Detecting and Mitigating LGBTQIA+ Bias in Large Norwegian Language Models

Selma Kristine Bergstrand and **Björn Gambäck**
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
selmakb@stud.ntnu.no, gamback@ntnu.no

Abstract

The paper aims to detect and mitigate LGBTQIA+ bias in large language models (LLMs). As the usage of LLMs quickly increases, so does the significance of the harms they may cause due to bias. The research field of bias in LLMs has seen massive growth, but few attempts have been made to detect or mitigate other biases than gender bias, and most focus has been on English LLMs. This work shows experimentally that LLMs may cause representational harms towards LGBTQIA+ individuals when evaluated on sentence completion tasks and on a benchmark dataset constructed from stereotypes reported by the queer community of Norway, collected through a survey in order to directly involve the affected community. Furthermore, Norwegian training corpora are probed for queer bias, revealing strong associations between queer terms and anti-queer slurs, as well as words related to pedophilia. Finally, a fine-tuning-based debiasing method is applied to two Norwegian LLMs. This method does not consistently reduce bias, but shows that queer bias can be altered, laying the foundation for future debiasing approaches. By shedding light on the severe discrimination that can occur through the usage of LLMs, this paper contributes to the ongoing fight for equal rights for the LGBTQIA+ community.

1 Introduction

Different bias types, like gender and racial bias, have been uncovered in a wide range of natural language processing (NLP) applications and resources, including large language models (LLMs) (Caliskan et al., 2017; May et al., 2019; Kurita et al., 2019; Nozza et al., 2021; Zhao et al., 2018a). Left untreated, bias in LLMs may reintroduce historical biases back into society, thereby erasing progress made to achieve equality and reduce discrimination. Bender et al. (2021) describe this issue as a *value-lock*, in which technology reliant on language models may reify older, less-inclusive understandings. The research field of bias in NLP aims to prevent this by introducing bias mitigation methods (Bolukbasi et al., 2016; Zhao et al., 2018b; Lauscher et al., 2021; Felkner et al., 2023). Despite these efforts, bias in LLMs remains a current and pressing issue.

A limitation of the current state of the research field is the primary focus being placed on gender bias (Talat et al., 2022). With a few notable exceptions (Nozza et al., 2022; Felkner et al., 2023), the effects that bias in LLMs may have on the LGBTQIA+ community remain largely unknown, constituting a major research gap. As the breakthroughs of the LGBTQIA+ rights movement are quite recent in most parts of the world, it is possible that negative attitudes and harmful language directed at the queer community[1] are present in training data of LLMs. Dodge et al. (2021) showed that efforts to filter web-based text corpora often remove text written by and about the LGBTQIA+ community, strengthening the hypothesis that LGBTQIA+ bias may be present in LLMs.

Furthermore, the development of LLMs has been dominated by the English language (Bender et al., 2021; Talat et al., 2022). As a result of this Anglocentrism, research on bias in LLMs tend to define social biases based on North American point-of-views, thereby not capturing the variations in attitudes and discrimination towards marginalized communities existing in other cultures. With only five million native speakers, Norwegian is classified as a low-resource language due to the difficulty to obtain high-quality corpora of a sufficient size for LLM training (Kummervold et al., 2022). Despite this, several Norwegian-only LLMs have been developed and released, such as NorBERT (Kutuzov et al., 2021) and NB-BERT (Kummervold et al., 2021), as well as NorMistral and NorBLOOM (Pyysalo et al., 2024), while some Scandinavian

[1] This paper uses the terms *LGBTQIA+* and *queer* interchangeably.

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 351–364
August 16, 2024 ©2024 Association for Computational Linguistics

language models, such as GPT-SW3 (Ekgren et al., 2024), are also trained on Norwegian data. Even though several researchers have assessed gender bias in these models (Touileb et al., 2022; Touileb and Nozza, 2022; Samuel et al., 2023), no other biases have been detected or removed.

The devastating terrorist attack in June of 2022, specifically targeting queer people at a gay bar in Oslo (NRK, 2024) reminded Norwegians that the fight for safety, rights and equality for the LGBTQIA+ community in Norway is certainly not finished. Detecting and removing LGBTQIA+ bias from LLMs is one of the ways in which the rights of the queer community can be protected. In their strategy for safe AI usage, the Norwegian government specifically points to control processes as a way of analyzing and mitigating bias in system decisions to ensure fairness and non-discrimination (KDD, 2020). Despite this, no such processes currently exist for LGBTQIA+ bias.

This paper employs an empirical research methodology, in which four experiments are conducted to detect or mitigate LGBTQIA+ bias in five Norwegian LLMs. The first experiment involves an analysis of output generated by the models in specific contexts, while the second utilizes a crafted benchmark dataset based on a survey sent to Norwegian LGBTQIA+ organizations. The third experiment evaluates bias in Norwegian training data through an analysis of the harmfulness of words associated with LGBTQIA+ terms, and the fourth aims to reduce the detected LGBTQIA+ bias through fine-tuning the models on a LGBTQIA+-focused dataset. Combined, the experiments fulfill the goals to detect, evaluate and mitigate LGBTQIA+ bias in Norwegian large language models, and to shed light on and minimize the harm caused by such models towards the queer community.

1.1 Disclaimer

Note that this paper contains examples of toxic, stereotypical and derogatory language towards the LGBTQIA+ community. This language does not represent the views or opinions of the authors, or of the Norwegian University of Science and Technology (NTNU).

To assess bias towards different identities of the LGBTQIA+ community, a subset of all queer identities are defined and included in the experiments. These identities are **not** included because they are more important than the identities excluded, but rather due to time and data restrictions.

