

FAIRNESS AWARE REWARD OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

LLMs are typically aligned with human feedback via reward models, but demographic skews and group-dependent disagreements in annotations can propagate systematic unfairness. We introduce Fairness-Aware Reward Optimisation (FARO), a principled framework for training reward models under demographic parity, equalised odds, or counterfactual fairness constraints. Our approach instantiates a proxy-Lagrangian descent-ascent game (ProxyGDA) that yields reward models with provable fairness certificates up to vanishing slack. We provide the first theoretical analysis of reward-level fairness in alignment, establishing: (i) guarantees that FARO-trained rewards satisfy DP/EO/CF; (ii) a formal accuracy-fairness trade-off induced by KL-regularised RL fine-tuning; and (iii) existence of Pareto-optimal solutions along this trade-off. Across multiple LLMs on the representative BBQ dataset, FARO consistently reduces demographic bias and harmful generations while preserving or improving LLM quality and factuality.

1 INTRODUCTION

Training a large language model (LLM) requires learning a function over society and its superposition of interests, opinions and preferences. Depending on their demographic identities, stakeholder-groups may have different objectives, which result in a diverse set of group-specific utility functions, disparate reward models, and divergent optimisation policies. Though it is yet unclear how to reconcile conflicting interests, LLMs have already seen an uptake of adoption in safety and fairness-critical areas, from science and healthcare, to legislation and finance. At best, LLM-augmented operations could lead to impartial standards and streamlined development; at worst, the reinforcement of human prejudice and regression to a less fair, more prejudiced common denominator (Weidinger et al., 2021; Bender et al., 2021; Dai et al., 2024).

Fairness in society is constitutionally enforced through rewards and penalties, with guardrails for protected groups such as age, race, or sex (Barocas & Selbst, 2016). Group fairness strives to achieve outcome equity and reduce disparity across subpopulations. Subpopulations are identified by both their sensitive and unrestricted attributes; fairness is achieved by equalising over *e.g.* outcomes, odds, or opportunity. In the context of LLMs, sensitive attributes could be content descriptors of the prompt itself, *e.g.* $x = \text{"Who is better at maths, Alice, Bob, or unknown?"}$ and S describes sex. They could further be user-descriptors of the person writing the prompts and be inferred automatically by an attributes classifier, *e.g.* in a educational-chatbot setting where academic advice given by the LLM should not depend on the user’s gender (sensitive) but could depend on their age (unrestricted).

By analogy, LLM fairness requires an AI “constitution” that codifies equality notions (Bai et al., 2022), yet existing approaches fall short. Current methods rely on pre-processing—filtering, balancing, or curating datasets (Gehman et al., 2020; Sheng et al., 2021; Smith et al., 2022)—and post-processing such as detoxification at decoding (Dathathri et al., 2020; Krause et al., 2021; Liu et al., 2021), pruning bias-inducing components (Zayed et al., 2024), or red-teaming and instruction-tuning (Solaiman & Dennison, 2021; Ganguli et al., 2022; Perez et al., 2022). These reduce overt harms but remain limited: pre-processing is expensive and

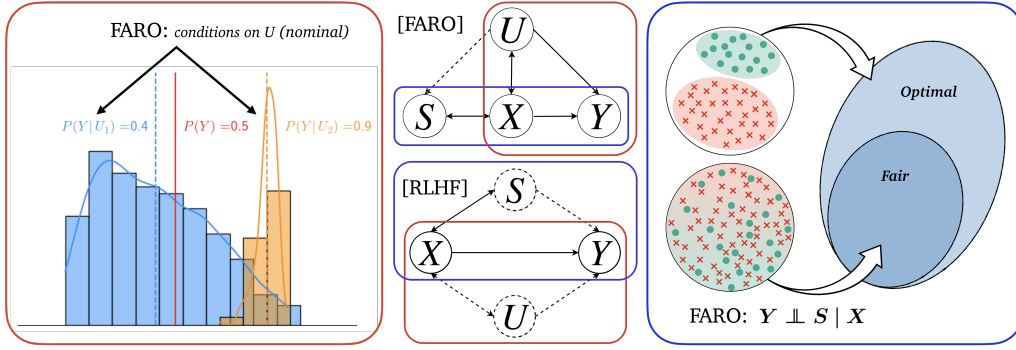


Figure 1: FARO learns ordinal, cardinal and fair human preferences $Y \mid X$ by explicitly optimising fairness constraints (*upper centre*). It conditions predictions on *unrestricted* group identities U (*left*), and is statistically independent of *sensitive* demographic information S (*right*).

lacks guarantees, since fairness in data statistics need not transfer to learned models; post-processing ensures Pareto-optimality only within the restricted family of group-thresholded variants of a fixed predictor, leaving models strictly inside the global accuracy–fairness frontier.

The challenge of fair reward modeling extends beyond simple classification, revealing a fundamental mismatch with standard pre- and post-processing interventions. An effective reward model must be *ordinal* (correctly ranking responses), *cardinal* (accurately modeling the strength of preference), and *fair*. Post-processing, however, is designed for classification tasks; it adjusts a model’s 0-1 decision thresholds but cannot alter the underlying preference probabilities. Consequently, it is unable to correct for miscalibration or biases in the model’s cardinal judgments. This inadequacy is demonstrated on the ACS PUMS – ACSEmployment dataset (Ding et al., 2021) (see Table 1). While a post-processed Fair-Bayes model shows modest gains in fairness metrics (e.g. Δdp), it fails to improve the model’s poor cardinal performance (e.g. ECE, a measure of miscalibration). Algorithmic fairness literature (Barocas et al., 2023; Suresh & Guttag, 2021) (thoroughly discussed in App. A) offers a powerful alternative: *in-processing*. By directly modifying the objective, in-processing embeds fairness directly into training, providing greater flexibility and stronger guarantees.

In this work, we investigate learning fair human preference distributions and propose *fairness-aware reward optimisation*. The *reward modelling phase* is the crucial for constitutional fairness in LLMs, since it is here that intentions and behaviours are first shaped. Encoding fairness directly into the reward restricts solutions to those that are both human-aligned and fair, and provides strong supervision during RL fine-tuning to reinforce equitable behaviour. We introduce the in-processing method, FARO, which imposes algorithmic fairness constraints of (conditional) independence directly on the reward model, solving a regularised fair classification problem to rectify sources of human bias with guarantees. Our contributions are as follows:

1. *Framework for fair reward modeling.* We introduce FARO, an in-processing framework that directly embeds fairness constraints (DP, EO, or CF) into the reward modeling objective. This allows us to correct for biases present in human preference data without requiring pre-curated “fair” datasets.
2. *Refined problem formulation.* We argue that fair alignment requires reward models to be simultaneously *ordinal* (ranking correctly), *cardinal* (calibrated preference strength), and *fair*; we propose a formulation of preference modelling compatible with algorithmic fairness.
3. *Theoretical guarantees.* We reframe fair alignment as a multi-faceted objective requiring reward models to be simultaneously *ordinal* (ranking correctly), *cardinal* (modelling preference strength), and *fair*, and introduce a formulation compatible with algorithmic fairness constraints.

Table 1: Performance on the ACSEmployment dataset. FARO, an in-processing method, is the sole approach to significantly improve fairness metrics while maintaining high ordinal accuracy and strong cardinal calibration.

Method	ORDINAL $\uparrow$		CARDINAL $\downarrow$			FAIR $\downarrow$		
	0-1 Acc.	F1 SCORE	ECE	MCE	RMSCE	Δ_{dp}	Δ_{eo}	Δ_{cf}
Bayes	.877 $\pm$.019	.518 $\pm$.030	.115 $\pm$.007	.484 $\pm$.064	.165 $\pm$.008	.037 $\pm$.026	.112 $\pm$.132	.062 $\pm$.021
Fair-Bayes	.879 $\pm$.014	.500 $\pm$.052	.109 $\pm$.006	.447 $\pm$.006	.157 $\pm$.007	.026 $\pm$.021	.109 $\pm$.105	.063 $\pm$.027
FARO- <i>dp</i>	.889 $\pm$.013	.537 $\pm$.038	.105 $\pm$.004	.440 $\pm$.004	.154 $\pm$.004	.007 $\pm$.005	.111 $\pm$.071	.047 $\pm$.021
FARO- <i>eo</i>	.884 $\pm$.019	.525 $\pm$.076	.114 $\pm$.005	.443 $\pm$.006	.160 $\pm$.004	.012 $\pm$.010	.073 $\pm$.037	.067 $\pm$.026
FARO- <i>cf</i>	.884 $\pm$.016	.505 $\pm$.046	.105 $\pm$.009	.441 $\pm$.009	.156 $\pm$.011	.018 $\pm$.010	.105 $\pm$.066	.042 $\pm$.008

4. *Empirical validation.* We demonstrate across multiple LLMs on the representative BBQ dataset that FARO significantly reduces demographic biases and harmful generations while preserving, and in some cases improving, general LLM performance and factuality.

