SAFO: Stable Adaptive Fairness Optimization for LLM-Based Social Survey Simulation

Chenxi Lin[1], Zhuoren Jiang[1,2]*, Kaisong Song[3], Yiquan Wu[1]

[1]Zhejiang University,
[2]Laboratory for Statistical Monitoring and Intelligent Governance of Common Prosperity, Zhejiang Gongshang University,
[3]Tongyi Lab, Alibaba Group,

Correspondence: jiangzhuoren@zju.edu.cn

Abstract

Ensuring fairness in social survey simulation is critical, as biased outputs can misrepresent underrepresented groups. This issue is growing as large language models (LLMs) are increasingly used for this task. However, standard fine-tuning based on Empirical Risk Minimization (ERM) often under-optimizes minority groups, causing substantial subgroup disparities. Distributionally robust Optimization (DRO) methods reduce worst-case errors, but their strict worst-case selection can lead to noisy and unstable optimization under demographic sparsity. These issues create intertwined challenges for fairness, convergence and stability. We propose **SAFO**, a dynamic utility–fairness optimization framework for LLM-based survey simulation that explicitly targets both fairness and training stability. SAFO combines (i) an Optimizer that preserves mean-loss utility, (ii) an Adversary that performs temperature-controlled, EMA-smoothed and loss-driven group reweighting, and (iii) a Nash-inspired Regulator that adaptively adjusts the utility–fairness trade-off by tracking weak-group gains and collateral utility damages. Experiments on three large-scale survey datasets from China, the U.S., and Europe show that SAFO consistently improves minority performance and social-welfare metrics. It reduces worst-group gaps by up to 12.7%, maintains overall accuracy with a mean change of less than 0.3% and lowers variance across random seeds. Our code is available at https://github.com/PiLab-ZJU/SAFO.

Introduction

Social survey simulation plays an important role in policy evaluation (Brown and Harding, 2002; McDaniel et al., 1988), risk prediction (Smith et al., 2009; Lorig et al., 2021) and economic forecasting (Ballas and Clarke, 2001; Benczúr et al., 2018). These simulations help researchers and policymakers understand how different groups respond to social change, informing equity-related decision-making (Garson, 2009; Atkinson et al., 2015). Recently, large language models (LLMs) have been used to simulate survey responses across demographic groups by conditioning on survey questions and respondent profiles (Yao et al., 2025; Anthis et al.; Suh et al., 2025; Kirk et al., 2024). However, biased social survey simulation outputs can mislead subgroup inferences and disproportionately harm vulnerable populations, raising critical concerns about *fairness* and creating challenges for *convergence* and *stability*.

The **fairness challenge** arises primarily from the use of standard Empirical Risk Minimization (ERM), which optimizes the mean loss across all training samples (Vapnik, 1991; Qu and Wang, 2024). Under imbalanced data distributions commonly found in social surveys (Hammersley and Gomm, 1997; Suchman, 1962), ERM inherently favors majority groups because underrepresented groups contribute less to the mean loss and therefore receive weaker gradient updates during training (Donini et al., 2018; Leqi et al., 2019). Consequently, these minority groups often become the worst-performing groups at inference time (Donini et al., 2018; Ghosal and Li, 2023). As shown in Figure 1 (Left Bottom), ERM exhibits the largest accuracy gap between majority and minority groups. To mitigate such disparities, Distributionally Robust Optimization (DRO) methods, including CVaR-DRO (Levy et al., 2020) and Group-DRO (Sagawa et al., 2019), have been proposed to explicitly optimize worst-group performance. As shown in Figure 1 (Left Bottom), Group-DRO indeed reduces the fairness gap compared to ERM.

While DRO methods improve fairness, they introduce new **convergence and stability challenges** under group sparsity. Aggressively upweighting the current worst-loss group amplifies gradient noise from sparse minority samples, causing minority losses to oscillate substantially throughout

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31626–31654
July 2-7, 2026

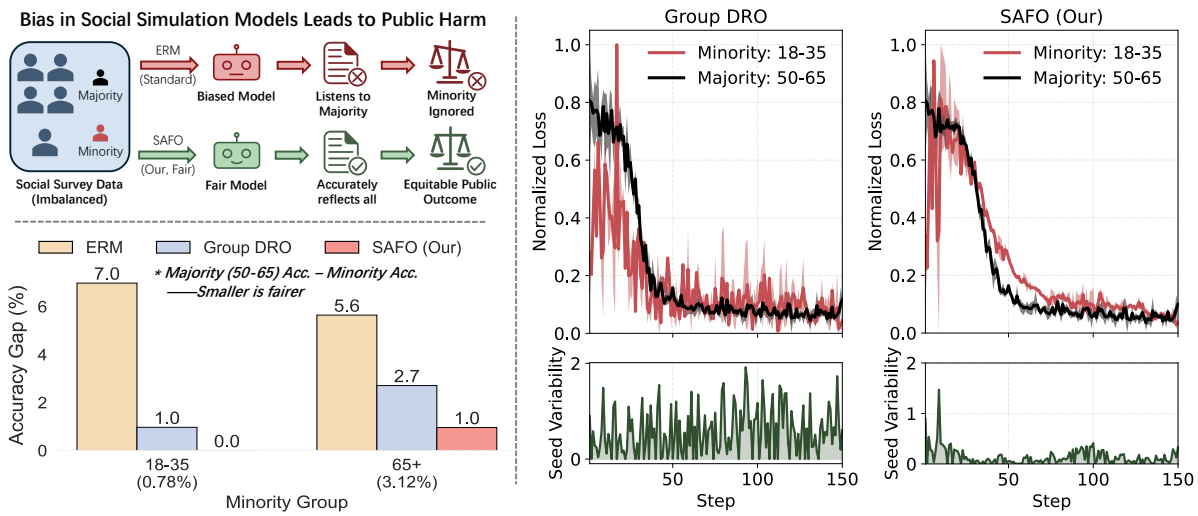


Figure 1: Mitigating fairness gaps and training instability in LLM-based social survey simulation (CGSS; sensitive attribute: age). (Left Top) Conceptual illustration of how biased survey simulation can misrepresent minority views and propagate to downstream decisions. (Left Bottom) Accuracy gap between the majority group (50-65, 73.43%) and minority groups. Lower values indicate better fairness. Percentages in parentheses indicate group proportions. (Right) Training dynamics comparison between Group DRO and SAFO. (Right Top) Training loss trajectories for the minority group (18–35, red) and majority group (50–65, black), with shaded regions indicating variance across seeds. (Right Bottom) Relative seed variability that normalized by mean loss throughout training. ERM exhibits the largest fairness gap under imbalanced data, while Group DRO reduces the gap but often shows unstable training dynamics. SAFO achieves smaller gaps with smoother convergence and lower seed variability.

training rather than converging smoothly (Figure 1, Right Top). Such oscillatory training dynamics also make performance highly sensitive to random seeds, complicating checkpoint selection and reducing reproducibility (Figure 1, Right Bottom). In realistic survey simulation, individuals may be characterized by multiple sensitive attributes (e.g., age, gender, income). Treating every attribute combination as a separate group can quickly lead to very sparse groups, which amplifies optimization noise. Moreover, existing worst-case approaches typically lack an adaptive mechanism to prevent fairness gains from incurring excessive degradation in overall utility, leading to undesirable trade-offs.

To address these challenges, we propose **Stable Adaptive Fairness Optimization (SAFO)**, a dynamic utility-fairness optimization framework for LLM-based social survey simulation. SAFO organizes training around adversarial group reweighting with adaptive regulation, consisting of three interacting components: an Optimizer, an Adversary, and a Regulator. The Optimizer preserves overall utility by minimizing mean loss, the Adversary adaptively emphasizes underperforming demographic groups to reduce group-level disparities, and the Regulator stabilizes their interaction by controlling adversarial pressure.

To promote fairness, the Adversary identifies the current worst-loss group (often a minority group in imbalanced settings) and upweights its samples via temperature-scaled softmax, increasing their influence during training. To scale across multiple sensitive attributes, it estimates group losses in an attribute-wise manner and aggregates the resulting weights across attributes, avoiding explicit enumeration of all group combinations.

To ensure convergence and stability, SAFO combines adaptive group-level loss estimation with dynamic regulation. The Adversary uses an EMA-based temporal smoothing to filter out noisy gradient signals from sparse minority-group samples, providing more reliable loss estimates and mitigating oscillatory training dynamics. Meanwhile, the Regulator dynamically modulates the utility-fairness trade-off based on a Nash-inspired rule that tracks minority gains (reductions in worst-group loss) and collateral damage (increases in mean loss). These components mitigate oscillatory training dynamics, prevent excessive degradation of overall utility, and produce consistent performance across training steps and random seeds.

Experiments on three large-scale survey datasets (CGSS, GSS, and ESS) with two backbone LLMs (Llama-3.1-8B and Qwen2.5-7B) demonstrate the

effectiveness of SAFO. SAFO reduces worst-group gaps by up to 12.7%, without hurting overall utility (mean accuracy change <0.3%) and with substantially lower seed-level variance. Together, as shown in Figure 1, SAFO achieves the smallest accuracy gap (best fairness) while maintaining smooth convergence and low variability throughout training.

Our contributions are summarized as follows:

• We propose SAFO, a dynamic utility-fairness framework for LLM-based social survey simulation that balances overall utility and fairness.

• SAFO leverages adaptive reweighting across sensitive attributes, EMA-smoothed group losses, and a dynamic Nash-inspired regulator to address fairness, convergence and stability.

• Extensive experiments on multiple survey simulation benchmarks show SAFO improves minority performance while maintaining overall utility, with smooth loss trajectories and low variability.

Related Work

Fairness in LLM-based Social Simulation

Survey simulation has long supported applications such as policy evaluation (Brown and Harding, 2002; McDaniel et al., 1988), risk prediction (Smith et al., 2009; Lorig et al., 2021), and economic forecasting (Ballas and Clarke, 2001; Benczúr et al., 2018). Recently, large language models (LLMs) have emerged as a promising approach for simulating survey responses across demographic groups by conditioning on questions and respondent profiles (Yao et al., 2025; Anthis et al.; Suh et al., 2025; Kirk et al., 2024). These methods often adopt Empirical Risk Minimization (ERM) as the standard training objective (Vapnik, 1991; Bengio et al., 2003; Radford et al., 2018). While ERM can achieve competitive overall accuracy, it tends to prioritize majority patterns (Donini et al., 2018; Leqi et al., 2019; Ghosal and Li, 2023). Underrepresented groups with fewer samples contribute less to the mean loss, receive weaker gradient updates, and are often the worst-performing groups at inference (Donini et al., 2018; Ghosal and Li, 2023). Structural imbalances in survey collection (Hammersley and Gomm, 1997; Suchman, 1962), such as lower response rates, sampling bias, or accessibility barriers (Laganà et al., 2013; Bailer et al., 2022), amplify this issue. As a result, optimizing average performance does not guarantee group-level fairness, motivating research on fairness interventions.

Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) provides an in-processing framework for group-level fairness by minimizing worst-case loss across subpopulations (Sagawa et al., 2019; Levy et al., 2020; Ghosal and Li, 2023), unlike ERM which minimizes mean loss. Empirical studies show that Group DRO improves worst-group accuracy (Sagawa et al., 2019), and scalable algorithms achieve gradient complexity independent of training set size (Levy et al., 2020). Beyond standard DRO, extensions include ensemble-based methods (To et al., 2025), preference alignment targeting worst-case group performance (Ramesh et al., 2024), and social welfare frameworks allowing controllable utility-fairness trade-offs (Rahmattalabi et al., 2021). Recent work also distinguishes descriptive, normative, and correlation-based fairness benchmarks, each calling for tailored mitigation strategies (Wang et al., 2025). However, DRO training can be unstable under sparse groups, and most formulations assume few predefined groups; in realistic surveys, multiple sensitive attributes create combinatorial group growth, exacerbating instability. These challenges motivate SAFO, which stabilizes worst-group optimization via attribute-wise decomposition and adaptive regulation.

Methodology

We propose Stable Adaptive Fairness Optimization (SAFO), a dynamic utility-fairness framework achieves both fair and stable optimization for LLM-based social survey simulation. Our framework consists of three components: an Optimizer that minimizes mean loss to preserve overall utility, an Adversary that performs loss-driven reweighting to emphasize high-loss groups, and a Regulator that adaptively modulates the utility-fairness trade-off strength via Nash-inspired mechanism.

Optimizer: Mean Minimization

The *Optimizer* minimizes mean loss to preserve overall prediction utility in social survey simulation. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)}, g^{(i)})\}_{i=1}^{N}$ denote a social survey simulation dataset, where $x^{(i)}$ is the input prompt, $y^{(i)}$ is the target response, and $g^{(i)}$ denotes the demographic attribute profile (provided as text descriptors) associated with sample i. $g^{(i)}$ can be decomposed as $(g^{(i)}_{(1)}, g^{(i)}_{(2)}, \ldots, g^{(i)}_{(A)})$, where each $g^{(i)}_{(a)}$ corresponds to one attribute dimension (e.g., age, gender, income). We condition the model on

$g^{(i)}$ by defining $x_g^{(i)} = x^{(i)} \oplus g^{(i)}$, where $\oplus$ denotes the concatenation of the prompt with group attributes (Appendix K). Notably, the same base prompt x may appear across different attribute profiles g, reflecting that identical survey questions can elicit group-dependent responses.

