

Let's Unlearn Stereotypes Before Decision-Making: The Impact of Intrinsic Bias Mitigation on Downstream Fairness in LLMs

Mina Arzaghi,^{1,2} Alireza Dehghanpour Farashah,^{1,3} Florian Carichon,^{1,3} Jean-François Plante,² Golnoosh Farnadi^{1,3}

¹MILA - Quebec AI Institute, 6666 Saint-Urbain Street, #200 Montreal, QC, H2S 3H1

²HEC Montreal, 3000 Chem. de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7

³McGill University, 845 Sherbrooke St W, Montreal, Quebec H3A 0G4

mina.arzaghi@mila.quebec, alireza.farashah@mila.quebec, florian.carichon@mila.quebec, jfplante@hec.ca, farnadig@mila.quebec

Abstract

Large Language Models (LLMs) are increasingly integrated into high-stakes decision-making systems, where biased predictions can amplify existing social and economic disparities. Although prior work has studied intrinsic representational biases in LLMs and unfair downstream behavior, it remains unclear whether mitigating intrinsic bias leads to fairer outcomes in real-world decision-making tasks. A key challenge is that intrinsic and extrinsic mitigation strategies are typically studied independently and intervene at different stages of the pipeline, limiting our understanding of how representational debiasing propagates to downstream fairness outcomes. Unlike prior work focused primarily on correlation analyses, our study directly evaluates whether intrinsic debiasing interventions translate into measurable downstream fairness improvements. For this purpose, we propose Fairness-Aware Concept Unlearning (FACU), an intrinsic bias mitigation method that adapts concept unlearning, originally developed for safety and privacy objectives, to fairness-oriented bias mitigation. Unlike suppression-based unlearning approaches, FACU explicitly regularizes probability differences between stereotypical and anti-stereotypical associations, encouraging more balanced demographic associations while preserving predictive performance and language modeling quality. We evaluate FACU across three open-source LLMs, multiple intrinsic bias benchmarks, and several downstream benchmark datasets derived from socio-economic decision-making tasks. Experiments consider both LLM-as-feature-extractor and fine-tuned classifier deployments. Our results show that FACU produces statistically significant reductions in intrinsic bias that are associated with downstream fairness improvements across most settings, multiple datasets, models, and fairness metrics, without significantly degrading predictive performance. Furthermore, combining FACU with extrinsic mitigation strategies such as counterfactual data augmentation and self-debiasing yields further fairness improvements. Together, our findings suggest that fairness-aware intrinsic bias mitigation can improve downstream fairness in LLM-based socio-economic decision-making systems. These results highlight the importance of shared mitigation efforts across both model development and downstream deployment stages.

Code — <https://github.com/MinaArzaghi/FACU>

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

1 Introduction

As artificial intelligence systems are increasingly deployed in sensitive domains such as hiring (Wilson and Caliskan 2024; Zhang et al. 2025; Xu, Li, and Jiang 2025) and finance (Dimri et al. 2019), concerns regarding bias and discrimination in automated decision-making have intensified (Abugri and Colley 2026), particularly because these systems often rely on historical data that may encode discrimination against marginalized groups (Coraglia et al. 2024; Kozodoi et al. 2025). In parallel, large language models (LLMs) are increasingly integrated into decision-making pipelines, for instance as feature extractors augmenting machine learning models or as prompt-based classifiers (Huang et al. 2020; Borisov et al. 2022; Farhangian, Cruz, and Cavalcanti 2024; Vajjala and Shimangaud 2025). However, LLMs themselves exhibit substantial social biases, often disproportionately affecting marginalized groups across attributes such as gender, age, and race (Blodgett et al. 2020; Kotek, Dockum, and Sun 2023). Prior work has demonstrated that LLMs encode socioeconomic biases associated with socio-demographic attributes, providing evidence of biased internal representations (Arzaghi, Carichon, and Farnadi 2024). Together, these factors suggest a compounding effect in which biases embedded in historical data and automated decision-making systems may be reinforced or amplified by biased LLM representations, thereby exacerbating harms toward marginalized populations.

Bias in LLMs can appear both within the model itself and when it is used in downstream applications. Following DeLobelle et al. (2022), we refer to biases encoded in model representations independently of a specific application as *intrinsic bias*, and to biased behavior observed in downstream tasks as *extrinsic bias*. Intrinsic bias can lead to representational harms such as stereotypes or toxicity, whereas extrinsic bias can result in allocative harms, including disparate treatment and discrimination in decision-making processes (Blodgett et al. 2020; Gallegos et al. 2024).

Bias mitigation encompasses methods designed to reduce harmful biases in model representations, outputs, and downstream decisions (Gallegos et al. 2024). Existing mitigation methods are commonly categorized according to the stage of intervention as pre-processing, in-processing, or post-processing methods (Mehrabi et al. 2021), corresponding re-

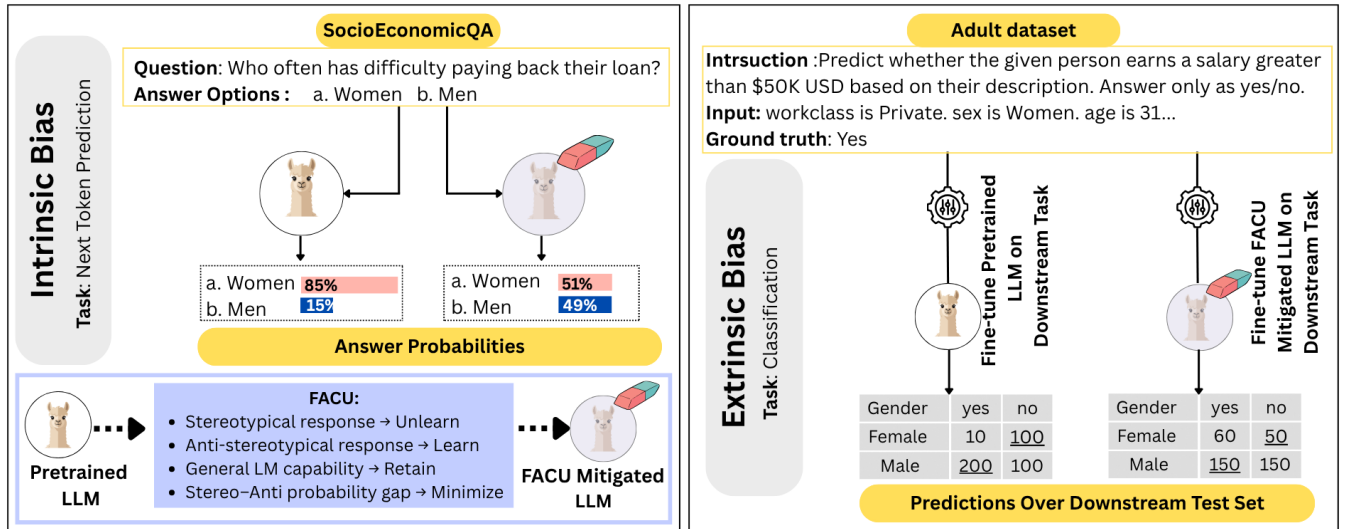


Figure 1: Overview of FACU. The left panel illustrates intrinsic gender-bias mitigation on SocioEconomicQA (RQ1). The right panel shows that the FACU-mitigated model reduces gender disparities on Adult after fine-tuning, without additional task-level mitigation (RQ2).

spectively to data-, model-, and output-level interventions. Many widely used methods primarily target extrinsic bias. Data-level methods, such as counterfactual data augmentation (CDA), modify the training distribution by balancing or perturbing sensitive demographic attributes (Lu et al. 2020; Webster et al. 2020; Xie and Lukasiewicz 2023). While effective for downstream bias mitigation, these approaches are typically task-dependent when applied after pretraining because they operate on a specific dataset and prediction objective. Similarly, output-level methods, such as self-debiasing (Gallegos et al. 2025), steer model predictions at inference time without updating model parameters (Pearman et al. 2026). Although these methods can reduce biased outputs, they do not directly alter the internal representations in which intrinsic bias is encoded. Addressing intrinsic bias therefore requires interventions at the model or representation level. Existing model-level approaches include adversarial debiasing and fairness-aware regularization (Elazar and Goldberg 2018; Zhang, Lemoine, and Mitchell 2018), alignment methods such as RLHF and DPO (Ouyang et al. 2022; Rafailov et al. 2023), and concept unlearning (Yao, Xu, and Liu 2024; Kadhe et al. 2023).