This paper uses *LGBTQIA+ bias* to refer to bias in large language models that adversely affect the LGBTQIA+ community; the correct term for this could arguably be *anti-LGBTQIA+ bias*. For simplicity and consistency with other bias types in the field (*e.g.*, gender bias, racial bias), the term *LGBTQIA+ bias* will be used as it is defined here.

1.2 Defining LGBTQIA+ Bias and Harms

Independent of technology, the term *discrimination* often conveys the same meaning as the definition of *bias* in the field of NLP. Amnesty International defines discrimination as differential treatment due to membership of a certain social group, often based on preconceived notions or prejudices held against said group. Such differential treatment may occur in policy, law or treatment.[2] Membership of a social group may occur based on certain protected characteristics. The Norwegian government specifies several such characteristics in the Equality and Anti-Discrimination Act of 2018, notably including gender, sexual orientation, gender identity and gender expression (KUD, 2022).

Defining the actual harms caused by bias in LLMs not only serves as a motivation for research on the topic, but also provides the framework for how bias can be evaluated. Crawford (2017) divided such harms into allocational and representational harms. *Allocational harms* concern the unfair allocation of resources among different social groups as a consequence of bias, while *representational harms* concern the unfair or discriminatory representation of certain social groups. Blodgett et al. (2020) create two categories of representational harms: stereotyping and disparate system performance. The second can further be divided into sub-categories, like derogatory and/or toxic language affecting only certain individuals, misrepresentation of queer identities and exclusionary norms erasing queer identities. Throughout this paper, the harms detected in LLMs will be categorized based on these representational harm types. Note, however, that what constitutes a representational harm is subjective — the categorization of harms in this paper is by necessity partially based on the subjective opinions of the authors, which is a limitation of this work.

This paper considers a model to contain LGBTQIA+ bias if the model causes one or more of the aforementioned harms to the LGBTQIA+ commu-

[2] www.amnesty.org/en/what-we-do/discrimination

nity, and will specifically consider representational harms rather than allocational harms. Previous definitions of gender bias in LLMs are often dependent on preferring one gender over another (as done by Caliskan et al., 2017; Bolukbasi et al., 2016; Touileb et al., 2022; Zhao et al., 2018a). However, the reason gender bias can be measured this way is due to the prevalence of gendered pronouns and words in natural language. This is not the case for LGBTQIA+-related terms. Consider, for instance, the words *heterosexual* and *cis-gender*. While these are used to describe a person who is *not* a part of the LGBTQIA+ community, they are very rarely used in a context that is independent of other LGBTQIA+ terms. This means that any bias a model holds against LGBTQIA+ individuals might also affect terms such as heterosexual and cis-gender. As a consequence, measuring the difference in LLM performance and harmfulness between two inputs, one using the term cis-gender and one using the term transgender, is likely not an accurate bias measure to assess the differences between the treatment of an actual cis-gendered and transgendered person.

Throughout this paper, bias and harms caused towards LGBTQIA+ individuals in LLMs are evaluated based on *only* LGBTQIA+ identity. However, as pointed out by Crenshaw (1989), discrimination and bias are affected by the intersection of multiple characteristics, such as sex, race, religion, etc. Fladmoe and Nadim (2019) showed this to be the case also in Norway, with individuals who are both queer and immigrants being much more likely to be targeted by hate speech than those who are only members of one of these groups. The lack of intersectionality is a significant limitation of this work.

2 Related Work

This section presents state-of-the-art methods of bias detection and mitigation, including the handful of methods proposed to evaluate LGBTQIA+ bias, as well as those concerning Norwegian LLMs specifically.

2.1 Detecting Bias in LLMs

State-of-the-art bias detection methods often belong to one of three categories: they can be embedding-based, benchmark-based or generated-text-based.

Bolukbasi et al. (2016) and Caliskan et al. (2017) both detected social bias in static word *embeddings*, using, respectively, the task of word analogy completions and the Word Embedding Association Test (WEAT). May et al. (2019) and Kurita et al. (2019) then adapted WEAT to contextual word embeddings, by using *semantic bleaching* in the form of sentence templates, showing different social biases were present there as well. Extending this approach, Nozza et al. (2021) crafted sentence templates specifically for prompting masked language models for occupations, descriptive adjectives and descriptive verbs related to gendered words (for instance: *the woman should work as a [MASK]*), naming their method the HONEST framework. To score bias, Nozza et al. calculate how many of the generated words are present in HurtLex, a multilingual lexicon of hurtful words (Bassignana et al., 2018). Later, Nozza et al. (2022) applied the same method to detect LGBTQIA+ bias in LLMs. By populating their sentence templates with queer terms, they showed that the quantity of detected harmfulness is greater for queer identities than for binary gender identities in the models tested. However, they point out that their scoring method is only able to measure word-level harmfulness, and does not capture harm that may arise at the sentence level. Further extending this framework, Touileb and Nozza (2022) translated the sentence templates to Norwegian to detect gender bias in Scandinavian LLMs, including NorBERT and NB-BERT, showing that both are more likely to generate toxic sentence completions for sentences containing female entities than male entities.

Another paradigm of bias detection methods uses the performance of LLMs on a *benchmark* dataset to assign bias scores to models. The CrowS-Pairs dataset (Nangia et al., 2020) is a widely-used benchmark created from stereotypes gathered through crowd-sourcing, containing 1508 sentence pairs, with one stereotypical and one anti-stereotypical sentence. The sentences are minimally distant; they are as semantically similar as possible. The dataset evaluates nine bias types, including gender, race and sexual orientation. Nangia et al. provide the following example of a sentence pair measuring bias against queer sexual orientations: *Danny had to force his partner (James / Judy) to watch football with him.* To quantify the bias of a model using the dataset, Nangia et al. (2020) define a scoring function. Each sentence, S, contains some unmodified tokens, $U = \{u_0, ..., u_m\}$, and some modified tokens, $M = \{m_0, ..., m_n\}$. For example, in the sentence pair shown earlier, the only modified tokens are *James* and *Judy*. The goal of the scoring function is to approximate the conditional probability that a model, given the modified tokens, will generate the unmodified tokens. Each unmodified token of a sentence is masked, one at a time, to calculate the probability that this token is generated given the rest of the sentence. The total bias score of a model is then defined as the percentage of sentence pairs in which the model is more likely to generate the stereotypical sentence than the non-stereotypical sentence. The ideal score is therefore 50, at which the probability of generating stereotypes and anti-stereotypes is equal.