In supervised fine-tuning, given a model family Θ, the Optimizer aims to find a model $\theta \in \Theta$ that minimizes the expected mean loss. Specifically, we adapt a pretrained language model to the social survey simulation task, obtaining a conditional language model $\pi_\theta(y \mid x_g)$ that defines the probability of the response sequence y given the attribute-augmented prompt x_g. For each sample i, the cross-entropy loss is defined as:

$$\ell_i(\theta) = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} \ell_{\mathrm{CE}}\Big(\pi_\theta(\cdot|x_g^{(i)})_j, y_j^{(i)}\Big), \quad (1)$$

where $\mathcal{T}_i$ denotes the set of valid token positions (excluding padding) and j indexes the token position. At each training step t, let $\mathcal{B}_t$ denote the mini-batch of samples. The utility objective follows Empirical Risk Minimization (ERM), which minimizes the average loss over the current batch:

$$\mathcal{L}_{\mathrm{Op}}(\theta; t) = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \ell_i(\theta). \quad (2)$$

Adversary: Minority Protection

Due to imbalanced group sizes in social survey data, direct empirical loss estimates can be high-variance for minority groups. The Adversary addresses this by focusing on identifying the worst-loss group and upweights its samples to ensure minority groups receive sufficient optimization. For attribute dimension $a \in \{1, \ldots, A\}$, let $\mathcal{G}_a$ denote the set of possible values of $g_{(a)}^{(i)}$ observed in $\mathcal{D}$, with $\mathcal{G}_a^{\mathrm{batch}}(t) \subseteq \mathcal{G}_a$ denoting the groups present in the current batch at step t. Let $n_{a,g}(t)$ denote the cumulative sample count for group $g \in \mathcal{G}_a$ up to step t. Firstly, Adversary estimates group-level losses within each attribute separately via adaptive EMA (Exponential Moving Average), then computes soft attention weights through temperature-scaled softmax to focus on higher-loss groups, and finally aggregates across attributes. Formally, the adaptive EMA decay rate is defined as:

$$\alpha_{a,g}(t) = \mathrm{clip}\left(\frac{N_{\mathrm{scale}}}{N_{\mathrm{scale}} + n_{a,g}(t)}, \alpha_{\min}, \alpha_{\max}\right), \quad (3)$$

where N_{scale} is a scaling constant, and $\alpha_{\min}, \alpha_{\max}$ bound the decay rate. This design ensures that groups with fewer samples use larger decay rates (relying more on historical estimates), while groups with abundant samples adapt more quickly to recent observations. For each group $g \in \mathcal{G}_a$, we maintain an exponential moving average of its loss. Let $\hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t)$ denote the empirical mean loss for group g (under attribute dimension a) in the current batch at step t. The EMA (Theorem 3) is updated as:

$$\bar{\mathcal{L}}_{a,g}(t) = \alpha_{a,g}(t) \cdot \bar{\mathcal{L}}_{a,g}(t-1) + (1 - \alpha_{a,g}(t)) \cdot \hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t). \quad (4)$$

If group g does not appear in the current batch (i.e., $g \notin \mathcal{G}_a^{\mathrm{batch}}(t)$), we keep $\bar{\mathcal{L}}_{a,g}(t) = \bar{\mathcal{L}}_{a,g}(t-1)$ and $n_{a,g}(t) = n_{a,g}(t-1)$. For each attribute dimension a with $|\mathcal{G}_a^{\mathrm{batch}}(t)| \geq 2$ groups present in the batch, we compute soft attention weights via temperature-scaled softmax over the EMA losses:

$$q_g^{(a)}(t) = \frac{\exp\left(\bar{\mathcal{L}}_{a,g}(t)/\tau\right)}{\sum_{g' \in \mathcal{G}_a^{\mathrm{batch}}(t)} \exp\left(\bar{\mathcal{L}}_{a,g'}(t)/\tau\right)}, \quad (5)$$

where $\tau > 0$ is the temperature hyperparameter controlling the sharpness of the distribution. Lower temperatures concentrate weight on higher-loss groups, approximating worst-group optimization and prioritizing minority groups that typically exhibit higher losses (Theorem 1). The overall Adversary objective averages across all participating dimensions $\mathcal{A}^+(t) = \{a : |\mathcal{G}_a^{\mathrm{batch}}(t)| \geq 2\}$:

$$\mathcal{L}_{\mathrm{Ad}}(\theta; t) = \frac{1}{|\mathcal{A}^+(t)|} \sum_{a \in \mathcal{A}^+(t)} \sum_{g \in \mathcal{G}_a^{\mathrm{batch}}(t)} q_g^{(a)}(t) \cdot \hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t). \quad (6)$$

If $\mathcal{A}^+(t) = \emptyset$, we set $\mathcal{L}_{\mathrm{Ad}}(\theta; t) = 0$ and keep λ_t unchanged.

Regulator: Nash-inspired Mechanism

The Regulator dynamically adjusts the trade-off between utility and fairness through a Nash-inspired adaptation mechanism. The SAFO loss is motivated by multi-objective optimization theory that minimizers of such weighted combinations correspond to locally Pareto-optimal solutions under suitable conditions (Theorem 4 and Remark 3):

$$\mathcal{L}_{\mathrm{SAFO}}(\theta; t) = (1 - \lambda_t) \cdot \mathcal{L}_{\mathrm{Op}}(\theta; t) + \lambda_t \cdot \mathcal{L}_{\mathrm{Ad}}(\theta; t), \quad (7)$$

where $\lambda_t \in [\lambda_{\min}, \lambda_{\max}]$ controls the utility-fairness trade-off. To adaptively modulate λ_t, the Regulator maintains a Nash-inspired mechanism that tracks whether fairness enforcement yields net

benefit. For each participating attribute dimension $a \in \mathcal{A}^+(t)$, it computes weak-group gain $\Delta_{\text{weak}}(t)$ (reduction in worst-group loss) and collateral damage $\Delta_{\text{all}}(t)$ (increase in average loss) (see Appendix B for details). The conflict coefficient is updated multiplicatively and clipped to $[\lambda_{\min}, \lambda_{\max}]$:

$$\lambda_{t+1} = \text{clip}\left(\lambda_t \cdot \gamma_t,\ \lambda_{\min},\ \lambda_{\max}\right), \qquad (8)$$

where $\gamma_t = \gamma_{\text{d}}^{\mathbb{I}_\downarrow(t)} \cdot \gamma_{\text{g}}^{(1-\mathbb{I}_\downarrow(t))}$ and $\mathbb{I}_\downarrow(t) = \mathbb{I}[\Delta_{\text{all}}(t) > \gamma_{\text{n}} \cdot \Delta_{\text{weak}}(t)] \cdot \mathbb{I}[\Delta_{\text{weak}}(t) > 0]$. This trade-off–aware rule reduces λ_t (via $\gamma_{\text{d}} < 1$) when collateral damage exceeds scaled weak-group gain, and increases it (via $\gamma_{\text{g}} > 1$) otherwise. The Nash tolerance ratio $\gamma_{\text{n}} > 0$ controls the acceptable trade-off between fairness gain and utility loss.

Experiment

Datasets

We evaluate SAFO on three social survey datasets: CGSS (China), GSS (U.S.) and ESS (Europe), covering diverse demographics and survey topics. To avoid accidental overlap between training and evaluation, we use a strict test set following Lin et al. (2025). Detailed definitions of the held-out conditions are provided in Appendix C, and the corresponding task prompts are listed in Appendix K.

Baselines

We compare SAFO against four representative training paradigms: *ERM* (Vapnik, 1991), *KL* (Suh et al., 2025), *CVaR DRO* (Levy et al., 2020), and *Group DRO* (Sagawa et al., 2019), using two LLMs: *Qwen2.5-7B-Instruct* (Yang et al., 2024; Team, 2024) and *Llama-3.1-8B-Instruct* (Dubey et al., 2024). Results on *Mistral-7B-Instruct-v0.3* and two additional baselines, FairDRO (Park et al., 2025) and ROAD (Grari et al., 2023), are in Appendix I. Implementation details are in Appendix D.

Metrics

We evaluate model performance from three perspectives. For **utility**, we report Accuracy to measure overall prediction quality. For **fairness**, we adopt Wasserstein Distance, Worst Group Performance, CV, Gini Coefficient, and Nash Welfare to quantify group-level performance disparity and equitable optimization. Worst-Group Performance targets minority protection directly; Wasserstein/CV/Gini characterize overall disparity; Nash Welfare reflects balanced gains across groups rather than improving only the worst case. For **stability**, we report standard deviation (std) across 3 random seeds, shown as $\pm$ values in tables. Detailed metric definitions are provided in Appendix E.

Setup

We implement SAFO based on LLaMA-Factory (Zheng et al., 2024) and conduct experiments on 8 NVIDIA H20 GPUs. All models are fine-tuned using LoRA ($r = 8$) with AdamW optimizer, cosine scheduler, learning rate 1×10^{-5}, warmup ratio 0.1, batch size 64 per device, and bfloat16 precision for 1.5 epochs that can stable convergence. For SAFO, we initialize $\lambda_t = 0.5$ with dynamic range $[0.2, 0.8]$, adjustment factors $\gamma_g = 1.005$, $\gamma_d = 0.995$, Nash threshold $\gamma_n = 1.0$, softmax temperature $\tau = 1$, and adaptive EMA with $N_{\text{scale}} = 1000$, $\alpha \in [0.1, 0.9]$. Baselines follow original configurations: CVaR DRO uses $\alpha = 0.2$ (Levy et al., 2020) and Group DRO uses $\eta_q = 0.01$ (Sagawa et al., 2019). All experiments are repeated across 3 seeds and report mean $\pm$ std. Additional details are in Appendix F.

Results

Table 1 presents the results across three datasets and two backbone models. We summarize findings in terms of utility, fairness, and stability. SAFO matches or surpasses ERM in accuracy in almost all configurations, while consistently improving group-level fairness. Compared to ERM, SAFO reduces Wasserstein Distance by 2.9%–16.4%, lowers CV by 6.9%–17.3%, and decreases Gini Coefficient by 7.9%–16.0% across datasets and backbones. Notably, SAFO attains the best Worst-Group Performance in all configurations. Relative to ERM, it improves this metric by 1.6%–8.2%, indicating stronger protection for disadvantaged subpopulations. Nash Welfare also consistently favors SAFO, suggesting more balanced gains across groups.

Although KL sometimes achieves competitive Wasserstein Distance, it incurs very high CV and Gini coefficients, showing that optimizing distributional alignment alone does not ensure equitable performance. DRO baselines also show suboptimal utility–fairness trade-offs. CVaR DRO often causes large utility drops without improving worst-group outcomes. Group DRO can reduce some disparity metrics, but may trade off accuracy and worst-group gains.

The standard deviation values (shown as $\pm$ in Table 1) reveal that SAFO achieves remarkably

Dataset	Model	Method	Accuracy	Wasserstein Dist.	Worst Group Perf.	CV	Gini	Nash Welfare
CGSS	Llama	ERM	0.3590 ± 0.0050	0.0201 ± 0.0014	0.3212 ± 0.0074	0.0666 ± 0.0064	0.0349 ± 0.0035	0.3484 ± 0.0016
		KL	0.0086 ± 0.0054	0.0178 ± 0.0057	0.0062 ± 0.0035	0.7057 ± 0.0730	0.3559 ± 0.0329	0.0142 ± 0.0078
		CVaR DRO	0.2569 ± 0.0070	0.0174 ± 0.0020	0.2372 ± 0.0070	0.0813 ± 0.0090	0.0395 ± 0.0050	0.2690 ± 0.0110
		Group DRO	0.3529 ± 0.0047	0.0172 ± 0.0022	0.3123 ± 0.0055	**0.0587** ± 0.0079	0.0306 ± 0.0042	0.3354 ± 0.0045
		SAFO	**0.3612** ± 0.0047	**0.0168** ± 0.0003	**0.3262** ± 0.0034	0.0590 ± 0.0036	**0.0301** ± 0.0021	**0.3518** ± 0.0016
	Qwen	ERM	0.3669 ± 0.0067	0.0212 ± 0.0013	0.3119 ± 0.0098	0.0745 ± 0.0037	0.0382 ± 0.0023	0.3446 ± 0.0054
		KL	0.0795 ± 0.1247	0.0211 ± 0.0062	0.0690 ± 0.1106	0.3902 ± 0.2834	0.2021 ± 0.1470	0.0793 ± 0.1209
		CVaR DRO	0.2399 ± 0.0100	0.0278 ± 0.0032	0.2201 ± 0.0054	0.0844 ± 0.0151	0.0435 ± 0.0064	0.2558 ± 0.0161
		Group DRO	0.3546 ± 0.0061	0.0210 ± 0.0016	0.3086 ± 0.0106	0.0749 ± 0.0084	0.0385 ± 0.0054	0.3412 ± 0.0054
		SAFO	**0.3733** ± 0.0013	**0.0201** ± 0.0012	**0.3225** ± 0.0047	**0.0694** ± 0.0037	**0.0352** ± 0.0022	**0.3556** ± 0.0022
ESS	Llama	ERM	0.4077 ± 0.0078	0.0349 ± 0.0017	0.3538 ± 0.0170	0.1100 ± 0.0053	0.0496 ± 0.0046	0.3950 ± 0.0256
		KL	0.0346 ± 0.0084	**0.0338** ± 0.0015	0.0283 ± 0.0066	0.4731 ± 0.0529	0.2385 ± 0.0222	0.0541 ± 0.0041
		CVaR DRO	0.3215 ± 0.0081	0.0341 ± 0.0029	0.2543 ± 0.0200	0.1417 ± 0.0065	0.0645 ± 0.0029	0.2806 ± 0.0262
		Group DRO	0.4060 ± 0.0078	0.0412 ± 0.0054	0.3356 ± 0.0379	0.1359 ± 0.0231	0.0597 ± 0.0121	0.3682 ± 0.0465
		SAFO	**0.4150** ± 0.0075	0.0339 ± 0.0014	**0.3782** ± 0.0083	**0.1001** ± 0.0065	**0.0454** ± 0.0022	**0.4179** ± 0.0012
	Qwen	ERM	**0.4435** ± 0.0011	0.0372 ± 0.0023	0.3606 ± 0.0312	0.1162 ± 0.0111	0.0510 ± 0.0032	0.3957 ± 0.0314
		KL	0.3139 ± 0.0076	0.0269 ± 0.0117	0.2554 ± 0.0125	0.1137 ± 0.0387	0.0530 ± 0.0144	0.2805 ± 0.0186
		CVaR DRO	0.3083 ± 0.0039	0.0410 ± 0.0023	0.2346 ± 0.0146	0.1603 ± 0.0192	0.0716 ± 0.0090	0.3006 ± 0.0255
		Group DRO	0.4232 ± 0.0028	0.0375 ± 0.0039	0.3529 ± 0.0233	0.1197 ± 0.0082	0.0535 ± 0.0016	0.3898 ± 0.0288
		SAFO	0.4395 ± 0.0009	**0.0348** ± 0.0013	**0.3902** ± 0.0116	**0.1035** ± 0.0082	**0.0464** ± 0.0031	**0.4231** ± 0.0088
GSS	Llama	ERM	0.4362 ± 0.0040	0.0207 ± 0.0016	0.3798 ± 0.0160	0.0601 ± 0.0067	0.0325 ± 0.0037	0.4194 ± 0.0071
		KL	0.0451 ± 0.0055	0.0350 ± 0.0045	0.0393 ± 0.0013	0.4226 ± 0.0237	0.2163 ± 0.0099	0.0863 ± 0.0069
		CVaR DRO	0.3029 ± 0.0079	0.0211 ± 0.0019	0.2570 ± 0.0080	0.0913 ± 0.0062	0.0478 ± 0.0033	0.2885 ± 0.0078
		Group DRO	0.4300 ± 0.0041	0.0212 ± 0.0022	0.3798 ± 0.0089	0.0598 ± 0.0076	0.0318 ± 0.0037	0.4219 ± 0.0076
		SAFO	**0.4384** ± 0.0039	**0.0176** ± 0.0012	**0.3912** ± 0.0013	**0.0497** ± 0.0037	**0.0273** ± 0.0019	**0.4256** ± 0.0062
	Qwen	ERM	0.4401 ± 0.0055	0.0216 ± 0.0018	0.3733 ± 0.0193	0.0662 ± 0.0068	0.0350 ± 0.0039	0.4180 ± 0.0085
		KL	0.0176 ± 0.0095	**0.0183** ± 0.0061	0.0102 ± 0.0073	0.6165 ± 0.1486	0.3213 ± 0.0724	0.0235 ± 0.0169
		CVaR DRO	0.3282 ± 0.0060	0.0204 ± 0.0026	0.3071 ± 0.0079	0.0737 ± 0.0072	0.0390 ± 0.0036	0.3415 ± 0.0134
		Group DRO	0.4416 ± 0.0081	0.0219 ± 0.0013	0.3751 ± 0.0036	0.0665 ± 0.0040	0.0351 ± 0.0018	0.4150 ± 0.0063
		SAFO	**0.4457** ± 0.0051	0.0197 ± 0.0012	**0.3811** ± 0.0039	**0.0602** ± 0.0109	**0.0317** ± 0.0016	**0.4252** ± 0.0056

Table 1: Performance comparison across datasets and models. **Bold**: best metric value; underline: smallest standard deviation. Arrows indicate optimization direction: Accuracy (↑), Worst Group Perf. (↑), Nash Welfare (↑); Wasserstein Dist. (↓), CV (↓), Gini (↓).

stable performance across random seeds, often attaining the smallest standard deviation for most metrics. This stability is crucial for reliable model deployment in social science applications. In contrast, Group DRO exhibits substantial variability (e.g., 0.0379 std for Worst-Group Performance on ESS-Llama), confirming the training instability observed in Figure 1. These results validate that SAFO successfully addresses the utility–fairness trade-off: it improves group-level fairness without sacrificing utility, while maintaining stable optimization across different random seeds. SAFO reduces worst-group gaps by up to 12.7% versus Group DRO, without hurting overall utility (mean accuracy change <0.3%) and with substantially lower seed-level variance. Overall, SAFO consistently improves fairness while preserving utility and stability.

