Bias Mitigation in Hiring-Related NLP: Interactions Between Masking, Rewriting, and Adversarial Debiasing

Alexandre Puttick
Bern University of Applied Sciences
Technik und Informatik
Quellgasse 21, 2501 Biel, Switzerland
alexandre.puttick@bfh.ch

Rami El-Wazzi
Bern University of Applied Sciences
Technik und Informatik
Quellgasse 21, 2501 Biel, Switzerland
rami.el-wazzi@students.bfh.ch

Abstract

AI-driven language technologies are increasingly used in hiring, but they may encode and reproduce harmful social stereotypes. Prior work often studies bias mitigation methods in isolation and outside realistic application settings. We examine the combined effects of data-level and model-level debiasing in a hiring-related context, using Norwegian-language academic bios and a proxy STEM/non-STEM classification task. Specifically, we study masking sensitive information, GenWriter-based rewrites (Soundararajan and Delany, 2025), and adversarial debiasing (Han et al., 2021, 2022). We evaluate these interventions using downstream task performance, group fairness metrics, intrinsic bias tests based on WEAT (Caliskan et al., 2017), and measures of gender leakage from hidden representations. We find that combining masking, GenWriter rewrites, and adversarial debiasing substantially reduces gender leakage while maintaining or improving downstream performance. However, effects on fairness gaps and intrinsic bias are mixed, underscoring the need for downstream, context-sensitive evaluation of bias mitigation methods in hiring-related NLP.

1 Introduction

AI-driven language technologies are increasingly used in hiring, where they may influence screening, ranking, and evaluation. These systems can encode and reproduce harmful social stereotypes, and a large body of literature has proposed methods for detecting and mitigating such biases. Common approaches are often grouped into *pre-processing* methods, such as counterfactual data augmentation (Zmigrod et al., 2019), *in-processing* methods, such as fairness-constrained optimization (Hardt et al., 2016), and *post-processing* methods, such as re-ranking (Zehlike et al., 2017) or embedding debiasing (Bolukbasi et al., 2016; Liang et al., 2020).

Despite this progress, limitations remain that are especially relevant for hiring-related NLP. First, many studies are weakly connected to concrete application contexts and to the ethical goals that motivate fairness interventions. Second, non-English languages remain underrepresented. Third, debiasing methods are often evaluated in isolation, even though real systems may combine interventions at multiple stages of the pipeline.

This paper addresses these gaps in a hiring-related setting as part of the EU Horizon-funded BIAS Project[1]. We study Norwegian-language academic bios from the Norwegian University of Science and Technology (NTNU) and use a proxy task of classifying bios as STEM or non-STEM. Although this is not a direct hiring task, it captures a recruitment-relevant scenario in which textual candidate information is mapped to a consequential category. Our aim is to study how different interventions affect not only downstream task performance, but also fairness-relevant properties of the model.

We take a downstream-first approach to debiasing. At the data level, we study *masking* of names and gendered pronouns and *GenWriter* rewrites (Soundararajan and Delany, 2025). Masking removes explicit gender markers, while GenWriter aims to reduce stylistic and other indirect gender signals through case-based rewriting. At the model level, we study *adversarial debiasing* (Zhang et al., 2018; Han et al., 2021, 2022), which is designed to reduce sensitive-attribute information in model representations while preserving task-relevant information.

Our analysis distinguishes between several related but non-identical notions. We treat *procedural fairness* as limiting the model's reliance on sensitive attributes and their proxies, which we assess through gender leakage from hidden representations. We treat *substantive fairness* as parity

[1] https://www.biasproject.eu/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 202–213

in predictive performance across gender groups, measured using performance gaps and equality-of-odds-related metrics. We also include *intrinsic bias* tests as diagnostic measures of stereotypical associations encoded in the model, while recognizing that these need not align with downstream fairness.

The following research questions guide the paper: **(RQ1)** How do masking, GenWriter rewrites, and adversarial debiasing interact in affecting downstream task accuracy, fairness gaps, and gender leakage? **(RQ2)** Do reductions in sensitive-attribute leakage correspond to reductions in intrinsic stereotype bias? **(RQ3)** Can a downstream-oriented mitigation strategy improve procedural fairness without large performance costs?

Our contributions are:

- We present a contextualized, downstream-first study of bias mitigation for hiring-related NLP in a Norwegian-language setting, using academic bios as a proxy recruitment task.
- We evaluate the interaction between masking, GenWriter rewrites, and adversarial debiasing across downstream performance, group fairness, intrinsic bias, and representation leakage.
- We show that combined interventions substantially reduce encoded gender information while maintaining or improving downstream task performance, whereas effects on fairness gaps and intrinsic bias are more mixed.

Bias Statement

We study *diversity bias* with a focus on gender. We distinguish between *intrinsic bias*, i.e. stereotypical associations encoded in language representations, and *extrinsic bias*, i.e. disparities in downstream task behavior across groups. In this work, extrinsic bias is operationalized as differences in STEM/non-STEM classification performance between gender groups. We therefore focus on accuracy parity and equality of odds rather than demographic parity, since the task is conditioned on ground-truth labels and those labels are not strongly aligned with gender stereotypes in the dataset (Table 1). We additionally treat low sensitive-attribute leakage as a procedural fairness objective.

