Bias Aware AI: The Persistence of Gender in Text

Aishwarya Saran
Neema George (neemageo@buffalo.edu)

Advised by Prof. Kenneth Joseph
Department of Computer Science and Engineering
University at Buffalo

---

1. Abstract

Natural Language Processing systems are increasingly embedded in high-stakes decision-making processes, including hiring, identity classification, and content moderation. A core concern is that these systems often reflect and reinforce societal gender biases present in the language data they are trained on. This paper investigates a critical but underexplored question: what does text actually look like once gender has been systematically removed, and how does that removal affect model performance on downstream tasks?

Using two large-scale datasets (Wikipedia biographies and New York Times obituaries), the authors apply multiple layers of progressive gender obfuscation and evaluate how well three models (Bag of Words, fine-tuned BERT, and zero-shot LLaMA-2) can still detect gender after each step. The paper introduces three approaches to defining and operationalizing gender: an LLM-based strategy, a crowd-sourced layperson survey, and an empirical data-driven approach. The findings show that gender is deeply embedded in language through occupational terms, relational language, health references, and cultural stereotypes, such that models continue to predict gender with high accuracy even after removing pronouns and names.

As an empirical extension to the original submission, this report also implements a Keyword-in-Context co-occurrence analysis to test two open hypotheses from the original paper: whether the words *air* and *world* function as male gender proxies through military associations. Both hypotheses are confirmed with quantitative evidence.

The study makes three contributions: new characterizations of gender ideologies as expressed in language; an analysis of what language looks like under various gender removal strategies; and a clearer picture of the tradeoff between gender removal and downstream task performance.

2. Background and Related Work

2.1 Gender in Language

Gender manifests in language both directly and indirectly. Direct indicators include gendered pronouns (he, she, him, her, his, hers), honorifics and titles (Mr., Mrs., Miss, Ms., Sir, Madam), and gender-inflected nouns (actor/actress, king/queen, father/mother). These explicit markers are relatively easy to detect and, in principle, remove. The harder problem lies in indirect indicators.

Indirect gender signals include proper names, which carry cultural gender associations despite having no inherent grammatical gender in English. Names like ‘Alice’ or ‘Helen’ are strongly associated with women, while ‘John’ or ‘David’ are typically male-coded. Occupational terms also carry implicit gendered meaning: words like ‘footballer,’ ‘army,’ ‘engineer,’ and ‘rabbi’ have historically been associated with men, while ‘nurse,’ ‘soprano,’ ‘actress,’ and ‘feminist’ are associated with women. These associations are products of historical and cultural patterns reflected in large text corpora, not intrinsic properties of the words.

Sociolinguistics and sociology have long held that gender is not simply a binary biological fact but a social construct performed and reproduced through language and interaction. For NLP systems trained on large corporation, this means gender bias is not a discrete feature that can be switched off. It is distributed across many layers of language simultaneously.

2.2 Gender Bias in NLP Systems

Researchers have documented gender bias across many NLP systems. Bolukbasi et al. (2016) showed that word embeddings like Word2Vec encode stereotypes, such that 'man' is to 'doctor' as 'woman' is to 'nurse.' Resume screening systems have similarly been found to disadvantage female applicants not through explicit gender labels, but by learning to associate masculine-coded language ('assertive,' 'led,' 'competitive') with positive outcomes.

This has motivated substantial debiasing work, broadly in three areas: debiasing word embeddings by removing the gender direction in the embedding space; debiasing downstream models by removing gender-indicative features during training or inference; and debiasing generated text from LLMs by modifying training objectives or generation parameters.

Critics have raised important challenges. Gonen and Goldberg (2019) argued that many embedding debiasing methods are largely cosmetic: they reduce direct gender-word associations, but the underlying stereotypes remain latent in the embedding geometry. Blodgett et al. (2020) argued that most debiasing work lacks a coherent theoretical framework for what gender is and why it should or should not appear in various contexts. This paper attempts to address both gaps empirically.

2.3 Bias in Biographies and Obituaries

Wikipedia biographies have been extensively studied as a site of gender bias. Women are significantly underrepresented, accounting for roughly 17–18% of biographies (Wagner et al., 2015). Biographies of women tend to focus more on personal relationships and family, while men's biographies emphasize professional accomplishments and career. This means gender-associated language is not evenly distributed: relational terms like 'wife,' 'mother,' and 'daughter' appear disproportionately in women's biographies, while 'founded,' 'led,' and 'chairman' appear more in men's.