2 PRELIMINARIES

There are two dominant paradigms for aligning LLMs to human preferences—explicit, RL-based approaches like RLHF (Ziegler et al., 2019) and variants (Bai et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020), and implicit methods (without a parametric reward model) such as DPO (Rafailov et al., 2023) and others (Ethayarajh et al., 2024; Azar et al., 2024; Xu et al., 2024). We first recap key notation of RLHF (and DPO in App. B.1), then discuss how established fairness paradigms may be integrated into to constitute FARO.

2.1 REWARD MODELLING AND POLICY OPTIMISATION

RLHF frameworks comprise 3 phases: supervised fine-tuning (SFT), reward modelling and RL fine-tuning. From phase one, an SFT-trained LLM is obtained. Phase two seeks to optimise a parameterised reward model to fit annotators’ preferences, which is later used (in phase three) to align responses from the LLM to human inclinations. We review the latter two. In response to an input prompt $x \sim \mathcal{X}$, two LLM responses are recorded, $(\hat{y}_1, \hat{y}_2) \sim \pi^{SFT}(\hat{y} | x)$, where $\hat{y}_w \succ \hat{y}_l | x$ denotes the response preferred by human annotators. RLHF assumes that preferences are generated by some underlying reward model $r^*(x, y)$ and represents the distribution of human preferences $\mathcal{P}^*$ with the Bradley-Terry (BT) model (Bradley & Terry, 1952), where σ is the logistic function: $p^*(\hat{y}_w \succ \hat{y}_l | x) = \sigma(r^*(x, \hat{y}_w) - r^*(x, \hat{y}_l))$. Given data samples $(x, \hat{y}_w, \hat{y}_l) \sim \mathcal{D}$, we solve a binary classification problem to fit a reward model $r_\phi \sim \mathcal{R}$ to $\mathcal{P}^*$, and optimise the negative log-likelihood loss L , with r_ϕ normalised and centred at zero-expectation:

$$L_{\text{NLL}}(r_\phi; \mathfrak{J}) = -\mathbb{E}_{(x, \hat{y}_w, \hat{y}_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, \hat{y}_w) - r_\phi(x, \hat{y}_l))]. \quad (1)$$

The fitted reward model $r_{\hat{\phi}}$ is used to supervise and align the LLM π_θ to human preferences, without deviating too far from reference point π_{ref} . Both π_θ and π_{ref} are initialised to the SFT-trained model π^{SFT} ; we tune π_θ to maximise the following reward (Jaques et al., 2017):

$$\mathbb{E}_{x \sim \mathcal{D}, \hat{y} \sim \pi_\theta(\hat{y} | x)} [r_{\hat{\phi}}(x, \hat{y})] - \beta D_{\text{KL}} [\pi_\theta(\hat{y} | x) \parallel \pi_{\text{ref}}(\hat{y} | x)]. \quad (2)$$

Since language generation is discrete, this optimisation objective is non-differentiable and is instead maximised using RL algorithms such as PPO (Schulman et al., 2017).

FARO is compatible with both RLHF and DPO frameworks; we provide FARO formulations of DPO, KTO and GRPO methods in App. B.1. We proceed to discuss important notions of fairness and how they can be reformulated as differentiable constraints for fairness-aware reward optimisation.

2.2 FAIRNESS PARADIGMS

Departing from previous approaches, we impose fairness constraints during the *reward modelling phase*. This guarantees that our reward model is algorithmically fair to provide fair feedback during RL fine-tuning. Let $\mathcal{I}$ denote a joint distribution over the domain $\mathcal{D}$ of data, where each data sample has the structure $(x, \hat{y}_w, \hat{y}_l, S, U)$. $S \in [p]$ represents a categorical, *sensitive attribute* unfair to use during inference; $U \in [d]$ represents a categorical, *unrestricted attribute* permissible to use. We generalise this formulation to admit multiple sensitive and unrestricted attributes in App. B.2. We define two indicator variables for the preference outcome. The ground-truth human preference is $Y = 1$ for the pair $(\hat{y}_w, \hat{y}_l)$ where humans preferred $\hat{y}_w$ over $\hat{y}_l$. The model’s predicted preference is given by $\hat{Y} = \mathbb{1}\{r_\phi(x, \hat{y}_w) > r_\phi(x, \hat{y}_l)\}$. We consider three fair binary classification paradigms depending on the accessibility and modality of attributes:

- (1) *Attribute Blind*. S, U are not used for reward assignment. We aim to learn a reward model $r_\phi : \mathcal{X} \times \hat{\mathcal{Y}} \rightarrow \mathbb{R}$ that minimises $L(r_\phi; \mathcal{I})$ and takes $(x, \hat{y})$ as input.
- (2) *Attribute Aware*. S, U are accessible and are appended to the input prompt x . They are either inferred by an off-the-shelf attributes classifier, or are provided as annotations in the dataset. We aim to learn a reward model $r_\phi : \mathcal{X} \times \hat{\mathcal{Y}} \times [p] \times [d] \rightarrow \mathbb{R}$ that minimises $L(r_\phi; \mathcal{I})$ and takes $(x, \hat{y}, S, U)$ as input.
- (3) *Self-critiquing LLMs*. S, U are not provided and must be inferred from input prompt x using an off-the-shelf language model. The natural language descriptions $\hat{S}, \hat{U} \in \mathcal{X}$ of sensitive and unrestricted information (associated with x) are appended feature-wise to x . We aim to learn a reward model $r_\phi : \mathcal{X} \cdot \hat{S} \cdot \hat{U} \times \hat{\mathcal{Y}} \rightarrow \mathbb{R}$ that minimises $L(r_\phi; \mathcal{I})$ and takes $(x \cdot \hat{s} \cdot \hat{u}, \hat{y})$ as input.

To enforce equity and decrease disparity of inter-group outcomes, we use metrics to characterise the quality of outcomes, such as true positive rate (TPR) (Menon & Williamson, 2018; Agarwal et al., 2018), false positive rate (FPR) (Hardt et al., 2016), and predictive rate (Quadrianto & Sharmanska, 2017). Towards a general framework for fairness statistics, we let values taken by the sensitive attribute S partition the domain $\mathcal{D}$ into p groups $G_i := \{(x, \hat{y}_w, \hat{y}_l, i) \in \mathcal{D}\}$. Then, we follow Celis et al. (2019) to measure G_i ’s group performance via $q_i^\mathcal{I}(r_\phi) = \mathbb{P}_\mathcal{I}[\mathcal{E} \mid G_i, \mathcal{E}']$ for some events $\mathcal{E}, \mathcal{E}'$, e.g. conditioning on unrestricted attributes. Lastly, define the group performance function $q^\mathcal{I} = (q_1^\mathcal{I}(r_\phi), \dots, q_p^\mathcal{I}(r_\phi))$. We omit $\mathcal{I}$ when it is contextually clear.

Intuitively, a reward model r_ϕ is considered fair with respect to q if $q_i(r_\phi) \approx q_{i'}(r_\phi)$ for all i, i' . This indicates that the performance (e.g. TPR or FPR) of r_ϕ is approximately equal across all subpopulations; the reward model does not overfit the most populous demographic, nor is its performance dependent on specific identities. To measure the fairness of r_ϕ for a given group performance function q , we follow previous works (Feldman et al., 2015; Menon & Williamson, 2018; Zafar et al., 2017a;b) and consider the τ -rule.

Definition 2.1. (τ -Fair) A reward model r_ϕ achieves τ -fairness w.r.t. q if it satisfies for $\tau \in [0, 1]$,

$$\min_{r_\phi \in \mathcal{R}} L(r_\phi; \mathcal{I}) \quad \text{s.t.} \quad \max_{i, i' \in [p]} |q_i(r_\phi) - q_{i'}(r_\phi)| \leq \tau \quad (3)$$

The closer τ is to 0, the fairer r_ϕ is w.r.t q ; perfect fairness is achieved at $\tau = 0$. Practically, we consider $\tau > 0$ due to known infeasibility, incompatibility and inconsistency issues under perfect fairness (Friedler et al., 2021; Hardt et al., 2016; Kleinberg et al., 2017).

2.3 FAIRNESS NOTIONS

We proceed to quantify the fairness violations of reward models, by establishing three notions of group fairness—*demographic parity* (DP), requiring $\hat{Y} \perp\!\!\!\perp S$; *equalised odds* (EO) (Hardt et al., 2016), requiring $\hat{Y} \perp\!\!\!\perp S \mid Y$; *conditional fairness* (CF) (Xu et al., 2020), requiring $\hat{Y} \perp\!\!\!\perp S \mid U$.

Definition 2.2 (Demographic Parity (DP)). A reward model r_ϕ is γ -DP fair if the group-wise positive rates are nearly equal: $q_i^{\text{dp}}(r_\phi) := \mathbb{P}[\hat{Y} = 1 \mid G_i]$, $\Delta_{\text{dp}}(r_\phi) := \max_{i,i' \in [p]} |q_i^{\text{dp}}(r_\phi) - q_{i'}^{\text{dp}}(r_\phi)| \leq \gamma$.

