LLM Bias Detection and Mitigation through the Lens of Desired Distributions

Ingroj Shrestha
University of Iowa
ingroj-shrestha@uiowa.edu

Padmini Srinivasan
University of Iowa
padmini-srinivasan@uiowa.edu

Abstract

Although prior work on bias mitigation has focused on promoting social equality and demographic parity, less attention has been given to aligning LLM's outputs to desired distributions. For example, we might want to align a model with real-world distributions to support factual grounding. Thus, we define bias as deviation from a desired distribution, which may be an equal or real-world distribution, depending on application goals. We propose a weighted adaptive loss[1] based fine-tuning method that aligns LLM's gender–profession output distribution with the desired distribution, while preserving language modeling capability. Using 3 profession sets—male-dominated, female-dominated, and gender-balanced—derived from U.S. labor statistics (2024), we assess both our adaptive method for reflecting reality and a non-adaptive variant for equality. Across three masked language models, bias is observed under both distributions. We achieve near-complete mitigation under equality and 30–75% reduction under real-world settings. Autoregressive LLMs show no bias under equality but notable bias under real-world settings, with the Llama Instruct models (3.2-3B, 3.1-8B) achieving a 50–62% reduction.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across a range of natural language processing (NLP) tasks. However, this success is tempered by the presence of social and representational biases (Gallegos et al., 2024; Guo et al., 2022; Kaneko and Bollegala, 2022; Nadeem et al., 2021). The computer science (CS) literature on LLMs bias typically considers any differences in association between attribute values (e.g., male and female for gender attribute) in a given context (e.g., profession) as an indication of bias. Psychology also views such differences as bias. However,

when a model reflects the real world, CS still sees this as bias, whereas psychology considers it an accurate reflection of reality. Thus, there are two bias viewpoints. Bias is (1) any deviation from an equal (50-50) distribution—often captured by fairness notions such as demographic parity, equalized odds, or equal opportunity (Gallegos et al., 2024; Mehrabi et al., 2021)— regardless of real-world distributions (2) deviation from real-world distributions. Less attention has been given in the CS bias literature to this second viewpoint, important in certain applications, such as healthcare, where genetic and biological predispositions (e.g., age, gender) make equality undesirable for LLMs in contexts like health chatbots, and precision medicine. The problem may be generalized as one where the aim is to align LLM to a user-specified distribution. We address the two specific cases of equal versus real-world distributions. It is of course possible for the two to be the same. In both cases, fairness can be achieved by adjusting the model's distribution to the desired distribution, which we achieve by fine-tuning.

Fairness with the first bias viewpoint has been explored extensively in CS (Gallegos et al., 2024; Delobelle et al., 2022; Guo et al., 2022; Stanczak and Augenstein, 2021). Prior work typically measures bias at a granular level using attribute–target combinations, often with or without templates, where gendered attribute words (e.g., male- and female-associated terms) are paired with target concepts (e.g., professions) to compute association scores that quantify the strength of gender–profession associations. These associations are estimated using either template-based probe sentences or corpus-based contextual embeddings (Shi et al., 2024; Yang et al., 2023; Guo et al., 2022; Limisiewicz and Mareček, 2022; Garimella et al., 2021), and evaluated on out-of-distribution bias benchmarks such as WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018) to test generalization be-

---

[1] Our code and data are available at https://github.com/IngrojShrestha/bias_through_lens_of_desired_distributions.

yond the training templates. In contrast, our work shifts focus to a coarser, distributional perspective: we analyze the gender distribution across professions, for example, for *dental assistant* it is: 8% male, 92% female, and aim to align these distributions to assess and mitigate bias.

Fairness with the second bias viewpoint requires more attention. When models do not reflect reality —whether factual or perceptual— they are likely to generate misleading information and hallucinations it is critically important to address such hallucinations (Sahoo et al., 2024; Niu et al., 2024; Su et al., 2024). Aligning LLMs with real-world trends also enhances fairness in fact-checking, thus, increasing model reliability and trustworthiness. Recent techniques such as Retrieval Augmentation Retrieval (RAG) (Lewis et al., 2020) and Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022; Ziegler et al., 2019) are designed to guide models toward producing outputs grounded in factual information and real-world context. Our second view of bias and fairness aligns with these.

Based on the selected bias viewpoint, we propose a bias mitigation strategy that fine-tunes the LLMs using a tailored loss function to recalibrate their output distribution towards a desired target. While existing methods often focus on reducing task-specific disparities, less attention is given to the foundational prediction behavior of pre-trained LLMs. In contrast, we align LLM's output distribution with a desired target distribution during fine-tuning while preserving performance—measured by *MLM loss* and downstream GLUE evaluation for masked language models (MLMs) and by *perplexity* and the LM Evaluation Harness on 5 benchmarks for auto-regressive language models (ALMs). To promote balanced and stable learning across profession groups (male-dominated, female-dominated, balanced), we introduce a weighted adaptive KL loss that dynamically adjusts updates based on group specific dynamics.

Contributions of our research:

1. In addition to debiasing towards *equality*, we also define debias by aligning a deviation from a *desired distribution*, even if this reflects inequality across groups. For equality, unlike prior fine-grained methods, our approach mitigates bias at a coarser-grained level across groups.

2. We propose a weighted adaptive loss-based approach to mitigate bias by aligning LLM's output distribution with a desired distribution, while preserving language modeling capability.

2 Method

We follow the standard template-based approach to estimate bias in MLMs (Gallegos et al., 2024; Cimitan et al., 2024; Nozza et al., 2022) and extend this approach to ALMs. A template refers to a sentence structure that includes *attribute* (a demographic group against which bias is studied), *target* (the context or domain in which the bias is analyzed), and other neutral words. We focus on a single attribute-target pair: gender (attribute) and profession (target). We chose profession as the target due to the availability of real-world gender distributions to ground our bias analysis.

We use six templates (Appendix A.1 Table 5) to analyze bias in relation to gender-profession distribution. The first five templates are adapted from Bartl et al. (2020), and we added the final one. This aligns with the common practice of using 2-5 templates as discussed in Shrestha et al. (2025).
Attributes: We used 11 pairs of binary gender-denoting words (Appendix Table 6), adapted from Bartl et al. (2020).
Targets ($\mathcal{R}$): We used 225 profession data from Bureau of Labor Statistics (2024) ($\geq$ 50k employed), grouped by female participation: *male-dominated* (DP_{male}: [0-30%]), *female-dominated* (DP_{female}: [70,100%]), and *balanced* ($DP_{balanced}$: [45,55%]).

Instead of extreme cutoffs (Bartl et al., 2020), we use [0, 30%] and [70,100%] to include both strongly and moderately dominated professions. The [45,55%] range approximates a nearly equal gender distribution, with ±5% margin to allow natural variations and maintain balanced representation.

We also shortened the profession titles, as in Bartl et al. (2020), to improve compatibility with the LLMs vocabulary by increasing the likelihood that profession terms would appear within it. Titles were lowercased, singularized, and simplified profession titles to reflect primary roles (e.g., *Railroad conductors and yardmasters* became *conductor*).

2.1 LLM's gender-profession distribution

For MLMs, we followed Shrestha et al. (2025) to measure the gender-profession association score. We mask the attribute (ideally, a gendered word) in the probe sentence derived from the template and compute the likelihood of predicting the original gendered word. To account for the possibility that LLM could be overly trained on a particular gender, we also compute a prior by masking both the attribute and target, and estimating the likelihood of

the attribute. Association score S is then obtained as the log-likelihood ratio of these two likelihoods.
Normalized gender-profession distribution: After computing the association score between each male/female gendered word and profession across templates (Appendix Table 5), we aggregate and normalize these scores such that the resulting gender distribution for each profession sums to 1.

Let $\mathcal{M}$ ($\mathcal{F}$) denote the set of male- (female-) gendered words. For gender $g \in \{\text{male}, \text{female}\}$, we define the corresponding word set as $\mathcal{G}_g = \mathcal{M}$ if $g = \text{male}$, and $\mathcal{G}_g = \mathcal{F}$ otherwise. The variable a denotes a gendered word selected from $\mathcal{G}_g$.