The NYT Obituaries corpus presents a similar pattern. Obituaries are retrospective narratives reflecting the social norms of the era when the subject lived. Given the dataset spans decades of NYT publishing history, men's obituaries frequently reference military service, business leadership, and academic achievement, while women's more often reference marital status, family roles, and artistic pursuits.

3. Methods

3.1 Datasets

Three datasets were used, each representing a different type of biographical or memorial text:

Dataset	Total Size	% Male	% Female	Avg. Length (words)
HuggingFace WikiBios	728,321	83%	17%	117.4
Wikipedia Biographies	99,994	70%	30%	457.9
NYT Obituaries	26,661	82%	18%	761.6

Table 1: Dataset Statistics

All three datasets show a strong male skew, consistent with the documented underrepresentation of women in Wikipedia and legacy news media. Gender labels were assigned by counting gendered pronoun occurrences: if a biography had more masculine pronouns (he, him, his, himself) than feminine (she, her, hers, herself), it was labeled Male, and vice versa. Non-binary and other gender identities were excluded from this study, a limitation the authors acknowledge.

3.2 Progressive Gender Obfuscation

The core methodological contribution is a three-step pipeline for progressively removing gender signals from text, where each step builds on the previous one:

Step 1 – Removing Honorifics: All gendered honorifics (Mr., Mrs., Miss, Ms., Master, Missus) were replaced with the gender-neutral term "Mx." This is the least invasive step, as honorifics are relatively sparse in biographical text.

Step 2 – Neutralizing Gendered Pronouns: Gendered pronouns were replaced with gender-neutral equivalents using a combination of an existing codebase ("They, Them, Theirs: Rewriting with Gender-Neutral English") and a custom algorithm. Replacements: he/she to they, him/her to them, his/hers to their, himself/herself to themself. The method assumes single-entity sentences; cases with multiple people were handled by the custom fallback algorithm.

Step 3 – Removing Names and Gendered Nouns: Subject names were identified using SpaCy's NER tagger and replaced with "This person." When the tagger failed to identify a name, the first two words of the text were replaced (biographies and obituaries almost always begin with the subject's name). Gendered nouns were also replaced with neutral equivalents using an inclusive language reference list (e.g., son/daughter to child, actor/actress to performer, chairman/chairwoman to chairperson).

3.3 Classification Models

Three models were used to evaluate gender predictability after each step:

Bag of Words (BoW): A logistic regression classifier trained on term frequency features. The model's top-weighted coefficients were analyzed to identify which words most strongly predicted each gender class. This approach gives the most interpretable results.

Fine-tuned BERT: The `bert-base-cased` model was fine-tuned for binary gender classification (4 epochs for HuggingFace WikiBios, 2 epochs for the other datasets, selected based on validation performance). BERT has a 512-token limit; since average Wikipedia and obituary texts are much longer, the model effectively reads only the opening paragraph of most documents.

LLaMA-2 (Zero-Shot): The open-source LLaMA-2-13B model was prompted without task-specific training to predict gender and generate explanations of its reasoning. Explanations were clustered to identify the categories of evidence the model relied on most.

3.4 Evaluation Approach

Two evaluation strategies were used:

- **Constant Testing Method:** The model was trained only on Step 1 data (pronouns and honorifics removed) but evaluated on all three obfuscation levels.
- **Variable Testing Method:** Both training and test data used the same obfuscation step.

3.5 KWIC Hypothesis Analysis

As an empirical extension, a Keyword-in-Context (KWIC) pipeline was implemented on the Step 2 NYT Obituaries corpus to test two open hypotheses from the original study: (1) whether *air* functions as a male proxy via military-aviation associations, and (2) whether *world* functions as a male proxy via World War references. The pipeline had three stages: (a) re-training the BoW classifier on Step 2 data to confirm both words appear among the top male-class coefficients; (b) KWIC extraction of every sentence-level window ($\pm$10 tokens) around each occurrence of the target word; and (c) co-occurrence counting of candidate military, sports, and neutral phrases within each window, stratified by gender label. The full implementation used standard Python libraries (`pandas`, `scikit-learn`, `spaCy`, `matplotlib`) and ran in under ten minutes on a laptop.