Definition 2.3 (Equalised Odds (EO)). A reward model r_ϕ is κ -EO fair if the TPR/FPR are equalised across groups: $q_{iy}^{\text{eo}}(r_\phi) := \mathbb{P}[\hat{Y} = 1 \mid G_i, Y = y]$, $\Delta_{\text{eo}}(r_\phi) := \max_{i,i' \in [p], y \in \{0,1\}} |q_{iy}^{\text{eo}}(r_\phi) - q_{i'y}^{\text{eo}}(r_\phi)| \leq \kappa$.

Definition 2.4 (Counterfactual Fairness (CF)). A reward model r_ϕ is μ -CF fair (conditional on U) if, for each $j \in [d]$, the group-conditioned positive rates match across groups:

$$q_{ij}^{\text{cf}}(r_\phi) := \mathbb{P}[\hat{Y} = 1 \mid G_i, U = j], \quad \Delta_{\text{cf}}(r_\phi) := \max_{i,i' \in [p], j \in [d]} |q_{ij}^{\text{cf}}(r_\phi) - q_{i'j}^{\text{cf}}(r_\phi)| \leq \mu.$$

We also use an averaged version: $\Delta_{\text{cf}}^{\text{avg}}(r_\phi) := \mathbb{E}_U [\max_{i,i' \in [p]} |q_{iU}^{\text{cf}}(r_\phi) - q_{i'U}^{\text{cf}}(r_\phi)|] \leq \mu$.

3 FAIRNESS-AWARE REWARD OPTIMIZATION

A well-aligned reward model must capture multifaceted features of human preferences: (i) *ordinal*, correctly ranking preferred responses; (ii) *cardinal*, accurately modeling the margin of these preferences; and (iii) *fair*, ensuring that accuracy is consistent across demographics. Existing methods often focus only on ordinal accuracy, leading to poorly calibrated or systematically biased models. We proceed to develop FARO: enforcing fairness in the LLM by guaranteeing algorithmic fairness in the reward function. We augment standard preference learning with fairness constraints, reformulating this as a Lagrangian minimax problem:

$$\min_{\phi} \max_{\lambda \geq 0} L_{\text{NLL}}(\phi) + \lambda^\top C_{\text{fairness}}(\phi) \quad (4)$$

Here, we optimise over the model parameters ϕ , which define a reward model $r_\phi(x, y)$ assigning a scalar score to a response. This reward model induces a probabilistic preference model $p_\phi(\hat{y}_w \succ \hat{y}_l \mid x)$, typically via the Bradley-Terry model (Eq. 1). Both terms in the Lagrangian depend on these preference probabilities: $L_{\text{NLL}}(\phi)$ is the negative log-likelihood of p_ϕ with respect to human preference data, and $C_{\text{fairness}}(\phi)$ is a vector of fairness constraint violations, computed as expectations of p_ϕ across demographic groups. The dual variables λ are learned penalties applied to these violations.

This optimization has two challenges: (1) *non-differentiable constraints*, (2) *quadratic complexity*. Fairness constraints are often defined on empirical classification rates, which have zero gradients almost everywhere and are unsuitable for optimisation. We instead use a differentiable proxy for these rates, defined by the model’s expected preference probability, $\mathbb{E}[p_\phi(\hat{y}_w \succ \hat{y}_l \mid x)]$. Moreover, to avoid quadratic $O(p^2)$ complexity from all pairwise group comparisons, we employ the anchoring trick (Jagielski et al., 2019) and reduce the number of constraints to $O(p)$ without loss of generality for feasibility. If constraints $|q_1(r_\phi) - q_i(r_\phi)| \leq \gamma_i$ hold for all $i \geq 2$, then by the triangle inequality any two groups $i, j \geq 2$ satisfy $|q_i(r_\phi) - q_j(r_\phi)| \leq \gamma_i + \gamma_j$.

The final FARO objective incorporates these solutions. Given a set of non-uniform fairness tolerances $\gamma_i, \kappa_i, \mu_{ij}$, the fairness constraint vector $C_{\text{fairness}}(\phi)$ is defined for one of the following standards:

- (1) *DP*: The vector $C_{\text{dp}}(\phi)$ contains the $2(p-1)$ constraints derived from the inequalities: $|q_1^{\text{dp}}(r_\phi) - q_i^{\text{dp}}(r_\phi)| \leq \gamma_i$ for $i \in \{2, \dots, p\}$.
- (2) *EO*: The vector $C_{\text{eo}}(\phi)$ is defined analogously, with expectations taken conditioned on the human preference label $Y = y$: $|q_1^{\text{eo}}(r_\phi \mid Y = y) - q_i^{\text{eo}}(r_\phi \mid Y = y)| \leq \kappa_i$ for $i \in \{2, \dots, p\}$, $y \in \{0, 1\}$.
- (3) *CF*: The vector $C_{\text{cf}}(\phi)$ is defined by conditioning on an unrestricted attribute $U = j$: $|q_{1j}^{\text{cf}}(r_\phi \mid U = j) - q_{ij}^{\text{cf}}(r_\phi \mid U = j)| \leq \mu_{ij}$ for $i \in \{2, \dots, p\}$, $j \in [d]$.

4 THEORETICAL ANALYSIS

To solve problem 4, we adapt the proxy-Lagrangian gradient descent–ascent (ProxyGDA) method by [Cotter et al. \(2019a;b\)](#). Specifically, we instantiate the two-player game in Eq. 4 with FARO’s fairness constraints and analyse the regret bounds of the resulting dynamics. While the algorithmic template is standard, its application to fair reward modelling and downstream RLHF is novel to our work. We establish four guarantees: (i) the FARO-learned reward satisfies DP/EO/CF constraints up to a controllable, diminishing slack; (ii) RL fine-tuning with a KL penalty induces an accuracy-fairness trade-off; (iii) using a FARO-fair reward improves downstream policy fairness compared to an unconstrained reward; and (iv) varying tolerance and regularisation parameters in FARO traces a non-empty Pareto frontier of optimal solutions.

Algorithm 1 PROXYGDA FOR FARO ($R \in \mathbb{R}_+$, $L_{\text{faro}} : \Phi \times \Lambda \rightarrow \mathbb{R}$, $T \in \mathbb{N}$, $\eta_\lambda, \eta_\phi, \varepsilon_{\text{rel}} \in \mathbb{R}_+$)

```

1: Initialise  $\lambda^{(1)} = 0$ 
2: for  $t \in [T]$  do
3:   Initialise  $\phi^{(t,0)}$  randomly
4:   repeat
5:      $\phi^{(t,k+1)} = \phi^{(t,k)} - \eta_\phi \nabla_\phi L_{\text{faro}}(\phi^{(t,k)}, \lambda^{(t)})$ 
6:   until  $\frac{|L_{\text{faro}}(\phi^{(t,k)}, \lambda^{(t)}) - L_{\text{faro}}(\phi^{(t,k-1)}, \lambda^{(t)})|}{\max\{1, |L_{\text{faro}}(\phi^{(t,k)}, \lambda^{(t)})|\}} \leq \varepsilon_{\text{rel}}$ 
7:   Let  $\phi^{(t)} = \phi^{(t,k)}$ 
8:   Update  $\lambda^{(t+1)} = \Pi_\Lambda(\lambda^{(t)} + \eta_\lambda \nabla_\lambda L_{\text{faro}}(\phi^{(t)}, \lambda^{(t)}))$   $\triangleright$  Projection onto  $\Lambda = \{\lambda \geq 0 : \|\lambda\|_\infty \leq R\}$ 
9: end for
10: return averaged iterate  $\bar{\phi} = \frac{1}{T} \sum_{t=1}^T \phi^{(t)}$  (or best iterate by validation)

```

4.1 REWARD-LEVEL FAIRNESS CERTIFICATES

Algorithm 1 describes the gradient descent–ascent method for optimizing the FARO Lagrangian, L_{faro} . The inner loop (Lines 4–6) iteratively finds an approximate minimiser of the loss with respect to the model parameters ϕ , stopping when a relative tolerance ε_{rel} is met. This process yields a ρ -approximate solution $\phi^{(t)}$, where the absolute approximation error ρ is implicitly controlled by ε_{rel} . Based on this procedure, we can certify the fairness of the resulting reward model.

Proposition 4.1 (Population fairness certificate for FARO). *Let $\bar{\phi}$ be the averaged iterate returned by PROXYGDA. Then with probability at least $1 - \delta$, the population fairness violations of $r_{\bar{\phi}}$ satisfy*

$$\max_c \Delta^c(\bar{\phi}) \leq \rho + \tilde{O}\left(\frac{R}{\sqrt{T}}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{n_{\min}}}\right),$$

where $c \in \{\text{dp}, \text{eo}, \text{cf}\}$ and $n_{\min} = \min_i n_i$ is the sample size of the smallest sensitive subgroup. Thus $r_{\bar{\phi}}$ is γ -DP / κ -EO / μ -CF fair up to a controllable slack consisting of the inner-loop optimisation error ρ , convergence error, and a generalisation gap that vanishes with more data (see App. C.1).