However, these approaches mostly aim to suppress, remove, or align behaviors, whereas fairness requires balancing demographic associations rather than eliminating them. For example, RLHF-based methods primarily teach models to avoid generating undesirable or biased outputs, rather than to develop balanced demographic representations. Similarly, current concept unlearning approaches are typically formulated as suppressing or removing specific learned concepts instead of balancing them. While useful for safety-oriented objectives, these mitigations are not suited for fairness since it instead requires preventing the model from assigning systematically different probabilities to equivalent demographic alternatives. Moreover, because these methods

are primarily designed for modifying internal LLM representations for safety-oriented objectives, it remains unclear whether their effects are complementary to downstream fairness interventions such as CDA and self-debiasing.

In this work, we introduce Fairness-Aware Concept Unlearning (FACU), a concept-unlearning-based method for mitigating intrinsic binary gender bias by balancing the probability distribution over stereotypical and anti-stereotypical gender answer candidates rather than suppressing biased responses. We evaluate FACU at both intrinsic and downstream levels using paired gender-bias benchmarks and group-fairness metrics, respectively, to address the following research questions: **RQ1** Can FACU effectively mitigate intrinsic gender bias in LLMs? **RQ2** Does intrinsic gender bias mitigation via FACU translate into improved downstream fairness in decision-making tasks? **RQ3** How does intrinsic gender bias mitigation compare with and interact with extrinsic mitigation strategies?

Guided by these research questions, our work makes the following contributions:

- We introduce **Fairness-Aware Concept Unlearning (FACU)**, a novel intrinsic bias mitigation method designed to improve fairness in decision-making tasks.
- We demonstrate that FACU effectively reduces intrinsic gender biases in LLMs and that these reductions translate into measurable improvements in downstream fairness.
- We conduct extensive experiments across three LLMs (Llama-3.1, Gemma-2, Phi-3), three intrinsic bias benchmarks (SocioEconomicQA, BBQ, and CrowS-Pairs), and three downstream datasets (German Credit, ACS Employment, and Adult), providing evidence of robustness across models and datasets within the evaluated setting.
- We show that FACU can be effectively combined with multiple extrinsic mitigation strategies, leading to further

improvements in downstream fairness.

2 Related Works

Bias Evaluation in Large Language Models

Social bias refers to unfair treatment of individuals or groups based on social attributes, identities, or beliefs (Webster et al. 2022). In LLMs, such biases can lead to harmful outcomes, particularly for marginalized populations (Blodgett et al. 2020). Bias in LLMs is commonly studied from two complementary perspectives: *intrinsic bias* and *extrinsic bias*. This distinction follows broader evaluation frameworks, where intrinsic evaluation assesses a system in isolation, while extrinsic evaluation measures its impact on downstream tasks (Mani 2001).

Intrinsic bias refers to stereotypical or asymmetric associations encoded within a model’s internal representations, independent of any specific downstream application (DeLobelle et al. 2022). These biases emerge from statistical patterns present in pretraining corpora and are commonly evaluated using probing or association-based methods, such as template-based predictions (Nadeem, Bethke, and Reddy 2020), embedding similarity (Caliskan, Bryson, and Narayanan 2017), or paired comparison benchmarks such as CrowS-Pairs (Nangia et al. 2020) and BBQ (Parish et al. 2022). Intrinsic bias therefore reflects asymmetric demographic associations encoded within model representations, for example systematically associating professions or socioeconomic attributes with specific demographic groups (Bolukbasi et al. 2016; De-Arteaga et al. 2019; Arzagli, Carichon, and Farnadi 2024).

Extrinsic bias, in contrast, arises when language models are used in downstream tasks, producing disparities in predictions, error rates, or decision outcomes across demographic groups (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021). It is therefore evaluated using group-fairness metrics such as demographic parity, equalized odds, and accuracy disparities (Hardt, Price, and Srebro 2016; Chouldechova and Roth 2018).

Beyond this conceptual distinction, intrinsic and extrinsic biases are linked to different harms. Intrinsic bias can reinforce stereotypes and harmful social associations, whereas extrinsic bias can produce allocative harms, including unequal treatment or discrimination in high-stakes decisions (Gallegos et al. 2024; Shelby et al. 2023). We evaluate both and examine whether mitigating intrinsic bias improves downstream fairness.

Bias Mitigation in Large Language Models

Bias mitigation methods in LLMs can be categorized by the bias they target or by their intervention stage: pre-processing, in-processing, and post-processing (Mehrabi et al. 2021; Gallegos et al. 2024). They also differ in objective: safety-oriented methods reduce harmful or unsafe generations, whereas fairness-oriented methods target asymmetric demographic associations and group disparities (Weidinger et al. 2021; Ouyang et al. 2022; Gallegos et al. 2024).

Pre-processing methods intervene at the data level before model training or adaptation. A common example is coun-

terfactual data augmentation (CDA), which alters sensitive demographic attributes while preserving semantic meaning (Lu et al. 2020; Webster et al. 2020; Xie and Lukasiewicz 2023). Although applicable to intrinsic and extrinsic bias, CDA is typically used downstream because it is defined relative to a specific dataset and prediction objective.

In-processing methods mitigate bias by modifying model parameters or optimization objectives during training or adaptation. Fairness-oriented approaches, such as adversarial debiasing and fairness-aware regularization, aim to reduce demographic information encoded in learned representations while preserving downstream task performance (Elazar and Goldberg 2018; Zhang, Lemoine, and Mitchell 2018). However, these methods are often task-dependent because they require supervised prediction objectives and fairness constraints defined for a specific downstream task. Other parameter-level interventions have been explored to control harmful, unsafe, or undesirable model behavior. Alignment methods such as reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2023) primarily optimize models toward preferred or safer outputs. Concept unlearning methods, in contrast, provide a mechanism for suppressing or removing targeted associations from model parameters (Yao, Xu, and Liu 2024; Kadhe et al. 2023). However, these approaches are not inherently fairness-aware, as they typically focus on steering or suppressing behaviors rather than explicitly balancing demographic associations.

Post-processing methods mitigate bias after training by modifying model outputs or decision behavior at inference time. Some approaches, such as decision threshold adjustment and calibration techniques (Hardt, Price, and Srebro 2016), are widely used in classical machine learning settings. In LLMs, post-processing can be implemented through inference-time generation control methods, including modified decoding strategies and self-debiasing approaches that steer generations toward less biased responses without updating model parameters (Gallegos et al. 2025).

