Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs

Anonymous ACL submission

Abstract

Gender bias is not only prevalent in Large Language Models (LLMs) and their training data. It is also firmly ingrained into the structural aspects of language itself. In this work we focus on gender-exclusive affixes in English, such as in *showgirl* or *man-cave*, which can perpetuate gender stereotypes and exclude association with non-binary genders. We use an LLM training dataset to extract a catalogue of 692 gender-exclusive words alongside gender-neutral variants. Our catalogue can aid in assessing gender skews in a given training corpus. We also use it to develop a fine-tuning dataset, the *Tiny Heap*, in which we replace gender-exclusive with gender-inclusive wording. We fine-tune three LLMs, observing an overall reduction in gender-stereotyping tendencies across the models. Our approach provides a practical method for enhancing gender inclusivity in LLM training data and contributes to the inclusion of queer-feminist linguistic activism in bias mitigation research in NLP.

1 Introduction

Large-language models have become ubiquitous in NLP due to their impressive capabilities in a variety of tasks. However, they also come with risks attached, because social biases contained in the training data are incorporated into models (Bender et al., 2021). Gender bias in LLMs is well-documented, demonstrating, among others, a reliance on gender stereotypes and lower performance for non-binary genders in pronoun resolution systems (Cao and Daumé, 2021), a higher likelihood of models to generate harmful and misgendering language if queer individuals are mentioned (Nozza et al., 2022; Ovalle et al., 2023), as well as "moderate to conservative" views of the category of gender itself as indicated by a prevalence of a binary model of gender (Watson et al., 2023).

However, harmful behavior such as a preference for male terminology, reliance on gender stereotypes and erasure of non-binary gender identities is not just a feature of trained language models, it is a feature of language itself. In English, linguistic constructions such as the use of *man* to mean all humans, the indication of only women's marital status through address terms (*Miss*, *Mrs*, *Ms*), or the marking of deviation from gendered norms (*male nurse*, *girl boss*) have a long history of reinforcing traditional gender roles and and the concept of male gender as the default (Mills, 2012). While these sexist and gender-exclusive constructions have been discouraged in official style guides (American Psychological Association, 2020) and their use has been declining (Baker, 2010), the slow and concurrent nature of language change, the large size of LLM training data, as well as distributional gender bias, favours the proliferation and reinforcement of traditional views of gender through LLMs.

As a way of mitigating gender bias, researchers have recently explored ways of using *gender-inclusive language*, which focuses on eliminating stereotyped and gender-specific associations, to fine-tune LLMs (Thakur et al., 2023). Data interventions with gender-inclusive text aim to reduce the frequency of mentions of binary gender terms in places where gender is irrelevant (for example, a *chairman* and *chairwoman* do the same job) and thereby allow for association of a term with all genders (*chairperson*), which is then transferred to the LLM during fine-tuning. However, the replacement of sexist and gender-exclusive terminology often relies on limited lists of gender-neutral terms.

In this research, we exploit structural elements of sexist language to expand the coverage of gender-neutral replacements. We extract nouns with gender-exclusive affixes from a common LLM training corpus, OpenWebText2 (Gao et al., 2020), demonstrating clear androcentric tendencies within the corpus and subsequently expand the list of extracted nouns with gender-neutral variants. We present a catalogue of 692 term pairs with gender-

exclusive suffixes and prefixes, which can be used to assess gender skew within LLM training corpora, as well as to replace gender-exclusive with gender-inclusive terminology. In the second part of our study, we create a small, multi-domain fine-tuning corpus, using our catalogue to replace gender-exclusive with gender-neutral words. We also use the NeuTralRewriter (Vanmassenhove et al., 2021) to replace gendered pronouns (*he*, *she*, *himself* etc.) with singular *they*. We use this corpus to fine-tune three different LLMs and demonstrate an overall tendency of reduction in gender-stereotyping exhibited by the models.