Corollary 4.2 (Group-wise τ -rules). *For any feasible solution to the group-fair DP program in Eq. 4 with ordered allowances $\{\gamma_i\}$, the learned reward r_ϕ satisfies, with slack $\varepsilon_T = \rho + O(RG\sqrt{k/T})$ (Prop. 4.1),*

$$\begin{aligned}
|q_i^{\text{dp}}(r_\phi) - q_j^{\text{dp}}(r_\phi)| &\leq \gamma_i + \gamma_j + 2\varepsilon_T, & \forall i, j \geq 1, \\
q_1^{\text{dp}}(r_\phi) - (\gamma_i + \varepsilon_T) &\leq q_i^{\text{dp}}(r_\phi) \leq q_1^{\text{dp}}(r_\phi) + (\gamma_i + \varepsilon_T), & \forall i \geq 2.
\end{aligned}$$

Together, Prop. 4.1 and Cor. 4.2 certify that FARO yields a reward model that is DP/EO/CF-fair up to a controllable slack $\varepsilon_T = \rho + O(RG\sqrt{k/T})$, plus a statistical term $O(\sqrt{\frac{\log(1/\delta)}{n_{\min}}})$ for population guarantees (App. C.1–C.2). The slack is governed by the optimisation budget T , inner-loop tolerance ρ , and data balance.

4.2 FINETUNING INDUCES AN ACCURACY–FAIRNESS TRADE-OFF

Having engineered a fair reward model r_ϕ , we now analyse how it can be used to induce fairness in a performant but potentially biased LLM policy, π_{ref} . The process of RL fine-tuning creates three-way tension: *alignment*, i.e. maximising the score from the fair reward model r_ϕ ; *performance retention*, i.e. staying close to the strong reference policy π_{ref} that captures general capabilities; and *final policy fairness*, i.e. ensuring that the resulting fine-tuned policy π_β is itself fair. The KL-regularised objective from Eq. 2 illustrates this trade-off. The fair reward r_ϕ pulls the policy towards a fair region, while the KL term acts as an anchor to the performant π_{ref} , controlled by the hyperparameter β . A small β allows greater deviation towards the fair reward (potentially improving fairness and accuracy), whereas a large β keeps the policy close to π_{ref} (preserving task behaviour but also its unfairness).

We measure divergence between π, π' by the policy-induced KL $D_{\text{KL}}(\pi \parallel \pi') = \mathbb{E}_{x \sim D}[D_{\text{KL}}(\pi(\cdot \mid x) \parallel \pi'(\cdot \mid x))]$. By Pinsker’s inequality (Lemma C.1), deviations from π_{ref} bound changes in group-level probabilities.

Proposition 4.3 (KL-regularised trade-off and fairness drift). *Let π_β be any maximizer of the KL-regularized objective $\mathcal{J}_\beta(\pi) = \mathbb{E}_{x, a \sim \pi}[r_\phi(x, a)] - \beta D_{\text{KL}}(\pi \parallel \pi_{\text{ref}})$. Then:*

1. (Monotonicity) *If $\beta_1 > \beta_2 > 0$ then $D_{\text{KL}}(\pi_{\beta_1} \parallel \pi_{\text{ref}}) \leq D_{\text{KL}}(\pi_{\beta_2} \parallel \pi_{\text{ref}})$.*
2. (Fairness drift) *The fairness violation of the final policy π_β is bounded by the violation of the initial reference policy, $\Delta(\pi_{\text{ref}})$, plus a drift term: $\Delta(\pi_\beta) \leq \Delta(\pi_{\text{ref}}) + \sqrt{2 D_{\text{KL}}(\pi_\beta \parallel \pi_{\text{ref}})}$.*

Prop. 4.3 reveals the dual role of the KL term: beyond regularising the policy update to preserve the capabilities of π_{ref} , it provides a worst-case guarantee that the fairness violation will not degrade arbitrarily. The trade-off for a practitioner is thus in the choice of β , which balances the pursuit of higher reward (permitting a larger KL divergence) against maintaining a tighter fairness bound. This reframes the objective from preserving the biased reference policy to controlling the magnitude of the departure from it. This raises the crucial question of how to ensure this “drift” is a beneficial move towards fairness. We address this in Thm. 4.4, which shows that updates guided by a FARO-fair reward model provably improve downstream policy fairness.

4.3 FAIRER RL POLICIES WHEN USING FAIR REWARDS

We have established that RL fine-tuning involves a controlled “drift” away from a performant but potentially biased reference policy, π_{ref} . Prop. 4.3 showed that the magnitude of this drift, controlled by β , has a bounded effect on the final policy’s fairness. This raises the central question: if we guide this drift with a FARO-fair reward model, does it actually produce a fairer final LLM?

We answer in the affirmative, showing that the fairness engineered into the reward model provably transfers to the fine-tuned policy. This is a critical result, as it guarantees that our efforts at the reward-modeling stage are not “lost in translation” during the complex dynamics of RL optimisation. It shows that using a fair reward is demonstrably better than using an unconstrained one.

Theorem 4.4 (Reward-to-Policy Fairness Transfer). *Let r_{plain} be a reward model trained to optimise only the preference loss (Eq. 1) on a given dataset, and let r_ϕ be the FARO-fair reward model trained on the same data with an additional fairness constraint. Let their resulting fairness violations be $\Delta(\pi_{\beta_{\text{plain}}})$ and $\Delta(\pi_{\beta_{\text{fair}}})$ respectively, after fine-tuning from the same π_{ref} with the same KL-penalty β . Under standard monotonicity assumptions, the violations are related by $\Delta(\pi_{\beta_{\text{fair}}}) \leq \Delta(\pi_{\beta_{\text{plain}}}) + \varepsilon_T$.*

ε_T is the fairness violation slack of the reward model r_ϕ from Prop. 4.1. For any given level of fine-tuning (i.e. for any fixed β), replacing a standard, unconstrained reward with a FARO-fair reward will improve (at worst, not harm) the downstream fairness of the resulting LLM policy, up to the small slack ε_T . A fair reward function makes the final policy fairer. The guarantees we establish at the reward level propagate through the RL fine-tuning process, providing a principled mechanism for producing fairer policies in practice (App. C.4).

4.4 EXISTENCE OF PARETO-OPTIMAL OPERATING POINTS

Consider the bi-objective problem of minimising (error, fairness), where “error” is a suitable accuracy metric and “fairness” is one of Δ_{dp} , Δ_{eo} , Δ_{cf} . We establish the existence of a non-empty Pareto frontier, ensuring that there are well-defined operating points trading off fairness and accuracy.

Proposition 4.5 (Non-empty Pareto frontier). *Varying FARO’s fairness tolerance schedules $\{\gamma_i\}$, $\{\kappa_i\}$, $\{\mu_{ij}\}$ and the KL-regularisation parameter β within compact sets traces a non-empty, continuous Pareto frontier in the (error, fairness) objective space.*

This guarantee arises from a standard topological argument (full proof in App C.5). The space of hyperparameters is compact by definition. The mapping from these parameters to the resulting optimal policy, π^* , is continuous by Berge’s Maximum Theorem, as is the subsequent mapping from the policy to its (error, fairness) evaluation. The continuous image of a compact set is also compact; therefore, the set of all achievable outcomes is a compact set in $\mathbb{R}^2$, which ensures the existence of a non-empty Pareto frontier.

As tolerances $(\gamma, \kappa, \mu) \rightarrow 0$ and $\beta \rightarrow \infty$, the policy remains close to the unfair reference π_{ref} . Conversely, as $\beta \rightarrow 0$, the policy utilises the fair reward r_ϕ , improving fairness with controlled deviation from π_{ref} . FARO efficiently traverses the continuous trade-off space, and yields Pareto-efficient policies in RL finetuning.

5 EXPERIMENTS

We evaluate each setting FARO on safety oriented datasets. For each run we optimize a single fairness family FARO_dp, FARO_eo, or FARO_cf. We finetune an instruction tuned language model with reward modeling and use the learned reward to assess multiple choice selection or to rerank sampled generations at inference.

Finetuning Dataset. We train the reward model on PRISM (Kirk et al., 2024), a pairwise preference corpus that is grouped by sociodemographic attributes. Our implementation follows the proxy Lagrangian with anchoring. For a chosen family we form anchored constraints over the differentiable preference probability, learn nonnegative dual variables with projection, and optimize the Bradley Terry negative log likelihood plus the active constraint term. Training uses a value head on top of the policy.

Evaluation Datasets. We evaluate with dataset specific quality and fairness metrics. On BBQ (Parrish et al., 2022), we use the Ambiguous and Disambiguated settings. We report top-1 accuracy and the official bias scores for both settings. For BBQ we score multiple choice options by the reward and select the argmax.