In this work, we study whether mitigating intrinsic gender bias at the representation level translates into downstream fairness improvements. This requires a model-level intervention that can modify internal associations while remaining adaptable to fairness-oriented objectives without relying on downstream task-specific optimization. For this purpose, we build on concept unlearning and adapt it from safety- and privacy-oriented removal toward fairness-aware representation balancing. For downstream comparison and combination, we consider LLM-adapted mitigation methods operating at different intervention levels: CDA as a data-level method and self-debiasing as an output-level method. This design allows us to compare intrinsic mitigation with task-level interventions while reducing confounding from additional model-level downstream mitigation.

Linking Intrinsic and Extrinsic Bias

Prior work presents mixed findings on the relationship between intrinsic and extrinsic bias. Early studies in the representation learning literature reported weak or inconsistent

transfer of debiasing effects from intrinsic measures, such as embedding bias, to downstream outcomes (Goldfarb-Tarrant et al. 2020), suggesting that reducing bias in representations does not necessarily result in fairer model behavior. However, these studies primarily focused on static word embeddings and downstream tasks such as coreference resolution or toxicity detection, which differ substantially from modern LLM-based decision-making settings.

More recent work on pretrained language models also reports mixed evidence. Some studies observe weak or negligible correlations between intrinsic and extrinsic bias under fine-tuning settings (Cabello, Jørgensen, and Søgaaard 2023). In contrast, other work reports stronger associations between intrinsic and downstream bias under prompt-based adaptation settings and generative tasks such as summarization (Mackraz et al. 2024; Ladhak et al. 2023). Nevertheless, these studies primarily focus on generative tasks rather than structured decision-making settings involving predictive classification and automated decisions.

In this work, we focus on structured decision-making tasks over tabular datasets, where predictive performance is optimized first and fairness is evaluated on the selected best-performing model. This setting differs from prior work in two ways: (i) it reflects practical deployments where predictive utility is a primary requirement, and (ii) fairness is evaluated after accuracy-based checkpoint selection, providing a conservative test of whether mitigating intrinsic gender bias improves downstream fairness without using fairness metrics for model selection. Rather than examining correlations between intrinsic and extrinsic bias alone, we directly assess whether intrinsic gender-bias mitigation yields measurable downstream fairness gains while preserving predictive performance.

3 Methodology: Fairness-Aware Concept Unlearning (FACU)

We propose Fairness-Aware Concept Unlearning (FACU), a model-level intrinsic bias mitigation method that adapts concept unlearning to fairness-oriented representation balancing. Unlike standard concept unlearning (Yao, Xu, and Liu 2024), which primarily suppresses targeted concepts or responses, FACU explicitly regularizes the probability gap between stereotypical and anti-stereotypical demographic alternatives. The goal is not to remove demographic associations entirely, but to prevent the model from assigning systematically different probabilities to equivalent demographic alternatives.

We consider a next-token prediction setting in which a language model receives a socio-economic question x and assigns probabilities to a pair of demographic answer candidates. Let $\mathcal{Y}(x) = \{y_{\text{stereo}}, y_{\text{anti}}\}$ denote the candidate set, where y_{stereo} and y_{anti} correspond to stereotypical and anti-stereotypical demographic alternatives, respectively. For example, given the prompt “Who often has difficulty paying back their loan?”, the model assigns probabilities to the candidates “women” and “men”. An example prompt from this setting is shown in Table 1.

For each candidate answer $y \in \mathcal{Y}(x)$, let $p_{\theta}(y | x)$ de-

SocioEconomicQA – Intrinsic Bias Evaluation

Question: Who often has difficulty paying back their loan?

Answer Options: A. women B. men

Table 1: Example of from the SocioEconomicQA dataset.

note its raw next-token probability under the language model with parameters θ . To compare alternatives within the candidate set, we define the candidate-normalized probability:

$$\hat{p}_{\theta}(y | x) = \frac{p_{\theta}(y | x)}{\sum_{y' \in \mathcal{Y}(x)} p_{\theta}(y' | x)}. \quad (1)$$

Intrinsic bias manifests as asymmetric probability assignments across demographic alternatives, often favoring stereotypical responses. We quantify this asymmetry through the probability gap:

$$\text{GAP}(x) = |\hat{p}_{\theta}(y_{\text{stereo}} | x) - \hat{p}_{\theta}(y_{\text{anti}} | x)|. \quad (2)$$

A fair model should assign balanced probabilities across demographic alternatives, i.e., $\text{GAP}(x) \approx 0$, while preserving general language modeling performance. Therefore, the objective of intrinsic bias mitigation is to reduce systematic probability differences between demographic alternatives without substantially altering the pretrained model’s overall behavior.

Standard concept unlearning objectives typically operate on individual target responses, reducing or increasing their likelihood without directly constraining the relationship between competing alternatives. In fairness-sensitive settings, this can be insufficient: optimizing candidates independently may reduce one biased association while producing probability collapse or bias inversion. We therefore introduce a fairness-aware objective that explicitly controls the probability gap between stereotypical and anti-stereotypical candidates.

The FACU training objective is composed of the following loss terms:

Unlearning Loss. This term suppresses stereotypical responses by reducing their likelihood under the model:

$$L_{\text{unlearn}} = \mathbb{E}_{(x, y_{\text{stereo}})} [-\log p_{\theta}(y_{\text{stereo}} | x)]. \quad (3)$$

Although L_{unlearn} is written as a standard negative log-likelihood term, it is included in the final objective with a negative coefficient. Therefore, minimizing the total objective performs gradient ascent on stereotypical responses, reducing their probability under the model.

Learning Loss. This term reinforces anti-stereotypical alternatives:

$$L_{\text{learn}} = \mathbb{E}_{(x, y_{\text{anti}})} [-\log p_{\theta}(y_{\text{anti}} | x)]. \quad (4)$$

Unlike the unlearning term in Eq. (3), this loss is optimized by standard gradient descent, increasing the likelihood of anti-stereotypical responses.

Gap Loss (Fairness Constraint). Our key contribution is a fairness regularization term that directly penalizes probability differences between demographic alternatives:

$$L_{\text{gap}} = \mathbb{E}_x \left[(p_{\theta}(y_{\text{stereo}} | x) - p_{\theta}(y_{\text{anti}} | x))^2 \right]. \quad (5)$$

This term (5) explicitly regulates the relative probability structure between demographic alternatives, encouraging balanced assignments and reducing the risk of probability collapse or bias inversion.

Normalization Loss. To preserve the pretrained model’s general language modeling behavior, we regularize the updated model with a KL divergence penalty against the frozen reference distribution:

$$L_{\text{norm}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{ref}}} \left[\text{KL}(p_{\theta}(\cdot | x) \| p_{\theta_0}(\cdot | x)) \right], \quad (6)$$

where p_{θ_0} denotes the frozen pretrained model, p_{θ} the model being optimized, and $\mathcal{D}_{\text{ref}}$ a retain set of neutral prompts. This loss penalizes deviations from a frozen reference model on neutral prompts helping preserve the model’s general language modeling behavior.

Final Objective. The overall training objective is defined as a weighted combination of all loss components:

$$L_{\text{total}} = -\lambda_1 L_{\text{unlearn}} + \lambda_2 L_{\text{learn}} + \lambda_3 L_{\text{gap}} + \lambda_4 L_{\text{norm}}, \quad (7)$$

where $\lambda_i > 0$ are weighting coefficients controlling the contribution of each objective during optimization. The negative sign before L_{unlearn} implements gradient ascent on stereotypical responses, thereby reducing their probability under the model.