While the CrowS-Pairs dataset claims to measure LGBTQIA+ bias, the dataset treats sexual orientation as a binary attribute, and does not include queer gender identities. Additionally, Blodgett et al. (2021) showed that CrowS-Pairs has several pitfalls weakening its quality — for instance, it is often not clear what stereotype a sentence pair measures, or why this is harmful. To address this, Felkner et al. (2023) created the WinoQueer dataset to measure queer bias in LLMs. In contrast to CrowS-Pairs, Felkner et al. gathered stereotypes only from members of the LGBTQIA+ community directly, by asking them what stereotypes they have experienced. This ensures the real-life relevance of all dataset entries, overcoming a significant limitation of the CrowS-Pairs dataset. WinoQueer follows the format and scoring function of CrowS-Pairs, but extend the metric of Nangia et al. by adding a separate scoring function for autoregressive language models. Felkner et al. specify that the individual sentence scores may not be comparable between the masked and autoregressive language models, but that the total bias scores are.

A third category of bias detection methods aim to analyze bias in the *generated* output of LLMs when instructed to perform a task. The previously discussed methods detect *intrinsic bias*, biases ingrained into a model through associations and assigned model probabilities. The methods of this category measure *extrinsic bias*: bias and harms that arise when a model is set to perform a task. Cheng et al. (2023) detect bias across the domains of race and gender using the concept of marked personas: by prompting an LLM to generate a description of a member of a given demographic group, the differences in outputs between *marked* and *unmarked* groups — assuming, for instance, that the unmarked group is white and male — reveal stereotyping and misrepresentation. Cheng et al. show that state-of-the-art LLMs such as GPT-3.5 and GPT-4 enforce common, stereotypical tropes for minority groups, such as the *strong black woman* stereotype. They also highlight how the descriptions of minority groups reflect *essentialism* (Rosenblum and Travis, 2003): rather than descriptions portraying the full range of humanity, the descriptions are reduced to a set of essential characteristics. This is also the case for non-binary identities, whose descriptions nearly always contained words such as *they, gender* and *identity* (Cheng et al., 2023). While the study does not consider the full range of marginalized identities, it highlights how LLM-generated content, despite not being toxic or negative in sentiment, enforces existing stereotypes in downstream tasks.

2.2 Mitigating Bias in LLMs

To remove the detected bias from LLMs, researchers have proposed several methods of debiasing. These often fall into one of three categories; augmenting the embeddings, augmenting the training data, and fine-tuning the model.

Bolukbasi et al. (2016) were the first to attempt debiasing static word embeddings, by defining a gender subspace in the vector space of all embeddings, and then placing all gender neutral words at the origin of this subspace. Removing gender association from all words might cause the modified word embedding to lose meaningful relationships though, for instance, for words related to social sciences or medicine. Zhao et al. (2018b) attempted to solve this problem by isolating the gender subspace from the rest of the word embedding by encoding all gender information into the last coordinate of each vector, so that it can easily be removed from embeddings as needed. However, the methods of Bolukbasi et al. and Zhao et al. both depend on selecting the correct gendered and neutral words, a difficult and time consuming process.

Another branch of debiasing methods aims to alter the training data of a model, in an attempt to address the root cause of bias. Two such methods are gender-swapping (Zhao et al., 2018a) and Counterfactual Data Augmentation (CDA; Lu et al., 2020). By swapping all gendered words in the training corpus of a model, such as *he* to *she* and *father* to *mother*, Zhao et al. and Lu et al. effectively double the size of their training corpora, and then retrain the models. Despite their promising results, these methods are difficult to generalize to LLMs due to the resources required to retrain such a model from scratch (as pointed out by Strubell et al., 2019 and Bender et al., 2021). Additionally, Lu et al. (2020) point out the difficulty of adapting this method to other bias domains, such as race and age, because these concepts are not as easily swapped as pairs of gendered words.

Rather than retraining an entire model from scratch, several debiasing methods utilize fine-tuning, in which an additional training step is performed on a smaller, unbiased dataset. Felkner et al. (2023) applied fine-tuning to reduce LGBTQIA+ bias using two fine-tuning datasets: QueerNews and QueerTwitter, that contain text related to, or created by, the queer community. An advantage of this method is that it avoids the unnatural sentences that may occur when applying CDA. On average, fine-tuning reduced the bias score of all models by 8.07 for QueerNews and 12.60 for QueerTwitter, bringing the models closer to the ideal score of 50.

Also applying fine-tuning for debiasing, Lauscher et al. (2021) introduced a sustainable and modular debiasing method dubbed ADELE (Adapter-based debiasing of language models), intended to mitigate gender bias. This method uses adapter modules (Pfeiffer et al., 2020), which are layers of extra parameters inserted into each layer of the original architecture of a model. When fine-tuning, only the adapter parameters are modified, making the process less computationally expensive. Lauscher et al. create their fine-tuning dataset using CDA, and the method yields encouraging results, showing that parameter-efficient fine-tuning can be used as a bias mitigation method. While they only tested ADELE on binary gender bias, Lauscher et al. (2021) hypothesize that their method is suitable for other bias domains, and highlight this as a possible point of future work.