2 Related Work

Fairness in NLP and algorithmic decision-making. A large body of literature studies bias detection and mitigation in NLP. At the same time, several authors argue that formal fairness metrics do not fully capture the social and normative dimensions of fairness, motivating the need to select metrics that align with explicit ethical aims (Fazelpour and Lipton, 2020; Tong et al., 2026; Majumder et al., 2023; Heidari et al., 2019; Mitchell et al., 2021).

Intrinsic bias, downstream bias, and leakage. A recurring finding in the literature is that intrinsic bias measures do not straightforwardly predict downstream unfairness. Gonen and Goldberg (2019) showed that embedding debiasing can reduce bias according to particular metrics while leaving recoverable gender structure in the representations. Goldfarb-Tarrant et al. (2021) found little relationship between intrinsic measures such as WEAT and downstream bias, and Delobelle et al. (2022) showed that many intrinsic bias metrics are highly sensitive to templates, target words, and representation choices. Tokpo et al. (2023) also show that some intrinsic debiasing approaches mainly reduce bias according to specific metrics rather than in downstream behavior. These findings motivate evaluating leakage, intrinsic bias, and downstream fairness separately.

Data-level and model-level debiasing. At the data level, masking explicit sensitive markers is a common baseline, though it does not remove more indirect proxies. We therefore also consider *GenWriter* (Soundararajan and Delany, 2025), a case-based rewriting framework that aims to reduce stylistic and demographic signals while preserving task-relevant content. At the model level, we study *adversarial debiasing* (Zhang et al., 2018; Han et al., 2021, 2022), which explicitly trains a model to perform its main task while removing sensitive-attribute information from hidden representations. Our contribution is not a new debiasing method, but an empirical study of how these interventions interact across downstream performance, fairness gaps, intrinsic bias, and representation leakage.

3 Methods

We describe our methods in this section. The prompts used for all prompt-related tasks can be found in the appendix Section A.2.

3.1 Data

NTNU bios. We collected a dataset of professional academic bios from faculty pages at the

Job Title	Male	Female	STEM	non-STEM
Professor	333(219)	195(94)	313	215
Associate Professor	252(149)	252(140)	289	215
PhD Candidate	151(115)	174(115)	230	95
Totals	736(483)	621(349)	832	525

Table 1: NTNU bios dataset composition. Numbers in parentheses indicate how many individuals in each gender group are in STEM-related fields.

Norwegian University of Science and Technology (NTNU). We restricted the dataset to the three most common job titles: *professor*, *associate professor*, and *PhD candidate*. The final dataset composition is shown in Table 1.

Automatic gender label inference. To obtain gender labels for analysis, 100 professor bios were manually labeled by native speakers using names and/or pronouns. Only one sample was explicitly non-binary; because this was insufficient for meaningful modeling, we restrict the present study to binary gender labels. We trained an `NbAiLab/nb-sbert-base` classifier on the manually labeled subset and used it to infer labels for the remaining bios. Five-fold cross-validation yielded a mean test accuracy of 98% with negligible variance.

STEM/non-STEM labels. The 51 NTNU departments represented in the dataset were manually classified as STEM or non-STEM based on their primary research focus and methods, using department descriptions and curricula from NTNU sources. This yielded 27 STEM departments (52.9%) and 24 non-STEM departments (47.1%). Interdisciplinary departments were classified conservatively and included as STEM only when their primary methods were quantitative and technical.

3.2 Data-level bias mitigation

Masking. As a simple data-level intervention, we mask features that explicitly reveal gender, namely names and gendered pronouns. This does not remove indirect proxies, but it provides a minimal fairness-through-unawareness baseline.

GenWriter rewrites. We evaluate *GenWriter* (Soundararajan and Delany, 2025), a case-based rewriting pipeline designed to reduce demographic signals while preserving job-relevant content. It first splits bios into sentences, classifies them into four categories (*Education*, *Demographics*, *Work*, and *Non-Professional*), converts them into template sentences with placeholders, retrieves similar templates from the opposite gender group, and uses an LLM to fill placeholders with details from the original bio. Sample templates and rewrites are shown in Appendix A.1.

Sentence labeling and case-base construction. To train the sentence classifier used in GenWriter, we manually labeled 222 sentences from professor bios and then labeled 900 additional sentences balanced across profession, gender, and STEM/non-STEM category with `GPT 5.1`, using category definitions and three human-labeled examples per category. Agreement between GPT and human labels was 88%. Because *Non-Professional* content was rare, we added 200 synthetic *Non-Professional* samples. The resulting label distribution is shown in Table 2. We then trained a Norwegian SBERT model, obtaining 5-fold cross-validation accuracy of $81.9 \pm 1.53\%$. Since many disagreements reflected short or category-ambiguous sentences, we constructed the case base using SBERT-predicted labels.

Table 2: Human- and GPT-labeled sentence data used for sentence classification.

Category	Num. of Samples
Work	813
Education	139
Demographics	119
Non-Professional	222

3.3 Model-level bias mitigation: adversarial debiasing

Our model-level intervention is adversarial debiasing (Zhang et al., 2018; Han et al., 2021, 2022). We use it because it directly targets the removal of sensitive-attribute information from model representations while preserving information relevant to the main task. In our setting, the main task is STEM/non-STEM classification from NTNU bios. The model consists of two components:

- a *main model*, which predicts the task label;
- a *discriminator*, which predicts gender from the main model's hidden representation.