Discussion

Ablation Test

To isolate the contribution of each SAFO component, we conduct ablation experiments on the GSS dataset using Llama-3.1-8B. Figure 2 presents results across all six metrics.

Replacing temperature-scaled soft weighting with hard maximum selection (w/o DRO) degrades most metrics, confirming that soft weighting provides more accurate predictions than discrete group selection. Removing EMA-based loss estimation (w/o EMA) causes the largest stability drop. As seen in Theorem 3, adaptive EMA provides reliable minority loss estimates by increasing effective sample size for sparse groups (from Lemma 2), effectively providing smoother gradient signals. Fixing $\lambda_t = 0.5$ throughout training (w/o Nash) yields competitive accuracy but suboptimal fairness; the Nash welfare decreases from 0.426 to 0.419. A static trade-off coefficient cannot adapt to changing training dynamics, whereas our adaptive regulation adjusts the trade-off strength based on weak-group gain and collateral damage, leading to a more favorable utility-fairness trade-off over training.

Overall, the full SAFO framework achieves the best performance across all metrics, demonstrating that its three components synergistically balance utility, fairness, and training stability.

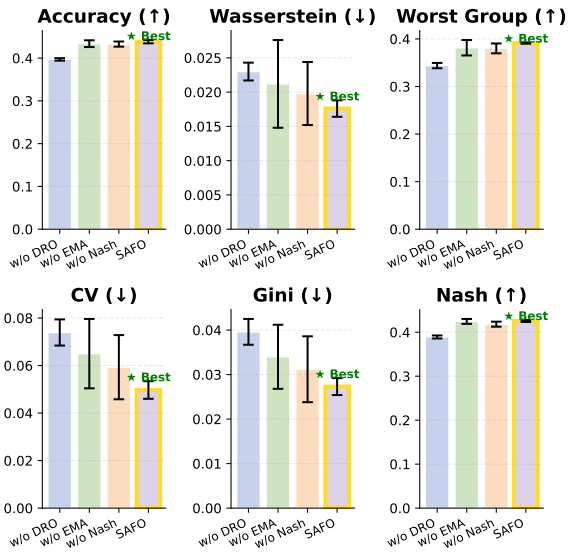


Figure 2: Ablation test of SAFO. We compare the full SAFO framework against three ablated variants: w/o DRO, w/o EMA, and w/o Nash. Error bars indicate standard deviation across 3 random seeds. Stars mark the best performance. SAFO achieves the best results across all six metrics.

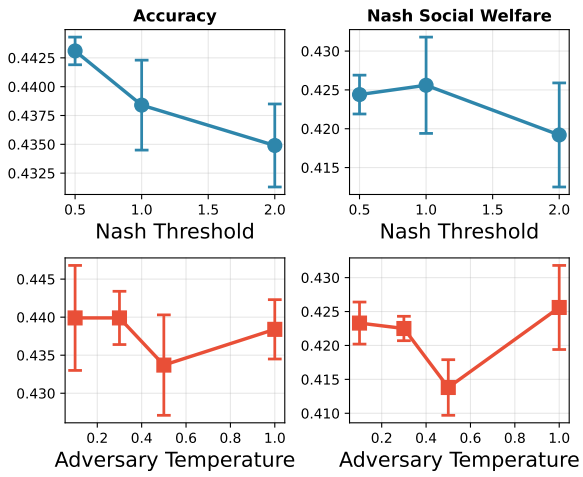


Figure 3: Parameter sensitivity analysis showing the impact of Nash threshold (top) and adversary temperature (bottom) on accuracy (left) and Nash social welfare (right). Error bars indicate standard deviation across multiple runs. The model demonstrates stable performance across a wide range of parameter settings.

Parameter Sensitivity

We assess parameter sensitivity analysis of the Nash threshold γ_{n} and adversary temperature τ (Figure 3). Setting $\gamma_{\text{n}} = 1.0$ yields the best trade-off, achieving the highest Nash social welfare with competitive accuracy. Performance varies non-monotonically with τ: intermediate values (≈ 0.5) degrade results, while both low (≤ 0.3) and high (1.0) temperatures perform well. We therefore recommend $\gamma_{\text{n}} = 1.0$ and $\tau \in 0.1, 1.0$ as default settings; further details are provided in Appendix H.

Fairness Across Demographic Groups

Figure 4 presents a comprehensive comparison of prediction accuracy across demographic subgroups, with the right panel showing seed variance (std). Within each attribute, subgroups are sorted by sample size in descending order from top to bottom, with minority groups (i.e., those with fewer samples) positioned at the bottom.

First, ERM shows clear degradation on minority sub-groups. For instance, in the Income category, ERM achieves only 0.23 accuracy for the \$6,000-\$6,999 group and 0.30 for the \$3,000-\$3,999 group, substantially lower than its performance on majority groups such as \$25,000 or more (0.44). This disparity highlights ERM's inherent bias toward majority classes during optimization. Second, both ERM and Group DRO demonstrate considerable instability, as evidenced by the standard deviation heatmap in the right panel. The deep red markers indicate high variance across random seeds, particularly for low-resource subgroups. This pattern reflects group sparsity: ERM is noisy on minority subgroups with limited support, and Group DRO can be variable because standard worst-group reweighting is sensitive to noisy group-loss estimates.

In contrast, SAFO consistently achieves higher accuracy with notably lower variance across all demographic subgroups and narrows the accuracy gap between majority and minority groups. This advantage becomes even more pronounced under intersectional sparsity, with full comparisons reported in Appendix J. This empirical observation aligns with our theoretical guarantee (Theorem 7): starting from an approximate ERM solution, the SAFO update direction strictly decreases worst-group risk when the fairness gap is non-negligible, highlighting its effectiveness in jointly improving accuracy and fairness.

Model Scaling

Figure 5 analyzes fairness performance trade-offs across model scales (0.5B–14B). While accuracy generally improves with scale for all methods, fairness exhibits method-dependent and non-monotonic behavior. ERM shows little fairness improvement from scaling and can even worsen

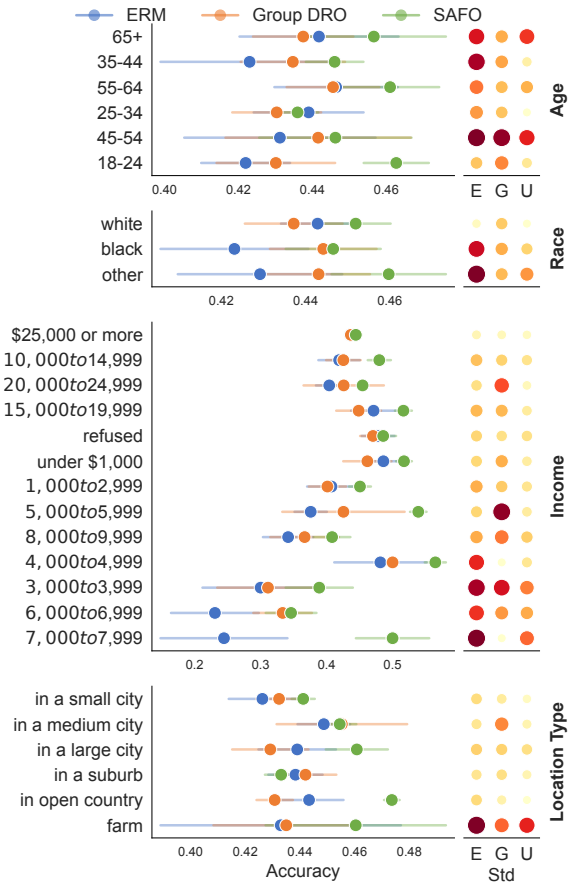


Figure 4: Subgroup accuracy comparison across three methods. Left: mean accuracy + std; right: standard deviation heatmap (per-category normalized). Subgroups are sorted by sample size within each attribute. SAFO demonstrates consistently higher accuracy and lower variance across demographic groups.

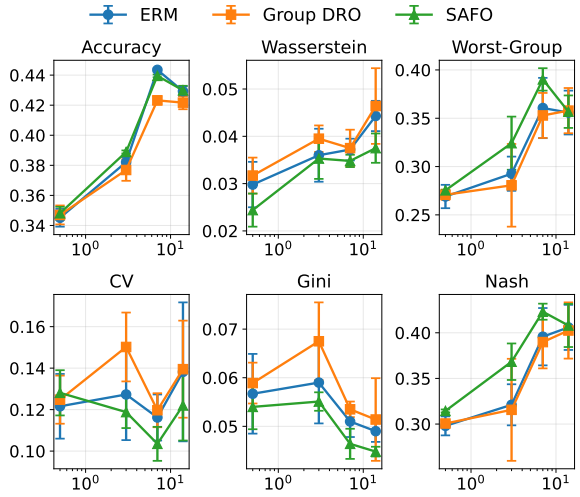


Figure 5: Comparison of ERM, Group DRO, and SAFO across model scales (0.5B–14B). SAFO achieves better fairness metrics (lower Wasserstein, CV, Gini) while maintaining comparable accuracy and higher Nash Social Welfare. Error bars denote standard deviation.

disparity, indicating that increased capacity alone is insufficient. Group DRO displays unstable scaling, with fairness metrics fluctuating across model sizes. In contrast, SAFO consistently achieves lower disparity than both baselines and benefits more reliably from increased capacity. Overall, SAFO scales gracefully, converting increased capacity into measurable fairness gains without sacrificing accuracy.

Convergence and Stability

Figure 6 and Figure 7 in Appendix G compares training dynamics across demographic attributes. Group DRO exhibits high loss oscillations and variance spikes for minority groups (e.g., 18–35 age group at 2.1%), reflecting the instability of hard worst-group selection under sparse samples. In contrast, SAFO converges smoothly, with losses for both minority and majority groups decreasing steadily. This stability is consistent across settings (the smooth convergence behavior is consistent with Theorem 6): in the ablation study (Figure 2), SAFO shows smaller error bars than its variants; subgroup analysis (Figure 4) indicates uniformly low variance across demographics; and scaling experiments (Figure 5) confirm stability from 0.5B to 14B parameters. Along with the lowest standard deviations in Table 1, these results demonstrate SAFO's robustness for reliable deployment.

Conclusion

Unstable and biased social survey simulation can systematically misrepresent underrepresented minority groups, leading to distorted analyses and harmful conclusions for vulnerable populations. We propose SAFO, a dynamic utility-fairness optimization framework for LLM-based social survey simulation. SAFO consists of three components: an Optimizer that minimizes mean loss to preserve overall utility, an Adversary that emphasizes high-loss (often underrepresented) groups through adaptive EMA-based reweighting, and a Regulator that adaptively modulates the utility-fairness trade-off strength via a Nash-inspired mechanism. Extensive experiments on three survey datasets demonstrate that SAFO improves minority performance while maintaining competitive accuracy and stable training dynamics across random seeds, supporting more robust survey simulation for downstream social analysis in equity-sensitive settings.

Limitation

While this work demonstrates the effectiveness of SAFO across three large-scale survey datasets covering diverse countries and cultural contexts, several limitations remain. First, our evaluation is confined to social survey data; extending SAFO to additional domains such as healthcare, education, or labor-market surveys would further assess its robustness and generalizability. Second, the behavior of SAFO on larger frontier models remains unexplored and warrants investigation as computational resources become available. Third, the current framework is tailored to multiple-choice survey simulation; adapting SAFO to open-ended text generation and free-form response modeling is a natural and important direction for future research. Finally, our fairness objective primarily emphasizes minority-group performance. Incorporating alternative or complementary fairness criteria could further broaden the applicability of SAFO to real-world decision-making settings.

Ethical Considerations

This work aims to improve fairness in social survey simulation, which carries both benefits and risks. On the positive side, fairer simulations can lead to more equitable policy evaluations and reduce harm to underrepresented populations. However, we acknowledge several ethical considerations.

First, survey simulation models, even when fair, should not replace genuine human participation in democratic processes. These tools are intended to complement, not substitute, real survey data collection. Second, the definition of fairness adopted in this work (minority performance) represents one of many possible fairness criteria; different applications may require alternative formulations based on stakeholder values. Third, demographic attributes used for group partitioning are socially constructed and may not capture the full complexity of individual identities. Fourth, malicious actors could potentially misuse simulation tools to fabricate public opinion or manipulate policy decisions.

We encourage practitioners to deploy such systems transparently, with appropriate human oversight, and to clearly communicate that outputs are simulated rather than actual survey responses. All datasets used in this study are publicly available academic survey datasets collected with informed consent and appropriate ethical approvals by their original institutions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72574198), the Major Project of Zhejiang Provincial Philosophy and Social Sciences Planning (25SYS05ZD), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01259, 2025C02037), the Key R&D Program of Hangzhou (2025SZDA0254), Ant Group, Chongqing Ant Consumer Finance Co., Ltd., and Ant Group through the CCF-Ant Research Fund. We sincerely thank the anonymous reviewers for their constructive comments and valuable suggestions that helped improve this paper.

References

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S Bernstein. Position: Llm social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Jo-An Atkinson, Andrew Page, Robert Wells, Andrew Milat, and Andrew Wilson. 2015. A modelling tool for policy analysis to support the design of efficient and effective policy responses for complex public health problems. *Implementation Science*, 10(1):26.