Let $S_{a,r,t}^{(g)}$ represent the association score for gender g, gendered word a, profession r, in template t. We compute the aggregated score $S_r^{(g)}$ across all templates and words associated with gender g. The normalized gender-probability distribution $p_{\text{pred}}^{(r)}(g)$ is then obtained by applying a softmax over genders for each profession $r \in \mathcal{R}$, as shown in Eq. 2

$$S_r^{(g)} = \sum_{t \in T} \sum_{a \in \mathcal{G}_g} \exp\left(S_{a,r,t}^{(g)}\right) \tag{1}$$

$$p_{\text{pred}}^{(r)}(g) = \frac{S_r^{(g)}}{\sum_{g' \in \{\text{male}, \text{female}\}} S_r^{(g')}} \tag{2}$$

For ALM, we use sentence loss as a proxy for association score, as in Hossain et al. (2023). Since higher loss indicates weaker association, we negate the exponent in Eq.1 and use $\exp\left(-S_{a,r,t}^{(g)}\right)$ instead of $\exp\left(S_{a,r,t}^{(g)}\right)$. This ensures that lower loss (stronger association) yields higher gender-profession distribution.

2.2 Bias detection

Bias is quantified using the Kullback–Leibler (KL) divergence between the predicted distribution $p_{\text{pred}}^{(r)}(g)$ and the desired distribution $p_{\text{true}}^{(r)}(g)$. $p_{\text{pred}}^{(r)}(g)$ denotes the predicted gender distribution for profession r obtained from LLM, and $p_{\text{true}}^{(r)}(g)$ is the corresponding desired distribution.

To detect bias, we compute the KL divergence between predicted and desired gender distributions for each profession r, averaging over male and female distributions. (Eq. 3[2]).

$$D_{\text{KL}}\left(p_{\text{true}}^{(r)} \parallel p_{\text{pred}}^{(r)}\right) = \frac{1}{2} \sum_{g \in \{\text{male}, \text{female}\}} p_{\text{true}}^{(r)}(g) \log\left(\frac{p_{\text{true}}^{(r)}(g)}{p_{\text{pred}}^{(r)}(g)}\right) \tag{3}$$

[2]Mathematically, $D_{KL}(P \parallel Q)$ measures how much the predicted distribution Q deviates from the true distribution P. In PyTorch, it quantifies how closely model predictions align with the target, effectively capturing the deviation of LLM outputs from the true distribution during fine-tuning.

Bias score: The final bias score is the average KL divergence across all professions (Eq. 4).

$$\text{BiasScore} = \frac{1}{|\mathcal{R}|} \sum_{r=1}^{|\mathcal{R}|} D_{\text{KL}}^{(r)} \tag{4}$$

where, $D_{\text{KL}}^{(r)}$ represents KL divergence for profession r. An ideal unbiased model is one with bias score close to zero.

2.3 Bias mitigation

We fine-tune gender-profession distribution in LLM to a desired distribution using templates, gendered words, and profession. This is done separately for two targets: *equal* (50-50) and *real-world* distributions, with one model for each. We used all three categories — DP_{male}, $\text{DP}_{\text{female}}$ and $\text{DP}_{\text{balanced}}$ for both fine-tuning and evaluation.

Non-adaptive KL Loss ($\mathcal{L}_{\text{KL,uniform}}$): To guide the model towards the desired distribution, we define loss as KL divergence of the LLM predicted gender-profession distribution ($p_{\text{pred}}^{(r)}(g)$) from the desired distribution ($p_{\text{true}}^{(r)}(g)$) across professions. Fine-tuning minimizes this loss. The overall loss, $\mathcal{L}_{\text{KL,uniform}}$ (Eq. 5), gives equal weight to each profession, regardless of its profession category.

$$\mathcal{L}_{\text{KL,uniform}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathcal{L}^{(r)} = \text{BiasScore} \tag{5}$$

$$\mathcal{L}^{(r)} = D_{\text{KL}}\left(p_{\text{true}}^{(r)} \parallel p_{\text{pred}}^{(r)}\right) \tag{6}$$

Weighted adaptive KL loss ($\mathcal{L}_{\text{KL,weighted\_adaptive}}$): To better balance learning across profession categories, we propose a weighted adaptive loss approach not previously explored in the context of bias mitigation. Instead of computing a uniform loss (Eq. 5), we make the loss computation profession-group aware. This design was motivated by validation-set analysis, where some groups (e.g., in BERT-base: DP_{male} : 0.232/0.087, $\text{DP}_{\text{female}}$: 0.038/0.001 and $\text{DP}_{\text{balanced}}$: 0.085/0.007) showed higher KL means and variances, indicating greater deviation and instability. This motivated the use of both adaptive loss and stability-aware weighting.

Adaptive loss: During tuning, we group each training batch $\mathcal{B}$ by profession category, $c \in \{\text{DP}_{\text{male}}, \text{DP}_{\text{female}}, \text{DP}_{\text{balanced}}\}$. We compute an adaptive loss for a profession category by dividing the current batch's KL divergence loss, $\mathcal{L}_{\text{cur}}^{(c)}$, by the exponentially updated moving average (shown in Eq. 8, where β refers to momentum parameter controlling how much weight is given to the old KL mean versus the current batch KL mean) of the

KL loss for that profession category, $\mu_{\text{KL,new}}^{(c)}$, as shown in Eq. 7. This KL mean —updated with each batch— captures how the model has historically performed on that group, including the current batch. Computing adaptive loss in this way ensures that groups with consistently high KL divergence (i.e., larger deviation from the desired distribution) do not disproportionately dominate the overall loss, thus promoting balanced learning across all groups.

$$\hat{\mathcal{L}}_{\text{cur}}^{(c)} = \frac{\mathcal{L}_{\text{cur}}^{(c)}}{\mu_{\text{KL,new}}^{(c)} + \alpha^{(c)}} \quad (7)$$

$$\mu_{\text{KL,new}}^{(c)} = \beta \cdot \mu_{\text{KL,old}}^{(c)} + (1 - \beta) \cdot \mathcal{L}_{\text{cur}}^{(c)} \quad (8)$$

Adaptive loss scaling: To further control update magnitude, we add a small constant $\alpha^{(c)}$ to the denominator (Eq. 7). α is set lower (higher) for the profession category with higher (lower) KL divergence from the target distribution, allowing for larger (smaller, more cautious) updates thus helping regulate how aggressively the model should adapt to each group's loss. Profession category with higher or lower KL are identified using validation set statistics before adjusting the distribution.

Stability aware weighting to adaptive loss: After normalization, we apply an adaptive weighting factor $\lambda^{(c)}$ (Eq. 10), computed from the variance of KL divergence for group c so far ($\text{Var}^{(c)}$), using Welford's online algorithm (Welford, 1962). This variance-based weight captures the stability of predictions—groups with higher variance, indicating less stability, are assigned lower weights, leading to slower, more conservative updates. In contrast, lower-variance (more stable) groups are weighted more heavily, enabling faster adaptation.

$$\text{VarFactor}^{(c)} = \frac{1}{1 + \text{Var}^{(c)}} \quad (9)$$

$$\lambda^{(c)} = \begin{cases} \max\left(\min(0.95 \cdot \mu \cdot V,\ 1.5),\ 0.8\right) & \text{if } c \in \text{high-KL group} \\ \max\left(\min(1.2 \cdot \mu \cdot V,\ 1.5),\ 1.0\right) & \text{otherwise} \end{cases} \quad (10)$$
$$\mu = \mu_{\text{KL,new}}^{(c)}, \quad V = \text{VarFactor}^{(c)}$$

Overall weighted adaptive loss $\mathcal{L}_{\text{KL,weighted\_adaptive}}$ (Eq. 11) ensures that model updates are fair across groups and responsive to each group's learning dynamics. Adaptive loss balances weight updates across categories, preventing domination by high-loss groups; adaptive loss scaling controls the magnitude of updates based on initial profession category loss; and stability-aware weighting adjusts update rate based on group stability. Overall loss is computed by averaging the weighted adaptive losses over all profession-group batches. Adaptive weighting is applied only during fine-tuning to guide learning.