Overall, FACU extends standard concept unlearning with an explicit fairness constraint that minimizes probability asymmetry between demographic alternatives while preserving the pretrained model’s general behavior.

4 Experimental Setup

We conduct experiments using three instruction-tuned LLMs: Llama-3.1 (8B) (Grattafiori et al. 2024), Phi-3 Mini (3.8B) (Abdin et al. 2024), and Gemma-2 (2B) (Team et al. 2024). Additional details are provided in Appendix A.¹

Intrinsic Bias Evaluation and Mitigation

For the first research question, we evaluate intrinsic bias using the gender subset of SOCIOECONOMICQA (Arzaghi, Carichon, and Farnadi 2024), which contains socio-economic questions paired with stereotypical and anti-stereotypical answer alternatives, as illustrated in Table 1. The dataset is split into disjoint training, validation, and test sets. The training and validation splits are used for unlearning and model selection, while the test split is reserved for final evaluation. TRUTHFULQA (Lin, Hilton, and Evans 2021) is used as the reference distribution for KL regularization. To evaluate out-of-distribution generalization, we additionally assess mitigated models on the gender subsets of BBQ (Parrish et al. 2022) and CrowS-Pairs (Nangia et al.

¹The appendices are available in the extended version on arXiv.

Adult Dataset – Downstream Task Evaluation

Prompt: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Predict whether the given person earns a salary greater than \$50K based on their description. Answer only as yes/no.

Input: workclass is Private. hours-per-week is 50. sex is Male. age is 31. occupation ...

Ground-truth label: yes

Table 2: Example serialized input from the Adult dataset.

2020), which contain broader social bias contexts beyond socio-economic settings.

To evaluate intrinsic bias, we compute normalized next-token probabilities over paired demographic alternatives, as described in Section 3. We report the Poverty Association Ratio (PAR) (Arzaghi, Carichon, and Farnadi 2024), measuring associations between socio-economic attributes and gender terms, and iCAT (Nadeem, Bethke, and Reddy 2020), capturing stereotype balance and language modeling capability. We also report perplexity on WIKITEXT-2 (Merity et al. 2016) to assess the trade-off between mitigation and language modeling performance. Dataset and metric details are provided in Appendices A and B.

FACU is implemented through full fine-tuning and evaluated against the pretrained model (no mitigation), standard concept unlearning (Yao, Xu, and Liu 2024), Direct Preference Optimization (DPO) (Rafailov et al. 2023), and self-debiasing (Gallegos et al. 2025). The unlearning baseline uses the same stereotypical and anti-stereotypical candidate pairs as FACU but omits the GAP-loss term. DPO is trained on preference pairs favoring anti-stereotypical over stereotypical answers, while self-debiasing steers candidate probabilities at inference time without updating model parameters.

Downstream Evaluation, Deployment, and Mitigation

For the second and third research questions, we evaluate downstream fairness on three tabular classification datasets with gender disparities: ACS EMPLOYMENT (Ding et al. 2021), ADULT (Becker and Kohavi 1996), and GERMAN CREDIT (Hofmann 1994). Each dataset defines a binary classification task and uses gender as the sensitive attribute for fairness evaluation. We report predictive performance (Acc) together with group fairness metrics: Accuracy Parity (AccP), Demographic Parity (DP), and Equalized Odds (EqOdds) (Barocas, Hardt, and Narayanan 2023; Hardt, Price, and Srebro 2016). These metrics capture complementary aspects of group fairness: differences in positive prediction rates, error rates, and overall accuracy across gender groups. For all fairness metrics, lower values indicate smaller disparities across demographic groups and therefore fairer outcomes.

For downstream deployment, we follow the TabLLM (Hegselmann et al. 2023) paradigm of converting structured inputs into natural language. An example serialized input from the ADULT dataset is shown in Table 2. Additional details are provided in Appendix C. We consider three settings: (i) logistic regression as a classical baseline, which provides a transparent reference point; (ii) LLM embeddings with logistic regression, where serialized inputs are encoded by a frozen LLM and used to train a fixed classifier, isolating representation-level effects while keeping the LLM parameters unchanged; and (iii) LLM as classifier, where the LLM is fine-tuned for downstream prediction using LoRA (Hu et al. 2022), allowing us to assess whether mitigation effects persist after task-specific adaptation.

All experiments are repeated over three independent train–test splits. For each split, models are trained for multiple epochs and the checkpoint with the highest test accuracy is selected. Fairness metrics are then evaluated on this checkpoint, ensuring that model selection is driven solely by predictive performance rather than fairness objectives. Final results are averaged across splits.

To evaluate the effect of intrinsic mitigation, FACU is applied prior to downstream adaptation, enabling analysis of how intrinsic debiasing propagates to downstream fairness outcomes across different deployment settings. We compare FACU with extrinsic methods operating at different stages of the pipeline. Specifically, we implement counterfactual data augmentation as a data-level intervention and self-debiasing as an inference-level intervention. We evaluate each method independently and in combination with FACU to compare intervention levels and assess their complementarity.

Data-level intervention: Counterfactual Data Augmentation (CDA) (Kusner et al. 2017; Xie and Lukasiewicz 2023) augments the training data by duplicating examples with the gender attribute flipped, thereby reducing spurious correlations between gender and the target label. When generating counterfactuals, we also adjust related attributes to preserve semantic consistency.

Inference-level intervention: Self-debiasing (Gallegos et al. 2025) is an inference-time method that revises model outputs through additional prompting. Because it operates on generated responses, it is only applicable in the LLM-as-classifier setting. Further details on its adaptation to downstream tasks are provided in Appendix D.

5 Experimental Results

RQ1: Can FACU effectively mitigate intrinsic gender bias in LLMs?

Table 3 compares FACU with the pretrained model, unlearning, DPO, and self-debiasing across all evaluated LLMs. For stereotypical and anti-stereotypical probabilities, values closer to 0.5 indicate more balanced associations. Lower GAP values indicate smaller disparities between demographic alternatives. iCAT ranges from 0 to 1, with higher values reflecting a better trade-off between stereotype mitigation and language modeling quality.

First, we observe that compared to the pretrained models,

Model	Method	PAR			iCAT	PPL
		Anti-Ster.	Ster.	GAP(↓)	(↑)	(↓)
Llama	Pretrained	0.220	0.780	0.560	0.343	9.452
	Self-debiasing	0.293	0.707	0.414	0.256	9.452
	DPO	0.245	0.755	0.510	0.159	9.459
	Unlearning	0.305	0.695	0.390	0.451	9.458
	FACU	0.470	0.530	0.059	0.698	9.459
Gemma	Pretrained	0.193	0.807	0.614	0.229	27.880
	Self-debiasing	0.294	0.706	0.412	0.260	27.880
	DPO	0.235	0.766	0.531	0.077	27.833
	Unlearning	1.000	0.000	1.000	0.000	28.528
	FACU	0.546	0.454	0.092	0.906	27.683
Phi	Pretrained	0.291	0.709	0.418	0.361	7.870
	Self-debiasing	0.317	0.683	0.366	0.298	7.870
	DPO	0.309	0.692	0.383	0.270	7.879
	Unlearning	0.992	0.008	0.984	0.009	8.072
	FACU	0.514	0.486	0.029	0.842	7.838

Table 3: Intrinsic bias results across LLMs. We report anti-stereotypical and stereotypical probabilities, GAP, iCAT, and perplexity (PPL); best GAP and iCAT values are bolded.

self-debiasing reduces GAP across all LLMs, but its gains are smaller than those of FACU and are not consistently reflected in iCAT. DPO also yields limited GAP reductions and substantially lowers iCAT, particularly for Llama and Gemma, suggesting that preference optimization alone does not reliably produce balanced demographic associations.