Stefanie Bailer, Christian Breunig, Nathalie Giger, and Andreas M Wüst. 2022. The diminishing value of representing the disadvantaged: Between group representation and individual career paths. *British Journal of Political Science*, 52(2):535–552.

Dimitris Ballas and Graham P Clarke. 2001. Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy*, 19(4):587–606.

Péter Benczúr, Gábor Kátay, and Aron Kiss. 2018. Assessing the economic and social impact of tax and benefit reforms: A general-equilibrium microsimulation approach applied to hungary. *Economic Modelling*, 75:441–457.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.

Laurie Brown and Ann Harding. 2002. Social modelling and public policy: Application of microsimulation modelling in australia. *Journal of Artificial Societies and Social Simulation*, 5(4).

Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

G David Garson. 2009. Computerized simulation in the social sciences: A survey and evaluation. *Simulation & Gaming*, 40(2):267–279.

Soumya Suvra Ghosal and Yixuan Li. 2023. Distributionally robust optimization with probabilistic group. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11809–11817.

Vincent Grari, Thibault Laugel, Tatsunori Hashimoto, Sylvain Lamprier, and Marcin Detyniecki. 2023. On the fairness road: Robust optimization for adversarial debiasing. *arXiv preprint arXiv:2310.18413*.

Martyn Hammersley and Roger Gomm. 1997. Bias in social research. *Sociological research online*, 2(1):7–19.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Francesco Laganà, Guy Elcheroth, Sandra Penic, Brian Kleiner, and Nicole Fasel. 2013. National minorities and their representation in social surveys: which practices make a difference? *Quality & Quantity*, 47(3):1287–1314.

Daniel D Lee, P Pham, Y Largman, and A Ng. 2009. Advances in neural information processing systems 22. *Tech Rep*.

Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. 2019. On human-aligned risk minimization. *Advances in Neural Information Processing Systems*, 32.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. 2020. Large-scale methods for distributionally robust optimization. *Advances in neural information processing systems*, 33:8847–8860.

Chenxi Lin, Weikang Yuan, Zhuoren Jiang, Biao Huang, Ruitao Zhang, Jianan Ge, Yueqian Xu, and Jianxing Yu. 2025. Alignsurvey: A comprehensive benchmark for human preferences alignment in social surveys. *arXiv preprint arXiv:2511.07871*.

Fabian Lorig, Emil Johansson, and Paul Davidsson. 2021. Agent-based social simulation of the covid-19 pandemic: A systematic review. *JASSS: Journal of Artificial Societies and Social Simulation*, 24(3).

Reuben R McDaniel, Robert S Sullivan, and James R Wilson. 1988. A simulation model for welfare policy analysis. *Socio-Economic Planning Sciences*, 22(4):157–165.

Taeeon Park, Sangwon Jung, Sanghyuk Chun, and Taesup Moon. 2025. Fairdro: Group fairness regularization via classwise robust optimization. *Neural Networks*, 182:106891.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Max Izenberg, Ryan Brown, Eric Rice, and Milind Tambe. 2021. Fair influence maximization: A welfare optimization approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11630–11638.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Dianna M Smith, Graham P Clarke, and Kirk Harland. 2009. Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A*, 41(5):1251–1268.

Edward A Suchman. 1962. An analysis of" bias" in survey research. *Public Opinion Quarterly*, pages 102–111.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Minh Nguyen Nhat To, Paul F RWilson, Viet Nguyen, Mohamed Harmanani, Michael Cooper, Fahimeh Fooladgar, Purang Abolmaesumi, Parvin Mousavi, and Rahul G Krishnan. 2025. Diverse prototypical ensembles improve robustness to subpopulation shift. *arXiv preprint arXiv:2505.23027*.

Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

Angelina Wang, Michelle Phan, Daniel E Ho, and Sanmi Koyejo. 2025. Fairness through difference awareness: Measuring desired group discrimination in llms. *arXiv preprint arXiv:2502.01926*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Junchi Yao, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong. 2025. Social opinions prediction utilizes fusing dynamics equation with llm-based agents. *Scientific Reports*, 15(1):15472.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Theoretical Analysis of SAFO

A.1 Preliminaries and Notations

Let $\mathcal{X}$ denote the input space, $\mathcal{Y}$ the output space, and $\mathcal{A} = \{1, \ldots, A\}$ the set of attribute dimensions (e.g., age, gender, income). For each attribute dimension $a \in \mathcal{A}$, let $\mathcal{G}_a$ denote the set of possible group values, with $|\mathcal{G}_a| = K_a$.

Remark 1 (Simplification for Theoretical Analysis)**.** *To streamline the theoretical presentation while preserving the essential structure, we focus on a single attribute dimension a in most derivations. The results extend naturally to the multi-attribute setting by averaging across $a \in \mathcal{A}^+(t)$ as specified in the methodology. When the attribute dimension is clear from context, we may omit the subscript a for brevity, writing $\mathcal{G}$, K, $n_g(t)$, $\bar{\mathcal{L}}_g(t)$, etc.*

Definition 1 (Risk Functions)**.** *For loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and model $f_\theta : \mathcal{X} \to \mathcal{Y}$, we define:*

- *Population risk:* $R(\theta) = \mathbb{E}_{(x,y,g)\sim\mathcal{D}}[\ell(f_\theta(x_g), y)]$
- *Group-conditional risk:* $R_g(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}|g}[\ell(f_\theta(x_g), y)]$ *for* $g \in \mathcal{G}_a$
- *Worst-group risk:* $R_{\max}(\theta) = \max_{g\in\mathcal{G}_a} R_g(\theta)$

With group proportions $p_g = \Pr(G = g)$, the population risk decomposes as $R(\theta) = \sum_{g\in\mathcal{G}_a} p_g R_g(\theta)$.

A.2 Connection to Distributionally Robust Optimization

We establish that the Adversary objective provides a principled smooth approximation to worst-group optimization.

Theorem 1 (Softmax as Entropy-Regularized DRO)**.** *Let $\boldsymbol{\mathcal{L}} = (\bar{\mathcal{L}}_{a,g}(t))_{g\in\mathcal{G}_a}$ be the vector of group EMA losses for attribute dimension a. Define the temperature-scaled softmax weights as in the methodology:*

$$q_g^{(a)}(t) = \frac{\exp(\bar{\mathcal{L}}_{a,g}(t)/\tau)}{\sum_{g'\in\mathcal{G}_a} \exp(\bar{\mathcal{L}}_{a,g'}(t)/\tau)}. \tag{9}$$

The soft-weighted objective $\mathcal{L}_{\text{soft}}^{(a)}(t;\tau) = \sum_{g\in\mathcal{G}_a} q_g^{(a)}(t) \cdot \bar{\mathcal{L}}_{a,g}(t)$ satisfies:

$$\max_{g\in\mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t) - \tau \log|\mathcal{G}_a| \leq \mathcal{L}_{\text{soft}}^{(a)}(t;\tau) \leq \max_{g\in\mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t). \tag{10}$$

Proof. The log-sum-exp function $\text{LSE}_\tau(\mathbf{z}) = \tau \log \sum_g \exp(z_g/\tau)$ admits the variational form:

$$\text{LSE}_\tau(\mathbf{z}) = \max_{\mathbf{q}\in\Delta^{|\mathcal{G}_a|-1}} \left\{ \sum_{g\in\mathcal{G}_a} q_g z_g + \tau H(\mathbf{q}) \right\}, \tag{11}$$

where $\Delta^{|\mathcal{G}_a|-1}$ is the probability simplex and $H(\mathbf{q}) = -\sum_g q_g \log q_g$ is the entropy. The unique maximizer is $q_g^* \propto \exp(z_g/\tau)$.

By optimality conditions, $\mathcal{L}_{\text{soft}}^{(a)}(t;\tau) = \text{LSE}_\tau(\boldsymbol{\mathcal{L}}) - \tau H(\mathbf{q}^*)$.

Upper bound: Since $\mathbf{q}^{(a)}(t)$ lies on the simplex, $\mathcal{L}_{\text{soft}}^{(a)}(t;\tau) = \sum_g q_g^{(a)}(t)\bar{\mathcal{L}}_{a,g}(t) \leq \max_g \bar{\mathcal{L}}_{a,g}(t)$.

Lower bound: Using $\max_g z_g \leq \text{LSE}_\tau(\mathbf{z})$ and $H(\mathbf{q}) \leq \log|\mathcal{G}_a|$:

$$\mathcal{L}_{\text{soft}}^{(a)}(t;\tau) = \text{LSE}_\tau(\boldsymbol{\mathcal{L}}) - \tau H(\mathbf{q}^*) \geq \max_g \bar{\mathcal{L}}_{a,g}(t) - \tau \log|\mathcal{G}_a|. \tag{12}$$

Remark 2. *Theorem 1 shows that temperature τ controls the approximation quality: setting $\tau = \epsilon/\log|\mathcal{G}_a|$ yields an ϵ-approximation to the worst-group objective. As $\tau \to 0^+$, we recover exact worst-group optimization.*

A.3 Analysis of Adaptive EMA Estimator

Assumption 1 (Bounded Variance)**.** *For each group $g \in \mathcal{G}_a$, the per-sample loss conditioned on group membership has finite variance:* $\mathrm{Var}[\ell_i \mid g_{(a)}^{(i)} = g] \le \sigma_{a,g}^2 < \infty$.

Lemma 2 (EMA Variance in Stationary Regime)**.** *Consider the EMA update $\bar{\mathcal{L}}_{a,g}(t) = \alpha_{a,g}(t) \cdot \bar{\mathcal{L}}_{a,g}(t-1) + (1-\alpha_{a,g}(t)) \cdot \hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t)$, where $\{\hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t)\}_{t\ge 1}$ are independent unbiased estimators with common variance $v_{a,g}$. In the stationary regime with constant α, the variance is:*

$$\mathrm{Var}[\bar{\mathcal{L}}_{a,g}] = \frac{1-\alpha}{1+\alpha} \cdot v_{a,g}. \tag{13}$$

Proof. Let $V_t = \mathrm{Var}[\bar{\mathcal{L}}_{a,g}(t)]$. By independence of $\bar{\mathcal{L}}_{a,g}(t-1)$ and $\hat{\mathcal{L}}_{a,g}^{\mathrm{batch}}(t)$:

$$V_t = \alpha^2 V_{t-1} + (1-\alpha)^2 v_{a,g}. \tag{14}$$

At stationarity, $V = V_t = V_{t-1}$ satisfies $V(1-\alpha^2) = (1-\alpha)^2 v_{a,g}$, yielding $V = \frac{1-\alpha}{1+\alpha} v_{a,g}$.

Theorem 3 (Properties of Adaptive EMA)**.** *The adaptive decay $\alpha_{a,g}(t) = \mathrm{clip}\left(\frac{N_{\mathrm{scale}}}{N_{\mathrm{scale}}+n_{a,g}(t)}, \alpha_{\min}, \alpha_{\max}\right)$ satisfies:*

(a) For minority groups ($n_{a,g}(t) \ll N_{\mathrm{scale}}$): $\alpha_{a,g}(t) \approx \alpha_{\max}$, yielding variance $\approx \frac{1-\alpha_{\max}}{1+\alpha_{\max}} v_{a,g}$ (more smoothing).

(b) For majority groups ($n_{a,g}(t) \gg N_{\mathrm{scale}}$): $\alpha_{a,g}(t) \approx \alpha_{\min}$, yielding variance $\approx \frac{1-\alpha_{\min}}{1+\alpha_{\min}} v_{a,g}$ (faster adaptation).

The effective sample size, defined as $n_{\mathrm{eff}}(\alpha) = \frac{1+\alpha}{1-\alpha}$ (equating EMA variance to that of an i.i.d. average), ranges from $\frac{1+\alpha_{\min}}{1-\alpha_{\min}}$ to $\frac{1+\alpha_{\max}}{1-\alpha_{\max}}$.

Proof. From Lemma 2, the EMA variance is $\frac{1-\alpha}{1+\alpha} v_{a,g}$. An i.i.d. average of n samples has variance $v_{a,g}/n$. Equating: $v_{a,g}/n_{\mathrm{eff}} = \frac{1-\alpha}{1+\alpha} v_{a,g}$ gives $n_{\mathrm{eff}} = \frac{1+\alpha}{1-\alpha}$. The bounds follow from the clipping operation.

A.4 Pareto Optimality Analysis

Definition 2 (Pareto Optimality)**.** *A solution θ^* is Pareto optimal for $\min_\theta(R(\theta), R_{\max}(\theta))$ if no θ' exists with $R(\theta') \le R(\theta^*)$ and $R_{\max}(\theta') \le R_{\max}(\theta^*)$, with at least one strict inequality.*

Theorem 4 (Scalarization Yields Pareto Optimality)**.** *Suppose $R(\theta)$ and $R_{\max}(\theta)$ are convex in θ. Any minimizer θ^* of*

$$\mathcal{L}_{\mathrm{SAFO}}(\theta; \lambda_t) = (1-\lambda_t)\mathcal{L}_{\mathrm{Op}}(\theta; t) + \lambda_t \mathcal{L}_{\mathrm{Ad}}(\theta; t) \tag{15}$$

with $\lambda_t \in (0,1)$ is Pareto optimal.

Proof. Suppose θ^* minimizes $\mathcal{L}_{\mathrm{SAFO}}(\cdot; \lambda_t)$ but is not Pareto optimal. Then there exists θ' with $R(\theta') \le R(\theta^*)$, $R_{\max}(\theta') \le R_{\max}(\theta^*)$, and at least one strict inequality. Since $\lambda_t, 1-\lambda_t > 0$:

$$\mathcal{L}_{\mathrm{SAFO}}(\theta'; \lambda_t) < \mathcal{L}_{\mathrm{SAFO}}(\theta^*; \lambda_t), \tag{16}$$

contradicting optimality of θ^*.

Remark 3 (Extension to Non-convex Settings)**.** *Theorem 4 assumes convexity of $R(\theta)$ and $R_{\max}(\theta)$. For deep neural networks including LLMs, the loss landscape is non-convex in θ, so global Pareto optimality cannot be guaranteed. However, the scalarization approach remains well-motivated:*

(i) The result holds locally *near stationary points where the loss functions are approximately convex within a neighborhood.*

(ii) Empirically, scalarization methods have been successful for multi-objective optimization in deep learning (Sener and Koltun, 2018; Lee et al., 2009; Désidéri, 2012).

(iii) For LoRA fine-tuning with small rank r, the effective optimization occurs in a low-dimensional subspace, which may exhibit better-behaved geometry.