2 Related Work

Large Language Models (LLMs) have been shown to encode a variety of social biases contained in their training data (Gupta et al., 2023; Salinas et al., 2023), among them gender bias (Stanczak and Augenstein, 2021). Due to the current prevalence of transfer learning in NLP, in which a pre-trained model is fine-tuned with task-specific data, transfer learning has recently also been adapted by works that aimed to reduce gender bias in LLMs (Lauscher et al., 2021; Ghanbarzadeh et al., 2023). In this approach, an LLM is fine-tuned with data that has undergone interventions to increase gender fairness. Supporting this approach, Steed et al. (2022) found that biases in fine-tuning data have a greater influence on downstream model behavior than biases in the pre-training data. Previous interventions to fine-tuning data include Counterfactual Data Augmentation (CDA), in which masculine and feminine pronouns and gendered nouns are swapped for the respective other (Ghanbarzadeh et al., 2023; Vashishtha et al., 2023; Fatemi et al., 2023). Another intervention replaces gendered words for gender-neutral words (*fire fighter* for *fireman*) or phrases containing both masculine and feminine genders (*he and she* for *he*; Thakur et al., 2023). This kind of intervention is not new: it rests upon a longstanding tradition of research and advocacy the field of feminist linguistics, which has been promoting changes in the lexicon to reduce gender stereotyping and masculine-default language since the 1970s (Kramer, 2016; Mills, 2012; Lakoff, 1973). More recently these changes to the language, which are also called *feminist language reform*, have incorporated ways of adapting language to include non-binary and trans gender identities, such as the third person singular (neo)pronouns (*they*, *xe*, *ze*, etc.). The usage and possible modelling of this extended lexicon of pronouns within the context of NLP was analyzed by Lauscher et al. (2022). Lund et al. (2023) also showed that training on data containing singular *they* can reduce gender bias in grammatical error correction. Furthermore, Vanmassenhove et al. (2021) and Sun et al. (2021) developed rule-based and neutral machine translation-based models to modify English text to render it gender-neutral. Vanmassenhove et al.'s (2021) NeuTralRewriter replaces gendered pronouns with singular *they* and a list of gendered nouns with neutral variants. However, while the amount of NLP research incorporating and exploring strategies of feminist language reform has grown, the queer-feminist linguistic research it is based on is, with some exceptions (Devinney et al., 2022; Piergentili et al., 2023a; Seaborn et al., 2023), rarely acknowledged and even less often informs the research itself.

Contributions This paper approaches gender-inclusive language from a linguistic vantage point. We exploit structural elements of English that relate to gender discrimination and exclusion in order to expand lists of words that are unnecessarily gendered and provide gender-neutral variants. Our method produces a catalogue of roughly triple the size of previously used word lists. Furthermore, we use our list, as well as the NeuTralRewriter (Vanmassenhove et al., 2021) to produce gender-neutral fine-tuning data. Fine-tuning with these data results in a gender bias reduction within LLMs that aligns with previous findings. We release our code, gender-neutral word catalogue, and fine-tuning datasets to the public upon publication.

3 Method

Gender bias in the English language is reflected in features such as masculine generics and is captured in datasets through, for examplke, skewed distributions of pronouns and profession words in the same context. However, it is also contained in structural elements of the language itself, such as gender-marking affixes. The most frequent are suffixes such as *-man* in *spokesman*, but gender can also be marked with a prefix, such as in *man-bun* or *girlboss*. Words marked with masculine suffixes have traditionally been used in a generic sense (e.g. *Madam Chairman*), however, with the emergence of feminist language reform, style guides have advised against their use (Piergentili et al., 2023b).

affix		round 1	round 2	round 3
prefix	*woman-*	10	4	4
	girl-	30	13	10
	man-	87	47	49
	boy-	59	11	7
	total	**186**	**75**	**70**
suffix	*-woman*	42	37	35
	-girl	47	24	14
	-man	271	238	180
	-boy	62	41	24
	-womanship	2	2	2
	-manship	53	32	30
	total	**477**	**342**	**285**
TOTAL		**663**	**417**	**355**
PERCENT		**100%**	**62.9%**	**53.54%**

Table 1: Number of singular nouns with gender-marking affixes extracted from subsection of OpenWebText2 corpus throughout verification process.

In English, the most common replacement strategy for gendered generics is neutralisation (*chairperson*), because all gender identities, not just male and female, can be referred to by gender-neutral nouns. In NLP, research using gender-neutral language in the context of English LLMs has mainly relied on lists of common gender-neutral replacements (Vanmassenhove et al., 2021; Thakur et al., 2023), without taking structural processes such as affixation into account in order to broaden the coverage of these lists.

In this section we first outline the process of extracting unnecessarily gendered words based on gender-marking affixes (§3.1). We then describe the gender-neutralizing interventions to our fine-tuning data (§3.2) as well as the models (§3.3) and bias measurements used (§3.4).

3.1 Word Catalogue

We extracted words with the suffixes *-man*, *-manship*, *-woman*, *-womanship*, *-boy*, *-girl* and words with the prefixes *man-*[1], *woman-*, *boy-* and *girl-*. We used a 200 million token random subsection of the OpenWebText2 corpus (Gao et al., 2020) for extraction. The words were extracted using regular expressions within Python. Besides fitting one of the ten affix-patterns, we additionally filtered the words to only include English singular nouns. We only filtered for singular nouns to reduce the amount of redundant extractions, and to simplify the dictionary verification later on. Plurals for all verified words were added after the third round of verification.

The **first round** of verification of extracted affixed terms generally followed a human-in-the-loop approach, meaning that after 20 files, each 1MB in size, the extracted words were manually checked for validity. This eliminated a variety of false positives such as words in which affixes did not denote gender (*german*, *ramen*), spelling errors (*camerman*, *sopkesman*), surnames (*zimmerman*), and other word creations (*heythereman*, *mrfredman*). In total, 663 words were extracted in the first round (ref. Table 1).