The optimization objective is:

$$\min_{M} \max_{A} \mathcal{L}_M(y, \hat{y}_M) - \lambda_{\text{adv}} \chi(g, \hat{g}_A), \quad (1)$$

where $\mathcal{L}_M$ is the main-task classification loss, χ is cross-entropy loss for the adversary, y and $\hat{y}_M$ are task labels and predictions, g and $\hat{g}_A$ are gender labels and predictions, and λ_{adv} controls the trade-off between task performance and debiasing.

3.4 Models and training details

Main model and adversary. Our main classifier is `NbAiLab/nb-bert-base`, fine-tuned end-to-end for the STEM/non-STEM task. The hidden representation passed to the adversary is the `[CLS]` token from the final hidden layer. The adversary is a 3-layer MLP. For simplicity, all reported experiments use a single discriminator, although prior work found somewhat stronger leakage reduction with multiple discriminators (Han et al., 2021).

Training setup. All models use the same 80/20 train-test split. For the downstream task, splits are stratified by gender, profession, and STEM/non-STEM label. We use Adam with a learning rate of 10^{-5} for both the main model and the discriminator. Following Han et al. (2021), training alternates between updating the main model and the adversary: each step consists of one epoch of main-task training followed by ten epochs of adversary training. We train for up to fifteen such steps with early stopping if validation accuracy does not improve for five epochs. We report the main model with the best validation performance, subject to discriminator validation accuracy remaining below 90%.

Prompted components. All prompt-based operations in the GenWriter pipeline were performed with `GPT 5.1` via the `OpenAI API` at temperature 0.7.

Leakage probing. To measure gender leakage, we train a LinearSVC probe on the hidden representations of the pretrained or fine-tuned main model and evaluate its ability to recover gender labels.

3.5 Experimental setup

We vary three dimensions in our experiments:

- *Dataset:* original bios vs. GenWriter rewrites
- *Masking:* masked vs. unmasked
- *Training:* standard fine-tuning vs. adversarial debiasing

For each condition, we fine-tune a STEM/non-STEM classifier and evaluate it along four dimensions: downstream task performance, group fairness, intrinsic bias, and gender leakage from hidden representations.

3.6 Evaluation

We distinguish between four types of evaluation.

Downstream task performance. We report classification accuracy on the STEM/non-STEM task.

Group fairness. We report accuracy gap, F1 gap, and equalized odds gap across gender groups. These quantify disparities in predictive performance and error behavior between groups.

Sensitive-attribute leakage. Leakage measures how well gender can be predicted from the model's hidden representations. Lower leakage indicates less recoverable gender information and is interpreted here as better procedural fairness.

Intrinsic bias. We evaluate intrinsic stereotypical associations using the *Word Embedding Association Test* (WEAT) (Caliskan et al., 2017) and the *Log-Probability Bias Score* (LPBS) (Kurita et al., 2019). These tests are applied to contextual representations derived from the `[CLS]` embedding. The test sets were translated and LPBS tests were adapted as described in [PROJECT CITATION].

The following WEAT-style tests were used:

Name	Description
WEAT6	career vs. family, male vs. female
WEAT7	math vs. art, male vs. female
WEAT8	science vs. art, male vs. female
GER1	male vs. female study choices
NO1	Iranian vs. Norwegian, negative vs. positive
NO2	Muslim vs. Christian, failure vs. success
NO3	Muslim vs. Christian, negative vs. positive
NO4	Iranian vs. Norwegian, failure vs. success
NO5	male vs. female, career vs. having children
NO6	single vs. in a couple, career vs. having children
NO7	Iranian vs. Norwegian, good sellers vs. not good sellers

Table 3: WEAT6–8 are from Caliskan et al. (2017); GER1 is from Kurpicz-Briki (2020). Norwegian bias tests developed within the BIAS Project from co-creation workshop discussions. Tests were designed for Norwegian culture specificity. All tests were translated by native speakers.

4 Results

4.1 Downstream performance, fairness, and leakage

Original bios. Table 4 reports downstream performance and group fairness results for models

trained on the original NTNU bios. Overall, fairness gaps were small across all settings. Standard fine-tuning on unmasked bios yielded the smallest accuracy, F1, and equality-of-odds gaps, but these differences were minor. Adversarial debiasing with unmasked data achieved the highest downstream accuracy, with somewhat larger gains for men than for women. Masking names and pronouns slightly reduced male accuracy under both standard and adversarial training.

Table 4: Original bios: STEM/non-STEM classification performance and fairness.

Method	Masked	Acc Gap	F1 Gap	EqOdds Gap	Acc F	Acc M
adversarial	No	0.012	0.009	0.012	**0.920**	**0.932**
adversarial	Yes	-0.008	-0.015	0.018	**0.920**	0.912
standard	No	**-0.000**	**-0.006**	**0.007**	0.912	0.912
standard	Yes	-0.015	-0.021	0.018	**0.920**	0.905

Table 5 shows gender leakage results for the same models. Relative to the pretrained unmasked baseline, masking alone produced a large reduction in recoverable gender information, reducing leakage from 0.912 to 0.676. Standard and adversarial fine-tuning without masking also reduced leakage substantially, to 0.728 and 0.732, respectively. The strongest reduction was obtained by combining masking with adversarial debiasing, which lowered leakage to 0.629. Thus, even when downstream fairness gaps changed only modestly, both masking and adversarial training reduced gender information in the learned representations.

Table 5: Original bios: gender leakage from hidden representations.