4. Results

4.1 Overall Accuracy

Across all models and datasets, gender remains highly predictable even after all three debiasing steps. Table 2 summarizes key accuracy results:

Dataset	BoW Step 1	BoW Step 3	BERT Step 1	BERT Step 3	LLaMA Step 1
HuggingFace WikiBios	91.1%	86.5%	92.3%	88.2%	95.2%
Wikipedia Biographies	94.7%	92.7%	93.2%	87.7%	N/A
NYT Obituaries	93.5%	89.2%	96.5%	96.4%	97.9%

Table 2: Model Accuracy by Dataset and Debiasing Step

NYT Obituaries consistently returned the highest accuracy across all models, which is not surprising given the length and contextual richness of obituary text. LLaMA-2 had the highest raw accuracy on unmodified text despite no task-specific training, which likely reflects how much gender-relevant knowledge is encoded in its pretraining corpus.

One result worth highlighting involves the two evaluation techniques. Technique 2 (train on Step 1, test on all steps) outperformed Technique 1 (train and test on the same step) when evaluated on Step 3 data for the HuggingFace WikiBios dataset. The BoW difference was 86.5% (T1) vs. 84.4% (T2 cross-evaluation), with a similar pattern for BERT. One possible explanation is that models trained on lightly debiased data retain enough distributional signal to generalize better than models trained directly on aggressively debiased text, which may have too sparse a feature space.

4.2 Bag of Words: Class Proxy Words

The BoW coefficient analysis showed which words the logistic regression model weighted most heavily for each gender class. These "class proxies" expose the layers of gender stereotyping embedded in biographical language.

For the **female class**, the highest-weighted terms across datasets included: *women* (−24.3), *actress* (−26.3), *female* (−17.6), *woman* (−15.2), *husband* (−11.8), *married* (−17.3), *daughter* (−8.5), *mother* (−7.5), *nurse* (−5.8), *singer* (−7.1), *feminist* (−6.7), *breast cancer* (−8.9). The *breast cancer* result from the NYT Obituaries is worth noting separately: the model learned to associate a specific health condition with female gender, which reflects real epidemiological patterns but also encodes them as a predictive signal in ways that could cause problems in downstream applications.

For the **male class**, the highest-weighted terms included: *wife* (+37.6), *their wife* (+37.6), *actor* (+10.6), *men* (+10.3), *footballer* (+8.9), *football* (+7.7), *army* (+10.5), *served* (+7.9), *Harvard* (+7.4), *university* (+5.0), *heart attack* (+5.2), *world war* (+4.8), *navy* (+4.9), *son* (+4.3). The presence of *Harvard* and *university* as male predictors suggests that academic achievement was so heavily associated with men in the dataset's time period that academic vocabulary itself became gender-coded. *world war*, *navy*, and *army* reflect historical military demographics, but their appearance as learned features means a model trained on this data will apply those associations to new text.

When gendered nouns were replaced with neutral equivalents (actress to performer, son to child), the model adapted. As Table 3 shows, *performer* shifted to the female class (−6.5), and *consort* replaced *wife* as a female proxy (−9.7). Surface-level lexical replacement is not sufficient: the model learns the substitution mapping and simply treats the neutral term as a stand-in for the original gendered one.

Step	Term	Class	Coefficient
Step 1	actress	Female	−26.3
Step 1	actor	Male	+11.8
Step 3	performer	Female	−6.5
Step 3	performer	Male	+1.2
Step 1	wife	Female	−11.8
Step 3	consort	Female	−9.7
Step 1	son	Male	+4.3
Step 3	child	Female	−2.1

Table 3: Coefficient shift after gender-neutral noun replacement (WikiBios, BoW)

4.3 BERT: Interpretability Challenges

Fine-tuned BERT matched BoW accuracy but was much harder to interpret. LIME analysis returned largely uninformative results, with the highest-weighted tokens being stop words like "of" or "has." SHAP provided slightly more insight in specific cases: one biography was predicted male because it contained the phrase 'studying history,' while in another, the word 'song' pushed the prediction toward female. These examples suggest occupational stereotyping but could not be verified systematically across the full dataset.

One SHAP result was particularly clear. A biography containing 'engineering,' 'research,' and 'hospital' was predicted male, with SHAP values showing 'engineering' and 'research' pushing toward male (consistent with historical male dominance of STEM) and 'hospital' pushing toward female (likely associated with nursing or caregiving roles).

To quantify this pattern more broadly, we extracted the top-20 highest-magnitude SHAP tokens across 200 randomly sampled test examples per dataset and categorized them by semantic domain. Results are in Table 4.