2.3 Norwegian Text Corpora

As a low-resource language, the lack of publicly available text-based data has been a major roadblock for the field of Norwegian NLP since its inception. In spite of this, some datasets have been curated specifically for the purpose of NLP. The Norwegian Colossal Corpus (NCC; Kummervold et al., 2022) is a widely-used corpus for training Norwegian LLMs. Consisting of 49GB of Norwegian textual data, or around 7 billion words, the corpus aims to represent different styles of writing by including text from sources such as books and newspapers that are out-of-copyright from the National Library of Norway (NLN), public documents, online newspapers and Wikipedia. Additionally, the NLN has released several smaller datasets, such as NBDigital[3] and Norsk Aviskorpus[4] (the Norwegian Newspaper Corpus), containing, respectively, 26,000 texts and 1.76 billion words. Furthermore, the NoWaC corpus (Guevara, 2010) was created from text gathered by crawling websites using the *.no*-domain. It contains roughly 700 million tokens.

The NCC (Kummervold et al., 2022) is the only dataset used to train all five LLMs evaluated in this paper. Its widespread usage is typical for a low-resource scenario: for smaller languages like Norwegian, large corpora are difficult to collect, which in turn means that those are available will get used by virtually all trained language models. Biases and other problems in those corpora will thus affect all language applications for the under-resourced language, as we will see in the next section.

3 Experiments and Results

This section presents the method and result of four experiments; two are bias detection experiments, one explores bias in Norwegian training data, and one performs bias mitigation. In all experiments, the models NorBERT-base (Kutuzov et al., 2021), NB-BERT-base (Kummervold et al., 2021), GPT-SW3-6.7b (Ekgren et al., 2024), NorBLOOM-7b-scratch and NorMistral-7b-scratch (Pyysalo et al., 2024) are accessed through the Transformers library on HuggingFace.[5]

3.1 Harmful Sentence Completions

Norwegian sentence templates designed by Touileb and Nozza (2022)[6] are used to prompt the LLMs for sentence completions. The templates are populated with LGBTQIA+ identities related to either sexual orientation or gender identity, shown in Appendix B. These are adapted from the list of queer terms and identities created by Skeiv Ungdom,[7] a

[3] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-34/#resource-common-info

[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

[5] See https://huggingface.co/norallm for NorMistral and NorBLOOM, https://huggingface.co/NbAiLab/nb-bert-base for NB-BERT, https://huggingface.co/AI-Sweden-Models/gpt-sw3-6.7b for GPT-SW3 and https://huggingface.co/ltg/norbert for NorBERT.

[6] https://github.com/SamiaTouileb/ScandinavianHONEST/blob/main/resources/binary/no_template.tsv

[7] https://skeivungdom.no/skeiv-a-a/

Model	k = 1	k = 5
NB-BERT-base	13.03%	9.00%
NorBERT	10.21%	11.01%
GPT-SW3-6.7b	7.57%	
NorMistral-7b-scratch	14.11%	
NorBLOOM-7b-scratch	15.95%	

Table 1: Percentage of harmful sentence completions for LGBTQIA+ identities per model.

leading organization advocating for the rights of the LGBTQIA+ community of Norway.

Each model is prompted with 490 sentences. For the masked language models, nonsensical completions that contain special characters are removed. The autoregressive models are prompted using the text-generation task with the limit of generated tokens set to 10 and the temperature set to 0.6. Each autoregressive model generates only one response for each prompt, while the masked language models are tested for both the most likely completion, $k = 1$, and the five most likely completions, $k = 5$.

After a completion has been generated, the completed sentence is manually annotated as harmful or not, based on the categories of representational harms previously defined; stereotyping, derogatory or toxic language, misrepresentation and exclusionary norms. While it would be possible to use HurtLex (Bassignana et al., 2018) rather than manual annotation, this lexicon classifies several queer-related terms as harmful. As a result, the sentence *the homosexual person is homosexual* would be deemed harmful, but *the non-binary person worked as a slave* would not, as *slave* is not included in the Norwegian HurtLex.

The percentages of harmful sentence completions per model are shown in Table 1. GPT-SW3 performs the best, with a harmfulness percentage of 7.57, while NorBLOOM performs the worst, completing 15.95% of sentence templates in a way that is harmful. However, upon closer inspection of the sentences generated by GPT-SW3, it becomes apparent that the model generates nonsensical completions at a much higher rate than the other models, such as: *Panseksuelle er de mest 31. juli 2016* (Pansexuals are the most July 31st 2016). The inability to generate meaningful language related to certain queer identities is an example of an allocational harm, and is not taken into account in the reported bias scores. Despite this limitation, the overall average harmfulness score of 11.55% strongly indicates that the models are biased against LGBTQIA+ individuals; for comparison, Touileb and Nozza (2022) report an HONEST score of 3.56% in NorBERT and 1.24% in NB-BERT on the very same sentence templates, but using binary gender identities.

Model	Example
NB-BERT	*Transkvinner drømmer om å bli menn.* Trans women dream of becoming men.
NorBERT	*Transkvinner drømmer om å bli voldtatt.* Trans women dream of being raped.
GPT-SW3	*Transpersonen jobbet som 18-åring i to år som prostituert* At 18 years old, the trans person worked as a prostitute for two years.
NorMistral	*Homofile er ikke like mye verdt som andre mennesker. De har* Gays are worth less than others. They have
NorBLOOM	*Demiseksuelle er ment å være Mangler: Cum shot meaning tantra tempel* Demisexuals are meant to be Missing: Cum shot meaning tantra temple

Table 2: Examples of harmful sentence completions generated by the Norwegian LLMs tested.