After extraction, the terms were verified in the **second round** using the API of the BabelNet encyclopedic dictionary (Navigli and Ponzetto, 2012). BabelNet was chosen due to its broad coverage of lexical resources; its search engine combines entries from WordNet, Wikidata and Wikipedia among others. Terms that did not return an entry in BabelNet were disregarded in order to eliminate less established terms, slang and sexually charged terminology. If a term contained a dash, such as in *man-bun*, but could not be found in BabelNet, we also searched for the term with a space instead of the dash to not disregard terms due to spelling differences. Table 2 shows the top ten words containing the four simple gender-marking suffixes and their frequency. The highest frequent words with gendered prefixes, and words with *-wo/manship* suffixes are shown in Table 6 and 7 in the Appendix, respectively.

Following the BabelNet verification, words were manually filtered in the **third round** to exclude words not related to gender (e.g. *boycott*, *boyne*), and proper names such as surnames or words related to pop culture (*batgirl*, *rainman*). Furthermore, terms that occurred with a feminine suffix (*noblewoman*) but did not have a masculine equivalent (*nobleman*) were added as their masculine variant to the list, because we treat gender-marking suffixes as exchangeable to mark a different gender. The third round left 353 singular affixed nouns, which is roughly half of the initially extracted 663 nouns.

[1]Words with *man-* prefixes were only included if they also had the dash (-) following *man*, because otherwise the false positive rate (*manager*, *mandate*, etc.) would have been too high.

-man	#	*-woman*	#	*-boy*	#	*-girl*	#
spokesman	18072	spokeswoman	5731	cowboy	523	showgirl	18
congressman	1702	congresswoman	163	playboy	163	fangirl	14
businessman	1588	businesswoman	101	fanboy	159	cowgirl	13
policeman	1155	policewoman	46	tomboy	55	supergirl	7
freshman	412	anchorwoman	19	busboy	37	batgirl	3
fisherman	376	forewoman	15	plowboy	31	dreamgirl	2
cameraman	375	gentlewoman	7	paperboy	28	bargirl	2
statesman	293	madwoman	6	homeboy	26	babygirl	1
defenseman	233	spokewoman	6	doughboy	6	tomgirl	1
madman	183	frontierswoman	6	sackboy	5	transgirl	1

Table 2: Top 10 words with gender-denoting suffixes after second round of verification and their frequencies within 200-million token subset of OpenWebText2

3.1.1 Gender-neutral variants

We then proceeded to add gender-neutral variants for all extracted words with gender-marking affixes. A single variant was added for all items in the list to simplify the replacement process.

Suffixes Some gender-marking suffix could simply be exchanged for one that is gender neutral, such as in the common neutralisation of *chairman/-woman* to *chairperson*. However, this simple replacement does not always work. For example, some frequent terms already have gender-neutral replacements such as *fire fighter* for *fireman* or *police officer* for *policeman*. In these cases, **fireperson* or **policeperson* would be ungrammatical[2]. A similar case can be made for less frequent words for which more elegant solutions are available than simply replacing *-man*/*-woman* with *-person*. One approach is to find more fitting suffixes or compound nouns, such as in the neutralisation of *crewman* with *crew member*. Another approach is to replace a word with a gender-neutral synonym, such as in the replacement of *hitman* with *assassin*. A third approach applies to words containing a verb as their root, such as the word *huntsman*, which has the root *hunt*. Here, the word can be replaced by a nominalisation: *hunter*. The final gender-neutral variants were agreed upon by the researchers.

Prefixes In the case of words with gender-marking prefixes, gender-neutral variants can be constructed by removing the prefix. For example, the word *man-crush* can be neutralised to *crush*.

Once the list of singular word pairs was fixed, the plural version of every word-pair was added to the final list. The plurals were obtained using the `inflect` library in Python (version 7.0.0). After adding plurals, we performed one last round of manual verification to ensure all plurals were formed correctly. The final list contains 692 term pairs. For comparison, Vanmassenhove et al. (2021) used a list of 91 term pairs. A sample of our final list can be found in Table 8 in the Appendix.

[2]As per linguistic convention we mark ungrammatical terms with a leading asterisk (*).

3.2 Fine-Tuning Data

		Heap	**Small Heap**	**Tiny Heap**
dataset	**original weight**	**# tokens**		
OWT2	50%	125M	25M	162k
CC-News	30%	75M	15M	240k
English Wikipedia	20%	50M	10M	112k
TOTAL	100%	250M	50M	514k

Table 3: Composition of Heap corpora; OWT2 = OpenWebText2, CC-News = Common Crawl News

To create a fine-tuning corpus with gender-neutral interventions, we first assembled a base corpus, which needed to have several features: (1) The configuration should be similar to current LLM pre-training data, meaning that it should contain a diverse set of sources. However, we excluded data that was too domain-specific, such as code and scientific publications, because we wanted to demonstrate methodology for general-purpose English. In the same line of reasoning, (2) the corpus should only contain English data, because the focus of this work is English, and the NeuTral Rewriter (Vanmassenhove et al., 2021), which replaces gendered pronouns with singular *they* does also only exist for

original sentence	He told *newsmen* at the scene that unknown criminals vandalised MD metres and armoured cables of the transformer.
after word replacement	He told *reporters* at the scene that unknown criminals vandalised MD metres and armoured cables of the transformer.
after rewriting and word replacement	*They* told *reporters* at the scene that unknown criminals vandalised MD metres and armoured cables of the transformer.