During evaluation, we compute the bias score using the original KL divergence formulation (Section 2.2), i.e., no adaptive weighting is applied during detection.

$$\mathcal{L}_{\text{KL,weighted\_adaptive}} = \frac{1}{|\mathcal{B}|} \sum_{c \in \mathcal{B}} \lambda^{(c)} \cdot \frac{\mathcal{L}_{\text{cur}}^{(c)}}{\mu_{\text{KL,new}}^{(c)} + \alpha^{(c)}} \quad (11)$$

MLM Loss: We combine KL divergence loss with an MLM loss as a secondary objective to retain masked language modeling ability while adjusting the MLM distribution. MLM loss is computed on probe sentences—derived from training templates, professions, and gendered words—by masking one token at a time and averaging the likelihood of the original tokens, following Salazar et al. (2020). Since most training probes are short (92.5% $\leq$ 8 words; only 7.5% have 9) we compute the loss across all tokens instead of masking 15% at random as in Devlin et al. (2019). The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{KL}} + \gamma \cdot \mathcal{L}_{\text{MLM}}$$

where γ is a hyperparameter that controls the relative importance of the MLM loss. Since KL divergence is our primary loss for bias mitigation, we control only the MLM loss via γ, using $\gamma \in \{0.001, 0.01, 0.1, 0.2, 0.5, 0.8, 1.0\}$

Unlike MLMs, where a separate MLM loss is added as a secondary objective to preserve language modeling ability during fine-tuning, ALMs inherently optimize for next-token prediction given prior context. In our setup, we use sentence loss as a proxy for association score, which already reflects the model's perplexity and thus captures its language modeling capability. As a result, there is no need to include a separate objective to preserve language modeling ability when fine-tuning ALMs.

3 Experiment Design

3.1 Models assessed

We evaluate bias across three MLMs: DistilBERT (Sanh et al., 2019) and two BERT variants (bert-base-uncased and bert-large-uncased) (Devlin et al., 2019), and two families of autoregressive Instruct models: Llama3 (3.1-8B, 3.2-3B, 3.3-70B) and Qwen2 (2.5-7B, 2.5-72B)[3]. We assess bias in all models but focus mitigation on Llama3.1-8B-Instruct and Llama3.2-3B-Instruct due to resource limitations. For consistency, we fine-tuned all models using case-insensitive probe sentences.

For MLMs, we performed full fine-tuning given their smaller sizes (66M–340M parameters). For

---

[3]Llama3 and Qwen2.5 are case-sensitive, but we used case-insensitive inputs during fine-tuning for consistency. This may slightly affect performance relative to its intended usage.

ALMs, we used parameter-efficient fine-tuning via LoRA (<7B) and QLoRA (≥7B, for memory efficiency at larger scales) given its significantly larger size. LoRA and QLoRA have been shown to perform well on Llama models (Xin et al., 2024), achieving competitive results with much lower memory and compute by updating only a small number of low-rank matrices instead of the full model weights. This makes them a practical and effective choice for large-scale ALMs.

3.2 Dataset split

We use a 65%-15%-20% stratified training-validation-testing split.

Attributes and Targets: We use the same set of gendered pairs across training, validation, and testing. However, since we shift profession distributions, we use distinct profession sets for each split (Appendix Table 7).

Templates: We use the same set of templates for training and validation, while a different set of templates for testing. We use 3 templates for each. We split the templates into training/validation and testing by balancing the selection of common or rare templates. Note that sentences derived from a template with lower pseudo-perplexity are common sentences. We use a cut-off of sentence pseudo-perplexity 15 to categorize templates.

T1 and T6 are rare, with fewer than half of their sentences having perplexity below 15, indicating less predictable language (Appendix Table 8). In contrast, T2–T5 are more common, with over 70% of sentences below the threshold, suggesting more natural, fluent patterns. We select T1-T3 for training/validation, and the rest for testing, balancing one rare and two common templates in each split.

3.3 Language modeling capability evaluation

To assess whether our bias mitigation impacts language modeling, we evaluated model performance on two external corpora: the GAP Corpus (Webster et al., 2018) and WikiText-103 (development and test sets) (Merity et al., 2017). For MLMs, we report MLM loss; for ALM, we report sentence perplexity, which reflects next token prediction quality.

To further assess preservation of language modeling capability, we also evaluated on downstream tasks: MLMs on GLUE tasks and ALMs using the LM Evaluation Harness (Gao et al., 2024) across 5 benchmarks—HellaSwag, LAMBADA (OpenAI), TruthfulQA (generation), MMLU, and GLUE—covering text generation, question answering, classification, and commonsense inference. While MLMs were fine-tuned and evaluated with case-insensitive inputs, ALM perplexity was computed case-sensitively, matching the model's original training setup for fair evaluation.

Model hyperparameters and selection methodology are detailed in Appendix A.4.

4 Baseline

Our method introduces a unique debiasing strategy, particularly for *real-world* distributions (Section 2.3) and adopts a stricter bias mitigation setting in specific contextual scenarios (e.g., profession), where fairness is defined with respect to distributional shifts. Since prior work instead defines fairness primarily under an *equal* distribution, this difference makes direct comparison under *real-world* distributions not possible. So, we only report baseline comparisons in the *equal* distribution setting, to ensure consistency with prior methods.

AttenD: Gaci et al. (2022) mitigates bias by fine-tuning the attention heads to equalize attention distribution across demographic word pairs (e.g., gender: he/she) in context. We trained AttenD using our training dataset. Three MLMs (DistilBERT, BERT-base, BERT-large) were evaluated with the hyperparameters reported in the paper, considering all attention heads, as their results showed this yields better performance.

Counterfactual Data Substitution (CDS): Bartl et al. (2020) mitigates bias by fine-tuning on a gender-swapped version of the GAP corpus, where gendered words and names are substituted to create balanced training data.

5 Results

Once we select the best configuration (using seed 42) based on the validation set, we provide the results averaged across five seeds as in Hansen et al. (2024). We adapt the seed values 42, 52, 62, 72, 82 from Zhou et al. (2025).

5.1 Results for debiasing MLMs

Tables 1 - 3 present our results on debiasing MLMs. *Base Model* refers to the original model without debiasing, while the rest of the rows represent the effect of using our loss functions to adjust MLM's distribution to a desired distribution.

For *equal* target distribution, we adjust MLM's distribution to a 50%-50% male-female distribution for each profession. So, we only apply non-adaptive loss, with equal weight across professions.

Desired Distribution	Model	Profession Category				MLM loss			Epoch	$\mathcal{B}$	β	γ
		DP_{male} (KL)	DP_{female} (KL)	$DP_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)				
equal	Base Model	0.020	0.045	0.029	0.032	0.477	0.419	0.423				
	AttenD	3.7E-4	3.8E-4	2.6E-4	3.6E-4	10.8	10.8	10.8				
	% drop	98.1%†	99.2%†	99.1%†	98.9%†	-2166%	-2490%	-2466%				
	+ $\mathcal{L}_{\mathbf{KL,uniform}}$	8.6E-5	1.4E-5	1.0E-5	1.1E-5	0.515	0.444	0.448	4	5	-	-
	% drop	99.6%†	99.7%†	99.6%†	99.6%†	-8.0%	-6.0%	-5.9%				
real-world	Base Model	0.189	0.066	0.028	0.107	0.477	0.419	0.423				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$	0.052	0.028	0.020	0.036	0.573	0.483	0.488	7	5	0.95	-
	% drop	72.4%†	57.0%†	27.5%†	66.4%†	-20.1%	-15.4%	-15.5%				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$ - α	0.051	0.031	0.019	0.037	0.583	0.489	0.494	8	5	0.95	-
	% drop	72.9%†	52.4%†	30.2%†	65.7%†	-22.2%	-16.7%	-16.8%				
	+ $\mathcal{L}_{\mathbf{KL,uniform}}$	0.049	0.037	0.022	0.039	0.510	0.440	0.444	8	5	-	-
	% drop	74.2%†	43.2%†	20.4%†	63.9%†	-6.9%	-5.0%	-5.1%				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$ + $\mathcal{L}_{\mathbf{MLM}}$	0.070	0.029	0.019	0.043	0.510	0.438	0.443	7	5	0.95	0.2
	% drop	63.2%†	56.4%†	31.3%†	59.8%†	-6.9%	-4.7%	-4.9%				

Table 1: Bias mitigation and language modeling performance: MLM loss, GLUE evaluation (Appendix Table 13). Results are averaged across five seed runs (**DistilBERT**). *Base Model* refers to pre-trained MLMs. % drop indicates reduction in bias or MLM loss of fine-tuned model relative to *Base Model*. **Baseline**: *AttenD*. † indicates a statistically significant bias reduction. ALL: includes professions from all three profession categories. $\mathcal{B}$: training batch size. β: weight is given to the old KL mean versus the current batch KL mean in adaptive loss, γ: relative importance to MLM loss. Values for epochs, $\mathcal{B}$ and γ are those that yielded the best validation dataset performance.