Method	Masked	Test Leakage	Test F1	Test ROC AUC
adversarial	No	0.732	0.729	0.801
adversarial	Yes	**0.629**	**0.627**	**0.658**
pretrained	No	0.912	0.911	0.978
pretrained	Yes	0.676	0.675	0.721
standard	No	0.728	0.727	0.815
standard	Yes	0.643	0.642	0.686

GenWriter-rewritten bios. Table 6 reports results for models trained on GenWriter-rewritten bios. Compared with the original-bio setting, rewritten bios consistently improved downstream accuracy for both groups. However, effects on fairness remained mixed. In the standard unmasked setting, the model performed slightly better for men than for women. Masking reduced this difference and yielded a near-zero accuracy gap, but at the cost of lower accuracy for both groups. Under adversarial training, the pattern reversed, with higher accuracy for women than for men; masking reduced but did not eliminate this gap. Overall, rewritten bios improved task performance more clearly than they improved group fairness.

Table 6: Rewritten bios: STEM/non-STEM classification performance and fairness.

Method	Masked	Acc Gap	F1 Gap	EqOdds Gap	Acc F	Acc M
adversarial	No	-0.039	-0.047	0.036	**0.959**	0.920
adversarial	Yes	-0.023	-0.032	0.030	0.950	0.927
standard	No	-0.017	-0.024	**0.023**	**0.959**	**0.942**
standard	Yes	**0.000**	**-0.006**	0.025	0.934	0.934

Leakage results for rewritten bios are shown in Table 7. Rewriting reduced leakage relative to the corresponding original-bio models in every setting. The strongest result again came from combining masking with adversarial debiasing, which reduced leakage to 0.566, close to chance-level prediction. Standard fine-tuning with masking also yielded lower leakage than its counterpart on the original bios. Taken together, these results suggest that GenWriter rewrites and adversarial debiasing are complementary for reducing recoverable gender information, even when improvements in group fairness metrics are less consistent.

Table 7: Rewritten bios: gender leakage from hidden representations.

Method	Masked	Test Leakage	Test F1	Test ROC AUC
adversarial	No	0.643	0.642	0.683
adversarial	Yes	**0.566**	**0.564**	**0.570**
standard	No	0.682	0.682	0.705
standard	Yes	0.628	0.628	0.664

4.2 Intrinsic bias

Table 8 summarizes all WEAT/LPBS test pairs that showed statistically significant bias in at least one model configuration.[2] For fine-tuned models, all results in this table are based on models trained on masked data.

Across models, intrinsic bias results were mixed and did not mirror the leakage results. The pretrained and standard fine-tuned models showed the fewest significant test outcomes overall, whereas other fine-tuned models often showed more significant effects. Several tests yielded broadly similar patterns across models. However, the direction and magnitude of changes varied by test. Overall, intrinsic bias measures did not improve uniformly

[2]LPBS-GER1 failed because several words were split into many subtokens, producing near-zero prior probabilities and numerical instability.

under the interventions that most strongly reduced leakage. To examine the role of masking more

Table 8: Intrinsic bias results for WEAT/LPBS test pairs. Boldface indicates significant bias ($ES > 0, p < 0.05$ or $ES < 0, p > 0.95$). All fine-tuned models were trained on masked data.

Test	Method	WEAT ES	WEAT p	LPBS ES	LPBS p
Ger1	pretrained	0.556	0.197	-	-
Ger1	adversarial	0.492	0.230	-	-
Ger1	adv. GW	**-1.518**	0.989	-	-
Ger1	standard	-0.640	0.815	-	-
Ger1	std. GW	-0.539	0.779	-	-
No2	pretrained	0.407	0.245	**0.894**	0.041
No2	adv. GW	0.503	0.185	**0.865**	0.049
No2	adversarial	0.612	0.137	**0.943**	0.035
No2	standard	0.521	0.184	**0.894**	0.041
No2	std. GW	0.475	0.206	**0.996**	0.024
No5	pretrained	**0.875**	0.029	0.610	0.124
No5	adversarial	**0.967**	0.013	0.636	0.111
No5	adv. GW	**0.921**	0.020	0.419	0.208
No5	standard	**0.985**	0.012	0.715	0.082
No5	std. GW	**0.919**	0.019	0.487	0.174
No6	pretrained	0.940	0.060	**0.922**	0.036
No6	adversarial	**1.076**	0.037	0.628	0.117
No6	adv. GW	**1.149**	0.017	**1.160**	0.011
No6	standard	**1.064**	0.043	**0.893**	0.037
No6	std. GW	**1.040**	0.046	**0.704**	0.088
WEAT6	pretrained	**0.771**	0.048	**1.053**	0.019
WEAT6	adversarial	**0.808**	0.039	0.722	0.085
WEAT6	adv. GW	**0.808**	0.038	**1.164**	0.011
WEAT6	standard	**0.817**	0.038	**1.074**	0.019
WEAT6	std. GW	**0.770**	0.050	0.725	0.085
WEAT8	pretrained	0.278	0.307	0.706	0.101
WEAT8	adversarial	0.161	0.384	**1.188**	0.013
WEAT8	adv. GW	0.565	0.137	**0.902**	0.047
WEAT8	standard	0.426	0.217	0.782	0.084
WEAT8	std. GW	0.096	0.438	**1.004**	0.034

directly, Table 9 presents masked and unmasked results for WEAT8, which was the only test that both detected significant bias in many cases and is closely related to the STEM/non-STEM classification setting. It is noteworthy that the pretrained and all *masked* fine-tuned models demonstrate some indication of bias, whereas p-values for *unmasked* fine-tuned models generally deviate widely from significance threshold values. This could be a result of the presence of gender-marking words in the dataset, which contains anti-stereotypical correlations that may mitigate stereotypical relations between the model's representations of gender-marking words and science/art words.