Table 4: Example of sentences in fine-tuning data at different stages of gender-neutral rewriting and replacement

English. (3) Finally, since we do not aim to worsen the performance of the LLM through fine-tuning, the corpus should only include high-quality text.

The final composition of our base corpus was inspired by the composition of GPT-3's training data (Brown et al., 2020) as well as The Pile corpus (Gao et al., 2020) and is shown in Table 3. Our original download has a size of 250 million tokens, which is approximately 1.5 GB of data. Since this is substantially smaller than The Pile (825GB), we are calling our dataset ***The Heap***. The dataset was downloaded using the Huggingface `datasets` library (version 1.18.3; Wolf et al., 2020) and tokenised with the `stanza` library (version 1.7.0; Qi et al., 2020).

The fine-tuning data were adjusted for gender-neutral wording in two rounds: first, we used our own list of extracted affixed words to replace sexist with gender-inclusive terms. Words that were part of named entities were not replaced. Second, feminine and masculine singular pronouns (*he*, *she*, *himself*, etc.) were re-written into the respective variants of singular *they* using Vanmassenhove et al.'s (2021) NeuTralRewriter. Table 4 illustrates this re-writing process and provides an example sentence within the different variants of the corpus: normal, with replacements, and rewritten with replacements.

After downloading this dataset, however, we realised that good fine-tuning results can be achieved with considerably less data (Thakur et al., 2023; Zhou et al., 2023), and fine-tuning a model with the entire corpus would have gone beyond computational resources available to us. Therefore, we first reduced the *Heap* corpus to a smaller dataset of 50 million tokens (the *Small Heap*, ~300MB), and finally only extracted lines containing word replacements. The composition of the final dataset, *Tiny Heap*, can be seen in Table 3.

3.3 Models and Fine-tuning

We ran our experiments on three models: GPT-2 (Radford et al., 2019), RoBERTa-large (Liu et al., 2019) and PHI-1.5 (Li et al., 2023). These models were chosen because they (1) cover both causal and masked language modelling architectures, (2) feature in previous research (GPT-2 and RoBERTa), and (3) have small parameter sizes meaning they require less resources to fine-tune. Microsoft's PHI-1.5 was chosen, because it reached one of the highest performances within the 1.5 billion parameter category of pre-trained models in Huggingface's OpenLeaderboard[3] at the time we conducted our experiments.

The models were fine-tuned for each one and three epochs (batch size 2) on an NVIDIA A100-SXM4-40GB GPU on Google Colaboratory, using 30 GPU hours in total for all models. The two fine-tuning datasets used were *Tiny Heap* with gender-neutral replacements (`tiny-heap-rep`) and gender-neutral replacements and rewriting (`tiny-heap-rep-neutral`). The learning rate was set to $2e{-}5$ with a weight decay of 0.01. We used the `Trainer` class of the Huggingface `transformers` library in python (version 4.38.0.dev0; Wolf et al., 2020) and kept all other hyperparameters at their default values.

3.4 Bias Evaluation Metrics

We chose three established metrics for quantifying bias. CrowS-Pairs (Nangia et al., 2020) and RedditBias (Barikeri et al., 2021) were chosen because they are not based on artificial templates but are crowdsourced and extracted from naturally occurring data, respectively. The third benchmark, HONEST (Nozza et al., 2021, 2022), was chosen as a extrinsic metric, because it relies on prompt completion. In addition to measuring bias along the binary male-female axis, both RedditBias and HONEST support gender bias evaluation in rela-

[3] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

tion to LGBTQ+ (Lesbian, Gay, Binary, Trans and Queer or Questioning) terminology.

CrowS-Pairs (Nangia et al., 2020) is a benchmark comprised of crowdsourced minimal sentence pairs differing in words that are related to a variety of social categories, such as race, ability and gender. Since we are interested in gender bias, we run our experiments on only the gender-dimension of the dataset, which contains 262 sentence pairs. We use Meade et al.'s (2022) implementation of the sentence scoring, which measures the likelihood of the changed, instead of the unchanged, tokens within a sentence. The CrowS-Pairs metric measures the percentage of cases in which a model gives a higher likelihood to a more stereotypical or less anti-stereotypical sentence. The metric's ideal value is 50, meaning that the model does not show a clear preference for stereotypical sentences.

RedditBias (Barikeri et al., 2021) also contains minimal sentence pairs expressing stereotypes for different demographic dimensions: *religion*, *race*, *gender* and *queerness*. Due to our focus on gender, we only calculate scores for the gender and queerness dimensions. The sentences in RedditBias were extracted from the Reddit social network forum and contain both a target term identifying a social demographic as well as an attribute term that expresses a (negative) stereotype related to the group. Minimal pairs differ either in the target or attribute term. Stereotyping in a model is quantified through calculating the perplexity of the model for the sentence pairs and performing the student's t-test on the perplexity pairs. Negative values of t indicate stereotypical bias in the model while p indicates statistical significance of the perplexity differences.