Desired Distribution	Model	Profession Category				MLM loss			Epoch	$\mathcal{B}$	β	γ
		DP_{male} (KL)	DP_{female} (KL)	$DP_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)				
equal	Base Model	0.046	0.164	0.060	0.096	0.474	0.438	0.446				
	AttenD	3.4E-4	3.2E-4	3.4E-4	3.3E-4	10.8	10.8	10.8				
	% drop	99.3%†	99.8.%†	99.4%†	99.7%†	-2176%	-2368%	-2325%				
	+ $\mathcal{L}_{\mathbf{KL,uniform}}$	5.2E-4	3.1E-4	6.0E-4	4.4E-4	0.496	0.449	0.457	3	5	-	-
	% drop	98.9%†	99.8%†	99.0%†	99.5%†	-4.7%	-2.4%	-2.3%				
real-world	Base Model	0.270	0.040	0.059	0.136	0.474	0.438	0.446				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$	0.063	0.040	0.016	0.044	0.554	0.491	0.497	6	5	0.60	-
	% drop	76.7%†	-0.3%	73.5%†	67.3%†	-17.1%	-12.2%	-11.5%				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$ - α	0.073	0.039	0.013	0.047	0.595	0.518	0.523	8	5	0.60	-
	% drop	73.1%†	2.4%	77.5%†	65.1%†	-25.6%	-18.2%	-17.1%				
	+ $\mathcal{L}_{\mathbf{KL,uniform}}$	0.065	0.044	0.020	0.048	0.505	0.458	0.465	8	5	-	-
	% drop	76.0%†	-10.3%	65.6%†	64.9%†	-6.6%	-4.6%	-4.2%				
	+ $\mathcal{L}_{\mathbf{KL,weighted\_adaptive}}$ + $\mathcal{L}_{\mathbf{MLM}}$	0.067	0.039	0.012	0.045	0.521	0.469	0.475	6	5	0.60	0.2
	% drop	75.1%†	2.3%	79.6%†	66.9%†	-10.0%	-7.1%	-6.5%				

Table 2: Bias mitigation and language modeling performance (**BERT-base**). See Table 1 for cell values and notation details.

For the *real-world* target distribution, where professions vary in gender dominance, we applied a weighted adaptive KL loss to mitigate bias.

5.1.1 Equal distribution

Across all MLMs, applying uniform KL loss results in a consistent and substantial bias reduction. In all cases, bias was almost completely removed, with reduction exceeding 98%, shifting the MLM's predicted gender distribution for each profession to a negligible deviation from the ideal 50%-50% male-female distribution. This mitigation was observed consistently across all three profession categories, and ALL (all profession categories). All the reductions are statistically significant (95% confidence level), as determined by *independent t-tests*.

Evaluating MLM loss on external corpora before and after bias mitigation, we find only small degradation, indicating that the language modeling capabilities were well preserved. Specifically, MLM loss increased by 2.3% to 14.5% (GAP: 4.7–14.5%, WikiText-103-dev: 2.3–11.5%, WikiText-103-test: 2.4–11.8%). For DistilBERT and BERT-base, degradation was minor ($<$ 8%), while BERT-large showed slightly higher degradation (around 11–14%). Additionally, across MLMs,

GLUE scores (Appendix Table 13 'debiased for equal' rows) remain consistent before and after debiasing, indicating preserved language modeling capabilities. Overall, the results demonstrate that bias mitigation through uniform KL loss achieves near-complete bias removal with minimal compromise to language modeling performance.

5.1.1.1 Baseline comparison

AttenD: Results are presented in the rows "AttenD" (Tables 1 - 3). Our bias mitigation approach performs comparably to AttenD, achieving near-complete bias mitigation across all MLMs, with slightly better results on BERT-large. However, our method preserves language modeling capability, whereas AttenD suffers a drastic MLM loss increase (over 1000%), despite maintaining downstream GLUE performance (their Table 5). Our debiased model also preserves GLUE performance, so while bias mitigation is similar, the difference in MLM loss is stark.

CDS: We do not re-run Bartl et al. (2020) in our setting as their debiasing relies on fine-tuning with the GAP corpus. In contrast, our method and AttenD operate directly on probe sentences derived from templates using profession and gendered words,

Desired Distribution	Model	Profession Category				MLM loss			Epoch	$\mathcal{B}$	β	γ
		DP$_{male}$ (KL)	DP$_{female}$ (KL)	DP$_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)				
equal	Base Model	0.073	0.095	0.046	0.076	0.983	0.888	0.893				
	AttenD	3.2E-3	3.2E-3	3.1E-3	3.2E-3	11.0	11.0	11.0				
	% drop	95.6%†	96.7%†	93.2%†	95.9%†	-1021%	-1138%	-1132%				
	+ $\mathcal{L}_{KL,uniform}$	2.8E-4	4.5E-4	8.2E-4	4.6E-4	1.125	0.993	0.996	3	5	-	-
	% drop	99.6%†	99.5%†	98.2%†	99.4%†	-14.5%	-11.8%	-11.5%				
real-world	Base Model	0.094	0.075	0.041	0.076	0.983	0.888	0.893				
	+ $\mathcal{L}_{KL,weighted\_adaptive}$	0.029	0.040	0.013	0.030	1.536	1.357	1.361	6	5	0.80	-
	% drop	69.3%†	46.2%†	66.8%†	59.9%†	-56.3%	-52.8%	-52.4%				
	+ $\mathcal{L}_{KL,weighted\_adaptive}$ - α	0.032	0.041	0.019	0.033	1.148	1.000	1.005	3	5	0.80	-
	% drop	66.6%†	45.4%†	52.9%†	56.7%†	-16.8%	-12.6%	-12.5%				
	+ $\mathcal{L}_{KL,uniform}$	0.026	0.038	0.022	0.030	1.129	0.990	0.997	5	5	-	-
	% drop	72.7%†	49.7%†	45.1%†	60.6%†	-14.9%	-11.4%	-11.6%				
	+ $\mathcal{L}_{KL,weighted\_adaptive}$ + $\mathcal{L}_{MLM}$	0.027	0.032	0.016	0.027	1.256	1.080	1.084	5	5	0.80	0.1
	% drop	71.6%†	57.0%†	60.2%†	64.6%†	-27.8%	-21.6%	-21.4%				

Table 3: Bias mitigation and language modeling performance (**BERT-large**). See Table 1 for cell values and notation details.

making CDS results not directly comparable. For relative context, we report their results (Appendix Table 10). They measured bias via association score differences, which capture disparities in the model's association between gender and profession, while we use distributional divergence. Despite different metrics, relative comparisons are insightful.

Bartl et al. (2020) reported > 90% mitigation for DP$_{male}$, 58% for DP$_{female}$, and 68% for DP$_{balanced}$. In contrast, our method achieves > 98% mitigation across all categories (Tables 1-3), showing consistent, robust, and near-complete mitigation for equality across MLMs, underscoring effectiveness.

Moreover, our approach supports the alignment of model attribute–target distribution to a desired distribution – extending beyond conventional emphasis on equity in prior work.

5.1.2 Real world distribution

We first present the results using the weighted adapted KL loss, which constitutes our main bias mitigation method. We then evaluate the effect of adding a secondary MLM loss to preserve language modeling capability, yielding the best trade-off between fairness and language modeling performance. Finally, we conduct ablation studies by independently removing adaptive loss scaling and weighting (i.e., using uniform KL loss across professions) to assess the importance of these components.