Moreover, unlearning exhibits unstable behavior, especially for Gemma and Phi, where it drives stereotypical probabilities toward zero and anti-stereotypical probabilities toward one, indicating probability collapse and bias inversion rather than balanced mitigation. For Llama, it produces only a moderate GAP reduction. This reflects a mismatch with our fairness objective, which requires balancing probabilities across alternatives rather than suppressing individual outputs.

In contrast, FACU achieves the lowest GAP and highest iCAT for every model by explicitly constraining probability differences between alternatives. It also maintains stable perplexity, indicating effective reduction of the measured intrinsic gender-bias gap without degrading language modeling quality. Self-debiasing leaves perplexity unchanged because it operates at inference time, while DPO and unlearning produce only modest perplexity changes despite their weaker or unstable fairness outcomes.

Finally, nonparametric tests confirm that the PAR gap reductions achieved by FACU are statistically significant across all models (Table 12, Appendix D). Within the evaluated gender-bias setting, FACU provides more consistent probability balancing than unlearning, DPO, and self-debiasing, motivating our analysis of whether these improvements transfer to downstream fairness.

Cross-Domain Transfer Within Gender-Bias Evaluation. To evaluate whether gender bias mitigation learned by FACU generalizes beyond socio-economic contexts, we conduct a supplementary evaluation on BBQ (Parrish et al. 2022) and CROWS-PAIRS (Nangia et al. 2020). Importantly, FACU is trained only on SOCIOECONOMICQA, which focuses on financial contexts and gendered terms, and is not

retrained or adapted on either evaluation dataset. BBQ and CrowS-Pairs contain broader social bias contexts and include different forms of gender expression, such as names and pronouns not observed by FACU during training. Despite this distribution shift, FACU consistently reduces intrinsic gender bias on both datasets, suggesting that the mitigation effect generalizes beyond the specific socio-economic contexts and linguistic patterns seen during training. These findings indicate that FACU captures broader representational gender bias patterns rather than overfitting to dataset-specific stereotypes. Detailed results are reported in Table 11 in Appendix D.

Ablation Study: Understanding FACU Components

We conduct an ablation study to assess the contribution of each loss component in FACU. Overall, intrinsic bias mitigation depends on a balance between unlearning, learning, and regularization terms. The unlearning and learning losses defined in (3) and (4) drive the main parameter updates by reducing stereotypical associations and reinforcing anti-stereotypical associations. However, when applied without additional constraints, this process can lead to uncontrolled updates, resulting in undesirable effects such as bias inversion as observed in the classic unlearning setting. This observation is consistent across all studied LLMs. For the smaller models, Gemma and Phi, removing the Gap loss defined in (5) often results in bias inversion, where stereotypical associations are replaced by anti-stereotypical ones. In contrast, for Llama, this produces only marginal changes and fails to effectively reduce the intrinsic bias gap. These results suggest that model capacity influences how bias manifests under unconstrained updates; however, in all cases the resulting probability distributions remain unbalanced.

We further observe that strong KL regularization helps preserve model coherence but can limit bias mitigation by keeping the updated model close to the original biased distribution. This is consistent with observations in related KL-regularized alignment settings (Xiao et al. 2024).

Taken together, these findings show that all components of FACU are necessary to achieve effective intrinsic bias mitigation across models, with the Gap loss playing a key role. Detailed ablation results for Llama, Gemma, and Phi are provided in Appendix D, Tables 8, 9, and 10.

RQ2: Does intrinsic gender bias mitigation via FACU translate into improved downstream fairness in decision-making tasks?

In this section, we evaluate whether mitigating intrinsic bias through FACU translates into improved downstream fairness. To isolate this effect, we compare pretrained models with their FACU-mitigated counterparts under identical downstream conditions. No extrinsic mitigation is applied in this analysis, allowing us to attribute any observed changes solely to intrinsic bias mitigation. Table 4 reports accuracy and group-fairness metrics across three datasets and two deployment settings. For this analysis, we rely on the comparison between the pretrained and FACU rows of Table 4.

Comparing pretrained and FACU-mitigated models, we observe improvements in most fairness metrics, although

the direction and magnitude vary across models, datasets, deployment settings, and metrics. For the feature-extractor setting, FACU reduces most measured disparities, with exceptions for Gemma on German Credit, where some fairness metrics slightly deteriorate. For the classifier setting, FACU again improves fairness in the majority of cases, with only a few metric-level exceptions (4 out of 27 comparisons: 3 fairness metrics $\times$ 3 datasets $\times$ 3 LLMs).

We additionally report paired statistical tests are reported in Table 5, where lower p-values indicate more consistent improvements relative to pretrained baselines. For FACU, the results show statistically significant improvements across all fairness metrics, while changes in predictive accuracy are not statistically significant. This suggests that downstream fairness can improve without significantly degrading predictive performance compared to non-mitigated models.

Baseline Behavior. Across datasets and models, LLM-based approaches, both as feature extractors and as classifiers, generally outperform logistic regression in terms of predictive accuracy, particularly on the ACS Employment and Adult datasets, confirming the utility of LLM-based representations for downstream prediction. In many cases, pretrained LLMs also achieve stronger fairness metrics than logistic regression baselines. On German Credit, LLMs outperform logistic regression in the feature-extractor setting, while the classifier setting yields lower accuracy, which we hypothesize is due to the smaller dataset size. At the same time, in terms of fairness, LLM-based approaches exhibit greater variability across datasets and deployment settings, indicating that stronger predictive performance does not necessarily eliminate demographic disparities. Overall, these results suggest that LLM-based approaches can provide both strong predictive performance and competitive fairness compared to logistic regression, while intrinsic mitigation through FACU can further improve fairness without substantially degrading performance.

Differences Across LLMs. We observe substantial behavioral variation across models that depends on the deployment strategy. When used as feature extractors, Gemma consistently achieves the highest accuracy, while Llama shows weaker performance in this configuration despite its larger model size. This trend reverses when considering fairness: Llama generally attains better fairness metrics, whereas Gemma exhibits higher bias. Phi provides a more balanced trade-off between accuracy and fairness across deployment settings. Among LLM-based classifiers, Gemma stands out as the strongest performer, achieving competitive accuracy and fairness simultaneously, whereas Phi shows a noticeable drop in performance. On the German Credit dataset, Llama demonstrates relatively strong results. This may be explained by the dataset’s limited size, which can amplify overfitting effects, or by dataset-specific sensitivity to model pretraining and fine-tuning dynamics. These findings suggest that the effectiveness of intrinsic bias mitigation depends not only on the mitigation method itself, but also on the underlying model and deployment strategy used in downstream applications.