HONEST differs from the first two measures in that it does not measure gender stereotyping but the presence of hurtful language in LLM sentence completions. The original HONEST benchmark consists of prompts containing binary masculine and feminine terms (Nozza et al., 2021). This was later extended with prompts containing LGBTQ+ terms (Nozza et al., 2022). The HONEST prompts were created for six different languages, however, since our work focuses on English specifically, we only use the English portion of the dataset. HONEST uses the HurtLex lexicon of harmful language (Bassignana et al., 2018) to measure the hurtfulness of words contained sentence completions. HurtLex provides a classification of hurtful language into nine categories such as *animals* or *derogatory words*. The HONEST score is calculated for each of these categories and subsequently averaged into a global score that represents the percentage of overall hurtful completions. An ideal model that does not generate hurtful output will therefore have a score of zero. For our experiments, we used $k = 20$ random sentence completions for GPT-2 and RoBERTa, keeping in line with the original paper, and $k = 5$ completions for PHI-1.5 in order to shorten the runs.

4 Results and Discussion

4.1 Gender-marking affixes

Table 1 illustrates the number of affixed word extractions for three rounds of verification. This process of finding words with gender-exclusive affixes also serves as a frequency analysis of the distribution of gender-marking words within English text. Overall, it can be clearly seen in Table 1 that gender-marking through suffixation is more common than prefixation. Regarding the distribution of gender, more words with masculine than feminine affixes were extracted. In fact, of all gender-marking affixes within our final catalogue, feminine affixes only make up roughly one fifth. This skewed distribution demonstrates a tendency within English text to over-represent masculine gender through, for example, masculine default forms. The over-representation of masculine gender is one of the origins of gender bias towards masculine forms in LLMs. Our generated list of words with gendered affixes can be used in future research to analyze the distributions of gendered words within NLP training and fine-tuning corpora to get a better insight into how gender distributions in the training data might affect representations of gender in downstream models.

4.2 Fine-tuning

Table 5 shows how fine-tuning impacted three different bias metrics for the three LLMs we tested. As can be seen in Table 5, each model was fine-tuned for one and three epochs, using (1) fine-tuning data with gender-exclusive replaced by gender-neutral wording using our own gender-neutral catalogue (cf. Section 3.1) and (2) gender-neutral rewriting (Vanmassenhove et al., 2021) in addition to the word replacement.

For **RedditBias** (Barikeri et al., 2021), we re-

model	epochs	FT	RedditBias		CrowsPairs			HONEST	
			t_{gender}	$t_{queerness}$	metric	stereo	anti-st.	binary	queer
GPT-2	0	baseline	-1.28	-1.65	56.87	53.46	62.14	0.140	0.146
	1	replacement	-2.01*	**-0.39**	54.96	51.57	60.19	**0.101**	**0.112**
		rep+neutral	**-0.77**	-0.69	54.96	58.94	**49.51**	0.107	0.119
	3	replacement	-1.54	-0.81	54.58	**49.69**	62.14	0.110	0.120
		rep+neutral	-1.54	-1.09	**54.2**	56.60	**50.49**	0.124	0.126
PHI-1.5	0	baseline	**-1.83**	**-0.34**	55.73	62.26	45.63	**0.079**	0.142
	1	replacement	-2.06*	-2.32*	51.15	**51.57**	50.49	0.109	**0.114**
		rep+neutral	-2.26*	-2.42*	**50.76**	55.35	**43.69**	0.123	0.154
	3	replacement	-2.72*	-2.87*	51.91	53.46	49.51	0.084	0.135
		rep+neutral	-2.71*	-2.16	51.91	55.97	45.63	0.093	0.129
RoBERTa	0	baseline	-0.50	1.50	60.15	72.15	42.16	0.035	0.05
	1	replacement	-0.56	1.42	50.19	58.23	**38.24**	0.044	0.066
		rep+neutral	-2.62*	-0.06	56.32	62.26	46.06	0.040	0.054
	3	replacement	-1.61	0.47	52.87	60.38	41.18	**0.012**	**0.035**
		rep+neutral	**0.22**	**2.18***	**49.04**	**54.72**	40.20	0.028	0.041

Table 5: Gender-stereotyping (RedditBias, CrowsPairs) and hurtful language generation (HONEST) results for different interventions to fine-tuning (FT) data, divided by baseline model, one, and three epochs of fine-tuning; RedditBias results marked * significant with $p < 0.05$. `rep+neutral` = gender-neutral replacements + neutral rewriting; anti-st = anti-stereotypical setting

port the values of the t-statistic for the Student's t-test. Negative values indicate higher perplexity of the model for sentence variants mentioning female/queer target terms, which indicates stereotypical bias in the model. The results illustrated in Table 5 show **binary gender bias** for all baseline LLMs in the binary gender setting. This bias can be reduced (increasing values of t) by fine-tuning in the case of GPT-2 and RoBERTa. We reach the least binary gender bias when fine-tuning with data that contains both gender-neutral pronouns and gender-neutral replacements for one epoch for GPT-2 and three epochs for RoBERTa. Fine-tuning PHI-1.5 achieves opposite results, increasing the binary bias metric.