Weighted adaptive KL loss: The rows "+ $\mathcal{L}_{KL,weighted\_adaptive}$" indicates the results of using weighted adaptive KL loss to reduce bias. Bias reductions are significant with one exception: in BERT-base, for DP$_{female}$, the divergence slightly increased by 0.3%. However, the initial bias was relatively low (0.04). Bias reduction for DP$_{male}$ ranged from 69%-77%, for DP$_{female}$ from 46%-57%, for DP$_{balanced}$ from 28%-74%, and for ALL from 59%-67%. These results demonstrate the effectiveness of the method across all profession categories. However, the improvements came with a

trade-off in degradation in MLM language modeling performance—MLM loss increased by 12%-56%. Notably, BERT-large exhibited the largest degradation (more than 50%).

Ablation: effect of adding adaptive loss scaling (α): Adaptive loss scaling controls the update magnitude. We assign a smaller α = 1e-6 to groups with larger KL divergence (larger updates) and a larger α = 1e-5 to groups with smaller KL divergence (smaller updates). Groups are defined using validation KL means (Appendix Table 9) of ALL: categories above the mean received a smaller α, and those below received a larger α. In all three MLMs, DP$_{male}$ formed a larger update group, while DP$_{female}$ and DP$_{balanced}$ formed smaller update groups.

We compare the effect of removing adaptive loss scaling ("+ $\mathcal{L}_{KL,weighted\_adaptive}$ - α") with "+ $\mathcal{L}_{KL,weighted\_adaptive}$". Removing scaling generally degrades bias mitigation. In BERT-large, removal worsens bias mitigation performance across all categories, though language modeling performance improves significantly (about 40% points). In BERT-base and DistilBERT, removing α slightly reduces bias mitigation for one of three categories and consistently across ALL slightly, alongside moderate degradation in MLM performance (1.3-8.5% points). Overall, adaptive loss scaling yields slight but consistent improvement in bias mitigation, with trade-offs in modeling performance.

Ablation: effect of weighted adaptive loss: We now compare weighted adaptive loss ($\mathcal{L}_{KL,weighted\_adaptive}$) with non-adaptive uniform KL loss ($\mathcal{L}_{KL,uniform}$).

Using uniform loss, impacts bias mitigation and language modeling with trade-offs across models. In DistilBERT, uniform weighting yields a very small improvement for DP$_{male}$, but leads to a larger drop for DP$_{female}$ (13.8% points), and a moderate drop for DP$_{balanced}$ (7.1% points), along with a slight overall drop in ALL. However, MLM loss

improves moderately (10.4%-13.2% points). Notably, DP_{male} has a larger KL divergence initially than DP_{female} and $DP_{balanced}$. Uniform loss emphasizes the group with a larger KL, contributing more to the overall loss and improving bias mitigation for the dominant group, but causing a noticeable drop for non-dominant ones.

This pattern persists in BERT-base and BERT-large. In BERT-base, there is a very small drop for DP_{male}, while DP_{female} and $DP_{balanced}$ experience moderate drops of 10.6% points and 7.9% points, respectively. Here, too, language modeling performance improves (7.3%-10.5% points), reflecting better preservation of MLM capability with uniform weighting. In BERT-large, uniform loss yields small improvements for both DP_{male} and DP_{female} but a noticeable degradation for $DP_{balanced}$ (21.7% points). Importantly, MLM loss improves substantially (about 41% points).

Weighted adaptive KL loss + MLM loss: Weighted adaptive loss improves bias mitigation but reduces MLM performance. To address this trade-off, we add a secondary MLM loss to preserve language modeling ability while minimizing deviation from the target distribution. Results are in the last rows of Tables 1-3. We compare with the row "+ $\mathcal{L}_{KL,weighted\_adaptive}$" (without MLM loss).

In DistilBERT, there is a modest reduction in bias mitigation (6.6-9.2% points) for DP_{male} and ALL, while stable or improved results for the rest. This fairness trade-off is accompanied by a substantial gain in language modeling, with MLM losses improving by 10.6-13.2% points across corpora. In BERT-base, bias mitigation performance remains largely stable or slightly improves (2%-6% points) after introducing MLM loss, while simultaneously achieving a 5-7% points improvement in MLM loss across both corpora. BERT-large exhibits the most favorable outcome: bias mitigation performance improves across profession categories, alongside a substantial improvement in MLM loss (reduction of 28-31% points). Downstream GLUE evaluation is maintained across all three MLMs, indicating preserved language modeling capabilities (Appendix Table 13 'debiased for real-world' rows). Overall, adding the MLM loss consistently improves language modeling performance (measured with MLM loss and maintained for GLUE evaluation), while modestly affecting bias mitigation performance (preserved or even improved slightly).

Summary: Weighted adaptive loss improves bias mitigation but trades off language modeling performance. Adding MLM loss reduces these trade-offs, maintaining (or only slightly reducing) bias mitigation while improving language modeling. Overall, weighted adaptive loss with MLM loss reduces MLM distribution deviation from the real-world target while preserving modeling performance.

5.2 Results for debiasing ALM

We present pre-debiasing results across Llama and Qwen Instruct models (3B - 72B), to examine how well ALM output distribution aligns with equality versus real-world target distribution across model sizes (Appendix Tables 11 - 12). KL divergence is near zero under the *equal* distribution, indicating no bias, but notable bias is observed under *real-world* distribution, with similar trends across model sizes and model type. Detailed discussions are provided in Appendix A.8. Now we will focus on debiasing results for two Llama instruct models (3.2-3B and 3.1-8B). Results are presented in Table 4.

For *equal* target distribution, the initial KL divergence is close to zero, indicating the absence of bias. However, for *real-world* target distribution, we find notable bias. We will now focus on bias mitigation for *real-world* as a desired distribution.

As a reminder, we use sentence loss as a proxy for measuring association, normalized to the gender-profession distribution. Since it also reflects ALM language modeling, we do not include additional language modeling preservation loss as in MLMs. As observed in MLMs, weighted adaptive loss has better performance compared to one without adaptive weighting, so we will only focus on weighted adaptive loss to mitigate bias in Llama3. As seen in Appendix Table 9, DP_{male} and DP_{female} show higher initial bias and thus receive larger adaptive loss scaling (α) than $DP_{balanced}$.

Applying the weighted-adaptive KL loss significantly improves bias mitigation across most profession categories. KL divergence drops by 50% - 62% across DP_{male}, DP_{female}, and ALL. For $DP_{balanced}$, although the KL divergence worsens by 255%/187%, the absolute value remains very small, indicating that the profession from this group is already closely aligned with the real-world distribution. Bias reductions are about the same for both models, with the larger model showing slightly better reduction for DP_{male} and slightly less reduction for DP_{female}. Irrespective of model size, bias mitigation remains a challenge for all profession categories, although it improves for larger models.

Meanwhile, degradation in language model-

| LLM | Desired Distribution | Model | Profession Category | | | | Perplexity | | | Epoch | $\mathcal{B}$ | β | γ |
			DP_{male} (KL)	DP_{female} (KL)	$DP_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)				
Llama3.2-3B-Instruct	equal	Base Model	4E-4	7E-4	1E-4	4E-4	30.6	16.9	17.1				
	real-world	Base Model	0.199	0.108	0.001	0.123	30.6	16.9	17.1				
		+ $\mathcal{L}_{KL,weighted\_adaptive}$	0.089	0.051	0.004	0.057	32.4	17.4	17.6	24	5	0.6	-
		% drop	55.1%†	52.7%†	-255%	53.8%†	-6.1%	-3.0%	-3.1%				
Llama3.1-8B-Instruct	equal	Base Model	2E-3	4E-4	2E-4	8E-4	25.6	12.5	12.7				
	real-world	Base Model	0.181	0.114	0.001	0.118	25.6	12.5	12.7				
		+ $\mathcal{L}_{KL,weighted\_adaptive}$	0.069	0.057	0.003	0.051	26.7	12.6	12.8	11	5	0.6	-
		% drop	61.9%†	49.5%†	-187.1%	56.7%†	-4.2%	-1.3%	-1.4%				

Table 4: Bias mitigation and language modeling performance: Perplexity, LLM Evaluation Harness (Appendix Table 14) for **Llama3 Instruct (3.1-8B and 3.2-3B)**. $lora_r = 64$, $\frac{lora_\alpha}{lora_r} = 0.25$. See Table 1 for explanation of cell values and notations.

ing, measured using perplexity, remains minimal. Larger model better preserves performance (e.g., Llama3.1-8B-Instruct: 1.3%, Llama3.2-3B-Instruct: 3% degradation on Wikitext-103-test). QLoRA fine-tuning on 8B-Instruct converged in fewer epochs than LoRA on 3B-Instruct, consistent with more expressive models requiring less training. Downstream performance (Appendix Table 14) across five benchmarks is maintained, showing language modeling ability is not compromised. Overall, weighted adaptive KL loss achieves successful bias mitigation while largely preserving language modeling performance.