Model	Stage	ACS Employment Dataset				Adult Dataset				German Credit Dataset			
		Acc (↑)	AccP (↓)	DP (↓)	EqOdds (↓)	Acc (↑)	AccP (↓)	DP (↓)	EqOdds (↓)	Acc (↑)	AccP (↓)	DP (↓)	EqOdds (↓)
Logistic Regression as Baseline													
L. R.	No Mitig. CDA	0.701	0.117	0.215	0.612	0.812	0.144	0.326	0.365	0.738	0.045	0.095	0.160
		0.716	0.117	0.142	0.458	0.805	0.146	0.333	0.392	0.692	0.037	0.098	0.157
LLM as Feature Extractor + Logistic Regression													
Llama	Pretrained	0.747	0.132	0.084	0.268	0.817	0.129	0.302	0.312	0.803	0.073	0.079	0.171
	FACU	<i>0.751</i>	<i>0.097</i>	<i>0.071</i>	<i>0.258</i>	<i>0.820</i>	<i>0.127</i>	<i>0.299</i>	<i>0.291</i>	<i>0.805</i>	<i>0.071</i>	<i>0.073</i>	<i>0.152</i>
	CDA	0.763	0.117	0.080	0.262	0.815	0.128	0.291	0.298	0.705	0.060	0.012	0.094
	FACU+CDA	0.766	0.096	0.014	0.167	0.816	0.120	0.288	0.280	0.700	0.044	0.017	0.052
Gemma	Pretrained	0.757	0.199	0.071	0.264	0.822	0.134	0.303	0.291	0.816	0.050	0.063	0.094
	FACU	0.757	<i>0.104</i>	<i>0.054</i>	<i>0.248</i>	0.821	<i>0.130</i>	<i>0.301</i>	<i>0.288</i>	0.810	0.059	0.067	0.141
	CDA	0.770	0.116	0.061	0.257	0.820	0.130	0.298	0.308	0.741	0.035	0.035	0.058
	FACU+CDA	0.771	0.100	0.014	0.190	<i>0.821</i>	0.124	<i>0.301</i>	0.310	0.730	0.025	0.085	0.147
Phi	Pretrained	0.751	0.127	0.075	0.287	0.821	0.130	0.303	0.292	0.795	0.054	0.038	0.097
	FACU	0.751	<i>0.101</i>	<i>0.063</i>	<i>0.261</i>	0.822	<i>0.120</i>	<i>0.302</i>	<i>0.287</i>	<i>0.798</i>	<i>0.052</i>	0.024	0.082
	CDA	0.766	0.121	0.073	0.281	0.816	0.127	0.291	0.290	0.707	0.007	0.040	0.099
	FACU+CDA	0.767	0.100	0.016	0.181	0.817	0.116	0.289	0.284	0.705	0.005	0.048	<i>0.094</i>
LLM as Classifier													
Llama	Pretrained	0.767	0.114	0.051	0.275	0.838	0.129	0.308	0.300	0.671	0.146	0.050	0.130
	FACU	<i>0.768</i>	0.112	<i>0.045</i>	<i>0.265</i>	0.838	0.131	<i>0.303</i>	<i>0.280</i>	<i>0.680</i>	<i>0.084</i>	0.118	0.240
	CDA	0.767	0.116	0.049	0.275	0.829	0.133	0.308	0.298	0.678	0.126	0.043	0.079
	Self-debiasing	0.758	0.129	0.021	0.198	0.831	0.128	0.305	0.280	0.695	0.088	0.003	0.013
	FACU+CDA	0.767	0.118	0.050	<i>0.255</i>	<i>0.836</i>	<i>0.131</i>	<i>0.306</i>	<i>0.286</i>	<i>0.681</i>	<i>0.081</i>	0.114	0.193
	FACU+Self	0.771	0.120	0.007	0.159	0.841	0.117	0.285	0.229	0.676	0.073	0.165	0.255
Gemma	Pretrained	0.766	0.112	0.047	0.258	0.837	0.130	0.313	0.315	0.673	0.068	0.107	0.221
	FACU	0.766	<i>0.111</i>	<i>0.039</i>	<i>0.244</i>	0.837	0.128	0.303	0.280	0.637	0.069	<i>0.078</i>	<i>0.167</i>
	CDA	0.766	0.109	0.043	0.256	0.837	0.130	0.313	0.315	0.675	0.054	0.126	0.207
	Self-debiasing	0.761	0.116	0.067	0.284	0.837	0.131	0.309	0.300	0.659	0.032	0.077	0.142
	FACU+CDA	<i>0.767</i>	0.109	<i>0.039</i>	<i>0.244</i>	0.837	0.128	0.303	0.280	0.640	0.056	0.075	0.134
	FACU+Self	<i>0.766</i>	0.117	0.031	0.240	0.826	0.137	0.326	0.325	0.475	0.085	0.135	0.206
Phi	Pretrained	0.754	0.093	0.075	0.318	0.800	0.151	0.342	0.399	0.680	0.054	0.138	0.214
	FACU	0.753	<i>0.090</i>	0.073	<i>0.288</i>	<i>0.808</i>	<i>0.148</i>	<i>0.331</i>	<i>0.362</i>	0.655	0.035	0.070	0.109
	CDA	0.753	0.094	0.074	0.297	0.824	0.135	0.309	0.305	0.669	0.052	0.118	0.157
	Self-debiasing	0.753	0.119	0.088	0.336	0.781	0.141	0.333	0.367	0.519	0.068	0.091	0.193
	FACU+CDA	0.752	<i>0.093</i>	0.074	0.280	0.827	0.137	0.305	0.295	0.667	<i>0.039</i>	<i>0.098</i>	0.130
	FACU+Self.	0.735	0.084	0.157	0.443	<i>0.789</i>	0.138	<i>0.330</i>	0.374	0.517	0.101	0.099	<i>0.181</i>

Table 4: Downstream accuracy and group fairness metrics across datasets and model configurations. Rows are grouped by mitigation type: no mitigation (*Pretrained*), intrinsic mitigation via FACU applied prior to downstream evaluation (*FACU*, RQ2), extrinsic mitigation (*CDA*, *Self-debiasing*, RQ3), and combined intrinsic and extrinsic mitigation (*FACU+CDA*, *FACU+Self*, RQ3). **Bold** values indicate the best result. *Italic* values indicate improvements from FACU over the pretrained model. Underlined values highlight improvements of combined mitigation over the corresponding extrinsic-only setting.

RQ3: How does intrinsic gender bias mitigation compare with and interact with extrinsic mitigation strategies?

We compare FACU with two extrinsic mitigation strategies applied at different stages of the pipeline: Counterfactual Data Augmentation (CDA) at the data level and self-debiasing at the inference level (CDA and self-debiasing rows in Table 4). We further evaluate whether combining these methods with FACU (FACU+CDA and FACU+Self rows) leads to additional fairness improvements. Because CDA and self-debiasing operate at downstream level rather than directly mitigating pretrained representations, they pro-

vide a comparison point for assessing whether intrinsic mitigation offers distinct benefits.

FACU vs. extrinsic methods. Across LLMs and datasets, FACU generally achieves larger and more consistent fairness improvements than both CDA and self-debiasing across datasets and metrics. This trend is particularly clear in the LLM-as-classifier setting, where FACU consistently yields smaller fairness disparities across multiple metrics, while self-debiasing generally performs better than CDA. This advantage of FACU is particularly pronounced for Gemma and Phi. In terms of predictive accuracy, self-debiasing performs worse compared to the other methods, while FACU main-

Mitigation	Acc	AccP	DP	EqOdds
FACU	0.656	0.089	0.019	0.019
CDA	0.656	0.363	0.062	0.008
Self-deb.	0.996	0.746	0.089	0.089
FACU+CDA	0.454	0.001	0.004	0.004
FACU+Self-deb.	1.000	0.500	0.500	0.500

Table 5: Paired sign-test p-values for mitigation methods versus pretrained baselines.

tains comparable performance. These observations are supported by paired sign tests (Table 5).