5 Discussion

5.1 Downstream performance and fairness trade-offs

Our experiments provide mixed evidence regarding the relationship between fairness interventions and

Table 9: WEAT8 intrinsic bias results for masked and unmasked training. Boldface indicates significant bias ($ES > 0, p < 0.05$ or $ES < 0, p > 0.95$). Italics indicate noteworthy measurements that may indicate bias ($p \approx 0.1$).

Masked	Method	WEAT ES	WEAT p	LPBS ES	LPBS p
-	pretrained	0.278	0.307	*0.706*	0.101
N	adv.	**-1.230**	0.993	0.175	0.403
Y	adv.	0.161	0.384	**1.188**	0.013
N	adv. GW	-0.175	0.636	0.275	0.309
Y	adv. GW	0.565	0.137	**0.902**	0.047
N	std.	-0.148	0.611	0.426	0.224
Y	std.	0.426	0.217	*0.782*	0.084
N	std. GW	0.475	0.187	**1.156**	0.012
Y	std. GW	0.096	0.438	**1.004**	0.034

downstream task performance. Across most settings, fairness gaps were already small in the baseline models, suggesting that the dataset itself does not strongly encode gender correlations with the STEM/non-STEM task. Consequently, masking and adversarial debiasing had only modest effects on group fairness metrics.

At the same time, these interventions often improved downstream accuracy. This superficially appears to contradict the commonly discussed *fairness–performance trade-off*. However, that trade-off typically arises in settings where fairness interventions explicitly alter outcome distributions relative to ground truth labels. In contrast, our methods aim to reduce the model's reliance on sensitive attributes while preserving task-relevant information. In this sense, the improvements we observe suggest that reducing gender information in model representations can be compatible with maintaining or even improving predictive performance.

GenWriter rewrites produced the largest gains in downstream accuracy. This may reflect a standardizing effect of rewriting, reducing stylistic variability across bios and may make task-relevant signals easier for the model to learn. However, improvements in performance did not translate into consistent improvements in group fairness metrics. In several settings, accuracy improvements were unevenly distributed across groups, resulting in small fairness gaps despite overall performance gains. This is likely due in part to the relatively weak correlation between gender and STEM/non-STEM labels in the NTNU dataset, which limits the magnitude of downstream disparities that can arise during training.

5.2 Procedural fairness and gender leakage

One of the clearest patterns in our results concerns gender leakage from hidden representations. Masking names and pronouns substantially reduced recoverable gender information, and adversarial training further reduced leakage. GenWriter rewrites provided an additional reduction, particularly when combined with adversarial debiasing.

These results support the idea that data-level and model-level mitigation methods can be complementary. Masking removes explicit gender markers, rewriting reduces stylistic or structural signals that may correlate with gender, and adversarial training discourages the model from encoding gender information in its representations. Together, these interventions produced near-random gender predictability in several configurations.

From the perspective of the research questions, this provides the strongest evidence for **RQ1** and **RQ3**. The combined mitigation strategy substantially reduces gender leakage while preserving or improving downstream performance, suggesting that procedural fairness objectives can be supported without large performance costs in this setting.

5.3 Intrinsic bias and downstream behavior

The intrinsic bias results present a different picture. Across the WEAT and LPBS test suites, bias signals were mixed and did not consistently decrease under the interventions that most strongly reduced gender leakage. In some cases, fine-tuning or masking even increased the magnitude of certain intrinsic bias signals.

This pattern aligns with prior work showing that intrinsic bias metrics often do not correlate with downstream fairness outcomes (Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022). Our results extend these findings by showing that intrinsic bias measures can also diverge from representation-level measures such as gender leakage. In other words, a model may encode less recoverable gender information in its representations while still exhibiting stereotypical associations under certain intrinsic tests.

One possible explanation is that intrinsic bias tests rely on specific lexical associations that are only weakly connected to the downstream task. However, our results show some evidence that stereotypical encodings can be weakened by unmasked fine-tuning, provided the dataset in question contains anti-stereotypical correlations connected to the corresponding WEAT test concepts. In any case, our results support the claim that intrinsic bias measures should be interpreted as diagnostic signals rather than direct indicators of downstream fairness.

5.4 Implications for bias mitigation in hiring-related NLP

Taken together, our findings suggest several broader implications for bias mitigation in hiring-related NLP systems.

First, evaluating bias mitigation methods in isolation may provide an incomplete picture. Our experiments show that combining interventions at different stages of the pipeline can produce stronger reductions in sensitive-attribute information than any single method alone.

Second, different bias measurements capture distinct aspects of model behavior. Leakage, intrinsic bias tests, fairness gaps, and predictive performance do not move together in a simple way. This reinforces the importance of evaluating bias mitigation across multiple dimensions, especially in high-stakes domains such as hiring.

Finally, the results highlight the value of grounding bias mitigation strategies in the intended downstream application. By focusing on reducing gender leakage while preserving task-relevant information, our approach targets a form of procedural fairness that is directly relevant to recruitment scenarios, where decisions should ideally depend on job-relevant attributes rather than demographic characteristics.