6 Related Works

Bias perspective: Prior bias work emphasizes equality (often framed as disparate impact, demographic parity, equalized odds, or equal opportunity (Gallegos et al., 2024; Mehrabi et al., 2021)), targeting fine-grained associations between demographic attributes and contexts (e.g., gendered term–profession pairs). These works often equalize association scores using static (Bolukbasi et al., 2016) embeddings or contextual embeddings from template-based probing (Shi et al., 2024; Guo et al., 2022; Garimella et al., 2021) and co-occurrence analysis (Dhamala et al., 2021; Bordia and Bowman, 2019). Literature enforces pairwise parity within a domain (e.g., profession), but often overlooks broader distributional alignment.

In contrast, we adopt a coarser, distributional approach—shifting gender ratios in professions to 50%-50%. Beyond promoting *equality*, we consider *real-world* alignment motivated by concerns about LLM hallucinations and the need for trustworthy fact-checking grounded in actual distributions. Depending on application, we frame bias relative to a desired distribution—equal (for social equity) or real-world (for factual grounding).

In-processing bias mitigation: Common approaches include modifying the model architecture (e.g., adding debiasing layers) (Xu et al., 2025; Kumar et al., 2023; Lauscher et al., 2021) or restricting training to certain parameters (e.g., attention head, adapter layers (Yang et al., 2025; Masoudian et al., 2024; Gaci et al., 2022; Attanasio et al., 2022)), to isolate and suppress biased representations while preserving general knowledge.

Another strategy modifies the loss function to encode fairness during training (Xu et al., 2025; Gallegos et al., 2024; Zheng et al., 2023; Yogarajan et al., 2023; Garimella et al., 2021). E.g., Guo et al. (2022) uses Jensen-Shannon Divergence to enforce uniformity in gender-conditioned output distribution. Woo et al. (2023) introduce KL-based regularization to reduce gender bias in stereotypical sentence representations, while preserving linguistic integrity on non-stereotypical content.

We propose a weighted adaptive KL loss that aligns LLM gender–profession distributions to a target, dynamically updating based on varying gender dominance to enable balanced bias mitigation across professions.

7 Conclusion

We present a new perspective of bias through the lens of *desired distribution*—either equal or real-world. We proposed a method that adjusts LLMs' gender-profession distribution toward a desired distribution (our primary objective), applied to 3 MLMs and 2 ALMs. To achieve this, we use a weighted adaptive KL loss combined with a secondary MLM loss for MLMs (not required for ALMs). Bias is measured via the KL divergence of the model's distribution from the desired distribution. We demonstrate the advantage of using this loss through two ablation studies. Adding the MLM loss improves language modeling with minimal impact on bias mitigation for MLMs. For ALMs, the KL loss alone effectively reduces bias across profession categories with minimal drop in language modeling performance. We evaluate across three profession categories—male-dominated, female-dominated, and balanced. Overall, we show that LLM output distributions can be effectively aligned with the desired distribution.

8   Limitations

Our bias analysis is limited to binary gender, reflecting the availability of real-world distribution data for binary gender categories only. We specifically address gender bias mitigation through distribution alignment within the context of professions, utilizing gender-profession data from the U.S. Bureau of Labor Statistics. We also acknowledge that our analysis is limited to US-based distributions and requires inclusive real-world distributions for global applicability. Further investigation is needed to extend this work to other demographic groups, such as race.

Our analysis is based on a limited set of 225 professions, suggesting that expanding to a broader range could yield additional insights. Similarly, we employed only six templates for bias mitigation. However, we balanced the selection of rare and common templates in the training and testing sets. Additionally, our analysis is limited to a template-based approach, which is a common approach used for bias mitigation in LLMs. Bias mitigation for open-ended generation introduces additional complexities, such as maintaining generation quality and ensuring contextual relevance. We acknowledge that for improved generalizability and comprehensiveness, it is important to evaluate in these settings reflecting real-world language use, where autoregressive LLMs are typically more suitable. However, given the scope of our work, we leave this exploration for future research. In addition, our analysis does not comprehensively evaluate out-of-distribution scenarios, such as template variation (e.g., altering the order of attributes and targets or using more diverse template structures), which could provide a stronger test of generalizability. We leave such exploration for future work.

Another limitation arises in the fine-tuning of masked language models (MLMs) using additional MLM loss: we relied on a small set of probe sentences derived from training templates to preserve language modeling capability. Using a larger external corpus during fine-tuning could better preserve the model's language modeling capabilities.

Finally, our bias mitigation efforts were focused on two variants of the Llama3 autoregressive language models (ALM). Extending this exploration to additional ALMs remains an important direction for future research.

9   Ethical Considerations

We limit our analysis to binary gender due to the availability of real-world distribution data for binary gender only.

Our bias viewpoint—where we align the model's output distribution with real-world data—can introduce or preserve existing social biases in LLMs. However, we motivate this choice by its positive impact on reducing hallucinations and improving fact-checking, especially in high-stakes domains where factual accuracy is critical. We emphasize that such bias introduction must not be misused to justify inequitable outcomes; rather, it should be applied with transparency and only when the application context prioritizes factual alignment. Importantly, real-world distributions often reflect underlying systemic inequalities, and their use as a bias mitigation target should be carefully justified and assessed on a case-by-case basis.

Moreover, our reliance on U.S.-based distributions may reinforce a geographically constrained perspective, underscoring the necessity of incorporating non-U.S. contexts for international deployment.

Future work should also aim to include non-binary and underrepresented gender identities to ensure a more inclusive and comprehensive fairness evaluation.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*,

pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Bureau of Labor Statistics. 2024. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. https://www.bls.gov/cps/cpsaat11.htm. Accessed: May 3, 2025.

Ana Cimitan, Ana Alves Pinto, and Michaela Geierhos. 2024. Curation of benchmark templates for measuring gender bias in named entity recognition models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4238–4246, Torino, Italia. ELRA and ICCL.

Pieter Delobelle, Ewoenam 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: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Victor Hansen, Atula Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Søgaard. 2024. The impact of differential privacy on group disparity mitigation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3952–3965, Mexico City, Mexico. Association for Computational Linguistics.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Michael Ibrahim. 2024. CUFE at StanceEval2024: Arabic stance detection with fine-tuned llama-3 model. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 807–810, Bangkok, Thailand. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational*

Linguistics: EMNLP 2021, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 17–29, Seattle, Washington. Association for Computational Linguistics.

Shahed Masoudian, Cornelia Volaucnik, Markus Schedl, and Navid Rekabsaz. 2024. Effective controllable bias mitigation for classification and retrieval using gate adapters. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2434–2453, St. Julian's, Malta. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Elvys Linhares Pontes, Carlos-Emiliano González-Gallardo, Mohamed Benjannet, Caryn Qu, and Antoine Doucet. 2024. L3iTC at the FinLLM challenge task: Quantization for financial text classification & summarization. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 141–145, Jeju, South Korea. -.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Bingkang Shi, Xiaodan Zhang, Dehan Kong, Yulei Wu, Zongzhen Liu, Honglei Lyu, and Longtao Huang. 2024. General Phrase Debiaser: Debiasing Masked Language Models at a Multi-Token Level. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6345–6349. IEEE.