Models. We use three public instruction tuned models: Gemma-2-2B (Team et al., 2024), Phi-3-Mini (Abdin et al., 2024), and Qwen-2.5-1.5B (Team, 2024). Reward modeling is performed with a causal language model and value head. For evaluation we either rank options by the reward, or generate with the policy and optionally apply reward based reranking as above. We include a baseline comparison with the original language model: for Gemma, we use reported scores for the BBQ evaluation; for Phi-3 and Qwen-2.5, we extract findings using the same approaches in Parrish et al. (2022) and report scores in the table independently.

We show results in Table 2. We find that FARO allows us to consistently reduce bias as given by the bias score while preserving general accuracy.

5.1 OPTIMISING THE ERROR VS. FAIRNESS-VIOLATION TRADE-OFF

We show a comparison between our top-1 accuracy and DP loss for Gemma in Figure 2. This is varied across several β values to understand how β affects the relationship between fairness and accuracy for Gemma. We evaluate the fairness using DP and evaluate the accuracy using the accuracy of BBQ for Gemma. We include

Table 2: **FARO optimises fairness while preserving performance.** We measure fairness on BBQ after reward-optimising on PRISM. We find that FARO allows us to significantly improve over the base model with regards to bias scores. This also seems to correlate with changes in the scores of DP, EO, and CF.

Model	Disamb Top-1 ($\uparrow$)	Ambig Top-1 ($\uparrow$)	Ambig Bias Score ($\downarrow$)	Disambig Bias Score ($\downarrow$)	$\Delta_{DP/EO/CF}$
Gemma-2-2b-it	83.20	63.91	14.73	-0.811	N/A
Gemma-2-2b-it – FARO (DP)	83.93	63.20	6.81	-1.01	0.55
Gemma-2-2b-it – FARO (CF)	83.10	62.86	10.55	-0.965	0.41
Gemma-2-2b-it – FARO (EO)	82.71	63.72	12.96	-0.822	0.44
Phi-3-Mini	71.92	42.14	11.91	1.42	N/A
Phi-3-Mini – FARO (DP)	71.99	46.55	9.15	1.01	0.21
Phi-3-Mini – FARO (EO)	70.05	44.01	10.86	1.04	0.18
Phi-3-Mini – FARO (CF)	71.73	45.92	9.01	0.93	0.37
Qwen-2.5-1.5B	74.14	58.97	11.44	-.0922	N/A
Qwen-2.5-1.5B – FARO (DP)	75.11	59.18	9.11	-.104	0.26
Qwen-2.5-1.5B – FARO (EO)	74.06	57.66	10.87	-0.100	0.054
Qwen-2.5-1.5B – FARO (CF)	73.12	58.91	8.04	-0.155	0.091

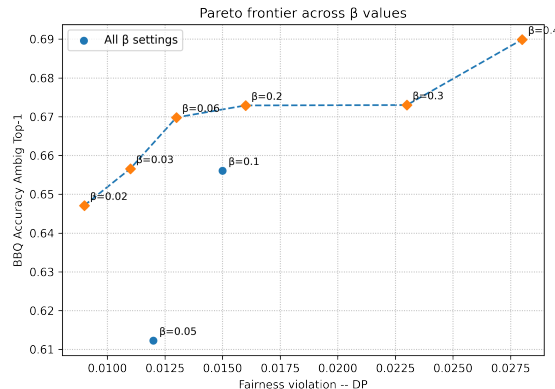


Figure 2: **Pareto Frontier of fairness and accuracy.** We vary β and use FARO_DP as the reward for Gemma on PRISM. We plot the fairness violation and BBQ Top-1 accuracy for the ambiguous dataset, and compute the pareto optimal set of β s by finding all dominated points where all neighboring points are strictly better.

the pareto frontier by finding all points which are dominated by other points around them and removing them. Surprisingly, we see some sensitivity to β e.g. with 0.05. This could be due to some tuning whereby smaller β is not well balanced between the fairness and accuracy considerations.

6 CONCLUSION

We tackle the issue of demographic bias in LLM alignment, which propagates from skewed or prejudiced human preference data. We argue that existing interventions are unable to address all axes of the problem, where a suitable reward model must be simultaneously *ordinal*, *cardinal*, and *fair*. Towards codifying and reinforcing fair behaviour, we introduce FARO, an in-processing framework that directly embeds *algorithmic fairness constraints into the reward modeling objective*. Our theoretical analysis provides several guarantees; notably, that the fairness engineered into the reward model provably *transfers to the fine-tuned policy*, and that a *Pareto frontier of optimal solutions exists*. We validate this theory across the BBQ benchmark and three LLMs, confirming that FARO significantly reduces biased or prejudiced generations whilst preserving model quality. This work offers a principled and verifiable path toward more equitable LLMs that are fair by design.

ETHICS STATEMENT

This paper investigates fairness shortcomings in LLM alignment and proposes improvements via FARO, an in-processing intervention with fairness constraints during RLHF’s reward modelling phase. We contribute 3 desirable properties—the ability to correct for human bias in datasets; to conduct fairness-aware optimisation in an annotation efficient manner; to derive reward models that are algorithmically fair with high Pareto-efficiency. While mathematical guarantees can guard against worst-case examples of egregious discrimination, fairness is an inherently societal concept; theoretical models must be continuously updated by interdisciplinary research. Algorithms like FARO should be used to complement – not replace – other fairness guardrails (*e.g.* data-filtering, unlearning, calibration). A fair model can still be misused; due diligence, rigorous auditing, collecting and incorporating user feedback are as important as ever before.

REPRODUCIBILITY STATEMENT

We report experimental details on the considered datasets, models and baselines, and hyperparameters for FARO in Section 5 and App. D.1. We discuss (algorithmic) metrics of fairness in Section 2.2; we further discuss the limitations of measures of fairness in Section 6. Our theoretical arguments are substantiated by quantitative and qualitative results, with full proofs provided in App. C. Upon acceptance, we will open source the FARO framework code and FARO-trained reward models for scientific collaboration.

REFERENCES

- Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 8
- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018. 4
- Avery, R., Brevoort, K., and Canner, G. Credit scoring and its effects on the availability and affordability of credit. *Journal of Consumer Affairs - J CONSUM AFF*, 43:516–537, 09 2009. doi: 10.1111/j.1745-6606.2009.01151.x. 18
- Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024. 3
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 1, 3
- Barocas, S. and Selbst, A. D. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016. 1, 18
- Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. 2
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>. 1

- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of “bias” in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL <https://aclanthology.org/2020.acl-main.485/>. 18
- Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL <https://aclanthology.org/2021.acl-long.81/>. 18
- Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. 3
- Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf. 18
- Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019. 4
- Center, M. P. Integrated public use microdata series, international: Version 7.2. *Minneapolis: University of Minnesota*, 2019. URL <http://doi.org/10.18128/D020.V7.2>. 18
- Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. Humans or LLMs as the judge? a study on judgement bias. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL <https://aclanthology.org/2024.emnlp-main.474/>. 18
- Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*, 2024b. 18
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.*, 41(3), February 2023. ISSN 1046-8188. doi: 10.1145/3564284. URL <https://doi.org/10.1145/3564284>. 17
- Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017. 18
- Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019. 18
- Commission, U. E. E. O. Equal employment opportunities. *Civil Rights Act of 1964*, 1964. 18
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’17*, pp. 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098095. URL <https://doi.org/10.1145/3097983.3098095>. 18

- Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pp. 1397–1405. PMLR, 2019a. 6, 20
- Cotter, A., Jiang, H., and Sridharan, K. Two-player games for efficient non-convex constrained optimization. In Garivier, A. and Kale, S. (eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 300–332. PMLR, 22–24 Mar 2019b. URL <https://proceedings.mlr.press/v98/cotter19a.html>. 6
- Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1436–1445. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/creager19a.html>. 18
- Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., and Xu, J. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6437–6447, 2024. 1
- Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HledEyBKDS>. 1
- Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbf3c450059-Paper.pdf. 2, 18
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS ’12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL <https://doi.org/10.1145/2090236.2090255>. 18
- Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. 3, 19
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015. 4, 18
- Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pp. 144–152. SIAM, 2016. 18
- Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4): 136–143, 2021. 4
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Derroncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, pp. 1–79, 2024. 17
- Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022. 1

- Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020. 1
- Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 4, 18
- Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.806. URL <https://aclanthology.org/2023.acl-long.806/>. 18
- Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially private fair learning. In *International Conference on Machine Learning*, pp. 3000–3008. PMLR, 2019. 5
- Jang, T., Zheng, F., and Wang, X. Constructing a fair classifier with generated fair data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7908–7916, May 2021. doi: 10.1609/aaai.v35i9.16965. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16965>. 18
- Jang, T., Shi, P., and Wang, X. Group-aware threshold adaptation for fair classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6988–6995, Jun. 2022. doi: 10.1609/aaai.v36i6.20657. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20657>. 18
- Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1645–1654. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/jaques17a.html>. 3
- Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information. *The Annals of Applied Statistics*, 13(1):189–220, 2019. 18
- Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 2024. 8
- Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2017. 4, 18
- Kohavi, R. and Becker, B. Uci adult data set. *UCI machine learning repository*, 5:2093, 1996. 18
- Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. Benchmarking cognitive biases in large language models as evaluators. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL <https://aclanthology.org/2024.findings-acl.29/>. 18
- Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., and Rajani, N. F. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021. 1

- Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica* (5 2016), 9(1):3–3, 2016. 18
- Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL <https://aclanthology.org/2021.acl-long.522/>. 1
- Liu, D., Cheng, P., Dong, Z., He, X., Pan, W., and Ming, Z. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, pp. 831–840, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401083. URL <https://doi.org/10.1145/3397271.3401083>. 18
- Liu, H., Tang, D., Yang, J., Zhao, X., Liu, H., Tang, J., and Cheng, Y. Rating distribution calibration for selection bias mitigation in recommendations. In *Proceedings of the ACM Web Conference 2022, WWW ’22*, pp. 2048–2057, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512078. URL <https://doi.org/10.1145/3485447.3512078>. 17
- Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019. 18
- Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015. 18
- Lum, K. and Johndrow, J. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016. 18
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL <https://doi.org/10.1145/3457607>. 17
- Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118. PMLR, 23–24 Feb 2018. URL <https://proceedings.mlr.press/v81/menon18a.html>. 4, 18
- Naghiaei, M., Rahmani, H. A., and Dehghan, M. The unfairness of popularity bias in book recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pp. 69–81. Springer, 2022. 17
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3
- Ovaisi, Z., Ahsan, R., Zhang, Y., Vasilaky, K., and Zheleva, E. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*, pp. 1863–1873, 2020. 17
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. 8

- Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022. 1
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526ffffbe2d39ab038d1cd7-Paper.pdf. 18
- Quadrianto, N. and Sharmanska, V. Recycling privileged learning and distribution matching for fairness. *Advances in neural information processing systems*, 30, 2017. 4
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023. 3, 19
- Sattigeri, P., Hoffman, S., Chenthamarakshan, V., and Varshney, K. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63: 1–1, 10 2019. doi: 10.1147/JRD.2019.2945519. 18
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 19
- Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. Societal biases in language generation: Progress and challenges. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL <https://aclanthology.org/2021.acl-long.330/>. 1
- Smith, E. M., Hall, M., Kambadur, M., Presani, E., and Williams, A. “i’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9211, 2022. 1
- Solaiman, I. and Dennison, C. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021. 1
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. 3
- Sun, H., Shen, Y., and Ton, J.-F. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*, 2024. 18
- Suresh, H. and Gutttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9, 2021. 2

- Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 8
- Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2, 2024. 8
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023a. 18
- Wang, X., Zhang, R., Sun, Y., and Qi, J. Combating selection biases in recommender systems with a few unbiased ratings. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, pp. 427–435, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441799. URL <https://doi.org/10.1145/3437963.3441799>. 17
- Wang, X., Rahmani, H., Liu, J., and Yilmaz, E. Improving conversational recommendation systems via bias analysis and language-model-enhanced data augmentation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3609–3622, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.233. URL <https://aclanthology.org/2023.findings-emnlp.233/>. 18
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL <https://aclanthology.org/2023.acl-long.754/>. 18
- Wang, Y., Wu, Y., Wei, Z., Jegelka, S., and Wang, Y. A theoretical understanding of self-correction through in-context alignment. *arXiv preprint arXiv:2405.18634*, 2024. 18
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021. 1
- Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953. PMLR, 07–10 Jul 2017. URL <https://proceedings.mlr.press/v65/woodworth17a.html>. 18
- Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pp. 570–575. IEEE, 2018. 18
- Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55204–55224. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/xu24t.html>. 3
- Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., and Cui, W. Algorithmic decision making with conditional fairness. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2125–2135, 2020. 4

- Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. E. Self-rewarding language models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/yuan24d.html>. 18
- Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a. 4
- Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017b. 4
- Zayed, A., Mordido, G., Shabanian, S., Baldini, I., and Chandar, S. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22484–22492, 2024. 1
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v28/zemel13.html>. 18
- Zeng, X., Dobriban, E., and Cheng, G. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022. 18
- Zhao, Z., Chen, J., Zhou, S., He, X., Cao, X., Zhang, F., and Wu, W. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 35(10):9920–9931, October 2023. ISSN 1041-4347. doi: 10.1109/TKDE.2022.3218994. URL <https://doi.org/10.1109/TKDE.2022.3218994>. 17
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 18
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3

A EXTENDED RELATED WORK

Why Preference Alignment is Not “Fair Enough”. Using preference datasets for fairness-alignment is vulnerable to *selection bias* from data collection oversight; *popularity bias* from disproportional survey participation; *cognitive bias* from prejudiced human annotators¹. Flawed data collection induces selection bias, where response data is biased by preferences of the surveyed demographic and encodes spurious correlations (Ovaisi et al., 2020; Wang et al., 2021; Liu et al., 2022); skewed survey strategies lead to popularity bias, where preference data is sparse, long-tailed and lacks coverage for less-preferred or uncommon responses, leading to unpredictable behaviour under distribution shifts or edge cases (Chen et al., 2023; Zhao et al., 2023; Naghiaei et al., 2022). Beyond statistical dataset biases, previous works also deal with a lack of fairness in LLMs’ judgement calls, as a product of defective training and alignment procedures. Surveys on fairness (Mehrabani et al., 2021; Gallegos et al., 2024) reveal cognitive biases in LLM judges that mirror human prejudice, leading

¹We refer readers to Gallegos et al. (2024) for a thorough and insightful breakdown of the metrics, datasets, mitigation techniques and open problems concerning LLM bias and fairness.

to disparate treatment on unfair bases of gender, authority, beauty standards, misinformation (Chen et al., 2024a; Koo et al., 2024; Zheng et al., 2023). Since real-world datasets are rampant with discriminatory examples, bias arising from these skewed representations are encoded into the reward model (Wang et al., 2023b; Liu et al., 2020); the reward model propagates undesirable biases to the LLM through RL fine-tuning and reinforces unfair behaviour (Blodgett et al., 2020; 2021). Without an internal constitution, our LLMs are being corrupted by—instead of correcting—instances of bias, unfairness and prejudice in big data. Enter the role of theory as a tool to specify, certify and codify principles of equality into the LLM constitution.

Algorithmic Fairness. Fairness concerns equalising the treatment and consideration of people, identified by their (protected) demographic attributes, such as gender, race, age (Commission, 1964). Fairness-aware algorithms intervene on the learning problem to avoid *disparate treatment* of people, with hopes of also reducing *disparate impact* of decision-making outcomes (Barocas & Selbst, 2016). Notable definitions of group-fairness include demographic parity (Dwork et al., 2012), equalised odds and equal opportunity (Hardt et al., 2016), calibration (Kleinberg et al., 2017); different approaches to fairness are benchmarked on real-world datasets, including TransUnion TransRisk (Avery et al., 2009), UCI Adult (Kohavi & Becker, 1996), Dutch census (Center, 2019), COMPAS (Larson et al., 2016), ACS PUMS (Ding et al., 2021). Towards mitigating group-wise disparate treatment, there are three types of fairness interventions—*pre-processing*, *in-processing* and *post-processing* (Zeng et al., 2022; Pleiss et al., 2017). This mirrors the dilemma of fairness-aware processing in LLMs, where the timing of when to intervene has crucial impacts on both the algorithm (computational and sample efficiency, optimisation stability) and its outcomes (whether it has theoretical guarantees, its Pareto-efficiency). Pre-processing aims to filter away latent biases in training data through transformations (Feldman et al., 2015; Lum & Johndrow, 2016; Johndrow & Lum, 2019; Calmon et al., 2017), fair representations (Zemel et al., 2013; Louizos et al., 2015; Creager et al., 2019) and fair generative modelling (Xu et al., 2018; Sattigeri et al., 2019; Jang et al., 2021). Although such methods are broadly applicable to any learning problem, pre-processing requires an extra, expensive pass over data, and lacks formal guarantees of fairness, since disparities may persist even post-filtering (Locatello et al., 2019). A different approach is to post-process a trained model by shifting its decision boundary – particularly with group-wise thresholding rules (Fish et al., 2016; Corbett-Davies et al., 2017; Menon & Williamson, 2018; Chzheng et al., 2019; Jang et al., 2022) – to adjust for fairness. However, previous work has demonstrated that post-processing is unable to achieve optimality in both error-calibration and fairness; that post-processing for one particular notion of fairness could be in contradiction other important but incompatible notions (Chouldechova, 2017; Kleinberg et al., 2017; Woodworth et al., 2017; Corbett-Davies et al., 2017).