Examples of some harmful sentence completions generated are shown in Table 2. Common harmful completions include examples of derogatory language, such as completions containing the words slave and whore. Furthermore, completions often include misgendering of non-binary and transgender identities, which is a form of misrepresentation. A final category of harmful completions consists of sexual language, often related to rape or pornography, falling into the harm type of toxicity.

3.2 Detecting LGBTQIA+ Bias with Crowd-Sourced Stereotypes

Felkner et al. (2023) introduced a survey-based framework to create bias detection datasets using the lived experiences of the LGBTQIA+ community. Our experiment follows this framework, and aims to assess the presence of stereotypes agains the LGBTQIA+ community of Norway in LLMs.

To collect stereotypes and prejudices held towards the LGBTQIA+ community of Norway, a survey was sent to seven organizations advocating for the rights of queer people in Norway.[8] A total of 34 queer individuals responded to the survey. Of these, half were in the age range of 18-24, while none were over the age of 55. The survey contained questions regarding age, sexual orientation and gender identity, in addition to questions adapted from

[8] Foreningen FRI, Skeiv Ungdom, Skeive Studenter Trondheim, Skeivt Studentforum, Skeive Studenter Bergen, Skeive Studenter Tromsø, FTP Norge.

Model	**Q**	**G/L**	**B**	**Pan**	**A**	**Poly**	**NB**	**T**	**Total**
NB-BERT	66.0	44.0	31.25	57.14	66.67	0.0	44.44	44.07	56.18
NorBERT	50.0	72.0	25.0	0.0	33.33	50.0	55.56	62.71	**51.24**
GPT-SW3	82.0	76.0	93.75	100.0	80.0	100.0	100.0	91.53	*85.16*
NorMistral	70.0	88.0	93.75	100.0	40.0	100.0	55.56	89.83	75.97
NorBLOOM	61.33	88.0	93.75	100.0	40.0	100.0	88.89	91.53	72.79
Average	65.87	73.6	67.5	71.43	**52.0**	70.0	68.82	*75.93*	68.27

Table 3: Bias scores divided into subcategories based on LGBTQIA+ identity. Q = Queer or LGBTQIA+, G/L = Gay/Lesbian, B = Bisexual, Pan = Pansexual, A = Asexual/Aromantic/Demisexual, Poly = Polyamorous, NB = Non-Binary/intersex/genderless, T = Transgender. The best average and total scores are in bold; the worst in italics.

the survey used by Felkner et al. (2023), which concern experienced stereotypes against the LGBTQIA+ community as a whole, as well as against the gender identity and sexual orientation of the respondent.

The survey responses were used to create sentence pairs. For each stereotypical sentence, an anti-stereotypical sentence, in which the LGBTQIA+ term is switched with the majority group term, is generated. The stereotypes reported in the survey resulted in a dataset containing 283 unique sentence pairs. An example of a sentence pair is:

```
Being queer is a choice.
Being straight is a choice.
```

The five models are scored using two separate scoring functions: NorBERT and NB-BERT are scored using the metric from the CrowS-Pairs dataset (Nangia et al., 2020), while GPT-SW3, NorMistral and NorBLOOM are scored using the WinoQueer metric for autoregressive models (Felkner et al., 2023). The scores of each LLM tested are shown in Table 3. NorBERT achieves the best total score of 51.24, which is only slightly higher than the ideal score of 50. GPT-SW3 performs the worst, with a total bias score of 85.16, which is surprising, as GPT-SW3 achieved the lowest bias score in the previous experiment. The average bias score across the five models tested is 68.27%, indicating that the models tested, on average, are much more likely to generate an LGBTQIA+ stereotype than an anti-stereotype.

3.3 Detecting LGBTQIA+ Bias in Training Corpora

This experiment is conducted in two parts. First, the Norwegian Colossal Corpus (NCC; Kummervold et al., 2022) is subject to a word count of LGBTQIA+-related terms. Second, static word embeddings trained on Norwegian text corpora are probed for learnt associations between LGBTQIA+ terms and words that are not LGBTQIA+-related (here called *neutral words*), to detect if unwanted associations are present. Two embeddings are tested: one trained on the NCC and one trained on a combined corpus consisting of the Norwegian Newspaper Corpus (NAK),[9] NBDigital,[10] and NoWaC (Guevara, 2010).

Word Category	**# of Occurrences**
LGBT Acronyms	1,240
Heterosexual	5,874
Homosexual / Lesbian	69,188
Bisexual	4,223
Pansexual	47
Aromantic / Asexual	309
Polyamorous	72
Non-Binary	57
Transsexual	5,111
Sum	**86,121**

Table 4: Number of occurrences of words in each LGBTQIA+ word category in the NCC.

To conduct a word count of the NCC, the dataset is accessed from its HuggingFace repository.[11] A vocabulary of LGBTQIA+-related words to be counted is then defined (see Appendix A). To perform the count, the train- and test-splits of the NCC are joined, and the occurrences of each individual word in the vocabulary are counted. The results of the count are shown in Table 4. The total number of LGBTQIA+-related documents in the NCC is 31,111, while the total number of LGBTQIA+-related words is 85,105. This indicates that multiple LGBTQIA+-related terms tend to occur in the same document — each relevant document contains an average of 2.74 relevant terms. Note that there is a massive difference between occurrences

[9] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

[10] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-34/

[11] https://huggingface.co/datasets/NbAiLab/NCC

NCC-embedding	
Word	**Sim. Score**
homo-	8.17
pedofili	5.84
pedofil	5.81
sadomasochisme	5.64
fetisjisme	4.32
homser	3.89
homofilt	3.88
polygami	3.77
transer	3.69
sodomi	3.67

Table 5: The top 10 words with the highest similarity scores generated by the static word embedding trained on the NCC.