Ingroj Shrestha, Louis Tay, and Padmini Srinivasan. 2025. Robust bias detection in MLMs and its application to human trait ratings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4842–4858, Albuquerque, New Mexico. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Qibin Wang, Xiaolin Hu, Weikai Xu, Wei Liu, Jian Luan, and Bin Wang. 2025. PMSS: Pretrained matrices skeleton selection for LLM fine-tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8841–8857, Abu Dhabi, UAE. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Barry Payne Welford. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chunlei Xin, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and Le Sun. 2024. Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2307–2317, Torino, Italia. ELRA and ICCL.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. 2025. BiasEdit: Debiasing stereotyped language models via model editing. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 166–184, Albuquerque, New Mexico. Association for Computational Linguistics.

Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P. Lalor, and Kar Yan Tam. 2025. Bias a-head? analyzing bias in transformer-based language model attention heads. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 276–290, Albuquerque, New Mexico. Association for Computational Linguistics.

Vithya Yogarajan, Gillian Dobbie, Te Taka Keegan, and Rostam J Neuwirth. 2023. Tackling bias in pre-trained language models: Current trends and under-represented societies. *arXiv preprint arXiv:2312.01509*.

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

Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1022–1040, Toronto, Canada. Association for Computational Linguistics.

Hao Zhou, Guergana Savova, and Lijing Wang. 2025. Assessing the macro and micro effects of random seeds on fine-tuning large language models. *arXiv preprint arXiv:2503.07329*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Templates

TID	Templates
T_1	[DET/PRONOUN] [attribute] is [ARTICLE] [target].
T_2	[DET/PRONOUN] [attribute] works as [ARTICLE] [target].
T_3	[DET/PRONOUN] [attribute] wants to become [ARTICLE] [target].
T_4	[DET/PRONOUN] [attribute] applied for the position of [target].
T_5	[DET/PRONOUN] [attribute], the [target] had a good day at work.
T_6	[DET/PRONOUN] [attribute] started a career as [ARTICLE] [target].

Table 5: Templates. TID: template id, attribute: genderedword, target: profession, DET: this, PRONOUN: my

A.2 Attribute

Male gendered words	Female gendered words
he, man, brother	she, woman, sister
son, husband, boyfriend	daughter, wife, girlfriend
father, uncle, dad	mother, aunt, mom
grandpa, grandfather	grandma, grandmother

Table 6: Attributes: Gendered words. These gendered words are preceded by DET: *this* for man/woman, no DET/PRONOUN for he/she, while for remaining, PRONOUN is *my* in the templates in Table 5.

A.3 Profession distribution

	Train	Valid	Test	Total
DP$_{male}$	59	13	18	90
DP$_{female}$	58	14	18	90
DP$_{balanced}$	29	7	9	45
ALL	146	34	45	225

Table 7: Distribution of professions (target).

A.4 Model hyperparameters and selection

Following Zhou et al. (2025), who observe that seed 42 often yields better performance in machine learning experiments, we fine-tune the model using seed 42 and select optimal hyperparameters based on validation KL divergence loss across epochs. We use the AdamW optimizer with a weight decay of 0.01. Final results are reported as the average over five seed runs using the selected configuration (see Section 5).

We implement fine-tuning using PyTorch, HuggingFace Transformers, and the PEFT library, running on NVIDIA A100 (80 GB) and A40 (45 GB) GPUs. On average, fine-tuning took 4 hours for MLMs and around 9 hours for the ALM.

Fine-tuning convergence criteria: We track the KL divergence loss across all professions on the validation set and stop fine-tuning if it fails to improve from the previous best by at least a threshold (0.0001 for *equal*, 0.001 for *real-world*) for $n = 5$ consecutive steps (patience). Note that adaptive weighting is not applied during validation. KL divergence is computed uniformly across professions as in standard bias detection.

Profession batch: We explore batch sizes of 5 and 8 for training and use a fixed batch size of 3 for validation. Note that we ensure each batch contains professions from the same profession group.

Learning rate: We evaluate the performance using a learning rate of 2e-5[4] adapted from (Devlin et al., 2019) for MLMs. For Llama3.2-3B-Instruct, we evaluate using 2e-5 and 2e-4, adapted from (Hu et al., 2022). For Llama3.1-8B-Instruct, we used the learning rate that performed best for Llama3.2-3B-Instruct.

---

[4]We did preliminary analysis using a small subset of professions adapted from Bartl et al. (2020) using a learning rate of 1e-5 and 2e-5. We found 2e-5 converges earlier and performs better. For the small model DistilBERT, we also explored a smaller learning rate of 5e-6, but it degraded the performance in at least one of the profession groups and ALL (combined profession groups).

Momentum weight for updating KL mean: We explored momentum weights $\beta \in \{0.60, 0.80, 0.95\}$

Configurations for Llama3 fine-tuning using LoRA and QLoRA: Following Hu et al. (2022), we fix attention dimension lora$_r$ = 64, which defines the rank of the low-rank adaptation and controls the number of trainable parameters. We vary the scaling factor lora$_\alpha \in \{16, 32, 64\}$, where the ratio $\frac{\text{lora}_\alpha}{\text{lora}_r} \in \{0.25, 0.50, 1\}$ controls the strength of the LoRA update. A LoRA dropout of 0.2 is applied for regularization.

We apply LoRA to the projection layers in the model, including query (q_proj), key (k_proj), value (v_proj), and output (o_proj) projections in the self-attention mechanism, as well as the gate (gate_proj), up (up_proj), and down (down_proj) projections in the MLP components. This setup allows LoRA to adapt both the attention and feedforward pathways, consistent with configurations shown to be effective in prior work on Llama-based instruction fine-tuning (Ibrahim, 2024; Pontes et al., 2024; Wang et al., 2025).

For models $\geq$7B parameters, such as Llama3.1–8B-Instruct, we apply QLoRA with 4-bit NF4 quantization (including double quantization and float16 compute) for both inference and bias mitigation. The same quantization configuration is used during pre-debiasing evaluation and QLoRA fine-tuning to ensure consistency. In contrast, for the smaller Llama3.2–3B-Instruct model, we use full-precision LoRA fine-tuning without quantization. Both models are fine-tuned for bias mitigation with identical hyperparameters.

Mean vs sum of KL divergence across gender: We observe that the MLM loss for male and female for a given profession varies, and to provide a balanced combined effect while computing the loss, we take the mean instead of the sum of individual divergence to get the total divergence for a profession.

Our preliminary analysis using 60 professions (adapted from Bartl et al. (2020); U.S. Bureau of Labor Statistics 2019) supports this. Irrespective of the *mean* or *sum* of KL divergence between the MLM-predicted and desired male-female distributions, convergence was similar for both equal and real-world distributions. However, 50% of KL loss with *sum* is achieved in the same epoch as *mean*, i.e., to achieve similar performance, the sum requires further tuning. So, taking the mean

divergence is optimal.

Selection of hyperparameter based on validation performance: We use the fine-tuned model (obtained using seed 42) to evaluate performance on the validation set, using the corresponding validation templates and professions. We select the configuration that achieves high overall performance across all profession groups, as well as ALL (combined all three profession categories) with consistent (less spread) performance across three profession groups. We will discuss the approach next.

We first compute the mean/standard deviation ($\frac{\mu}{\sigma}$), denoted as R, of performance improvement (relative to the Base model) across male-dominated, female-dominated, and balanced groups, aiming for a high mean with low variability, indicating strong and stable results. Runs are ranked by R in descending order, and we select the top two whose ALL improvement exceeds the average ALL (allowing a small offset to include runs within 1 point of the mean). The final selection is based on the higher median improvement across the three profession groups, ensuring robustness to outliers and consistently strong improvement in at least half the groups.