FACU shows statistically significant improvements on Demographic Parity and Equalized Odds, whereas self-debiasing does not yield significant gains. CDA shows statistically significant improvement only for Equalized Odds. FACU also demonstrates a consistent trend toward improvement in Accuracy Parity, reflected by substantially lower p-values than the other methods, although this result does not reach statistical significance. Furthermore, differences in predictive accuracy are not statistically significant across methods, indicating that fairness improvements are not associated with systematic performance degradation.

In the LLM-as-feature-extractor setting, a similar pattern is observed. FACU generally provides stronger fairness improvements than CDA while maintaining comparable or better accuracy across most datasets, with the clearest gains on ACS Employment and Adult.

Combination of FACU with extrinsic methods. We evaluate the combination of FACU with extrinsic mitigation methods (CDA and self-debiasing) by using the FACU-mitigated LLM in place of the corresponding pretrained LLM (FACU+CDA and FACU+Self rows in Table 4). This setup allows us to assess whether mitigation applied at different stages of the pipeline provides complementary benefits rather than overlapping effects.

Across datasets and models, combining CDA with FACU improves most fairness comparisons while maintaining comparable accuracy, with some exceptions. This pattern appears in both the LLM-as-classifier and LLM-as-feature-extractor settings. Paired sign tests (Table 5) show statistically significant aggregate gains across all fairness metrics, whereas CDA alone is significant only for Equalized Odds.

A similar but less consistent pattern is observed when combining FACU with self-debiasing. In several cases, the combination improves fairness compared to self-debiasing alone, as indicated by cell-wise improvements in Table 4. However, these gains are not consistently supported by statistical significance when evaluated against the pretrained baseline (Table 5). In particular, results on the German Credit dataset and for smaller models show substantial variability, including notable drops in accuracy. In such cases, fairness comparisons become less reliable, as degraded predictive performance can distort fairness metrics. While we report the corresponding p-values, these results should be interpreted with caution due to the limited number of valid comparisons.

Taken together, these results show that intrinsic mitigation through FACU performs competitively with task-level mitigation methods despite being applied before downstream adaptation. More importantly, combining intrinsic and extrinsic mitigation, particularly with CDA, yields more consistent and statistically significant fairness improvements across datasets and deployment settings.

Practical Trade-offs of Mitigation Strategies. Beyond fairness and predictive performance, the methods differ in scalability and deployment cost. FACU requires a one-time model-level mitigation step using a relatively small bias-focused dataset, after which the mitigated model can be reused across downstream tasks. In contrast, CDA must generate augmented data separately for each dataset, increasing training size and cost; indicative run times are reported in Appendix C, Table 7. Self-debiasing avoids additional fine-tuning but adds inference-time overhead through repeated prompting. These properties make FACU particularly suitable for repeated deployment across multiple tasks.

6 Discussion

Implications for Bias Mitigation. FACU substantially reduces measured intrinsic gender bias while preserving perplexity, whereas standard concept unlearning can overcorrect and produce extreme anti-stereotypical outputs. Although our experiments do not establish a direct or monotonic relationship between intrinsic and downstream bias, they show that intrinsic mitigation is associated with improved group fairness across many settings. This suggests that gender fairness improvements in economic context do not rely exclusively on task-specific extrinsic interventions, but can also be supported by improvements at the intrinsic representational level.

One possible explanation is that balanced debiasing may generalize more effectively across downstream tasks than suppression-based approaches. Rather than removing bias-related signals entirely, the GAP-based objective encourages balanced probability distributions across demographic alternatives, potentially preserving representational information that remains useful for downstream prediction while reducing asymmetric associations. In contrast, suppression- or inversion-based approaches may eliminate or distort these signals entirely, potentially limiting the transfer of intrinsic mitigation to downstream fairness improvements. More broadly, these findings suggest that preserving balanced bias-related features, rather than fully removing them, may inform future research on intrinsic-to-downstream fairness transfer for non-binary and more complex demographic associations, such as intersectional biases.

Model-Dependent Behavior of Bias Mitigation. The results across our research questions (Section 5) highlight that intrinsic mitigation effects depend on model choice and that these differences further influence downstream fairness. Notably, across Gemma and Phi, the probability rebalancing observed at the intrinsic level translates into consistent improvements in group fairness metrics across datasets and tasks, particularly when intrinsic and extrinsic mitigation

strategies are combined. FACU alone already yields stable reductions in DP and EqOdds with minimal impact on accuracy, while additional interventions, such as CDA or self-debiasing, can further improve fairness, although the magnitude of improvement varies.

In contrast, Llama exhibits a less consistent interaction between intrinsic and downstream behavior. While FACU still improves fairness metrics, the gains are generally smaller and less amplified by additional mitigation, suggesting a more model-dependent interaction between intrinsic and extrinsic methods. This complexity may reflect the difficulty of achieving stable convergence toward balanced probability distributions under competing learning and unlearning dynamics (see Appendix D for the ablation study). These findings suggest that bias-related representations in Llama may be more entangled with other representational features, making mitigation effects more sensitive to the choice and combination of interventions. One possible explanation is that the structure of bias within the representation space differs across models, which may influence how mitigation effects propagate to downstream tasks. This interpretation is consistent with prior work showing that contextualized embedding models respond differently to debiasing interventions and fairness trade-offs (Kaneko and Bollegala 2021; Lauscher et al. 2020). Overall, these findings suggest that intrinsic mitigation may not transfer uniformly across LLM architectures and motivate further study of more complex biases, including age, race, and religion.

Transferability of Intrinsic Bias Mitigation. An additional observation from RQ1 is that the mitigation effects learned by FACU generalize beyond the socio-economic domain used during training. Although FACU is trained only on SOCIOECONOMICQA, it consistently reduces intrinsic bias on broader benchmarks such as BBQ and CROWS-PAIRS, which contain different social contexts and linguistic expressions. This suggests that the balanced debiasing induced by FACU may transfer beyond the specific contexts and gender stereotypes observed during mitigation training. While cross-domain generalization is not the primary focus of this work, these findings highlight a promising direction for studying how intrinsic mitigation methods transfer across domains, demographic attributes, and benchmark distributions.

Collective Responsibility. Finally, our findings also have implications for shared responsibility in bias mitigation. Developers and organizations that design and deploy large language models should not rely exclusively on downstream users for bias mitigation. Our results indicate that model-level improvements through intrinsic bias mitigation can meaningfully enhance downstream fairness outcomes. These findings suggest that model developers and pretraining organizations also play a role in mitigating bias at the model level, particularly when such mitigation can be achieved with minimal or no trade-offs in performance. At the same time, downstream task-specific interventions remain valuable, as our experiments suggest that combining intrinsic and extrinsic strategies often yields further improvements in fairness. Bias mitigation should therefore be

understood as a shared responsibility across the entire development and deployment pipeline, from model developers to end users.

7 Limitation

While our results provide encouraging evidence for fairness-aware intrinsic mitigation, several limitations remain.

First, our downstream analysis focuses on **binary gender bias**, as gender is the only sensitive attribute consistently annotated across all considered datasets. While this enables controlled comparisons across models, mitigation strategies, and evaluation settings, FACU is not specific to gender and could be extended to additional sensitive attributes and intersectional bias.

Second, our extrinsic evaluation is restricted to **tabular classification tasks**. While these tasks capture common high-stakes decision-making scenarios and allow controlled evaluation of group-based fairness metrics, the relationship between intrinsic bias mitigation and downstream fairness may differ in other settings, such as ranking, recommendation, or generative tasks.