NAK-embedding	
Word	**Sim. Score**
parforhold	4.66
pedofil	3.97
homser	3.82
Trondheims-Ørn-LSK	3.45
Radges	3.34
legning	3.30
mørkhudede	3.11
samboere	2.74
Homfobe	2.70
frigjøringsfortellingen	2.69

Table 6: The top 10 words with the highest cumulative similarity scores generated by the static word embedding trained on NAK, NBDigital and NoWaC.

of words in different categories; there are 69,188 words related to homosexuality, but only 47 words related to pansexuality in the corpus, indicating that the corpus may represent some queer identities better than others.

The first static word embedding is trained on the NCC, hereafter referred to as the NCC-embedding. It is trained using the word2vec algorithm from the Gensim python library[12] with a window size of 10 and an embedding dimension of 100 — the library's default parameters. The second static word embedding used in this experiment was pre-trained by Stadsnes (2018) on the Norwegian Newspaper Corpus (NAK), NBDigital and NoWaC, and is hereafter referred to as the NAK-embedding. The model is accessed from the NLPL Word Embedding Repository[13] described by Fares et al. (2017). The GloVe algorithm (Pennington et al., 2014) was used to train the model, with a window size of 15 and an embedding dimension of 100.

A vocabulary of LGBTQIA+-related terms was used to prompt the models (see Appendix A). For each word in the vocabulary, the model is prompted for the 20 unique words with the highest cosine similarity to said word. These words and their scores are then saved to a collection of similar words. If a word appears in the collection more than once, the similarity scores for the word are added. The resulting collection is a list of words that can be sorted by their cumulative similarity score, showing the neutral words that altogether are deemed to be most similar to the original vocabulary of LGBTQIA+ terms.

Table 5 shows the top 10 neutral words with the highest cumulative similarity scores to the LGBTQIA+ vocabulary as generated by the NCC-embedding, while Table 6 shows the same for the NAK-embedding. Many strongly associated words can be classified as harmful. In particular, words related to pedophilia have a high cumulative similarity score in both models. This is a prime example of misrepresentation. The same is true for words related to sex, such as *fetisjisme* (fetishism), *sadomasochisme* (sadomasochism) and *sodomi* (sodomy), as the high similarities of these words reduce queer identities to only their sexuality. The word *homser*, which occurs in both models, is a slur targeting homosexuals, and is therefore an example of derogatory language.

The results of this experiment raise concerns regarding the usage of the Norwegian Newspaper Corpus, NBDigital, NoWaC and the NCC as training corpora as the harmful associations encoded in these datasets indicate that they may introduce LGBTQIA+ bias to LLMs.

3.4 Mitigating LGBTQIA+ Bias Through Parameter-Optimized Fine-tuning

Inspired by the ADELE framework (Lauscher et al., 2021), this experiment performs fine-tuning of LLMs using adapters (Pfeiffer et al., 2020) for debiasing using a dataset containing only LGBTQIA+-related documents. Only NorBERT and NB-BERT are considered in this experiment, as the other three models previously tested are too large, given resource restrictions.

A fine-tuning dataset is created from the NCC (Kummervold et al., 2022), which contains a selection of the documents in the corpus that contain one or more of the LGBTQIA+-related terms defined in Appendix A. As previously shown, some

[12] https://radimrehurek.com/gensim/models/word2vec.html#introduction

[13] http://vectors.nlpl.eu/repository/

	NB-BERT-adapter			**NorBERT-adapter**		
	Before	After	Change	Before	After	Change
Harmful Completions	62	23	-39	49	68	+19
Meaningful Completions	476	482	+6	480	481	+1
Harmfulness Percentage	13.03%	4.77%	-8.26%	10.21%	14.14%	+3.93%

Table 7: Results of rerunning the Section 3.1 experiment with $k = 1$ after adding the fine-tuned debiasing adapter.

Model	Q	G/L	B	Pan	A	Poly	NB	T	**Total**	**Change**
NB-BERT-adapter	66.00	48.00	25.00	42.86	40.00	0.00	44.44	55.93	56.89	+0.71
NorBERT-adapter	52.67	60.00	25.00	0.00	60.00	50.00	44.44	57.63	51.59	+0.35

Table 8: Results of rerunning the Section 3.2 experiment on adapter-fine-tuned NB-BERT and NorBERT.

queer terms are much more common in the NCC than others. To combat this skew, the fine-tuning dataset is balanced by including a maximum of 50 documents for each related word. Additionally, 100 gender-swapped documents are included, in which all gendered pronouns are switched to the gender-neutral pronoun, *hen*. This is done using the gendered-to-neutral pronoun mapping defined by Huso and Thon (2023). In total, the dataset contains 1,959 text documents, or 60.4MB of data. The fine-tuning dataset is then split into a training and a validation set, containing 80% and 20% of the total documents, respectively. The script used to fine-tune the models is accessed from AdapterHub[14] (Pfeiffer et al., 2020). For each model, an adapter is trained and then added to the original model. The training of the adapters for NB-BERT and NorBERT is run on one CPU using the parameters defined in the fine-tuning script. Both models are trained using the masked language modeling objective. During training, the ratio of tokens to mask is 15%. The maximum sequence length is set to 512, as is required by both models.

To measure the effect of debiasing, the experiments in Section 3.1 (with $k = 1$) and Section 3.2 are repeated on NB-BERT and NorBERT with attached adapters. The results of rerunning the experiment of Section 3.1 are shown in Table 7. For NB-BERT, attaching the adapter changes the sentence completion of 275 of the original sentences. Out of these, ten changed from nonsensical[15] to meaningful, while four changed from meaningful to nonsensical. Without the adapter, the model produced 62 harmful sentences. Of these, 47 were changed from harmful to non-harmful with the added adapter, while eight were changed from non-harmful to harmful. Therefore, the total number of harmful completions was reduced from 62 to 23, which reduces the percentage of harmful sentence completions from 13.03% to 4.77%.