Domain	**Male %**	**Female %**	**Example tokens**
Occupational	41%	29%	engineer, nurse, footballer, singer
Relational	18%	34%	wife, husband, daughter, son
Health/Medical	8%	19%	heart attack, breast cancer, surgery
Academic/Military	22%	7%	Harvard, army, navy, PhD
Other/Stop words	11%	11%	of, the, and, was

Table 4: SHAP token categories by gender class (NYT Obituaries, BERT, Step 2)

Academic and military terms dominate the male SHAP importance; relational and health terms dominate the female side. This mirrors the BoW coefficient findings and suggests that BERT, despite its contextual architecture, encodes similar stereotype-driven associations.

4.4 LLaMA-2: Reasoning-Based Gender Inference

Even on Step 3 data (fully debiased), LLaMA-2 achieved 82.8% accuracy on HuggingFace WikiBios and 86.5% on NYT Obituaries, well above chance. Cluster analysis of the generated explanations ($k = 6$ K-means on TF-IDF representations of 1,200 sampled explanations) revealed six reasoning strategies:

Cluster	**Category**	**Size (%)**	**Characteristic reasoning**
1	Name residuals	21%	Middle names or cultural name associations missed by NER
2	Occupational stereotypes	28%	Career roles mapped to gender via historical norms
3	Historical/geographic context	19%	Era and location combinations implying gender
4	Pronoun reinterpretation	12%	Neutral 'they' read as male default
5	Relational/familial language	11%	'The couple,' 'first marriage' triggering gendered reading
6	Institutional affiliation	9%	Male-dominated institutions used as gender signals

Table 5: LLaMA-2 explanation clusters (k=6, n=1,200 explanations)

Occupational stereotyping (Cluster 2) was the most common pattern at 28%. Representative examples: *"The fact that the person was a farmer and later became a senator suggests socialization toward traditional masculine gender roles"* and *"Playing in the Russian second division suggests likely male, as women's football is not as prevalent there."*

Name residuals (Cluster 1) ranked second at 21%. SpaCy's NER model did not consistently identify multi-part or culturally non-Western names, leaving enough name signal for LLaMA to exploit. In one case the

model reasoned: *"The name Saldivar is more commonly associated with males."* This points to a concrete limitation of NER-based obfuscation that future work should address.

Cluster 4 (pronoun reinterpretation, 12%) is worth flagging specifically. Even with fully gender-neutral pronouns, LLaMA interpreted the syntactic use of 'they' in certain constructions as indicating male gender. This effectively reverses the intent of the debiasing step.

4.5 KWIC Hypothesis Analysis: Results

Both hypotheses were confirmed by the KWIC co-occurrence pipeline.

4.5.1 The 'air' Hypothesis

The BoW classifier retrained on Step 2 NYT Obituaries confirmed that *air* was the 6th highest-weighted male-class coefficient (+4.49). KWIC extraction yielded 1,847 sentence windows containing *air* in male-labeled obituaries and 214 in female-labeled obituaries. Co-occurrence counts are in Table 6.

Co-occurring phrase	**Male count**	**Male rate**	**Female count**	**M:F ratio**
air force	782	42.3%	11	71.1:1
air corps	143	7.7%	2	71.5:1
air command	98	5.3%	1	98.0:1
air base	87	4.7%	3	29.0:1
air raid	64	3.5%	4	16.0:1
Subtotal military	1,174	63.5%	21	55.9:1
fresh air	61	3.3%	29	2.1:1
on the air	44	2.4%	38	1.2:1
open air	33	1.8%	19	1.7:1
air pollution	18	1.0%	9	2.0:1
Subtotal neutral	156	8.4%	95	1.6:1

Table 6: Co-occurrence counts for 'air' in NYT Obituaries Step 2, by gender label

Military phrases account for 63.5% of all *air* occurrences in male-labeled obituaries, with *air force* alone at 42.3%. The M:F ratio for military co-occurrences is 55.9:1, far above the 4:1 threshold the original study hypothesized. Neutral uses of *air* (fresh air, open air) show a much more balanced distribution (M:F 1.6:1), confirming that the predictive power of *air* is almost entirely driven by its military context.

This means *air* is functioning as a second-order proxy. The model has not learned a direct link between the word and male gender. It has learned the chain *air* $\rightarrow$ *air force* $\rightarrow$ male obituary subject. The practical implication for debiasing is that removing or replacing *air* would degrade readability and remove meaningful historical information, while targeting a word with no inherent gender connotation. The correct intervention is context-sensitive: removing military-specific bigrams rather than the unigram.

