviewer is confident but not absolutely certain that the evaluation is correct

## Official Review of Submission3 by Reviewer 96L6

Official Review by Reviewer 96L6 📅 02 Mar 2025, 10:12 (modified: 07 Mar 2025, 18:51)

👁 Program Chairs, Reviewers Submitted, Reviewer 96L6, Authors 📄 Revisions (/revisions?id=IWqdZgX8FT)

### Summary:

This paper analyses methods that attempt to tackle gender bias in Large Language Models, by drawing on commentaries on binaries in Queer Studies. It demonstrates how current formulations of binaries in LLMs treat binaries as excisable, as categorical, and as being able to be levelled. This argument serves to support the author's conclusion, which is that the conventional understanding of the binary can be reformulated to proliferate rather than eliminate bias.

### Quality And Clarity:

The paper draws on a lot of other work to illustrate its point, but overly relies on such work to build its argument. In general, the author's voice (and therefore argument) is drowned out by other voices. For example, the author lists out different definitions of bias in NLP research, but does not go onto say how it is defined in this paper. As a reader, it felt as though I had to string the information that was presented by myself, just to understand what the paper was trying to get at. This impression of mine can also be linked to the author's overuse of direct quotations. This habit is compounded with the author's tendency to introduce things and then move on, which led to quotes being introduced at the end of paragraphs without further elaboration. For example, the author ends the second section with a quote that they don't even (claim to) address until the paper's conclusion. Moreover, the quote is not even reintroduced; while the content of the conclusion is linked to the quote, this is a connection that readers have to make by themselves (if they even remember it from the second section). The paper does contain a few strong sentences but they are difficult to weed out and therefore fail to stand out amongst everything else.

### Impact And Relevance:

I think it is hard to judge this considering my critique (I.e., that I felt as though I had to string the information that was present), as I am unsure if what I understood was what the author was trying to get at argument wise. The evidence is clear, but what point and overall argument it is intended to support is not as obvious. I think the paper could benefit from a clearer motivation as well as an illustration of the paper's urgency and necessity. Information for this can be taken from later sections of the paper that felt misplaced (e.g., how "the underlying strategy of using word embeddings continues to influence a distinct trajectory of development," how "equity is not the same as equality"). I also feel like the author's debunking of the three myths does not directly link to the conclusion; an explanation of how the traditional binary is inadequate is not enough evidence to claim another way is better. Additionally, the author illustrates this by once again relying on other works, rather than spelling out the argument themselves.

### Changes And Additional Comments:

Replacing direct quotes with paraphrasing will serve to clarify how referenced concepts are relevant. If being used,

direct quotes should be followed up with an explanation of how it is linked with the discussion at hand. I also noticed a few typos so another proofread is necessary.

**Rating:** 2: Weak Reject

**Confidence:** 3: The reviewer is fairly confident that the evaluation is correct

## Official Review of Submission3 by Reviewer aD53

Official Review by Reviewer aD53 📅 26 Feb 2025, 20:48 (modified: 07 Mar 2025, 18:51)

👁 Program Chairs, Reviewers Submitted, Reviewer aD53, Authors 📄 Revisions (/revisions?id=J36INDgYF5)

### Summary:

The paper articulates critiques from queer studies of the gender binary. It also criticises common methods of 'gender debiasing,' articulating three main areas where this fails: common debiasing methods 1) fail to take into account context, including where offensive terms are reclaimed; 2) reduce bias to a binary yes/no, flattening social meanings; and 3) miss gendered associations to supposedly 'neutral' words. It concludes by calling for more flexible methods to address gender bias.

### Quality And Clarity:

It's great to see more work applying a queer lens to computing in general and NLP in particular, and I can see that the author(s) here have engaged with some key authors in queer studies, especially Sedgwick, Halberstam and Butler. All the 'myths' that the author(s) engage with are well worth critiquing!

However, I think this paper would benefit from a much clearer linking of the queer studies literature to the three areas it is critiquing. The critique of Myth 1 holds for reclaimed language from any axis of discrimination (for example, reclamation of the n-word by Black people in some communities); and of Myth 2, for any examination of bias that reduces to 'biased/not biased. It's not clear what the queer angle is here, or why the author(s) think that a gendered or queer reading adds to the critique. For Myth 3, the gender angle is clear, but again, it's not clear why a queer reading is useful.

It's also not clear to me that the author(s) have engaged with sections of the relevant literature. Critiques of the gender binary in society form a large part of the field of Gender Studies (which is oddly not mentioned in the introduction), and this paper does not engage with feminist analyses of gender bias in language and in society.

I am also not clear what this paper adds to the literature. The conclusion points to some alternatives to gender debiasing, and provides some additional queer theory, but does not explain why the alternatives benefit from the theory.

### Impact And Relevance:

There are indeed problems with commonly-used debiasing techniques, and some of these reinforce gender discrimination, which affects members of LGBTQIA+ communities. However, the critiques in this paper do not specifically address the situations of LGBTQIA+ people, and, as outlined above, do not specifically articulate a connection to queer theory or understandings. This paper might introduce readers from an NLP background to some key thinkers from queer theory, which might be of value, depending on the other accepted papers in the workshop.

### Changes And Additional Comments:

Thank you for the chance to offer feedback on this paper. I do think there's some interesting ideas here, but there needs to be more work to draw these out, and to link the theory and conclusion to the content of the analysis.

- I recommend the author(s) more clearly articulate why your critiques (which are valid!) benefit from a queer understanding. What, precisely, does this add to the existing critique of these methods? How is a queer understanding different from a more general non-discrimination perspective which is relevant to lots of different marginalised groups?
- I recommend the authors consider perspectives on gender bias and discrimination from the Gender Studies field: the book Data Feminism by Catherine D'Ignazio and Lauren Klein is an introduction to some of these ideas from a

critical data perspective.

- There are some informative papers on critiques of gender binaries in computing which I think would be worth considering in your analysis, including for example: Keyes O, 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition' (2018) 2 Proceedings of the ACM on Human-Computer Interaction, 88:1; Tsika N, 'CompuQueer: Protocological Constraints, Algorithmic Streamlining, and the Search for Queer Methods Online' (2016) 44 Women's Studies Quarterly 111

**Rating:** 2: Weak Reject

**Confidence:** 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/  
Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[Frequently Asked Questions \(https://](https://docs.openreview.net/getting-started/frequently-asked-questions)

[docs.openreview.net/getting-started/frequently-  
asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

[Feedback](#)

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review with legal nonprofit status.

We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors). © 2025 OpenReview