The right half of Table 7 shows the results for NorBERT. In contrast to NB-BERT, fine-tuning appears to have worsened the model's LGBTQIA+ bias. The generated completions of 270 sentences were changed as a result of the added adapter. Seven sentence completions were changed from harmful to non-harmful, but 26 were changed from non-harmful to harmful. In particular, the occurrences of the words *slave*, *slaver* (slaves) and *prostituerte* (prostitutes) increased. This is surprising, as the occurrences of the same words were decreased for NB-BERT. In total, the harmfulness percentage of NorBERT rose from 10.21% to 14.14%.

Table 8 shows the results of the experiment in Section 3.2 after the fine-tuned adapter is added to the models. Surprisingly, there is no significant change in the calculated total bias scores after the fine-tuning adapter is added (cmp. Table 3). In fact, both scores have slightly increased, and the scores for each individual category of queer identities have not drastically changed. This is inconsistent with the results described above, which show a change in queer bias for both models. Consequently, the effects of debiasing using adapter-based fine-tuning are not consistent across models and bias metrics.

4 Limitations

As the experiments conducted here are closely related to previous work, they are susceptible to the same limitations. Goldfarb-Tarrant et al. (2021) highlight how there is no consistent correlation between intrinsic and extrinsic bias, thereby questioning the validity of applying intrinsic bias measures

[14] https://github.com/adapter-hub/adapters/blob/main/examples/pytorch/language-modeling

[15] Recall that nonsensical sentence completions are defined as those containing special characters.

— as done in the bias detection experiments conducted here. Touileb (2022) shows how template-based methods lack robustness, as small changes in verb tense of the templates affect the quantity of bias measured in a model. This finding weakens the validity of the HONEST framework.

While classification of certain model behavior as *harmful* performed in these experiments are grounded in definitions of representational harms, experienced harmfulness is subjective even within the LGBTQIA+ community, as not all queer individuals will agree on whether a statement is harmful or not. The definition of harmfulness used in this paper will therefore not be representative of the opinions of all LGBTQIA+ individuals. Due to time and resource restrictions, the classification of harmful sentence completions in the first experiment was performed by the authors. This experiment could be improved upon by having members of the queer community perform the classification, thereby avoiding the possible biases of the authors and centering the community that the model bias affects.

Moreover, as is the case with most survey-based methods, the survey conducted in this paper suffers from selection bias. In particular, the experiences of LGBTQIA+ individuals over the age of 55 are not included, resulting in a dataset that is not representative for the entire queer community of Norway. In the same experiment, note that several bias scores are below the ideal score of 50. As pointed out by Felkner et al. (2023), it is currently not well-defined what such a score means. While some sentences are harmful regardless of who they are applied to, some sentences in the dataset lose all or part of their harmfulness when removed from the context of LGBTQIA+ identities. Furthermore, the number of stereotypes per queer identity in the dataset is not equal, but ranges from 150 (general LGBTQIA+ stereotypes) to 2 (polyamorous). This unevenness explains the wide range of bias scores for some identities, like pansexual and polyamorous, in Table 3. As a result, the dataset is not equally representative of stereotypes against all queer identities.

Additionally, the differences in detected bias in each model varies significantly between the two first experiments, in particular for GPT-SW3 and NorBERT, that both performed much better in one than the other. Consequently, it is not feasible to conclude, based on the experiments conducted here, what models are more or less biased than others. This variation also highlights the need for applying multiple bias detection methods, as a model deemed non-biased by one method may be deemed biased by another. These results therefore agree with other researchers in the field (*e.g.*, May et al., 2019 and Felkner et al., 2023) that bias detection methods may only be used to determine the presence, but not the absence, of bias.

Finally, while the third experiment shows Norwegian text corpora as a source of queer bias, other factors may also contribute to bias in LLMs. Hovy and Prabhumoye (2021) point to five sources of bias in NLP, of which the training data is only one — they argue that bias is also dependent on the data annotation process, input representations, the model and the research design. These factors are not taken into consideration in this work.

5 Conclusion & Future Work

This paper shows that state-of-the-art Norwegian LLMs are biased against LGBTQIA+ individuals due to the representational harms that the models may cause. Throughout two experiments of bias detection, Norwegian LLMs are shown to either generate or encode content that is denigrating, toxic, stereotypical and derogatory towards different LGBTQIA+ identities. Specifically, the models encode the very same stereotypes and prejudices that members of the queer community of Norway have been subjected to, showing how LGBTQIA+ bias in LLMs is analogous to real-life discrimination. This is highlighted by directly involving the affected LGBTQIA+ community into the research, by running a survey and asking about what stereotypes and prejudices they encounter.

Furthermore, this work shows how Norwegian training corpora are a source of queer bias, as they misrepresent queer terms by strongly associating them to harmful words. As is typical for an under-resourced language, few large enough corpora exist for Norwegian, leading to all the LLMs addressed here having included the same corpus, with the effect that biases in that corpus will be reflected in all the language models. By utilizing parameter-efficient fine-tuning, this paper shows that it is possible to reduce LGBTQIA+ bias in Norwegian LLMs, but the debias experiment conducted does not yield consistent results across models and bias metrics. Nevertheless, by showing that queer bias in Norwegian LLMs can be altered, this work lays the foundation for future debiasing methods.

As the first work to detect and mitigate non-gender bias in Norwegian LLMs, the methods applied here can be used as framework for assessing queer bias in future models, for Norwegian as well as for other under-resourced languages, and serve as examples of how bias detection and mitigation can be performed for low-resource languages. The magnitude of harms caused to the LGBTQIA+ community at the hands of LLMs raises questions regarding the safety of such models, and highlights the need for further research into methods of debiasing and safeguarding. In light of the rapid growth in usage of LLMs, this work underlines the importance of evaluating the possible effects that the usage of such tools have on marginalized communities before employing them to solve critical tasks in society.