6 Conclusion

This paper investigated how data-level and model-level debiasing methods interact in a hiring-related NLP setting. Using Norwegian-language academic bios and a STEM/non-STEM proxy task, we evaluated masking, GenWriter rewrites, and adversarial debiasing across downstream performance, fairness metrics, intrinsic bias tests, and gender leakage.

Regarding **RQ1**, we find that the interventions interact in complementary ways. Masking and adversarial training substantially reduce gender leakage, and GenWriter rewrites further strengthen this effect. The combination of all three methods reduces recoverable gender information in model representations to near-random levels while maintaining, and in some settings improving, downstream task performance.

For **RQ2**, reductions in gender leakage do not consistently correspond to reductions in intrinsic bias as measured by WEAT or LPBS. Intrinsic bias signals remain mixed across models and tests, reinforcing prior findings that intrinsic bias metrics are not reliable indicators of downstream fairness behavior.

Finally, for **RQ3**, our results suggest that a downstream-oriented mitigation strategy can improve *procedural fairness*—interpreted here as limiting the model's reliance on gender information—without large performance costs. In several cases, debiasing interventions improved task accuracy while reducing gender leakage.

Overall, these findings highlight that different bias measurements capture distinct aspects of model behavior. Effective evaluation of bias mitigation in hiring-related NLP therefore requires a multi-dimensional approach that considers downstream performance, group fairness, representation leakage, and intrinsic bias simultaneously. Our results suggest that combining data-level and model-level interventions can support fairer and more robust systems, particularly in multilingual and context-specific recruitment settings.

Limitations

Several limitations should be considered when interpreting the results.

Proxy task and dataset scope. Our experiments use a proxy task—classification of academic bios into STEM and non-STEM domains—rather than real hiring decisions. Although this task captures a recruitment-related scenario in which textual candidate information is mapped to a consequential category, it does not reflect the full complexity of hiring processes, such as multi-stage evaluation, human judgment, or institutional policies. In addition, the dataset is drawn from a single institution (NTNU), which may limit generalizability to other organizations, sectors, or cultural contexts.

Limited representation of sensitive attributes. This study focuses primarily on gender and uses a binary gender representation. Because the collected dataset contained only a very small number of non-binary examples, non-binary identities could not be meaningfully modeled. This is an important limitation, as real-world hiring systems must account for gender diversity beyond binary categories. We also do not examine intersectional fairness involving attributes such as ethnicity, nationality, religion, or disability. Future work should prioritize datasets that support intersectional analysis and include underrepresented groups.

Model scope. Our experiments focus on BERT-based encoder models rather than modern large language models (LLMs). This choice reflects the classification-oriented nature of the task and enables the use of established adversarial debiasing techniques that operate on hidden representations. Encoder models also remain common in practical pipelines due to efficiency, cost, and data governance considerations. Nevertheless, LLMs are increasingly used in hiring-related applications, and future research should investigate how the mitigation strategies studied here translate to generative and instruction-tuned systems.

Evaluation limitations. Although we evaluate multiple dimensions—including downstream performance, group fairness gaps, intrinsic bias tests, and representation leakage—these metrics capture only partial aspects of fairness. Automated labeling steps used in the dataset construction may also introduce noise. More comprehensive evaluation would require larger annotated datasets and human-centered studies that examine how such systems affect real hiring workflows and stakeholder perceptions of fairness.

Acknowledgments

This work is part of the Europe Horizon project BIAS, grant agreement number 101070468, funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

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

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American*

Chapter of the Association for Computational Linguistics, pages 1693–1706. Association for Computational Linguistics.

Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63.

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.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. *arXiv preprint arXiv:2101.10001*.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Towards equal opportunity fairness through adversarial learning. *arXiv preprint arXiv:2203.06317*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. In *Proceedings of 5th SwissText and 16th KONVENS Joint Conference 2020*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5502–5515.

Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. 2023. Fair enough: Searching for sufficient measures of fairness. *ACM Transactions on Software Engineering and Methodology*, 32(6):1–22.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163.

Shweta Soundararajan and Sarah Jane Delany. 2025. Genwriter: Reducing gender cues in biographies through text rewriting. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 347–357.

Ewoenam Kwaku Tokpo, Pieter Delobelle, Bettina Berendt, and Toon Calders. 2023. How far can it go? on intrinsic gender bias mitigation for text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3418–3433.

Schrasing Tong, Minseok Jung, Ilaria Liccardi, and Lalana Kagal. 2026. Measuring perceptions of fairness in ai systems: The effects of infra-marginality. *arXiv preprint arXiv:2603.05889*.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix

A.1 Sample Bios and Rewrites

The following samples were created using ChatGPT-generated fictional bios, which were run through our implementation of the Genwriter pipeline.

Sample 1:

Gender: Female
Profession: Associate Professor, non-STEM

Original Bio: *Dr. Ingrid Halvorsen er førsteamanuensis ved Institutt for kunst- og medievitenskap ved NTNU. Hun forsker på visuell kultur, digital estetikk og kritisk medieteori, med særlig*

fokus på hvordan digitale plattformer og algoritmer påvirker produksjon og tolkning av kunst og medier. Hun underviser i visuell analyse og digital kultur, veileder studenter på master- og ph.d.-nivå og publiserer i internasjonale tidsskrifter innen medievitenskap og visuell kultur.