Measuring **queerness bias**, GPT-2 exhibits the most stereotypical bias, followed by PHI-1.5, which actually shows a low negative value of $t_{queerness}$, indicating that the model might not be as biased towards the LGBTQ+ community as GPT-2. Even further, baseline RoBERTa shows a positive value for $t_{queerness}$ (1.5). Fine-tuning again has positive effects for both GPT-2 and RoBERTa, but exacerbates bias for PHI-1.5. Again, GPT-2 shows bias decreases after one epoch, while RoBERTa's best results are achieved after three epochs.

For **CrowS-Pairs** (Nangia et al., 2020), we report the percentage of cases in which a model assigns higher likelihood to gendered target terms within a sentence expressing a stereotype ('stereotype' column in Table 5) or a lower probability to target terms in sentences expressing an anti-stereotype ('anti-st.' column in Table 5). The 'metric' column contains the overall stereotype score. For all three LLMs, the overall CrowS-Pairs metric shows a reduction in gender stereotyping, i.e. results that are lower than the baseline and approach a value of 50. This result is mostly in line or goes beyond of what Thakur et al. (2023) reported for their methods of fine-tuning with gender-inclusive text; they showed a maximum reduction of the CrowS-Pairs score of approximately 0.03 for RoBERTa-base. Our RoBERTa-large model trained for 3 epochs on data with gender-neutral pronouns and replacements shows the largest reduction (difference of 0.11) to a value even less than the ideal of 50 percent likelihood of preferring a stereotyped sentence. GPT-2 shows the best result (54.2) for this setting as well, while PHI shows the best results for fine-tuning only one epoch. Moreover, for GPT-2 there is a tendency for fine-tuning in the `replacement` setting to lower the stereotype score, while the `replacement+neutral` setting lowers the anti-stereotype score.

The **HONEST** scores contain the percentage of sentence completions for sentences containing a term referring to binary or queer gender were completed with hurtful language. The two baseline

causal LLMs GPT-2 and PHI-1.5 generate hurtful sentence completions around 15% of the time in the queer setting, while RoBERTa has a much lower starting point with only 5% hurtful completions. Table 5 shows that our method method of fine-tuning language models can be used to reduce the number of hurtful completions. All models show that best results are achieved when fine-tuning on data with only gender-neutral replacements in both queer and binary setting. However, depending on the model and the setting (binary vs. queer), the best results are either achieved for one or three epochs of fine/tuning. Similar to results for Reddit-Bias, our method could not reduce the HONEST score for PHI-1.5 in the binary setting.

Overall, our results echo Aribandi et al. (2021) who found that bias metrics within the NLP literature often do not correlate: while we could demonstrate a reduction in stereotyping as measured by CrowS-Pairs as well as a reduction in the generation of hurtful language, the RedditBias metric did not show a bias reduction for all models. Moreover, the fact that different models proved to be susceptible to bias reduction in different settings, such as level of gender-neutralisation in fine-tuning data or number of fine-tuning epochs, additionally shows that model specifications such as architecture and model size need to be taken into account when choosing a bias mitigation strategy. For instance, RoBERTa generally shows a larger bias reduction when fine-tuning for three epochs, while the best number of epochs for PHI-1.5 and GPT-2 depends on the fine-tuning data. Furthermore, we demonstrated that a newer model, PHI-1.5, which was released in 2023 (Li et al., 2023) as opposed to RoBERTa and GPT-2 in 2019 (Liu et al., 2019; Radford et al., 2019), was less susceptible to gender bias reduction through fine-tuning. However, the baseline PHI-1.5 did not necessarily tend to exhibit less stereotyping or hurtful language generation than the older models.

5 Conclusion

Gender-inclusive language has a long history of development and advocacy within the field of feminist linguistics, but it has only recently entered gender bias research in NLP. In this paper, we presented a way of semi-automatically extracting gender-exclusive nouns based on the presence of gender-marking affixes. We then extended this list with gender-neutral variants, presenting a catalogue of 692 gender-exclusive vs. -inclusive pairs, which we make available for future research.

We then performed fine-tuning experiments on three LLMs. To create a fine-tuning corpus we used our catalogue to replace gender-exclusive with gender-neutral nouns and, in an additional step, re-wrote gendered pronouns with the respective variants of singular *they*. Fine-tuning with gender-neutral data showed an overall reduction in gender stereotyping as measured by likelihood of gendered word generation in stereotyped settings, as well as a reduction in the generation of harmful language when prompted with sentences containing words related to binary gender as well as the LGBTQ+ community. However, we also showed that optimal bias reduction is dependent on model architecture and number of fine-tuning epochs, which need to be considered in deployment. We hope that our work will inspire further research into the effects of gender-inclusive terminology within large language models.