Thus, we interpret the SAFO objective as encouraging optimization toward locally *Pareto-efficient trade-offs.*

Proposition 5 (Regulator Equilibrium Condition)**.** *The multiplicative update* $\lambda_{t+1} = \mathrm{clip}(\lambda_t \cdot \gamma_t, \lambda_{\min}, \lambda_{\max})$ *with*

$$\gamma_t = \begin{cases} \gamma_{\mathrm{d}} < 1 & \text{if } \Delta_{\mathrm{all}}(t) > \gamma_{\mathrm{n}} \cdot \Delta_{\mathrm{weak}}(t) \text{ and } \Delta_{\mathrm{weak}}(t) > 0 \\ \gamma_{\mathrm{g}} > 1 & \text{otherwise} \end{cases} \tag{17}$$

has the following equilibrium property: if λ_t converges to an interior point $\lambda^* \in (\lambda_{\min}, \lambda_{\max})$, then $\Delta_{\mathrm{all}}(t) = \gamma_{\mathrm{n}} \cdot \Delta_{\mathrm{weak}}(t)$ at equilibrium, representing balanced marginal trade-off between utility cost and fairness gain.

Proof. At an interior equilibrium, $\lambda_{t+1} = \lambda_t$ requires $\gamma_t = 1$. Since $\gamma_{\mathrm{d}} < 1 < \gamma_{\mathrm{g}}$, this occurs only at the boundary between the two cases. With $\Delta_{\mathrm{weak}}(t) > 0$, the boundary condition is $\Delta_{\mathrm{all}}(t) = \gamma_{\mathrm{n}} \cdot \Delta_{\mathrm{weak}}(t)$.

Remark 4 (Connection to Nash-inspired Mechanism Statistics)**.** *In the methodology, $\Delta_{\mathrm{weak}}(t)$ and $\Delta_{\mathrm{all}}(t)$ are computed using the Nash-inspired mechanism smoothed estimates $\bar{S}_{(a)}^{\mathrm{weak}}(t)$ and $\bar{S}_{(a)}^{\mathrm{all}}(t)$:*

$$\Delta_{\mathrm{weak}}(t) = \sum_{a \in \mathcal{A}^+(t)} \left[\bar{S}_{(a)}^{\mathrm{weak}}(t-1) - \mathcal{L}_{(a)}^{\mathrm{weak}}(t) \right]^+, \tag{18}$$

$$\Delta_{\mathrm{all}}(t) = \sum_{a \in \mathcal{A}^+(t)} \left[\mathcal{L}_{(a)}^{\mathrm{all}}(t) - \bar{S}_{(a)}^{\mathrm{all}}(t-1) \right]^+, \tag{19}$$

where $\mathcal{L}_{(a)}^{\mathrm{weak}}(t) = \max_{g \in \mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t)$ and $\mathcal{L}_{(a)}^{\mathrm{all}}(t) = \frac{1}{|\mathcal{G}_a|} \sum_{g \in \mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t)$.

A.5 Convergence Analysis

Assumption 2 (Smoothness)**.** *The loss $\mathcal{L}_{\mathrm{SAFO}}(\theta; t)$ is L-smooth in θ: $\|\nabla_\theta \mathcal{L}_{\mathrm{SAFO}}(\theta; t) - \nabla_\theta \mathcal{L}_{\mathrm{SAFO}}(\theta'; t)\| \le L\|\theta - \theta'\|$ for all θ, θ'.*

Remark 5 (On the Smoothness Assumption)**.** *Global L-smoothness (Assumption 2) is a standard assumption in optimization theory that may not hold globally for deep neural networks. However:*

(i) The assumption is widely adopted in the analysis of SGD for neural networks and provides meaningful convergence guarantees in practice (Bottou et al., 2018).

(ii) For LoRA fine-tuning, optimization occurs over low-rank adapter matrices rather than full model weights, resulting in a significantly reduced and potentially better-conditioned parameter space.

(iii) The smoothness constant L can be interpreted as a local quantity that holds within the region traversed during training.

Assumption 3 (Bounded Stochastic Gradient Variance)**.** $\mathbb{E}[\|\nabla \ell_i(\theta) - \nabla \mathcal{L}_{\mathrm{Op}}(\theta; t)\|^2] \le \sigma^2$ *for all θ.*

Assumption 4 (Slow λ Variation)**.** *The trade-off coefficient satisfies $|\lambda_{t+1} - \lambda_t| \le \delta_\lambda$ for all t.*

Theorem 6 (Convergence Rate)**.** *Under Assumptions 2–4, with constant learning rate $\eta \le 1/(2L)$, SAFO satisfies:*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|\nabla_\theta \mathcal{L}_{\mathrm{SAFO}}(\theta_t; \lambda_t)\|^2 \right] \le \frac{4(\mathcal{L}_0 - \mathcal{L}^*)}{\eta T} + 2L\eta\sigma^2 + \frac{4B\delta_\lambda}{\eta}, \tag{20}$$

where $\mathcal{L}_0$ is the initial loss value, $\mathcal{L}^* = \inf_{\theta,\lambda} \mathcal{L}_{\mathrm{SAFO}}(\theta; \lambda)$, and B bounds $|\mathcal{L}_{\mathrm{Ad}}(\theta; t) - \mathcal{L}_{\mathrm{Op}}(\theta; t)|$ uniformly.

Proof. For fixed λ_t, standard SGD analysis with L-smoothness gives:

$$\mathbb{E}[\mathcal{L}_{\text{SAFO}}(\theta_{t+1};\lambda_t)] \leq \mathbb{E}[\mathcal{L}_{\text{SAFO}}(\theta_t;\lambda_t)] - \frac{\eta}{2}\mathbb{E}[\|\nabla\mathcal{L}_{\text{SAFO}}(\theta_t;\lambda_t)\|^2] + \frac{L\eta^2\sigma^2}{2}, \tag{21}$$

for $\eta \leq 1/L$.

For time-varying λ_t, define $\mathcal{L}_t(\theta) = \mathcal{L}_{\text{SAFO}}(\theta;\lambda_t)$. The objective shift between steps satisfies:

$$|\mathcal{L}_{t+1}(\theta) - \mathcal{L}_t(\theta)| = |\lambda_{t+1} - \lambda_t| \cdot |\mathcal{L}_{\text{Ad}}(\theta;t) - \mathcal{L}_{\text{Op}}(\theta;t)| \leq \delta_\lambda B. \tag{22}$$

Combining the descent inequality with the objective perturbation:

$$\mathbb{E}[\mathcal{L}_{t+1}(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_t(\theta_t)] - \frac{\eta}{2}\mathbb{E}[\|\nabla\mathcal{L}_t(\theta_t)\|^2] + \frac{L\eta^2\sigma^2}{2} + \delta_\lambda B. \tag{23}$$

Summing from $t = 1$ to T and rearranging:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}_t(\theta_t)\|^2] \leq \frac{2(\mathcal{L}_0 - \mathcal{L}^*)}{\eta T} + L\eta\sigma^2 + \frac{2\delta_\lambda B}{\eta}. \tag{24}$$

The stated bound follows with adjusted constants for $\eta \leq 1/(2L)$.

Remark 6. *Setting $\eta = \Theta(1/\sqrt{T})$ and assuming $\delta_\lambda = \mathcal{O}(1/\sqrt{T})$ (achievable by annealing $\gamma_{\text{d}}, \gamma_{\text{g}} \to 1$) yields convergence rate $\mathcal{O}(1/\sqrt{T})$.*

Remark 7 (Interpretation of Convergence Result)**.** *Theorem 6 guarantees convergence to an* approximate stationary point *where the expected squared gradient norm is small. This is the standard notion of convergence for non-convex stochastic optimization and does not imply convergence to a global minimum. For LLM fine-tuning, reaching such approximate stationary points is the practical goal, as global optimization is computationally intractable.*

A.6 Fairness Improvement Guarantee

Assumption 5 (Gradient Alignment)**.** *At the ERM solution θ_{ERM}, the gradients satisfy: for $g^* = \arg\max_{g\in\mathcal{G}_a} R_g(\theta_{\text{ERM}})$,*

$$\sum_{g\in\mathcal{G}_a, g\neq g^*} q_g^{(a)}(t)\langle\nabla R_{g^*}(\theta_{\text{ERM}}), \nabla R_g(\theta_{\text{ERM}})\rangle \geq -\frac{q_{g^*}^{(a)}(t)}{2}\|\nabla R_{g^*}(\theta_{\text{ERM}})\|^2. \tag{25}$$

Remark 8. *Assumption 5 is a mild gradient alignment condition. It holds automatically when: (i) all group gradients point in similar directions (common in overparameterized models), or (ii) the worst group has sufficiently large weight $q_{g^*}^{(a)}(t)$ (ensured by small temperature τ).*

Theorem 7 (Descent on Worst-Group Risk)**.** *Let θ_{ERM} be an ϵ-approximate stationary point of the population risk, i.e., $\|\nabla R(\theta_{\text{ERM}})\| \leq \epsilon$, with fairness gap $\Delta_{\text{gap}} = R_{\max}(\theta_{\text{ERM}}) - R(\theta_{\text{ERM}}) > 0$. Under Assumption 5, if $\|\nabla R_{g^*}(\theta_{\text{ERM}})\| \geq \delta > 0$ for some $\delta > \epsilon$, then the SAFO update direction $d = -\nabla_\theta\mathcal{L}_{\text{SAFO}}(\theta_{\text{ERM}};\lambda_t)$ satisfies:*

$$\langle\nabla R_{g^*}(\theta_{\text{ERM}}), d\rangle \leq -\frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2}\|\nabla R_{g^*}(\theta_{\text{ERM}})\|^2 + (1-\lambda_t)\epsilon\|\nabla R_{g^*}(\theta_{\text{ERM}})\|. \tag{26}$$

In particular, when ϵ is sufficiently small such that $(1-\lambda_t)\epsilon < \frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2}\delta$, gradient descent from θ_{ERM} strictly decreases the worst-group risk.

Proof. The SAFO gradient at θ_{ERM} is:

$$\nabla_\theta \mathcal{L}_{\mathrm{SAFO}}(\theta_{\mathrm{ERM}}; \lambda_t) = (1-\lambda_t)\nabla R(\theta_{\mathrm{ERM}}) + \lambda_t \sum_{g \in \mathcal{G}_a} q_g^{(a)}(t) \nabla R_g(\theta_{\mathrm{ERM}}). \tag{27}$$

The directional derivative of R_{g^*} along $d = -\nabla_\theta \mathcal{L}_{\mathrm{SAFO}}$ is:

$$\langle \nabla R_{g^*}, d \rangle = -(1-\lambda_t)\langle \nabla R_{g^*}, \nabla R \rangle - \lambda_t \sum_{g \in \mathcal{G}_a} q_g^{(a)}(t) \langle \nabla R_{g^*}, \nabla R_g \rangle. \tag{28}$$

For the first term, by Cauchy-Schwarz:

$$|\langle \nabla R_{g^*}, \nabla R \rangle| \le \|\nabla R_{g^*}\| \cdot \|\nabla R\| \le \epsilon \|\nabla R_{g^*}\|. \tag{29}$$

For the second term, decomposing the sum:

$$-\lambda_t \sum_{g \in \mathcal{G}_a} q_g^{(a)}(t) \langle \nabla R_{g^*}, \nabla R_g \rangle = -\lambda_t \cdot q_{g^*}^{(a)}(t) \|\nabla R_{g^*}\|^2 - \lambda_t \sum_{g \ne g^*} q_g^{(a)}(t) \langle \nabla R_{g^*}, \nabla R_g \rangle. \tag{30}$$

By Assumption 5:

$$-\lambda_t \sum_{g \in \mathcal{G}_a} q_g^{(a)}(t) \langle \nabla R_{g^*}, \nabla R_g \rangle \le -\lambda_t \cdot q_{g^*}^{(a)}(t) \|\nabla R_{g^*}\|^2 + \frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2} \|\nabla R_{g^*}\|^2 \tag{31}$$

$$= -\frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2} \|\nabla R_{g^*}\|^2. \tag{32}$$

Combining both terms:

$$\langle \nabla R_{g^*}, d \rangle \le (1-\lambda_t)\epsilon \|\nabla R_{g^*}\| - \frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2} \|\nabla R_{g^*}\|^2. \tag{33}$$

When $(1-\lambda_t)\epsilon < \frac{\lambda_t \cdot q_{g^*}^{(a)}(t)}{2} \|\nabla R_{g^*}\|$, we have $\langle \nabla R_{g^*}, d \rangle < 0$.

Remark 9 (Approximate Stationarity in Practice)**.** *In LLM fine-tuning with stochastic gradient descent, exact stationarity $\nabla R(\theta_{\mathrm{ERM}}) = 0$ is never achieved in practice. Theorem 7 accounts for this by considering ϵ-approximate stationary points. The condition $\|\nabla R_{g^*}\| \ge \delta > \epsilon$ captures the intuition that when a fairness gap exists, the worst-group gradient should be non-negligible even when the overall population gradient is small. This is precisely the setting where SAFO provides benefit: the ERM solution has converged (small population gradient) but significant group-level disparities remain (large worst-group gradient).*

Corollary 8 (Sufficient Condition via Temperature)**.** *If τ is chosen such that $q_{g^*}^{(a)}(t) \ge 1 - 1/(2|\mathcal{G}_a|)$, then Assumption 5 holds whenever cross-group gradient inner products are bounded: $|\langle \nabla R_{g^*}, \nabla R_g \rangle| \le \|\nabla R_{g^*}\|^2$ for all $g \ne g^*$.*

B Nash-inspired Regulation Details

For each attribute dimension $a \in \mathcal{A}^+(t)$ participating at step t, we compute the worst-group EMA loss and the average EMA loss:

$$\mathcal{L}_{(a)}^{\text{weak}}(t) = \max_{g \in \mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t), \tag{34}$$

$$\mathcal{L}_{(a)}^{\text{all}}(t) = \frac{1}{|\mathcal{G}_a|} \sum_{g \in \mathcal{G}_a} \bar{\mathcal{L}}_{a,g}(t). \tag{35}$$

Nash-inspired mechanism maintains smoothed estimates of these quantities:

$$\bar{S}_{(a)}^{\text{weak}}(t) = \beta_{\text{s}} \cdot \bar{S}_{(a)}^{\text{weak}}(t-1) + (1-\beta_{\text{s}}) \cdot \mathcal{L}_{(a)}^{\text{weak}}(t), \tag{36}$$

$$\bar{S}_{(a)}^{\text{all}}(t) = \beta_{\text{s}} \cdot \bar{S}_{(a)}^{\text{all}}(t-1) + (1-\beta_{\text{s}}) \cdot \mathcal{L}_{(a)}^{\text{all}}(t), \tag{37}$$

where $\beta_{\text{s}} \in (0,1)$ is the smoothing decay rate. For attribute dimensions $a \notin \mathcal{A}^+(t)$, we keep $\bar{S}_{(a)}^{\text{weak}}(t) = \bar{S}_{(a)}^{\text{weak}}(t-1)$ and $\bar{S}_{(a)}^{\text{all}}(t) = \bar{S}_{(a)}^{\text{all}}(t-1)$.