Translation: *Dr Ingrid Halvorsen is an associate professor at the Department of Art and Media Studies at NTNU. Her research focuses on visual culture, digital aesthetics and critical media theory, with a particular emphasis on how digital platforms and algorithms influence the production and interpretation of art and media. She teaches visual analysis and digital culture, supervises Master's and PhD students, and publishes in international journals in the fields of media studies and visual culture.*

Retrieved Template Bio: *Dr. [Navn på personen] er [Tittel] ved [Institutt] ved [Universitet]. [Navn på personen]s forskning sentrerer seg rundt [Felt] og [Felt], samt anvendelser av disse teknologiene i [Felt], [Felt] og [Felt]. [Navn på personen] underviser i [Fagområde], [Fagområde] og [Fagområde], samt veileder [Studienivå]- og [Studienivå]studenter.*

Translation: *Dr. [Name of person] is [Title] at [Department] at [University]. [Name of person]'s research focuses on [Field] and [Field], as well as applications of these technologies in [Field], [Field] and [Field]. [Name of person] teaches [Subject area], [Subject area] and [Subject area], and supervises [Level of study] and [Level of study] students.*

Rewritten Bio: *Dr. Ingrid Halvorsen er førsteamanuensis ved Institutt for kunst- og medievitenskap ved NTNU. Ingrid Halvorsens forskning sentrerer seg rundt visuell kultur og digital estetikk, samt anvendelser av disse teknologiene i kritisk medieteori, [Felt] og [Felt]. Ingrid Halvorsen underviser i visuell analyse, digital kultur og [Fagområde], samt veileder master- og ph.d.-studenter.*

Translation: *Dr Ingrid Halvorsen is an associate professor at the Department of Art and Media Studies at NTNU. Ingrid Halvorsen's research focuses on visual culture and digital aesthetics, as well as the applications of these technologies in critical media theory, [Field] and [Field]. Ingrid Halvorsen teaches visual analysis, digital culture and [Subject area], and supervises Master's and PhD students.*

Sample 2:

Gender: Male
Profession: PhD Candidate, STEM

Original Bio: *Erik Lundgaard er ph.d.-stipendiat ved Institutt for fysikk ved NTNU. Forskningen hans omhandler kvantematerialer og hvordan elektroniske interaksjoner gir opphav til nye kollektive fenomener i faste stoffer. Han arbeider særlig med beregningsbaserte metoder for å modellere kvantemekaniske systemer. Lundgaard deltar også i undervisning i grunnleggende fysikk og samarbeider med forskere i internasjonale prosjekter.*
Translation: *Erik Lundgaard is a PhD fellow at the Department of Physics at NTNU. His research focuses on quantum materials and how electronic interactions give rise to new collective phenomena in solids. He works in particular with computational methods for modelling quantum mechanical systems. Lundgaard also teaches basic physics and collaborates with researchers on international projects.*

Retrieved Template Bio: *[Navn på personen] er [Yrke] ved [Institutt] ved [Universitet]. Forskningstema [Navn på personen] er opptatt av er [Felt], [Felt], [Felt] og [Felt]. Jobber med [Felt] og modellering av [Felt] for [Felt]. [Navn på personen] har særlig forsket og undervist på temaer knyttet til [Felt], [Felt] og [Felt], og [Han/Hun] har deltatt i ulike forskningsprosjekter med nasjonal og internasjonal finansiering innenfor disse områdene.*
Translation: *[Name of person] is a [Occupation] at [Department] at [University]. The research topics [Name of person] is interested in are [Field], [Field], [Field] and [Field]. Works with [Field] and modelling of [Field] for [Field]. [Name of person] has conducted research and taught on topics related to [Field], [Field] and [Field], and [He/She] has participated in various nationally and internationally funded research projects within these areas.*

Rewritten Bio: *Erik Lundgaard er ph.d.-stipendiat ved Institutt for fysikk ved NTNU. Forskningstema Erik Lundgaard er opptatt av er kvantematerialer, elektroniske interaksjoner, kollektive fenomener i faste stoffer og kvantemekaniske systemer. Jobber med beregningsbaserte metoder og modellering av kvantemekaniske systemer for å studere kollektive fenomener i faste stoffer. Erik Lundgaard har særlig forsket og undervist på temaer knyttet til kvantematerialer, grunnleggende fysikk og kvan-*

temekaniske systemer, og han har deltatt i ulike forskningsprosjekter med nasjonal og internasjonal finansiering innenfor disse områdene.
Translation: *Erik Lundgaard is a PhD fellow at the Department of Physics at NTNU. Erik Lundgaard's research interests include quantum materials, electronic interactions, collective phenomena in solids and quantum mechanical systems. He works with computational methods and modelling of quantum mechanical systems to study collective phenomena in solids. Erik Lundgaard has conducted research and taught on topics related to quantum materials, fundamental physics and quantum mechanical systems, and he has participated in various nationally and internationally funded research projects in these areas.*

A.2 Prompts

Category labels for sentences: "You are classifying Norwegian sentences extracted from professional academic bios into the following categories:

- Education: Degrees, institutions attended, and academic achievements.

- Demographics: Age, gender, nationality, current job title, location and other personal background information.

- Work: Professional experience, job titles, affiliations, research interests and discussion, publications and contributions to the field.

- Non-Professional: Personal hobbies and interests, and other non-professional information.

Here are some examples of sentences and their corresponding categories:

[EXAMPLES]

Please classify the following sentence into one of the categories mentioned above:

[SENTENCE]

Please provide only the category name as your answer."