6 Limitations

This study is limited by four main factors:

Firstly, our study is **limited to English** specifically. Gender-inclusive language strategies differ depending on the language and might be complicated by aspects such as grammatical gender marking (Piergentili et al., 2023a). Therefore, while our general approach could be applied to other languages in future research, the resources we developed and utilised, i.e. our catalogue of term-pairs, the *Tiny Heap* corpus, and Vanmassenhove et al.'s (2021) NeuTral Rewriter, are monolingual.

Secondly, we performed **naive replacements** within our fine-tuning data: words that were found within our catalogue of gendered words were replaced with gender-neutral variants without regard for the sentence context. The only restriction we posed was that the word was not part of a named entity. This might have created ungrammatical or nonsensical constructions, impacting the quality of the text, which in turn could have impacted model performance. Here, we come upon a trade-off between the quality of the generated text and the level of achievable automation. This is an important consideration when scaling up our method to larger amounts of data. Additionally, words were only replaced gender-exclusive terms by a single neutral term, however for some words several variations are possible, such as *chairperson* or *chair*

for *chairman/-woman*. Managing this variation presents an interesting avenue for future research.

Thirdly, there is an increasing number of **bias metrics** to measure gender bias, and a growing body of work critiquing them (Goldfarb-Tarrant et al., 2023; Orgad and Belinkov, 2022; Goldfarb-Tarrant et al., 2021). For example, Blodgett et al. (2021) found several pitfalls in the CrowS-Pairs benchmark (Nangia et al., 2020), which we used in this paper. Therefore, we would like to point out that just because our metrics report a reduction in stereotyping in the models, this does not ensure a bias-free model but should rather be interpreted as a tendency toward decreased stereotyping. We tried to pick a diverse range of metrics that would measure gender bias without relying solely on a binary conceptualisation of gender. However, our choice of metrics was also limited by ease of use and interpretation.

Lastly, our study was limited to **language models of relatively small size**. The largest models we used (GPT-2 and PHI-1.5) each have 1.5 billion parameters, which is significantly smaller than for example the smallest (7 billion) model in the Llama suite of LLMs (Touvron et al., 2023), which reaches state-of-the-art performance using an open-source approach. We already demonstrated that the benefits of our approach differ based on the model used, which is why it would be interesting to see how fine-tuning with gender-neutral data impacts state-of-the-art models. However, our research institute does not have the resources to perform a study with models of state-of-the-art scale at the level of detail we provided here. Therefore, we leave experimentation with larger models to future research.

References

American Psychological Association. 2020. *Publication Manual of the American Psychological Association: the Official Guide to Apa Style*, 7th edition. Book, Whole. American Psychological Association.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How Reliable are Model Diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

Paul Baker. 2010. Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, 4(1):125–149.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *CEUR Workshop Proceedings*, volume 2253. Accademia University Press.

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

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Yang Trista Cao and Hal Daumé, III. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *ACM FAccT Conference 2022, Conference on Fairness, Accountability, and Transparency, Hybrid via Seoul, Soth Korea, June 21-14, 2022*.

Zahra Fatemi, Chen Xing, Wenhao Liu, and Caiming Xiong. 2023. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1249–1262, Toronto, Canada. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. ArXiv:2101.00027 [cs].

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring \textlessmask\textgreater: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J. Passonneau. 2023. Survey on Sociodemographic Bias in Natural Language Processing. ArXiv:2306.08158 [cs].

Elise Kramer. 2016. Feminist Linguistics and Linguistic Feminisms. In Ellen Lewin and Leni M. Silverstein, editors, *Mapping Feminist Anthropology in the Twenty-First Century*, page 65. Rutgers University Press.

Robin Lakoff. 1973. Language and Woman's Place. *Language in Society*, 2(1):45–80. Publisher: Cambridge University Press.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. *arXiv:2202.11923 [cs]*. ArXiv: 2202.11923.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report. ArXiv:2309.05463 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Gunnar Lund, Kostiantyn Omelianchuk, and Igor Samokhin. 2023. Gender-inclusive grammatical error correction through augmentation. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 148–162, Toronto, Canada. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Sara Mills. 2012. *Gender matters : feminist linguistic analysis*. Equinox Publishing Ltd, London.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language

Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1246–1266, New York, NY, USA. Association for Computing Machinery.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023b. Gender Neutralization for an Inclusive Machine Translation: from Theoretical Foundations to Open Challenges. ArXiv:2301.10075 [cs].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ArXiv:2003.07082 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Katie Seaborn, Shruti Chandra, and Thibault Fabre. 2023. Transcending the "Male Code": Implicit Masculine Biases in NLP Contexts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Hamburg Germany. ACM.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv:2112.14168 [cs]*. ArXiv: 2112.14168.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv:2102.06788 [cs]*. ArXiv: 2102.06788.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada. Association for Computational Linguistics.

Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023. What social attitudes about gender does BERT encode? leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6790–6809, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. ArXiv:2305.11206 [cs].