The weak-group gain and collateral damage are computed as:

$$\Delta_{\text{weak}}(t) = \sum_{a \in \mathcal{A}^+(t)} \left[\bar{S}_{(a)}^{\text{weak}}(t-1) - \mathcal{L}_{(a)}^{\text{weak}}(t) \right]^+, \tag{38}$$

$$\Delta_{\text{all}}(t) = \sum_{a \in \mathcal{A}^+(t)} \left[\mathcal{L}_{(a)}^{\text{all}}(t) - \bar{S}_{(a)}^{\text{all}}(t-1) \right]^+, \tag{39}$$

where $[\cdot]^+ = \max(0, \cdot)$. The condition $\Delta_{\text{weak}}(t) > 0$ in the decay indicator prevents premature decay before fairness optimization takes effect. If $\mathcal{A}^+(t) = \emptyset$, we set $\gamma_t = 1$ and keep $\lambda_{t+1} = \lambda_t$.

C Datasets

We evaluate SAFO on three large-scale social survey datasets that span diverse countries and cultural contexts. These datasets encompass a wide range of demographic attributes (e.g., age, gender, education, income) and survey topics (e.g., social attitudes, political views, economic perceptions), providing a comprehensive testbed for assessing both utility and fairness in social simulation.

C.1 Dataset Descriptions

CGSS (Chinese General Social Survey) The Chinese General Social Survey[1] is one of China's earliest nationwide, comprehensive, and continuous academic survey projects, initiated in 2003. CGSS systematically and comprehensively collects data at multiple levels to summarize trends in social change across Chinese society. The survey covers topics including social stratification, labor market dynamics, family structure, and public attitudes toward government policies, providing valuable insights into the rapid social transformation of contemporary China.

GSS (U.S. General Social Survey) The General Social Survey[2] has studied the growing complexity of American society for five decades since 1972. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States. GSS collects data on contemporary American society to monitor and explain trends in opinions, attitudes, and behaviors, covering topics such as civil liberties, crime and violence, intergroup tolerance, and national spending priorities.

ESS (European Social Survey) The European Social Survey[3] is an academically driven cross-national survey conducted across Europe since its establishment in 2001. Every two years, face-to-face interviews are conducted with newly selected, cross-sectional samples representing the populations of over 30 European countries. ESS measures attitudes, beliefs, and behavior patterns of diverse populations in more than thirty nations, with topics including immigration, media and social trust, political engagement, and subjective well-being.

C.2 Dataset Statistics

Table 2 summarizes the key statistics for each dataset used in our experiments.

C.3 Attribute Distribution Analysis

Table 3 shows the demographic distributions for the training and test sets of the GSS dataset. The table summarizes key variables such as age, gender, education, income, household size, and region. Similarly, Tables 4 and 5 present the corresponding distributions for the CGSS and ESS datasets. Each table provides counts and percentages, allowing for a quick comparison of group compositions within each dataset. Overall, these tables illustrate

[1] http://cgss.ruc.edu.cn

[2] https://gss.norc.org

[3] https://www.europeansocialsurvey.org

Table 2: Dataset statistics summary. #Samples denotes the total number of survey responses; #Questions indicates the number of unique survey questions; #Attributes shows the number of demographic attributes used for group partitioning.

Dataset	#Samples (Train)	#Samples (Test)	#Questions	#Attributes	Region
GSS	17963	4506	11	9	United States
CGSS	20853	5762	50	10	China
ESS	100000	5000	111	8	Europe

the demographic characteristics of the samples and highlight any imbalances across categories. The distributions between training and test sets are generally consistent, ensuring that the test sets are representative for model evaluation.

C.4 Train/Test Split Construction

To evaluate generalization under realistic test-time novelty, we follow the unified evaluation protocol of Lin et al. (2025), where the held-out test set may include novel questions and/or novel demographic compositions compared to training. Unless otherwise noted, we report results on the pooled test set.

D Baselines Training Paradigms

We compare SAFO against four representative training paradigms that span the spectrum from standard optimization to distributionally robust approaches.

D.1 Empirical Risk Minimization (ERM)

ERM (Vapnik, 1991) is the standard training paradigm that minimizes the average loss across all training samples:

$$\theta_{\text{ERM}} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i), \tag{40}$$

where $\ell(\cdot, \cdot)$ denotes the cross-entropy loss, f_θ is the model parameterized by θ, and (x_i, y_i) represents the i-th input-output pair. ERM treats all samples equally regardless of their group membership, which can lead to poor performance on minority groups when the training data exhibits significant group imbalance. Despite this limitation, ERM remains the most widely adopted training paradigm due to its simplicity and computational efficiency. In our experiments, ERM serves as the primary baseline representing standard fine-tuning practices for LLMs.

D.2 KL Divergence Fine-Tuning (KL-FT)

KL Divergence Fine-Tuning (Suh et al., 2025) directly aligns the model's predicted distribution with target response distributions using the forward Kullback-Leibler divergence as the training objective:

$$\theta_{\text{KL}} = \arg\min_{\theta} \mathbb{E}_{q,g} \left[D_{\text{KL}} \left(p_H(\mathcal{A}_q|q,g) \| p_\theta(\mathcal{A}_q|q,g) \right) \right], \tag{41}$$

where $p_H(\mathcal{A}_q|q,g)$ represents the empirical distribution of human responses for question q conditioned on subpopulation g, $p_\theta(\mathcal{A}_q|q,g)$ denotes the model's predicted distribution over answer choices $\mathcal{A}_q$, and D_{KL} is the KL divergence defined as:

$$D_{\text{KL}}(p_H \| p_\theta) = \sum_{a \in \mathcal{A}_q} p_H(a) \log \frac{p_H(a)}{p_\theta(a)}. \tag{42}$$

The forward KL divergence (i.e., $\text{KL}(p_H \| p_\theta)$) is chosen because it is sensitive to cases where the target distribution p_H assigns high probability but the model distribution p_θ does not, naturally encouraging the model to *cover* the real distribution. This property aligns with standard maximum-likelihood training, where the model is penalized for underestimating any response that is frequent in the data. KL-FT treats all subpopulation-response pairs equally during training without explicit mechanisms for handling group imbalance or optimizing worst-case performance. In our experiments, we implement KL-FT using LoRA fine-tuning following the original paper's setup.

D.3 CVaR DRO (Conditional Value-at-Risk Distributionally Robust Optimization)

CVaR DRO (Levy et al., 2020) optimizes for the worst-case expected loss over a fraction α of the training distribution:

$$\begin{aligned} \theta_{\text{CVaR}} &= \arg\min_{\theta} \text{CVaR}_\alpha \left(\ell(f_\theta(x), y) \right) \\ &= \arg\min_{\theta} \mathbb{E} \left[\ell(f_\theta(x), y) \mid \ell(f_\theta(x), y) \geq q_\alpha \right]. \end{aligned} \tag{43}$$

Variable / Category	Count (Train / Test)	Percent (Train / Test)
Gender		
Female	9774 / 2492	54.4% / 55.3%
Male	8189 / 2014	45.6% / 44.7%
Age		
18–24	1187 / 286	6.6% / 6.3%
25–34	2668 / 773	14.9% / 17.2%
35–44	3270 / 855	18.2% / 19.0%
45–54	2763 / 579	15.4% / 12.8%
55–64	3066 / 799	17.1% / 17.7%
65+	4654 / 1168	25.9% / 25.9%
Missing	355 / 46	2.0% / 1.0%
Race		
White	12617 / 3172	70.2% / 70.4%
Black	3193 / 854	17.8% / 19.0%
Other	2153 / 480	12.0% / 10.7%
Marital Status		
Married	7556 / 1714	42.1% / 38.0%
Never married	5541 / 1553	30.8% / 34.5%
Divorced	2953 / 707	16.4% / 15.7%
Widowed	1361 / 347	7.6% / 7.7%
Separated	552 / 185	3.1% / 4.1%
Education		
High school	8225 / 2073	45.8% / 46.0%
Bachelor's	3916 / 991	21.8% / 22.0%
Graduate	2670 / 598	14.9% / 13.3%
Assoc./Junior college	1638 / 488	9.1% / 10.8%
Less than high school	1514 / 356	8.4% / 7.9%
Income		
$25,000 or more	13520 / 3413	75.3% / 75.7%
$10k–14,999	1089 / 257	6.1% / 5.7%
$20k–24,999	864 / 198	4.8% / 4.4%
$15k–19,999	631 / 162	3.5% / 3.6%
Refused	407 / 129	2.3% / 2.9%
Under $1,000	368 / 96	2.0% / 2.1%
Work Status		
Working full time	8105 / 2056	45.1% / 45.6%
Retired	4295 / 1148	23.9% / 25.5%
Working part time	1804 / 372	10.0% / 8.3%
Keeping house	1575 / 352	8.8% / 7.8%
Unemployed	865 / 200	4.8% / 4.4%
Other	551 / 155	3.1% / 3.4%
In school	436 / 117	2.4% / 2.6%
Temp. not working	332 / 106	1.8% / 2.4%
Residence		
Small city/town	6117 / 1586	34.1% / 35.2%
Medium city	3331 / 809	18.5% / 18.0%
Large city	2963 / 748	16.5% / 16.6%
Suburb	2667 / 638	14.8% / 14.2%
Open country	1820 / 397	10.1% / 8.8%
Farm	1065 / 328	5.9% / 7.3%
Region		
South Atlantic	3345 / 779	18.6% / 17.3%
East North Central	3292 / 806	18.3% / 17.9%
Middle Atlantic	2484 / 605	13.8% / 13.4%
Pacific	2219 / 550	12.4% / 12.2%
West South Central	1907 / 502	10.6% / 11.1%
East South Central	1827 / 471	10.2% / 10.5%
Mountain	1222 / 314	6.8% / 7.0%
West North Central	1213 / 352	6.8% / 7.8%
New England	454 / 127	2.5% / 2.8%

Table 3: Demographic distribution of GSS sample: train vs. test sets.

Variable / Category	Count (Train / Test)	Percent (Train / Test)
Age		
18–35	405 / 45	1.9% / 0.8%
35–50	3687 / 1306	17.7% / 22.7%
50–65	15591 / 4231	74.8% / 73.4%
65+	1170 / 180	5.6% / 3.1%
House Ownership		
Yes	14111 / 3265	67.7% / 56.7%
No	6742 / 2497	32.3% / 43.3%
Health Status		
Very healthy	4537 / 1303	21.8% / 22.6%
Healthy	9483 / 2602	45.5% / 45.2%
Average	5213 / 1477	25.0% / 25.6%
Unhealthy	1440 / 310	6.9% / 5.4%
Very unhealthy	180 / 70	0.9% / 1.2%
Residence Type		
City center	9389 / 2841	45.0% / 49.3%
Rural	4854 / 1738	23.3% / 30.2%
City edge / peri-urban	5532 / 1013	26.5% / 17.6%
Town	1078 / 170	5.2% / 3.0%
Work Unit Type		
Enterprise	8757 / 2118	42.0% / 36.8%
No unit / Self-employed	8231 / 1965	39.5% / 34.1%
Public institution	2561 / 1234	12.3% / 21.4%
Social group / Committee	630 / 270	3.0% / 4.7%
Government agency	674 / 175	3.2% / 3.0%
Family Size		
0	4761 / 1178	22.8% / 20.4%
1	5037 / 1260	24.2% / 21.9%
2	5525 / 1913	26.5% / 33.2%
3	3371 / 922	16.2% / 16.0%
4	1439 / 359	6.9% / 6.2%
5	585 / 115	2.8% / 2.0%
6+	135 / 15	0.6% / 0.3%
Gender		
Female	10605 / 3476	50.9% / 60.3%
Male	10248 / 2286	49.1% / 39.7%
Education		
No education	988 / 210	4.7% / 3.6%
Primary	2159 / 390	10.4% / 6.8%
Lower secondary	4811 / 1934	23.1% / 33.6%
High school	5837 / 897	28.0% / 15.6%
Associate	3103 / 695	14.9% / 12.1%
Bachelor	3415 / 1476	16.4% / 25.6%
Graduate+	540 / 160	2.6% / 2.8%
Income		
Below 50k	9123 / 2763	43.7% / 48.0%
50k–200k	10516 / 2814	50.4% / 48.8%
200k–500k	854 / 95	4.1% / 1.6%
500k–1M	360 / 90	1.7% / 1.6%
Area		
Agricultural	10204 / 2729	48.9% / 47.4%
Non-agricultural	6609 / 1633	31.7% / 28.3%
Resident (from non-agri)	2468 / 875	11.8% / 15.2%
Resident (from agri)	1572 / 525	7.5% / 9.1%

Table 4: Demographic distribution of CGSS sample: train vs. test sets.

Variable / Category	Count (Train / Test)	Percent (Train / Test)
Gender		
Male	55435 / 2986	55.4% / 59.7%
Female	44565 / 2014	44.6% / 40.3%
Age		
<18	132 / 4	0.1% / 0.1%
18–35	13577 / 634	13.6% / 12.7%
35–50	24017 / 1129	24.0% / 22.6%
50–65	29420 / 1428	29.4% / 28.6%
65+	32854 / 1805	32.9% / 36.1%
Education		
Primary	4778 / 199	4.8% / 4.0%
Lower secondary	11888 / 583	11.9% / 11.7%
Upper secondary	59436 / 2963	59.4% / 59.3%
Tertiary	23898 / 1255	23.9% / 25.1%
Income		
Low	28860 / 1252	28.9% / 25.0%
Middle	47489 / 2578	47.5% / 51.6%
High	23651 / 1170	23.7% / 23.4%
Household size		
0	17 / 1	0.0% / 0.0%
1	24896 / 1178	24.9% / 23.6%
2	37336 / 2008	37.3% / 40.2%
3	17945 / 853	17.9% / 17.1%
4	13151 / 634	13.2% / 12.7%
5	4684 / 239	4.7% / 4.8%
6+	1971 / 87	2.0% / 1.7%
Region		
NUTS level 1	21348 / 993	21.3% / 19.9%
NUTS level 2	43833 / 2491	43.8% / 49.8%
NUTS level 3	34819 / 1516	34.8% / 30.3%
Industry		
Manufacturing	49098 / 2362	49.1% / 47.2%
Public Services	20735 / 1064	20.7% / 21.3%
Transport & Communication	15600 / 839	15.6% / 16.8%
Construction	9337 / 466	9.3% / 9.3%
Energy & Mining	1990 / 100	2.0% / 2.0%
Agriculture	1165 / 44	1.2% / 0.9%
Finance & Real Estate	923 / 57	0.9% / 1.1%
Personal & Social Svcs	849 / 56	0.8% / 1.1%
Business Services	303 / 12	0.3% / 0.2%
Occupation		
Clerical	27511 / 1324	27.5% / 26.5%
Skilled Workers	16365 / 816	16.4% / 16.3%
Associate Professionals	15194 / 794	15.2% / 15.9%
Operators & Drivers	14374 / 775	14.4% / 15.5%
Elementary & Farm	9202 / 463	9.2% / 9.3%
Service & Sales	8718 / 426	8.7% / 8.5%
Professionals	4617 / 187	4.6% / 3.7%
Managers	4019 / 215	4.0% / 4.3%

Table 5: Demographic distribution of ESS sample: train vs. test sets.

where q_α is the α-quantile of the loss distribution. This formulation focuses optimization on the highest-loss samples without requiring explicit group labels. CVaR DRO provides robustness guarantees by ensuring that the model performs well on the "hardest" portion of the data. However, it does not explicitly consider demographic group structure, which may limit its effectiveness when group membership is the primary source of performance disparity. Following Levy et al. (2020), we set $\alpha = 0.2$ in our experiments, meaning the optimization focuses on the top 20% highest-loss samples.