Synthetic non-professional examples: "Du hjelper til med å utvide et datasett bestående av profesjonelle akademiske biografier fra NTNU-ansatte.

Skriv en setning som er litt annerledes enn følgende eksempel:

[EKSEMPEL]

Oppgi kun setningen som svar."

Translation: "You are helping to expand a dataset consisting of professional academic biographies of NTNU staff.

Write a sentence that is slightly different from the following example:

[EXAMPLE]

Please provide only the sentence as your answer."

Creation of generic sentence templates: "Du er en streng omskriver. Gjør om enhver gitt setning til en generell mal ved å konsekvent erstatte ALLE entiteter og pronomener med norske plassholdere som [Navn på personen], [Yrke], [Spesialisering], [Universitet], [År], [Antall år], [Sykehus], [Sted], [Han/Hun], osv., nøyaktig som vist i eksemplene under. Ikke erstatt førstepersonspronomen som «jeg» eller førstepersons eieform som «min».

Setningene kommer fra profesjonelle biografier, og målet er å erstatte detaljer som er spesifikke for den enkelte med plassholdermerker. Enheter som ikke er relatert til dette målet, trenger ikke å erstattes.

Regler:

- Returner KUN den omskrevne setningen. Ingen forklaringer, ingen ekstra tekst, ingen "Input" eller "Output".

- Hvis input er tom, kun navn/tittel, eller <5 ord, returner input uendret.

- Bruk konsistente plassholdere gjennom hele teksten.

- Hvis noe ikke matcher, behold ordet hvis det ikke er entitet, ellers utelat.

- Behold alltid setningens grammatikk og struktur.

Eksempler:

Dr. Ole Nilsen er en ortopedisk kirurg spesialisert innen artroskopisk kirurgi for kneet.

→ Dr. [Navn på personen] er en [Yrke] spesialisert innen [Spesialisering].

Dr. Hansen fullførte medisinstudiet ved Universitetet i Oslo i 1995 og har vært i praksis i 28 år.

→ Dr. [Navn på personen] fullførte [Studium] ved [Universitet] i [År] og har vært i praksis i [Antall år].

Hun arbeider ved Akershus universitetssykehus med sitt team i Lørenskog, Oslo universitetssykehus i Oslo og har erfaring fra St. Olavs hospital i Trondheim.

→ [Han/Hun] arbeider ved [Sykehus] med sitt team i [Sted], [Sykehus] i [Sted] og har erfaring fra [Sykehus] i [Sted].

Jeg har erfaring med prosjekter innen digital læring.

→ Jeg har erfaring med prosjekter innen [Felt]."

Translation: "You are a strict rewriter. Convert any given sentence into a general template by consistently replacing ALL entities and pronouns with Norwegian placeholders such as [Name of person], [Occupation], [Specialisation], [University], [Year], [Number of years], [Hospital], [Location], [He/She], etc., exactly as shown in the examples below. Do not replace first-person pronouns such as 'I' or first-person possessives such as 'my'.

The sentences are taken from professional biographies, and the aim is to replace details specific to the individual with placeholder tags. Entities not related to this aim do not need to be replaced.

Rules:

- Return ONLY the rewritten sentence. No explanations, no extra text, no 'Input' or 'Output'.
- If the input is empty, contains only a name/title, or has fewer than 5 words, return the input unchanged.
- Use consistent placeholders throughout the text.
- If something does not match, retain the word if it is not an entity, otherwise omit it.
- Always retain the sentence's grammar and structure.

Examples:

Dr Ole Nilsen is an orthopaedic surgeon specialising in arthroscopic knee surgery.

→ Dr [Name of person] is a [Profession] specialising in [Specialisation].

Dr Hansen completed his medical degree at the University of Oslo in 1995 and has been practising for 28 years.

→ Dr [Name of person] completed [Degree] at [University] in [Year] and has been practising for [Number of years].

She works at Akershus University Hospital with her team in Lørenskog, Oslo University Hospital in Oslo, and has experience from St. Olavs Hospital in Trondheim.

→ [He/She] works at [Hospital] with his/her team in [Place], [Hospital] in [Place] and has experience from [Hospital] in [Place].

I have experience with projects in digital learning.

→ I have experience with projects in [Field]."

Rewriting bios: "Gitt følgende biografi og mal, utfør følgende trinn:

1. Forstå biografien og malen: Les og analyser biografien og malen nøye for å forstå konteksten, plassholdere og tilgjengelig informasjon.
2. Erstatt plassholdere: Erstatt hver plassholder i malen med passende verdier utledet fra biografien. Bruk følgende regler når du erstatter plassholdere: - Hold formatet og strukturen til malen uendret. - Hvis en plassholder ikke kan erstattes på grunn av utilstrekkelig informasjon i biografien, behold plassholderen som den er.
3. Utdata: Gi kun den endelige utfylte malen med plassholdere erstattet der det er mulig.
4. Rett opp avvik."

Translation: "Given the following biography and template, carry out the following steps:

1. Understand the biography and template: Read and analyse the biography and template carefully to understand the context, placeholders and available information.
2. Replace placeholders: Replace each placeholder in the template with appropriate values derived from the biography. Use the following rules when replacing placeholders: - Keep the format and structure of the template unchanged. - If a placeholder cannot be replaced due to insufficient information in the biography, leave the placeholder as it is.
3. Output: Provide only the final completed template with placeholders replaced where possible.
4. Correct discrepancies."