A Appendix

man-	**#**	***woman-***	**#**	***boy-***	**#**	***girl-***	**#**
man-made	200	womankind	53	boyscout	9	girllove	6
man-child	27	womanism	11	boyar	7	girlfight	5
man-eating	21	womanist	9	boyism	4	girldom	2
man-eater	15	womanly	2	boysgirls	3	girlification	2
man hater	12			boying	1	girlcott	2
man-boobs	11			boyishly	1	girlfag	1
manpower	11			boytoy	1	girlvinyl	1
man-crush	10					girlishly	1
man-ape	9					girlpower	1
manpack	8						

Table 6: Top 10 words with gender-denoting prefixes after second round of verification and their frequencies within 200-million token subset of OpenWebText2; empty rows indicate that < 10 instances were found.

-manship	**#**
chairmanship	693
craftsmanship	424
workmanship	174
sportsmanship	155
statesmanship	154
showmanship	149
marksmanship	149
gamesmanship	147
brinkmanship	119
upmanship	118
salesmanship	105
brinksmanship	73
penmanship	62
seamanship	31
swordsmanship	28
airmanship	21
draftsmanship	13
horsemanship	12
craftmanship	6
draughtsmanship	5
-womanship	**#**
stateswomanship	2
workwomanship	2

Table 7: Top 20 words with *-manship* suffix and the two words with *-womanship* suffix after second round of verification and their frequencies within 200-million token subset of OpenWebText2

suffix: -woman
ambulancewoman::emergency medical technician, anchorwoman::anchorperson, anti-woman::misogynist, antiwoman::misogynist, bogeywoman::monster, bondwoman::slave, businesswoman::businessperson, cavewoman::caveperson, charwoman::cleaner, congress-woman::congressperson, craftswoman::craftsoerson, everywoman::ordinary person, fisher-woman::fisher, forewoman::foreperson, frontierswoman::explorer, frontwoman::frontperson, gentlewoman::refined person, hitwoman::assassin, horsewoman::equestrian, madwoman::maniac
suffix: -womanship
stateswomanship::statespersonship, workwomanship::workpersonship
suffix: -girl
babygirl::baby, ballgirl::ball person, bargirl::bartender, callgirl::sex worker, cavegirl::caveperson, cowgirl::cow herder, fangirl::fan, farmgirl::farm worker, papergirl::newspaper delivery person, play-girl::player, showgirl::performer, slavegirl::slave, snowgirl::snowperson, tomgirl::timid child
suffix: -man
adman::advertiser, almsman::medical social worker, ambulanceman::emergency medical technician, anchorman::anchorperson, artilleryman::cannoneer, assemblyman::assembly member, ass-man::assperson, backwoodsman::explorer, bagman::travelling salesperson, bargeman::barge operator, barman::bartender, baseman::baseperson, batsman::batter, bellman::bellhop, binman::garbage collector, bluesman::bluesperson, boatman::boater, bogeyman::monster, bondman::slave, bondsman::slave
suffix: -manship
airmanship::aerial skill, batsmanship::batting skill, brinkmanship::extreme strategy, brinks-manship::extreme strategy, chairmanship::chairpersonship, churchmanship ::churchpersonship, craftmanship::craftpersonship, craftsmanship::craftspersonship, draftsmanship::draftspersonship, draughtsmanship::draughtspersonship, foremanship::forepersonship, gamesmanship::unsporting tactic, gentlemanship::refinedness, grantsmanship::grant acquisition expertise, handcraftsman-ship::handcraftspersonship, horsemanship::equestrian skill, journeymanship::artisanship, man-ship::courage, marksmanship::sharpshooting skill, oarsmanship::rowing skill
suffix: -boy
ballboy::ball person, batboy::bat person, bellboy::bellhop, busboy::restaurant attendant, callboy::sex worker, copyboy::junior newspaper worker, cowboy::cow herder, doughboy::foot soldier, fanboy::fan, farmboy::farm worker, femboy::effeminate person, fisherboy::young fisher, fratboy::fraternity member, headboy::student leader, homeboy::fellow member, houseboy::domestic worker, ladyboy::genderqueer person, nancyboy::nancy, newsboy::newspaper delivery person, paperboy::newspaper delivery person
prefix: woman-
womanism::feminism, womanist::feminist, womankind::humankind, womanly::feminine
prefix: girl-
girldom::feminine sphere, girlfag::woman attracted to gay men, girlfight::fight, girlfriend::partner, girlification::feminization, girliness::femininity, girlish::feminine, girlishly::childishly, girllove::love, girlpower::power
prefix: man-
man cave::sanctuary, man hater::hater, man hating::misandry, man hug::pound hug, man hunt::organized search, man magnet::attractive person, man marking::marking, man servant::servant, man up::adult up, man-ass::ass, man-bag::handbag, man-boobs::boobs, man-cave::sanctuary, man-cession::recession, man-child::child, man-crush::crush, man-eater::cannibal, man-eating::human-eating, man-friend::friend, man-hater::hater
prefix: boy-
boyband::band, boyfriend::partner, boyish::childish, boyishly::childishly, boyism::childism, boyscout::scout, boytoy::toy

Table 8: Example terms (SG) from catalogue of gender-exclusive terms and gender-inclusive replacements; each category contains 20 example pairs or the number of pairs in the catalogue if there are < 20 singular pairs