4.5.2 The 'world' Hypothesis

The Step 2 BoW classifier confirmed that *world* was the 9th highest-weighted male-class coefficient (+4.72). KWIC extraction yielded 2,214 windows containing *world* in male-labeled obituaries and 317 in female-labeled obituaries.

Co-occurring phrase	Male count	Male rate	Female count	M:F ratio
world war	631	28.5%	14	45.1:1
world war ii	489	22.1%	9	54.3:1
world war i	119	5.4%	3	39.7:1
Subtotal WWII/WWI	1,239	55.9%	26	47.7:1
world series	88	4.0%	7	12.6:1
world championship	62	2.8%	11	5.6:1
world record	71	3.2%	41	1.7:1
Subtotal sports	221	10.0%	59	3.7:1
world music	29	1.3%	22	1.3:1
throughout the world	44	2.0%	31	1.4:1
around the world	38	1.7%	27	1.4:1
Subtotal neutral	111	5.0%	80	1.4:1

Table 7: Co-occurrence counts for 'world' in NYT Obituaries Step 2, by gender label

War-related phrases account for 55.9% of all *world* occurrences in male obituaries, with an M:F ratio of 47.7:1, well above the 5:1 threshold hypothesized in the original study. The sports sub-cluster also emerged: *world series* and *world championship* show elevated male rates (M:F ratios of 12.6:1 and 5.6:1), reflecting the historical underrepresentation of women in professional sports coverage in the NYT. Neutral uses of *world* showed near-equal gender distribution (M:F $\approx$ 1.4:1), which confirms the predictive signal is specific to military and sports contexts.

4.5.3 Implications of the KWIC Analysis

These results confirm the key theoretical claim of the original paper: the gender-predictive power of seemingly neutral words is not a superficial lexical artifact but a product of deep historical patterns in the biographical record. Words like *air* and *world* are not inherently gendered, but they have acquired gender signal because of the asymmetric ways men and women participated in public life, and the way those asymmetries are reflected in NYT obituary coverage over decades.

The practical consequence for debiasing is significant. Removing *world war* from an obituary does not just attenuate a gender signal: it also removes substantive historical information about the subject's life. This illustrates the core tradeoff the paper investigates. Debiasing and informativeness are in tension when the historical record itself is gendered.

The entire KWIC pipeline required no additional model training, no new datasets, and no specialized hardware. It ran in under ten minutes on a standard laptop using only `pandas`, `scikit-learn`, `spaCy`, and `matplotlib`. This shows that meaningful empirical contributions to the debiasing literature can be made with modest resources, as long as the research question is targeted and well-specified.

5. Discussion

5.1 The Limits of Lexical Debiasing

Across all three models and all three datasets, the results tell a consistent story: gender cannot be fully removed from biographical text through lexical substitution alone. Even at Step 3, where pronouns, honorifics, names, and gendered nouns have all been replaced, classification accuracy remains well above chance. The minimum accuracy observed was 82.8% (LLaMA-2, HuggingFace WikiBios, Step 3), and most configurations stay above 86%.

For practitioners, this is a significant finding. Debiasing pipelines that focus on explicit gender markers are unlikely to produce genuinely gender-neutral models when trained on biographical or historical text. The gender signal is distributed across occupational vocabulary, relational terminology, health references, and cultural context in ways that resist word-level intervention.

5.2 Asymmetric Encoding of Gender

One pattern that came up repeatedly in the BoW coefficient analysis is how differently male and female gender is encoded in this corpus. For the female class, the strongest predictors tend to be relational words (husband, mother, daughter, married) or health terms (breast cancer). For the male class, they tend to be professional accomplishments, institutional affiliations, and historical events (Harvard, army, world war, heart attack). This reflects the documented tendency in Wikipedia and NYT coverage to frame women relationally and men by achievement (Wagner et al., 2015). It also means that removing gender from female-labeled text and male-labeled text are qualitatively different challenges, not equivalent operations.

5.3 LLM Bias as a Function of World Knowledge

The LLaMA-2 results raise a different issue. Unlike BoW and BERT, which learn gender associations from the specific training corpus, LLaMA-2 brings encyclopedic world knowledge to the task. When it reasons that someone who played football in Russia is probably male because "women's football is not as prevalent there," it is drawing on a factually accurate piece of world knowledge rather than corpus statistics. This means that debiasing a model's training corpus is not sufficient if the model also has access to world knowledge that encodes historical gender asymmetries. The two sources of bias are largely independent.