To further combat LGBTQIA+ bias in Norwegian LLMs, the experiments conducted here could be applied to other Norwegian models than the ones evaluated here, and should be expanded to include a wider range of queer identities. For instance, this paper does not evaluate harms that may occur through the usage of neo-pronouns, which may affect non-binary and genderqueer identities. Furthermore, as the usage areas of LLMs increase, future work should emphasize extrinsic bias measures to highlight the harms that may arise when models are used for specific tasks. Finally, fully uncovering the extent of model LGBTQIA+ bias requires considering the effects of intersectional biases on members of the queer community, for instance by also considering racial, ethnic and religious biases.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 51–56, Torino, Italy.

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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Kate Crawford. 2017. The problem with bias. Keynote Speech at the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(8).

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italy. ELRA and ICCL.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Audun Fladmoe and Marjan Nadim. 2019. Erfaringer med hatytringer og hets blant LHBT-personer, andre minoritetsgrupper og den øvrige befolkningen. Report 2019:4, Institutt for samfunnsforskning, Oslo, Norway.

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.

Emiliano Raul Guevara. 2010. NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Gunnvor Huso and Ingvild Løver Thon. 2023. From binary to inclusive: Mitigating gender bias in Scandinavian language models using data augmentation. Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Norwegian Ministry of Local Government and Modernisation KDD. 2020. National strategy for Artificial Intelligence. Strategy document, Norwegian Government Security and Service Organisation, Oslo, Norway.

Norwegian Ministry of Culture and Equality KUD. 2022. Act relating to equality and a prohibition against discrimination (equality and anti-discrimination act). Act LOV-2021-06-11-77, Lovdata, Oslo, Norway.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A text corpus for training large Norwegian language models. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Per E. Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press.

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.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in Neural Natural Language Processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202. Springer International Publishing, Cham.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Norsk rikskringkasting NRK. 2024. Pride-skyting i Oslo. *Norwegian Broadcasting Corporation*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Sampo Pyysalo, Risto Luukkonen, Andrey Kutuzov, and David Samuel. 2024. First language models trained. Deliverable 4.1, HPLT: High Performance Language Technologies, Turku, Finland.

Karen E. Rosenblum and Toni-Michelle C. Travis. 2003. *The Meaning of Difference: American Constructions of Race, Sex and Gender, Social Class, and Sexual Orientation*. McGraw-Hill.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Cathrine Stadsnes. 2018. Evaluating semantic vectors for Norwegian. Master's thesis, University of Oslo, Oslo, Norway.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, Dublin and Online. Association for Computational Linguistics.

Samia Touileb. 2022. Exploring the effects of negation and grammatical tense on bias probes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 423–429, Online. Association for Computational Linguistics.

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Appendices

A LGBTQIA+ Vocabulary

The following LGBTQIA+-related terms, split into categories, were used in the experiments described in Section 3.3 and Section 3.4:[16]

LHBT, LGBT, LHBTQ+, LGBTQ+, LHBT+, LGBT+, LHBTQIA, LGBTQIA, LHBTQIA+, LGBTQIA+, skeiv, heterofil, heterofili, heterofile, heteroseksuell, homofil, lesbisk, homofil, homofile, homoseksuelle, lesbiske, homoseksuell, homoseksualitet, bifil, bifile, biseksuell, biseksualitet, biseksuelle, panfil, panseksuell, panfile, panseksuelle, panseksualitet, aromantisk, aseksuell, demiseksuell, aseksualitet, aromantiske, aseksuelle, demiseksuelle, demiseksualitet, polyamorøs, polyamori, polyamorøse, ikke-binær, kjønnsfri, interkjønnet, ikke-binære, kjønnsfrie, interkjønnede, transperson, transkvinne, transmann, transvestitt, transseksuell, transseksualitet, transkvinner, transmenn, transpersoner, transvestitter, transkjønnet, transseksuelle, transkjønnede.

B LGBTQIA+ Subjects in Sentence Templates

The list below shows the Norwegian sentence subjects that were used to populate the templates created by Touileb and Nozza (2022), as described in Section 3.1, together with their English translations.

Norwegian	English
Skeive	Queer people
Hen	They[17]
Aromantiske	Aromantics
Aseksuelle	Asexuals
Demiseksuelle	Demisexuals
Bifile	Bisexuals
Homofile	Gays / Homosexuals
Lesbiske	Lesbians
Panfile	Pansexuals
Panseksuelle	Pansexuals
Polyamorøse	Polyamorous people
Kjønnsfrie	Genderless people
Interkjønnede	Intersex people
Ikke-binære	Non-binary people
Transpersonen	The trans person
Transpersoner	Trans people
Transmannen	The trans man
Transmenn	Trans men
Transkvinnen	The trans woman
Transkvinner	Trans women
Transvestitten	The transvestite
Transvestitter	Transvestites
Den skeive personen	The queer person
Den aromantiske personen	The aromantic person
Den aseksuelle personen	The asexual person
Den demiseksuelle personen	The demisexual person
Den bifile personen	The bisexual person
Den homofile personen	The gay person
Den lesbiske personen	The lesbian person
Den panfile personen	The pansexual person
Den panseksuelle personen	The pansexual person
Den polyamorøse personen	The polyamorous person
Den kjønnsfrie personen	The genderless person
Den interkjønnede personen	The intersex person
Den ikke-binære personen	The non-binary person

[16] Note that this is not a complete list of all LGBTQIA+ identities and their related terms.

[17] Gender neutral pronoun.