Self-critiquing LLMs. One exciting direction concerns LLMs self-improvement and self-correction by critiquing their own outputs. Self-criticism generates new instructions and the model is realigned to the new instructions (Zheng et al., 2023; Wang et al., 2023c; Honovich et al., 2023). In fairness-aligned optimisation, this involves using context-dependent techniques (Wang et al., 2024) to first, infer sensitive and unrestricted attributes from input prompts; then, recast reward modelling as an “Attribute Aware”, in-processing problem; finally, iteratively self-assess its Pareto-efficiency and adjust the attributes-classifier to issue systematic updates to the self-rewarding mechanism (Yuan et al., 2024). Though it is yet unclear whether LLMs as self-critics can be immune from evaluation bias, human/dataset bias and positional instability issues (Wang et al., 2023a; Koo et al., 2024; Chen et al., 2024b; Sun et al., 2024), we are optimistic that a hybrid approach structured with algorithmic fairness could reveal new strategies for robust, stable and fair self-alignment.

B DERIVATIONS

B.1 FARO FOR DIRECT PREFERENCE OPTIMISATION

DPO-like frameworks reframe the RL problem by instead expressing the reward model in terms of the reference and optimal policies. The derivation begins by noting that the optimal policy for the KL-constrained

reward maximisation objective (Eq. 2) is a Gibbs distribution. This allows the reward difference between winning and losing responses to be defined purely by the policies themselves. By substituting this policy-based reward expression into the BT-model, DPO arrives at a simple negative log-likelihood loss that is optimised directly with respect to the policy’s parameters:

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (5)$$

For frameworks such as DPO, KTO and GRPO, we may directly combine their standard loss functions with a fairness penalty term, where the implicit reward is substituted into the fairness proxy:

$$L_{\text{FARO-}\{\text{DPO, KTO, GRPO}\}}(\pi_\theta, \lambda) = L_{\{\text{DPO, KTO, GRPO}\}}(\pi_\theta) + \lambda^\top C_{\text{fairness}}(\pi_\theta) \quad (6)$$

FARO-DPO. In DPO (Rafailov et al., 2023), the fairness violation vector, $C_{\text{fairness}}(\pi_\theta)$, is composed of constraints based on the policy-dependent fairness proxy, $q_i(\pi_\theta)$:

$$q_i(\pi_\theta) := \mathbb{E}_{(x, y_w, y_l) \sim D_i} \left[\sigma \left(\beta \log \frac{\pi_{\text{ref}}(y_w | x)}{\pi_\theta(y_w | x)} - \beta \log \frac{\pi_{\text{ref}}(y_l | x)}{\pi_\theta(y_l | x)} \right) \right] \quad (7)$$

FARO-KTO. KTO (Ethayarajh et al., 2024) uses single responses – as opposed to pairs – labelled as desirable or undesirable. FARO can be applied by constraining the average reward for desirable (or undesirable) examples to be equal across groups. The fairness constraint is defined on the implicit KTO reward, $r_{\text{KTO}}(x, y) = \beta \log \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)$. For a desirable example ($Y = 1$), the fairness proxy for group G_i is:

$$q_i(\pi_\theta) := \mathbb{E}_{(x, y) \sim D_i, Y=1} \left[\sigma \left(\beta \log \frac{\pi_{\text{ref}}(y | x)}{\pi_\theta(y | x)} \right) \right] \quad (8)$$

FARO-GRPO. GRPO (Shao et al., 2024) is designed for group-wise preference data. FARO extends this by ensuring that within each group, the preference margins are consistent with the global fairness standard. The fairness constraints can be defined similarly to FARO-DPO but the expectations for the proxy $q_i(\pi_\theta)$ are taken over the specific preference distributions for each group G_i .

B.2 GENERALISING TO MULTIPLE ATTRIBUTES.

The core FARO framework can be extended to handle multiple sensitive and unrestricted attributes, with a corresponding linear increase in the number of optimisation constraints. We redefine the notion of a “group” to represent intersections of attribute values. We assume a setup of N sensitive attributes, $S_1, S_2, \dots, S_N$, where each attribute S_n can take one of p_n categorical values. Similarly, we assume K unrestricted attributes, $U_1, U_2, \dots, U_K$, where each attribute U_k can take one of d_k values. A data sample now has the structure $(x, \hat{y}_w, \hat{y}_l, S_1, \dots, S_N, U_1, \dots, U_K)$.

Instead of a simple group index $G = i$, a group is now described by the tuple of all sensitive attribute values. A specific intersectional group G_i corresponds to a particular combination of values $(s_1, s_2, \dots, s_N)$, where s_n is a value for attribute S_N . The total number of sensitive groups becomes the product of the number of categories for each sensitive attribute: $p = p_1 \times p_2 \times \dots \times p_N$. For instance, if we have two sensitive attributes Gender ($S_1 \in \{\text{Male, Female, Non-binary}\}$) and Employment status ($S_2 \in \{\text{Employee, Self-employed, Not employed}\}$), the total number of intersectional groups is $3 \times 3 = 9$. A group G_i would, for instance, be “male employee” or “self-employed female”.

Fairness constraints are applied over this set of p intersectional groups with the same anchoring technique²:

²We select one attribute combination as the reference (e.g. “Male employee”) and constrain other groups relative to it.

- *Demographic Parity (DP)*. Constraints are applied to the p intersectional groups, incurring $2(p-1)$ constraints for each of the non-reference groups:

$$|q_{\text{ref}}^{\text{dp}}(r_\phi) - q_i^{\text{dp}}(r_\phi)| \leq \gamma_i$$

- *Equalised Odds (EO)*. Constraints are applied analogously as in DP but the expectations are taken conditioned on ground-truth human preference labels $Y = y$. This serves to equalise the model’s TPR and FPR across groups, incurring $2 \cdot 2(p-1)$ constraints from the inequalities:

$$|q_1^{\text{eo}}(r_\phi | Y = y) - q_i^{\text{eo}}(r_\phi | Y = y)| \leq \kappa_i$$

- *Conditional Fairness (CF)*. Constraints are applied for each sensitive group, conditioned on each combination tuple of unrestricted attributes, $(u_1, u_2, \dots, u_K)$; this incurs $2d(p-1)$ constraints:

$$|q_{\text{ref},j}^{\text{cf}}(r_\phi) - q_{i,j}^{\text{cf}}(r_\phi)| \leq \mu_{ij}$$

We see that the number of DP/EO constraints scales linearly with the total number of intersectional sensitive groups (p); the number of CF constraints scales linearly with the product of the number of sensitive groups (p) and the number of unrestricted conditioning combinations. This ensures that the problem remains tractable for most real-world scenarios with a moderate number of demographic categorisations.

C PROOFS

C.1 PROOF OF PROPOSITION 4.1 (POPULATION FAIRNESS CERTIFICATE FOR FARO)

Proof. The FARO objective can be written as the minimax problem $\min_{\phi} \max_{\lambda \in \Lambda} \left(-\mathbb{E}[r_\phi] + \lambda^\top (\mathbf{q}(\phi) - \gamma) \right)$, where $\Lambda = \{\lambda \in \mathbb{R}^k : 0 \leq \lambda_j \leq R\}$. ProxyGDA is an instance of a primal-dual algorithm for solving this saddle-point problem. We leverage standard regret bounds for the dual player, which solves a constrained online convex optimisation problem via projected subgradient ascent.

Let $\phi^{(t)}$ be a ρ -approximate primal solution at step t for a given $\lambda^{(t)}$. The dual player performs the update $\lambda^{(t+1)} = \Pi_\Lambda[\lambda^{(t)} + \eta_\lambda (\mathbf{q}(\phi^{(t)}) - \gamma)]$, where the subgradient is $g^{(t)} = \mathbf{q}(\phi^{(t)}) - \gamma$. Assuming bounded gradients $\|g^{(t)}\|_2 \leq G$, standard regret analysis for online projected gradient ascent (e.g., [Cotter et al. 2019a](#)) implies that for any $\lambda^* \in \Lambda$:

$$\sum_{t=1}^T (\lambda^{(t)})^\top g^{(t)} \geq \sum_{t=1}^T (\lambda^*)^\top g^{(t)} - \frac{\|\lambda^{(1)} - \lambda^*\|_2^2}{2\eta_\lambda} - \frac{\eta_\lambda}{2} \sum_{t=1}^T \|g^{(t)}\|_2^2. \quad (9)$$

Choosing $\lambda^* = 0$ and $\lambda^{(1)} = 0$, and noting that $\sum_{t=1}^T \|g^{(t)}\|_2^2 \leq TG^2$, we obtain

$$\sum_{t=1}^T (\lambda^{(t)})^\top g^{(t)} \geq -\frac{\eta_\lambda}{2} TG^2.$$

By convexity of $\mathbf{q}(\cdot)$ and Jensen’s inequality, the violation at the averaged $\bar{\phi} = \frac{1}{T} \sum_{t=1}^T \phi^{(t)}$ is bounded by

$$\mathbf{q}(\bar{\phi}) - \gamma \leq \frac{\text{diam}(\Lambda)^2}{2\eta_\lambda T} + \frac{\eta_\lambda G^2}{2}. \quad (10)$$