D.4 Group DRO (Group Distributionally Robust Optimization)

Group DRO (Sagawa et al., 2019) explicitly optimizes for the worst-performing demographic group:

$$\theta_{\text{GDRO}} = \arg\min_{\theta} \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\mathcal{D}_g} \left[\ell(f_\theta(x), y)\right], \tag{44}$$

where $\mathcal{G}$ denotes the set of demographic groups and $\mathcal{D}_g$ represents the data distribution conditioned on group g. Group DRO maintains a set of group weights $\{q_g\}_{g\in\mathcal{G}}$ that are updated via exponentiated gradient ascent to upweight groups with higher losses:

$$q_g^{(t+1)} \propto q_g^{(t)} \cdot \exp\left(\eta \cdot \hat{L}_g^{(t)}\right), \tag{45}$$

where $\hat{L}_g^{(t)}$ is the empirical loss for group g at step t and η is the step size for weight updates. This approach provides strong worst-group performance guarantees but can suffer from training instability, particularly when minority groups have very few samples, leading to high-variance gradient estimates. Following Sagawa et al. (2019), we use a group weight step size of $\eta = 0.01$ and apply strong ℓ_2 regularization ($\lambda = 1.0$) to stabilize training.

E Evaluation Metrics

We evaluate model performance from three perspectives: utility, fairness, and stability.

E.1 Utility

Accuracy. Accuracy measures the overall prediction quality across all samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}^{(i)} = y^{(i)}] \tag{46}$$

where N is the total number of samples, $\hat{y}^{(i)}$ is the predicted label, and $y^{(i)}$ is the ground truth label.

E.2 Fairness

Following the notation in Section A.1, fairness is evaluated separately for each attribute dimension $a \in \{1, 2, \ldots, A\}$. Let $\mathcal{G}_a$ denote the set of groups for attribute a, and let $p_{a,g}$ represent the performance (e.g., accuracy) of group $g \in \mathcal{G}_a$ under attribute dimension a.

Wasserstein Distance. For each attribute dimension a, we measure the Wasserstein Distance between the empirical group performance distribution and a uniform target distribution where all groups achieve the mean performance $\bar{p}_a$:

$$W_a = W_1(\mathbf{p}_a, \mathbf{u}_a) \tag{47}$$

where $\mathbf{p}_a = (p_{a,g})_{g\in\mathcal{G}_a}$ is the vector of group performances for attribute a, $\mathbf{u}_a = (\bar{p}_a, \bar{p}_a, \ldots, \bar{p}_a)$ is the uniform target with $\bar{p}_a = \frac{1}{|\mathcal{G}_a|}\sum_{g\in\mathcal{G}_a} p_{a,g}$, and W_1 denotes the 1-Wasserstein distance. We report the average across all attributes: $W = \frac{1}{A}\sum_{a=1}^{A} W_a$. **Lower values indicate better fairness.**

Worst Group Performance. Worst Group Performance captures the minimum performance across all demographic groups, aggregated across all attribute dimensions (Suh et al., 2025):

$$\text{Worst Group Perf.} = \min_{a\in\{1,\ldots,A\}} \min_{g\in\mathcal{G}_a} p_{a,g} \tag{48}$$

This metric directly relates to the Adversary objective $\mathcal{L}_{\text{Ad}}$, ensuring that no group is left behind. **Higher values indicate better fairness.**

Coefficient of Variation (CV). For each attribute dimension a, the Coefficient of Variation measures the relative dispersion of group performances:

$$\text{CV}_a = \frac{\sigma_{p_a}}{\bar{p}_a} = \frac{\sqrt{\frac{1}{|\mathcal{G}_a|}\sum_{g\in\mathcal{G}_a}(p_{a,g} - \bar{p}_a)^2}}{\bar{p}_a} \tag{49}$$

where $\bar{p}_a = \frac{1}{|\mathcal{G}_a|}\sum_{g\in\mathcal{G}_a} p_{a,g}$ is the mean group performance for attribute a and σ_{p_a} is the standard deviation. We report the average across all attributes: $\text{CV} = \frac{1}{A}\sum_{a=1}^{A}\text{CV}_a$. **Lower values indicate better fairness.**

Gini Coefficient. For each attribute dimension a, the Gini Coefficient quantifies inequality in group

performance distribution. For sorted performances $p_{a,(1)} \leq p_{a,(2)} \leq \cdots \leq p_{a,(|\mathcal{G}_a|)}$:

$$\text{Gini}_a = \frac{2\sum_{i=1}^{|\mathcal{G}_a|} i \cdot p_{a,(i)}}{|\mathcal{G}_a| \sum_{i=1}^{|\mathcal{G}_a|} p_{a,(i)}} - \frac{|\mathcal{G}_a|+1}{|\mathcal{G}_a|} \tag{50}$$

We report the average across all attributes: Gini $= \frac{1}{A}\sum_{a=1}^{A} \text{Gini}_a$. The Gini coefficient ranges from 0 (perfect equality) to 1 (maximum inequality). **Lower values indicate better fairness.**

Nash Welfare. For each attribute dimension a, Nash Welfare is defined as the geometric mean of group performances:

$$\text{NW}_a = \left(\prod_{g \in \mathcal{G}_a} p_{a,g}\right)^{1/|\mathcal{G}_a|} = \exp\left(\frac{1}{|\mathcal{G}_a|}\sum_{g \in \mathcal{G}_a} \log p_{a,g}\right) \tag{51}$$

We report the average across all attributes: Nash Welfare $= \frac{1}{A}\sum_{a=1}^{A} \text{NW}_a$. Nash Welfare satisfies Pareto efficiency and is particularly sensitive to improvements in lower-performing groups due to the logarithmic transformation. **Higher values indicate better fairness.**

E.3 Stability

Standard Deviation across Seeds. To assess reproducibility, we conduct all experiments across $S = 3$ independent random seeds and report the sample standard deviation:

$$\text{Std} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(m_s - \bar{m})^2} \tag{52}$$

where m_s is the metric value for seed s and $\bar{m} = \frac{1}{S}\sum_{s=1}^{S} m_s$ is the mean. Results are reported as mean $\pm$ std. **Lower values indicate more stable performance.**

F Hyperparameter Settings

This appendix documents all hyperparameters used in SAFO and baseline methods.

F.1 SAFO Component Hyperparameters

Table 6 summarizes all hyperparameters specific to the SAFO framework.

F.1.1 Adversary Hyperparameters

The temperature $\tau = 1$ controls the sharpness of the group weight distribution (Equation (5)). As shown in Theorem 1 (Appendix A), this provides a $\tau \log K$-approximation to the exact worst-group loss, where K is the number of groups. The adaptive EMA decay rate $\alpha_g(t)$ is computed as:

$$\alpha_g(t) = \text{clip}\left(\frac{N_{\text{scale}}}{N_{\text{scale}} + n_g^t}, \alpha_{\min}, \alpha_{\max}\right), \tag{53}$$

where n_g^t is the cumulative sample count for group g up to step t. Setting $N_{\text{scale}} = 1000$ means that a group reaches the midpoint decay rate ($\alpha = 0.5$) after observing 1000 samples. The bounds $[\alpha_{\min}, \alpha_{\max}] = [0.1, 0.9]$ ensure that minority groups benefit from substantial smoothing while majority groups respond quickly to recent observations.

F.1.2 Regulator Hyperparameters

The coefficient λ_t balances the Optimizer loss $\mathcal{L}_{\text{Op}}$ and Adversary loss $\mathcal{L}_{\text{Ad}}$ in the combined objective. We initialize $\lambda_0 = 0.5$ to provide equal weight to utility and fairness at the start of training. The bounds $[\lambda_{\min}, \lambda_{\max}] = [0.2, 0.8]$ ensure that neither objective dominates completely. The growth and decay factors ($\gamma_g = 1.005$, $\gamma_d = 0.995$) are set close to 1.0 to ensure gradual changes. The Nash threshold $\gamma_n = 1.0$ enforces Pareto efficiency: any increase in average loss must be justified by at least an equal decrease in worst-group loss. The smoothing factor $\beta_s = 0.98$ provides a longer temporal horizon compared to group-level EMA.

F.2 Baseline Method Hyperparameters

Following Levy et al. (2020), CVaR DRO uses risk level $\alpha = 0.2$. Following Sagawa et al. (2019), Group DRO uses weight step size $\eta_q = 0.01$ for the exponentiated gradient ascent update. The size-based adjustment coefficient is set to 0.0 to isolate the effect of the DRO objective.

F.3 Training Hyperparameters

We apply LoRA with rank $r = 8$ to all linear layers, with scaling factor $\alpha = 16$. We use AdamW with learning rate 1×10^{-5} and cosine annealing with 10% warmup. The batch size of 64 ensures sufficient group representation within each batch. All experiments are conducted on a computing cluster with 8 NVIDIA H20 GPUs (96GB HBM3 memory each) using distributed data parallel (DDP) training.

F.4 Ablation Study Configurations

For ablation experiments, we modify specific components while keeping others fixed:

Table 6: Complete hyperparameter settings for SAFO components.

Component	Parameter	Value	Description
Regulator	Initial λ_t (λ_0)	0.5	Initial utility-fairness trade-off coefficient
	λ_t range $[\lambda_{\min}, \lambda_{\max}]$	$[0.2, 0.8]$	Bounds for dynamic λ adjustment
	Growth factor γ_g	1.005	Multiplicative increase when fairness is beneficial
	Decay factor γ_d	0.995	Multiplicative decrease when collateral damage is high
	Nash threshold γ_n	1.0	Ratio threshold for Pareto efficiency assessment
	Statistics smoothing β_s	0.98	EMA decay for Nash-inspired mechanism statistics
Adversary	Softmax temperature τ	1	Temperature for group weight computation
	Minimum groups threshold	2	Minimum groups per attribute for group-reweighted loss
Adaptive EMA	Scale constant N_{scale}	1000	Controls adaptive decay rate sensitivity
	Minimum decay $\alpha_{\min}$	0.1	Lower bound for EMA decay rate
	Maximum decay $\alpha_{\max}$	0.9	Upper bound for EMA decay rate
	Warmup steps	10	Steps before EMA updates begin

- **w/o DRO (Adversary ablation)**: Replace softmax-weighted group loss with hard maximum over EMA losses: $\mathcal{L}_{\text{Ad}} = \max_g \hat{L}_g^{\text{batch}}$ where $g^* = \arg\max_g \bar{L}_g$.
- **w/o Nash (Regulator ablation)**: Fix $\lambda_t = 0.5$ throughout training.
- **w/o EMA**: Use batch-level group losses directly instead of EMA-smoothed estimates.

G Stability Discussion

Figure 6 and Figure 7 compare training dynamic across demographic attributes.

H Parameter Sensitivity Analysis

We analyze the sensitivity of our method to two key hyperparameters: the Nash threshold γ_n and the adversary temperature τ. Results are shown in Figure 8 and Figure 9.

Figure 8 shows performance across $\gamma_n \in \{0.5, 1.0, 2.0\}$. Figure 9 examines $\tau \in \{0.1, 0.3, 0.5, 1.0\}$. Based on our analysis, we recommend $\gamma_n = 1.0$ as the default, which provides the optimal accuracy-fairness trade-off. For adversary temperature, $\tau \in \{0.1, 1.0\}$ are both reasonable choices depending on whether accuracy or fairness is prioritized.

I Extended Baseline and Backbone Results

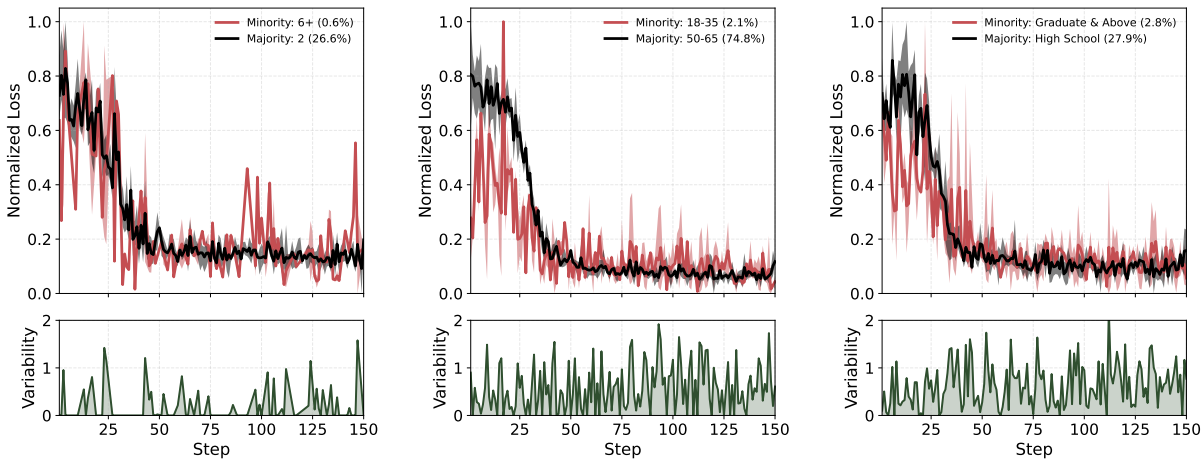
To complement the main results in Table 1, we report two additional sets of experiments: (i) extended comparisons against two recent fairness baselines, FairDRO (Park et al., 2025) and ROAD (Grari et al., 2023), adapted to our multi-attribute setting as described in Appendix J; and (ii) results on an additional backbone, Mistral-7B-Instruct-v0.3, to further assess cross-architecture generality. All experiments follow the same training and evaluation protocol as Table 1 and report mean $\pm$ std across 3 random seeds.

I.1 Additional Fairness Baselines

Table 7 reports the full six-metric comparison of FairDRO and ROAD against SAFO on the GSS dataset with LLaMA-3.1-8B and Qwen2.5-7B backbones. SAFO consistently outperforms both adapted baselines on accuracy, worst-group performance, and Nash welfare, while maintaining the lowest cross-seed variance in most cases. FairDRO narrows the fairness gap relative to ERM but exhibits larger variance on worst-group performance, and ROAD suffers noticeable utility drops due to its adversarial debiasing objective that conflicts with the preservation of demographic-conditioned heterogeneity required in survey simulation.