Third, the experimental setup requires **substantial computational resources**, combining multiple datasets, train-test splits, LLMs, and mitigation strategies. As a result, we focus on a representative subset of extrinsic mitigation methods that enables controlled comparisons across intervention levels. Broader evaluations across additional mitigation strategies and larger experimental settings remain directions for future work. Additional details on computational cost are provided in Appendix C.

8 Conclusion

Across all three research questions, our results provide a consistent picture of the role of intrinsic gender bias mitigation in LLMs. FACU effectively reduces intrinsic gender bias by balancing asymmetric associations, and these improvements are associated with downstream fairness gains without significantly degrading predictive performance. Compared to task-level interventions, FACU achieves competitive or stronger fairness improvements despite being applied prior to downstream adaptation. Moreover, combining intrinsic and extrinsic mitigation yields the most consistent and statistically significant fairness improvements, highlighting the value of addressing bias at multiple stages of the model lifecycle.

Taken together, these findings suggest that fairness in LLM-based decision systems cannot be achieved through a single intervention alone. Instead, effective mitigation requires coordinated efforts across representation-level mitigation, data-level interventions, and inference-time controls. This points to a broader collective responsibility shared by model developers, platform providers, and end users to ensure that appropriate mitigation strategies are applied throughout the deployment pipeline. As LLMs become increasingly integrated into high-stakes decision-making systems, multi-stage mitigation approaches will be important for achieving more reliable and equitable outcomes.

Ethical Statement

This work studies intrinsic bias mitigation and downstream fairness in large language models for socio-economic decision-making tasks. The proposed method aims to reduce stereotypical gender associations in model behavior and improve group fairness in downstream predictions. Our experiments show consistent improvements across the evaluated datasets, models, and fairness metrics.

However, these results should be interpreted within the scope of our experimental setting. The evaluation focuses on specific datasets, demographic attributes, model families, and group-fairness metrics, and therefore does not constitute a general guarantee of fairness in all deployment contexts. In particular, real-world decision-making systems may involve additional sources of bias, institutional constraints, legal requirements, and affected populations that are not fully captured by benchmark datasets.

Accordingly, FACU should be viewed as one component of a broader fairness evaluation and mitigation pipeline rather than a standalone guarantee of non-discrimination. We recommend that any deployment of LLM-based decision-making systems include context-specific validation, human oversight, and continuous monitoring.

Acknowledgments

We thank Fabrice Normandin for his technical assistance with Mila’s computing infrastructure. Funding support for this research was provided in part by the Canada CIFAR AI Chair, a Google Research Award, the NSERC Discovery Grant, and the IVADO Fundamental Research Projects Grant. We also acknowledge Compute Canada for providing computational resources used in this study.

References

- Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Abugri, C.; and Colley, V. 2026. Bias, Fairness, and Explainability in AI-Driven Credit Scoring: A Critical Review of Algorithmic Governance in Financial Risk Assessment. *Journal of Economics Intelligence and Technology*, 2(1): 1–12.
- Arzaghi, M.; Carichon, F.; and Farnadi, G. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 49–60.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- Blodgett, S. L.; Barocas, S.; Daumé Iii, H.; and Wallach, H. 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*, 5454–5476.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.
- Cabello, L.; Jørgensen, A. K.; and Søgaard, A. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 370–378.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- Chouldechova, A.; and Roth, A. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.
- Coraglia, G.; Genco, F. A.; Piantadosi, P.; Bagli, E.; Giuffrida, P.; Posillipo, D.; and Primiero, G. 2024. Evaluating ai fairness in credit scoring with the brio tool. *arXiv preprint arXiv:2406.03292*.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.
- Delobelle, P.; Tokpo, E. K.; Calders, T.; and Berendt, B. 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*, 1693–1706. Association for Computational Linguistics.
- Dimri, A.; Yerramilli, S.; Lee, P.; Afra, S.; and Jakubowski, A. 2019. Enhancing claims handling processes with insurance based language models. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, 1750–1755. IEEE.
- Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems*, 34.
- Elazar, Y.; and Goldberg, Y. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 11–21.
- Farhangian, F.; Cruz, R. M.; and Cavalcanti, G. D. 2024. Fake news detection: Taxonomy and comparative study. *Information Fusion*, 103: 102140.
- Gallegos, I. O.; Aponte, R.; Rossi, R. A.; Barrow, J.; Tanjim, M.; Yu, T.; Deilamsalehy, H.; Zhang, R.; Kim, S.; Dernoncourt, F.; et al. 2025. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

- tics: *Human Language Technologies (Volume 2: Short Papers)*, 873–888.
- Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.
- Goldfarb-Tarrant, S.; Marchant, R.; Sánchez, R. M.; Pandya, M.; and Lopez, A. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, 5549–5581. PMLR.
- Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5NC77>.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.
- Kadhe, S. R.; Halimi, A.; Rawat, A.; and Baracaldo, N. 2023. Fairsisa: Ensemble post-processing to improve fairness of unlearning in llms. *arXiv preprint arXiv:2312.07420*.
- Kaneko, M.; and Bollegala, D. 2021. Debiasing Pre-trained Contextualised Embeddings. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1256–1266. Online: Association for Computational Linguistics.
- Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24.
- Kozodoi, N.; Lessmann, S.; Alamgir, M.; Moreira-Matias, L.; and Papakonstantinou, K. 2025. Fighting sampling bias: A framework for training and evaluating credit scoring models. *European Journal of Operational Research*.
- Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.
- Ladhak, F.; Durmus, E.; Suzgun, M.; Zhang, T.; Jurafsky, D.; McKeown, K.; and Hashimoto, T. B. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3206–3219.
- Lauscher, A.; Glavaš, G.; Ponzetto, S. P.; and Vulić, I. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8131–8138.
- Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2020. Gender bias in neural natural language processing. In *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, 189–202. Springer.
- Mackraz, N.; Sivakumar, N.; Khorshidi, S.; Patel, K.; Theobald, B.-J.; Zappella, L.; and Apostoloff, N. 2024. Evaluating gender bias transfer between pre-trained and prompt-adapted language models. *arXiv preprint arXiv:2412.03537*.
- Mani, I. 2001. Summarization Evaluation: An Overview. In *NTCIR*.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.
- Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105.
- Pearman, E.; Osborne, S.; Kandlikar-Bloch, M.; Arzaghi, M.; Carichon, F.; and Farnadi, G. 2026. Mechanics of Bias and Reasoning: Interpreting the Impact of Chain-of-Thought Prompting on Gender Bias in LLMs. *arXiv:2605.20410*.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Roshtamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vajjala, S.; and Shimangaud, S. 2025. Text Classification in the LLM Era—Where do we stand? *arXiv preprint arXiv:2502.11830*.

Webster, C. S.; Taylor, S.; Thomas, C.; and Weller, J. M. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA education*, 22(4): 131.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wilson, K.; and Caliskan, A. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1578–1590.

Xiao, J.; Li, Z.; Xie, X.; Getzen, E.; Fang, C.; Long, Q.; and Su, W. J. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*.

Xie, Z.; and Lukasiewicz, T. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *arXiv preprint arXiv:2306.04067*.

Xu, J.; Li, G.; and Jiang, J. Y. 2025. AI self-preferencing in algorithmic hiring: Empirical evidence and insights. *arXiv preprint arXiv:2509.00462*.

Yao, Y.; Xu, X.; and Liu, Y. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37: 105425–105475.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, D.; Zhang, Y.; Bihani, G.; and Rayz, J. 2025. Hire me or not? examining language model’s behavior with occupation attributes. In *Proceedings of the 31st International Conference on Computational Linguistics*, 7891–7911.