A.5 Common vs Rare templates

TID	% sentences with ppl < 15		
	DistilBERT	**BERT-base**	**BERT-large**
T1	41%	48%	46%
T2	63%	70%	71%
T3	77%	74%	80%
T4	73%	86%	87%
T5	85%	98%	99%
T6	40%	46%	39%

Table 8: Percentage of sentences with pseudo-perplexity (ppl) below 15

A.6 Bias on Validation set

Model	Profession category							
	DP_male		**DP_female**		**DP_balanced**		**ALL**	
	μ_{KL}	σ^2_{KL}	μ_{KL}	σ^2_{KL}	μ_{KL}	σ^2_{KL}	μ_{KL}	σ^2_{KL}
DistilBERT	0.166	0.019	0.036	0.003	0.025	3E-04	0.083	0.013
BERT-base	0.232	0.087	0.038	0.001	0.085	0.007	0.122	0.042
BERT-large	0.131	0.015	0.033	0.002	0.021	0.001	0.068	0.009
Llama3.2-3B	0.184	0.006	0.099	0.003	0.002	4E-06	0.112	0.008
Llama3.1-8B	0.172	0.005	0.102	0.003	0.002	8E-06	0.108	0.007

Table 9: Validation performance (KL mean: μ_{KL} and KL variance: σ^2_{KL}) across LLMs for **real world** as desired distribution (Before Debiasing)

A.7 Bartl et al. (2020) bias mitigation result

Profession Category	Gender	Pre Mean	Post Mean	|m-f|_Pre	|m-f|_Post	$\triangle$|m-f|_Post-Pre	% Bias Reduction
DP_balanced	f	-0.35	0.20	0.40	0.13	0.27	67.5%
	m	0.05	0.07				
DP_female	f	0.50	0.36	1.18	0.50	0.68	57.6%
	m	-0.68	-0.14				
DP_male	f	-0.83	0.13	0.99	0.08	0.91	91.9%
	m	0.16	0.21				

Table 10: Results are directly adapted from Table 4 in Bartl et al. (2020). *Pre* refers to the association scores between gender and profession before debiasing, and *Post* refers to the scores after debiasing, based on the templates used in their paper.

A.8 Bias detection across ALMs

A.8.1 Equal distribution

Table 11 shows the pre-debiasing results using equality (50-50 gender-profession distribution) as the desired distribution across ALMs.

The initial KL divergence values remain close to zero for all ALMs irrespective of size. This indicates that there is no bias even in larger LLMs. In terms of language modeling capability, perplexity consistently decreases with model size, aligning with expectations—larger LLMs generally demonstrate stronger modeling performance. On average, perplexity drops (Llama 3B → 8B → 70B, Qwen 7B→72B) by approximately 16.9% on the GAP corpus, 31.7% on WikiText-103-test, and 28.6% on WikiText-103-dev as model size increases.

A.8.2 Real world distribution

Table 12 presents the pre-debiasing result using the real-world distribution as the desired distribution across ALMs.

We find similar levels of bias across model sizes. Consistently, male-dominated profession groups (DP_male) show higher bias than female-dominated ones (DP_female) and ALL profession categories, and there is no bias in the balanced (DP_balanced) category. Bias magnitude does not vary significantly with model size, thus, we observe a consistent trend with increasing size. The perplexity finding remains consistent with those discussed in the equal distribution section earlier.

LLM	Profession Category				Perplexity		
	DP_{male} (KL)	DP_{female} (KL)	$DP_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)
Llama3.2-3B-Instruct	4E-4	7E-4	1E-4	4E-4	30.6	16.9	17.1
Llama3.1-8B-Instruct	2E-3	4E-4	2E-4	8E-4	25.6	12.5	12.7
Llama3.3-70B-Instruct	1E-3	6E-4	4E-4	9E-4	21.0	8.3	8.9
Qwen-2.5-7B-Instruct	4E-3	5E-4	1E-3	2E-3	26.8	13.3	14.0
Qwen-2.5-72B-Instruct	3E-3	5E-4	6E-4	2E-3	22.4	8.6	9.8

Table 11: Pre-debiasing results across ALMs (*equal* distribution) on test set

LLM	Profession Category				Perplexity		
	DP_{male} (KL)	DP_{female} (KL)	$DP_{balanced}$ (KL)	ALL (KL)	GAP corpus	WikiText-103 (test)	WikiText-103 (dev)
Llama3.2-3B-Instruct	0.199	0.108	0.001	0.123	30.6	16.9	17.1
Llama3.1-8B-Instruct	0.181	0.114	0.001	0.118	25.6	12.5	12.7
Llama3.3-70B-Instruct	0.185	0.110	0.001	0.118	21.0	8.3	8.9
Qwen-2.5-7B-Instruct	0.164	0.127	0.002	0.117	26.8	13.3	14.0
Qwen-2.5-72B-Instruct	0.169	0.120	0.001	0.116	22.4	8.6	9.8

Table 12: Pre-debiasing results across ALMs (*real-world* distribution) on test set

A.9 Language modeling capability evaluation on downstream tasks

MLM	Desired Distribution	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-(m/mm)	QNLI	RTE	Average
DistilBERT	Base Model	0.49	0.91	0.90	0.86	0.87	0.82/0.82	0.89	0.60	0.79
	debiased for equal	0.49	0.91	0.89	0.86	0.87	0.82/0.82	0.88	0.61	0.79
	debiased for real-world	0.50	0.91	0.89	0.86	0.87	0.82/0.82	0.89	0.60	0.80
BERT-base	Base Model	0.56	0.93	0.88	0.88	0.88	0.84/0.85	0.91	0.62	0.82
	debiased for equal	0.56	0.93	0.89	0.89	0.88	0.85/0.85	0.91	0.65	0.82
	debiased for real-world	0.56	0.93	0.89	0.89	0.88	0.85/0.85	0.91	0.64	0.82
BERT-large	Base Model	0.61	0.94	0.89	0.88	0.89	0.86/0.87	0.92	0.71	0.84
	debiased for equal	0.59	0.94	0.90	0.90	0.88	0.86/0.86	0.92	0.71	0.84
	debiased for real-world	0.61	0.94	0.90	0.90	0.88	0.87/0.86	0.92	0.73	0.84

Table 13: GLUE dev results across MLMs before and after mitigation (using weighted adaptive loss). Accuracy is reported for SST-2, MNLI, QNLI, and RTE; Spearman correlation for STS-B; Matthews correlation for CoLA; and F1 for MRPC and QQP. *Base Model* refers to the pretrained model before debiasing, while *debiased for equal* refers to models debiased by fine-tuning using non-adaptive KL loss, and *debiased for real-world* refers to models debiased by fine-tuning using weighted adaptive loss combined with MLM loss. For each task, we report the average performance evaluated across five debiased models (obtained using random seeds 42, 52, 62, 72, and 82) for both equal and real-world target distributions.

LLM	Model	HellaSwag	LAMBADA OpenAI	MMLU	TruthfulQA	CoLA	SST-2	MRPC	QQP	MNLI-(m/mm)	QNLI	RTE	GLUE Average
Llama3.1	Base Model	0.59	3.68/0.71	0.66	0.68/0.66/0.66	0.06	0.88	0.81	0.55	0.52/0.52	0.51	0.70	0.57
	+ $\mathcal{L}_{KL,weighted\_adaptive}$	0.59	3.68/0.70	0.66	0.67/0.65/0.65	0.07	0.88	0.81	0.56	0.53/0.52	0.51	0.70	0.57
Llama3.2	Base Model	0.52	3.67/0.71	0.60	0.66/0.64/0.64	0.03	0.83	0.83	0.56	0.52/0.52	0.54	0.74	0.58
	+ $\mathcal{L}_{KL,weighted\_adaptive}$	0.53	3.67/0.71	0.60	0.67/0.65/0.65	0.03	0.83	0.82	0.56	0.53/0.52	0.53	0.73	0.58

Table 14: LM Evaluation Harness for Llama3.2-3B-Instruct (Llama3.2) and Llama3.1-8B-Instruct (Llama3.1) before and after debiasing (using weighted adaptive KL loss). Metrics used — Accuracy: [HellSwag, MMLU, SST-2, MNLI, QNLI, RTE]; Perplexity/Accuracy: [LAMBADA OpenAI]; BLEU/ROUGE-1/ROUGE-L: [TruthfulQA_Gen]; Matthews Correlation: [CoLA]; F1: [MRPC, QQP].