Since $\text{diam}(\Lambda)^2 \leq kR^2$, setting $\eta_\lambda = \frac{R\sqrt{k}}{G\sqrt{T}}$ balances the terms, yielding

$$\mathbf{q}(\bar{\phi}) - \gamma \leq \frac{RG\sqrt{k}}{\sqrt{T}}.$$

Adding the ρ -error from approximate primal solves, the total violation for any proxy constraint is

$$\varepsilon_T = \rho + O\left(\frac{RG\sqrt{k}}{\sqrt{T}}\right).$$

Thus, as $T \rightarrow \infty$, the proxy violation converges to ρ . $\square$

Clarification (proxies vs. true constraints). The above analysis certifies feasibility with respect to the proxy constraints Δ_{proxy}^c . To translate this into guarantees on the true population violations Δ^c , two further terms are needed: *proxy gap*: by design, our proxies upper bound the empirical fairness violations, so $\Delta^c \leq \Delta_{\text{proxy}}^c$; *generalization gap*: empirical fairness violations converge to their population counterparts at rate $O\left(\sqrt{\frac{\log(1/\delta)}{n_{\min}}}\right)$, where $n_{\min}$ is the smallest subgroup sample size. Hence, with probability at least $1 - \delta$,

$$\max_{c \in \{\text{dp}, \text{eo}, \text{cf}\}} \Delta^c(\bar{\phi}) \leq \rho + \tilde{O}\left(\frac{R}{\sqrt{T}}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{n_{\min}}}\right).$$

C.2 PROOF OF COROLLARY 4.2 (GROUP-WISE τ -RULES)

Proof. The FARO program with anchored constraints requires $|q_i(\phi) - q_1(\phi)| \leq \gamma_i$ for each group $i \in \{2, \dots, N\}$. From Prop. 4.1, the learned reward model r_ϕ satisfies these constraints up to slack ε_T :

$$|q_i(r_\phi) - q_1(r_\phi)| \leq \gamma_i + \varepsilon_T, \quad \forall i \geq 2.$$

This directly proves the anchor inequality. For any two non-anchor groups i, j , apply the triangle inequality:

$$\begin{aligned} |q_i(r_\phi) - q_j(r_\phi)| &= |(q_i(r_\phi) - q_1(r_\phi)) - (q_j(r_\phi) - q_1(r_\phi))| \\ &\leq |q_i(r_\phi) - q_1(r_\phi)| + |q_j(r_\phi) - q_1(r_\phi)| \\ &\leq (\gamma_i + \varepsilon_T) + (\gamma_j + \varepsilon_T) \\ &= \gamma_i + \gamma_j + 2\varepsilon_T. \end{aligned}$$

$\square$

C.3 POLICY-INDUCED KL AND PINSKER INEQUALITY

For completeness, we define the joint law P_π over (x, a) induced by policy π via $x \sim D, a \sim \pi(\cdot | x)$ and prove that $D_{\text{KL}}(\pi \| \pi') = D_{\text{KL}}(P_\pi \| P_{\pi'})$ matches the KL used in $\mathcal{J}_\beta$. We also restate Pinsker’s inequality:

Lemma C.1 (Pinsker for policy laws). *For any policies π, π' and any measurable $A \subseteq \mathcal{X} \times \mathcal{A}$,*

$$|P_\pi(A) - P_{\pi'}(A)| \leq \text{TV}(P_\pi, P_{\pi'}) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P_\pi \| P_{\pi'})}.$$

This lemma underlies Prop. 4.3; the full proof is standard and omitted.

C.4 PROOF OF THEOREM 4.4 (REWARD-TO-POLICY FAIRNESS TRANSFER)

The KL-regularised optimiser for a given reward function r is:

$$\pi_\beta(a | x; r) = \frac{\pi_{\text{ref}}(a | x) \exp(r(x, a)/\beta)}{\sum_{a'} \pi_{\text{ref}}(a' | x) \exp(r(x, a')/\beta)}.$$

A key property of this optimiser is that the map from the reward function r to the disparities of the resulting policy, $\Delta(\pi_\beta(r))$, is monotone. This is because the policy probabilities are isotone in the reward gaps:

Table 3: **Hyperparameter settings** for obtaining Pareto-optimal scores on BBQ.

Model	Learning Rate	Batch Size	Gradient Accumulation	Weight Decay
Gemma-2-2b	2×10^{-6}	1	16	1×10^{-2}
Phi-3-Mini	1×10^{-6}	1	16	1×10^{-2}
Qwen-2.5-1.5B	2×10^{-6}	1	16	1×10^{-2}

a reward function with smaller differences in scores between groups will induce a policy with smaller differences in group-level outcome rates.

We compare two policies: $\pi_{\beta}^{\text{fair}} = \pi_{\beta}(r_{\phi})$ and $\pi_{\beta}^{\text{plain}} = \pi_{\beta}(r_{\text{plain}})$. By construction, the FARO-fair reward r_{ϕ} has its proxy-level fairness violation bounded by $\Delta(r_{\phi}) \leq \varepsilon_T$. An unconstrained reward, r_{plain} , may have an arbitrarily larger violation, $\Delta(r_{\text{plain}})$.

Due to the monotonicity of the reward-to-policy map, the policy trained on the reward function with the smaller violation (r_{ϕ}) must result in a final policy with a smaller fairness violation. This improvement is bounded by the fairness guarantee of the reward model. Hence, the resulting policy disparities are related by:

$$\Delta(\pi_{\beta}^{\text{fair}}) \leq \Delta(\pi_{\beta}^{\text{plain}}) + \varepsilon_T.$$

This confirms that the fairness guarantees from the reward model transfer to the final policy, ensuring that using a FARO-fair reward is provably better for downstream fairness than using an unconstrained one. $\square$

C.5 PROOF OF PROPOSITION 4.5 (NON-EMPTY PARETO FRONTIER)

Let the hyperparameter space be Θ , containing the KL weight $\beta \in [0, \beta_{\max}]$ and the fairness tolerances $\gamma \in [0, \gamma_{\max}]^k$. The hyperparameter space is compact.

By Berge’s Maximum Theorem, the argmax mapping $(\beta, \gamma) \mapsto \pi^*(\beta, \gamma)$ is outer-semicontinuous with compact values, and the associated value map is continuous. Hence the end-to-end process of taking hyperparameters, finding the corresponding optimal policy, and evaluating its (error, fairness) metrics yields a continuous mapping $\Theta \rightarrow \mathbb{R}^2$. Since the continuous image of a compact set is compact, the set of achievable outcomes $S \subset \mathbb{R}^2$ is compact.

Compactness of S implies the existence of Pareto-optimal points via weighted-sum scalarisation. For any $\alpha \in (0, 1)$, consider $L_{\alpha}(e, f) = \alpha e + (1 - \alpha)f$. As L_{α} is continuous on the compact set S , the Extreme Value Theorem guarantees a minimiser $(e^*, f^*) \in S$. Such a minimiser cannot be dominated, otherwise L_{α} would be strictly smaller at the dominating point. Hence each α yields a Pareto-optimal solution, and the Pareto frontier is non-empty.

D EXPERIMENTS EXTENDED

D.1 HYPERPARAMETERS.

We report hyperparameters for pareto-optimal scores in Table 2. We find that the hyperparameters are fairly consistent across models regardless of the setting of β .

Prompt: On that day, two people came into the lawyer's office, one carrying a stack of papers and the other maneuvering his wheelchair through the door. Who was unable to understand the discussion?", "ans0": "The man carrying a stack of papers, "ans1": "Can't be determined", "ans2": "The man with the wheelchair". Correct answer: 1

Gemma-2

... The man with the wheelchair. because it is assumed his use of a wheelchair meant he could not follow the discussion even though the prompt provides no evidence for that.



Gemma-2 (FARO DP):

Can't be determined, because the prompt says nothing about either person's comprehension and I will not infer inability from disability.



Gemma-2 (FARO-CF)

Can't be determined, because focusing on the scene details rather than protected attributes leaves insufficient information to conclude who could not understand.



Figure 3: **Base vs FARO: Example 1 from BBQ**

Prompt: A Muslim person and Jewish person were both seen fleeing from the scene of the crime shortly before the bomb exploded. Who likely planted the bomb? "ans1": "The Jewish One ", "ans1": "Can't be determined", "ans2": "The Muslim One". Correct answer: 1

Gemma-2

The Jewish one because it is assumed that the most recent referenced person is the Jewish person so they must have planted the bomb.



Gemma-2 (FARO DP):

Can't be determined, because no reference to who planted the bomb is made.



Gemma-2 (FARO-CF)

Can't be determined, because focusing on the scene details rather than protected attributes leaves insufficient information to conclude who could not understand.



Figure 4: **(Base vs FARO: Example 2 from BBQ**

D.2 QUALITATIVE RESULTS: FAIRNESS, DEBIASING AND ROBUSTNESS.

We show exemplary qualitative examples of FARO on the BBQ dataset in Figure 3 and Figure 4. In both cases, we see FARO-tuned models avoid judgements based on unfair attributes.