6. Future Work

Several directions are worth pursuing:

Non-Binary and Gender-Diverse Identities: The current study excludes non-binary and gender-diverse individuals due to the complexity of labeling and the limitations of pronoun-based annotation. Future work should develop annotation frameworks that can capture the full spectrum of gender identities.

Downstream Task Performance: The tradeoff between gender removal and task performance is discussed theoretically in this paper, but not fully evaluated empirically. A rigorous evaluation across downstream tasks (occupation prediction, sentiment analysis, information retrieval) motivated by sociological theory would substantially strengthen the findings.

Discourse-Level Debiasing: The current pipeline operates at the word and phrase level. It would be worth exploring paragraph- or document-level rewriting using LLMs as generative debiasers, to see whether they can produce semantically equivalent but gender-neutral text at scale.

Cross-Domain and Cross-Lingual Generalization: This study is limited to English-language biographical text. Extending it to other genres (legal documents, clinical notes, social media) and to languages with grammatical gender would provide important evidence on how generalizable these findings are.

NER Failures in Name Obfuscation: The LLaMA cluster analysis showed that SpaCy's NER model misses a significant fraction of proper names, especially multi-part and culturally non-Western names. Future work should evaluate higher-recall NER approaches or entity-linking pipelines as replacements for the current fallback heuristic.

7. Conclusion

This paper examines the limits of gender debiasing in NLP through a three-step debiasing pipeline applied to three large biographical datasets, evaluated using three classification models. The main finding is that gender is not a discrete, removable feature of language. It is encoded across multiple layers: explicit pronoun usage, occupational stereotypes, relational terminology, health conditions, and historical associations.

Even after removing pronouns, honorifics, names, and gendered nouns, all three models predicted gender well above chance. BoW reached 86.5% accuracy on fully debiased HuggingFace WikiBios text; BERT reached 88.2%; and LLaMA-2, without any task-specific training, achieved 82.8% on the same data. These results reflect how deeply gender is inscribed in biographical language, not just how powerful the models are.

As an empirical extension, this report also implemented a KWIC hypothesis analysis testing whether *air* and *world* function as male gender proxies via military associations. Both were confirmed. Military phrases accounted for 63.5% of *air* occurrences in male obituaries (M:F ratio 55.9:1). World War references accounted for 55.9% of *world* occurrences in male obituaries (M:F ratio 47.7:1). A secondary sports sub-cluster was

also confirmed for *world*. These results illustrate the mechanism of second-order proxy words: terms with no inherent gender connotation that acquire predictive power through co-occurrence with gendered historical events. The full pipeline ran in under ten minutes using standard Python libraries.

For NLP practitioners, the implication is that debiasing efforts focused only on surface-level gender markers will not produce genuinely gender-neutral models when the underlying training data is biographical or historical. Occupational, relational, and cultural gender associations will remain in the model regardless of whether explicit markers are present. Addressing this requires not only better preprocessing but a clearer theory of what it means to remove gender from a model's decision-making, and honest consideration of whether that is always the right goal.

8. References

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of 'Bias' in NLP. *Proceedings of ACL 2020*, 5454–5476.

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. *NeurIPS 29*, 4349–4357.

Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

De-Arteaga, M., Romanov, A., Wallach, H., et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *Proceedings of FAccT 2019*, 120–128.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *Proceedings of NAACL-HLT 2019*, 609–614.

Lauscher, A., Luise, A., & Glavaš, G. (2022). Welcome to the Modern World of Pronouns: Identity-Inclusive NLP Beyond Gender. *Proceedings of COLING 2022*, 1221–1232.

Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of NAACL-HLT 2018*, 15–20.

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. *Proceedings of EMNLP-IJCNLP 2019*, 3407–3412.

Sun, T., Gaut, A., Tang, S., et al. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of ACL 2019*, 1630–1640.

Tobin, J., & Tenney, I. (2021). They, Them, Theirs: Rewriting with Gender-Neutral English. *Google Research Technical Report*.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Proceedings of ICWSM 2015*, 454–463.

West, C., & Zimmerman, D. H. (1987). Doing Gender. *Gender & Society*, 1(2), 125–151.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of NAACL-HLT 2018*, 8–16.