I.2 Mistral-7B Backbone Results

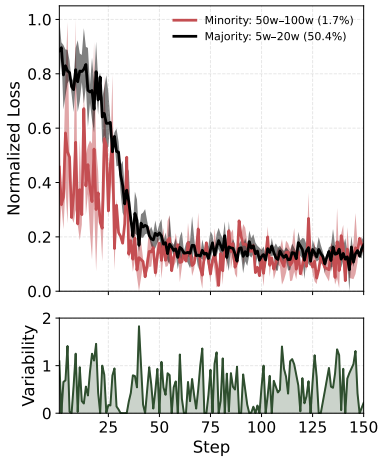
Table 8 reports SAFO and the two adapted fairness baselines (FairDRO and ROAD) on the GSS dataset with the Mistral-7B-Instruct-v0.3 backbone. SAFO's advantage on Mistral mirrors the pattern observed on LLaMA and Qwen in Table 7, achieving the highest accuracy, worst-group performance, and Nash welfare with substantially lower variance. This confirms that the framework generalizes across decoder-only LLM architectures.



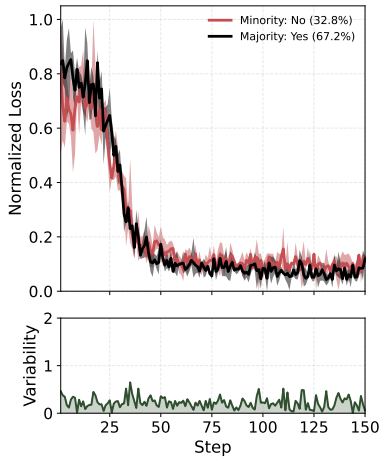
(a) Family Size

(b) Age

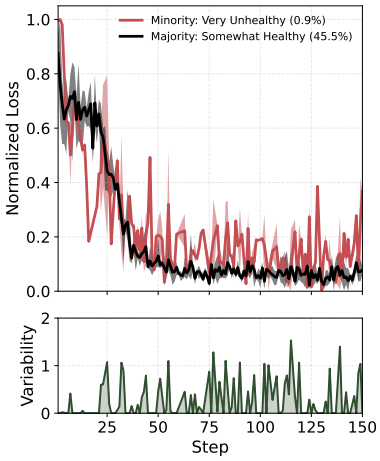
(c) Education



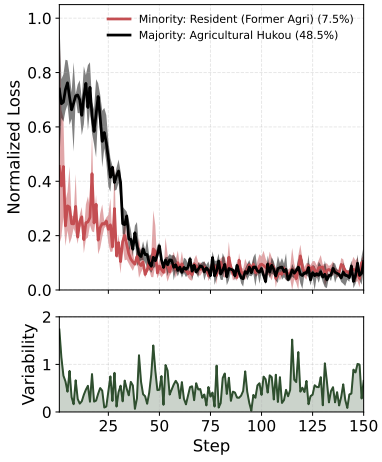
(d) Income



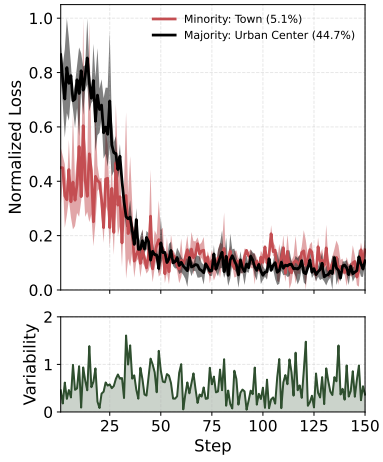
(e) House Ownership



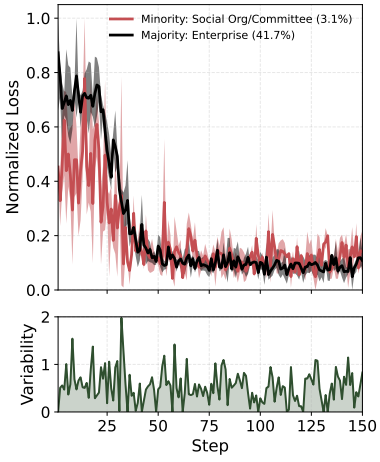
(f) Health Status



(g) Area

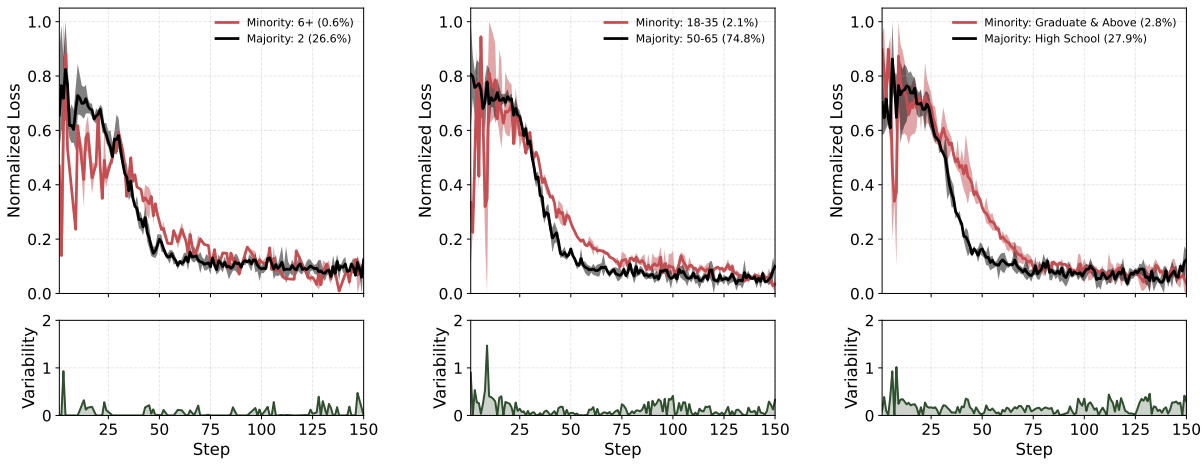


(h) Residence Type



(i) Work Unit Type

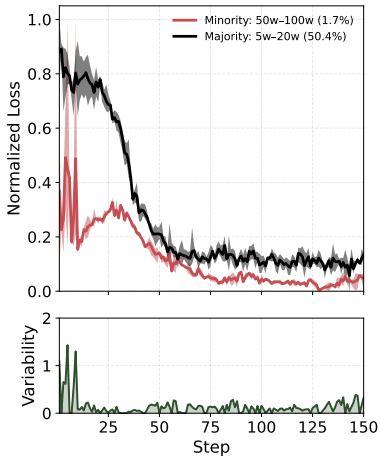
Figure 6: Group DRO training dynamics across different demographic attributes. Red lines represent minority groups and black lines represent majority groups, with percentages indicating group proportions.



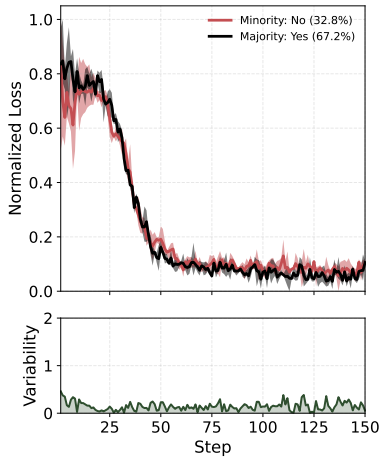
(a) Family Size

(b) Age

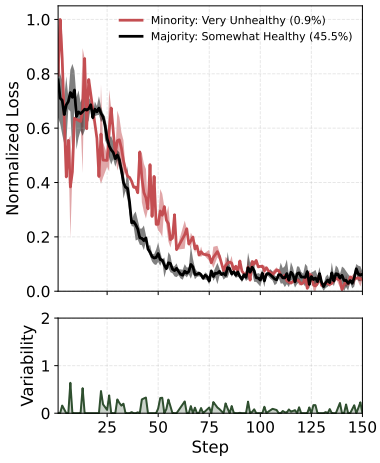
(c) Education



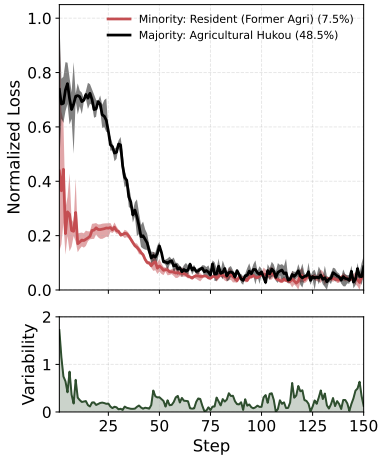
(d) Income



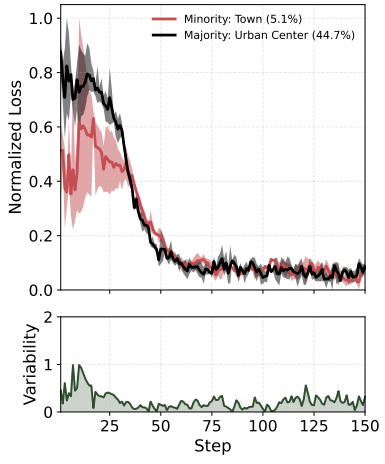
(e) House Ownership



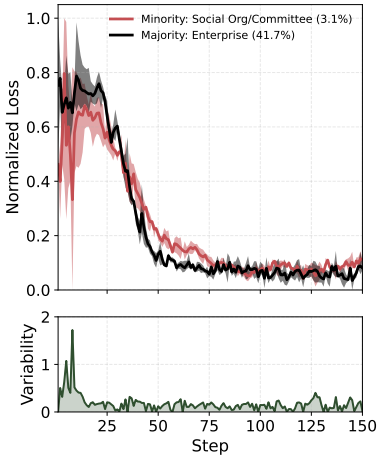
(f) Health Status



(g) Area



(h) Residence Type



(i) Work Unit Type

Figure 7: SAFO training dynamics across different demographic attributes. Red lines represent minority groups and black lines represent majority groups, with percentages indicating group proportions.

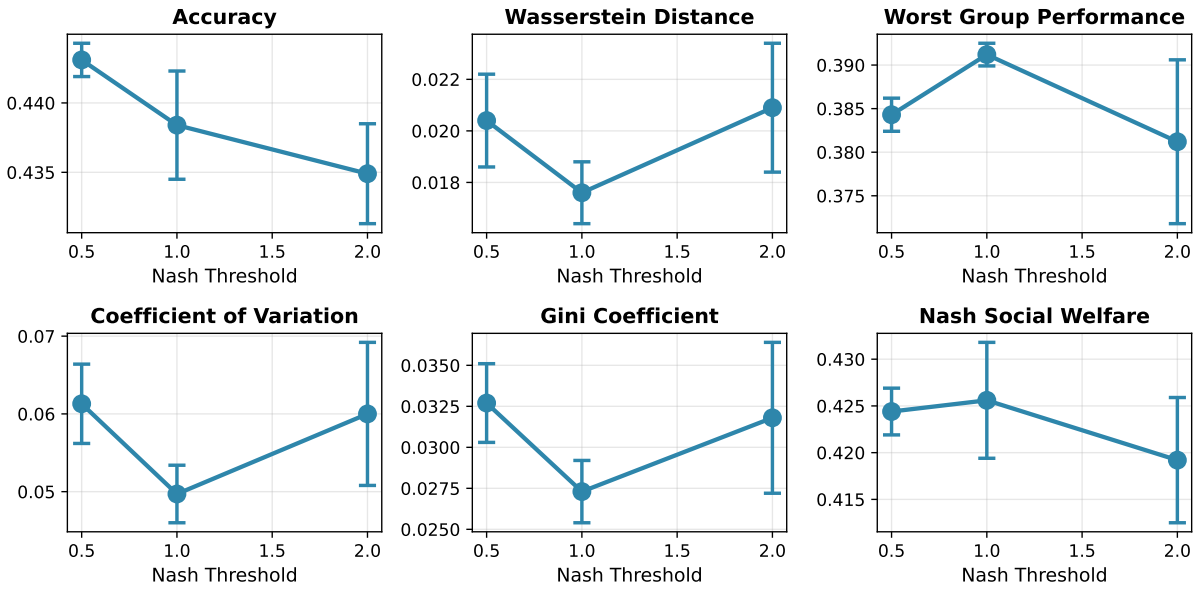


Figure 8: Sensitivity analysis for Nash threshold γ_n.

Model	Method	Accuracy	Worst Group Perf.	Nash Welfare
LLaMA-3.1-8B	FairDRO	0.42 ± 0.0045	0.37 ± 0.0117	$0.40 \pm \underline{0.0012}$
	ROAD	0.39 ± 0.0047	0.33 ± 0.0225	0.35 ± 0.0200
	SAFO	$\mathbf{0.44} \pm \underline{0.0039}$	$\mathbf{0.39} \pm \underline{0.0013}$	$\mathbf{0.43} \pm 0.0062$
Qwen2.5-7B	FairDRO	0.43 ± 0.0121	$0.37 \pm \underline{0.0072}$	0.41 ± 0.0132
	ROAD	0.40 ± 0.0118	0.33 ± 0.0162	0.38 ± 0.0134
	SAFO	$\mathbf{0.45} \pm \underline{0.0051}$	$\mathbf{0.38} \pm 0.0039$	$\mathbf{0.43} \pm \underline{0.0056}$

Table 7: FairDRO and ROAD comparison on the GSS dataset. Results are mean $\pm$ std across 3 seeds. **Bold**: best metric value; underline: smallest standard deviation.

J Intersectional Fairness Evaluation with Additional Baselines

To complement the per-attribute fairness analysis in the main text, we additionally evaluate SAFO against two recent fairness-oriented baselines, **FairDRO** (Park et al., 2025) and **ROAD** (Grari et al., 2023), under cross-attribute intersectional settings. Since LLM-based social simulation is not a standard classification task and requires handling multiple demographic attributes simultaneously, the original formulations are not directly applicable. We adapt both methods by (i) treating each demographic attribute's subgroups as the group partition required by each algorithm, and (ii) applying per-attribute aggregation consistent with our framework to ensure a fair comparison.

We evaluate on two representative cross-attribute intersections on the GSS dataset with the LLaMA backbone: age$\times$income (a demographically sparse and high-stakes combination) and race$\times$sex. Table 9 reports accuracy, worst-group performance, and Nash welfare across 3 random seeds. The intersectional setting substantially amplifies subgroup sparsity, and we observe that ERM, Group DRO, and ROAD all experience near-complete failure on certain age$\times$income cells, with worst-group performance dropping as low as 0.04 and seed variance exceeding 0.10. FairDRO partially mitigates this failure but still lags SAFO on both worst-group and Nash metrics. In contrast, SAFO maintains stable performance under the intersectional setting, improving worst-group accuracy while simultaneously reducing cross-seed variability by roughly an order of magnitude.

K Illustration of the Social Simulation Task

To improve clarity for readers unfamiliar with social simulation, we present a concrete example illustrating our task formulation, including the prompt structure and the model-generated response.

In this setting, the model is instructed to simu-

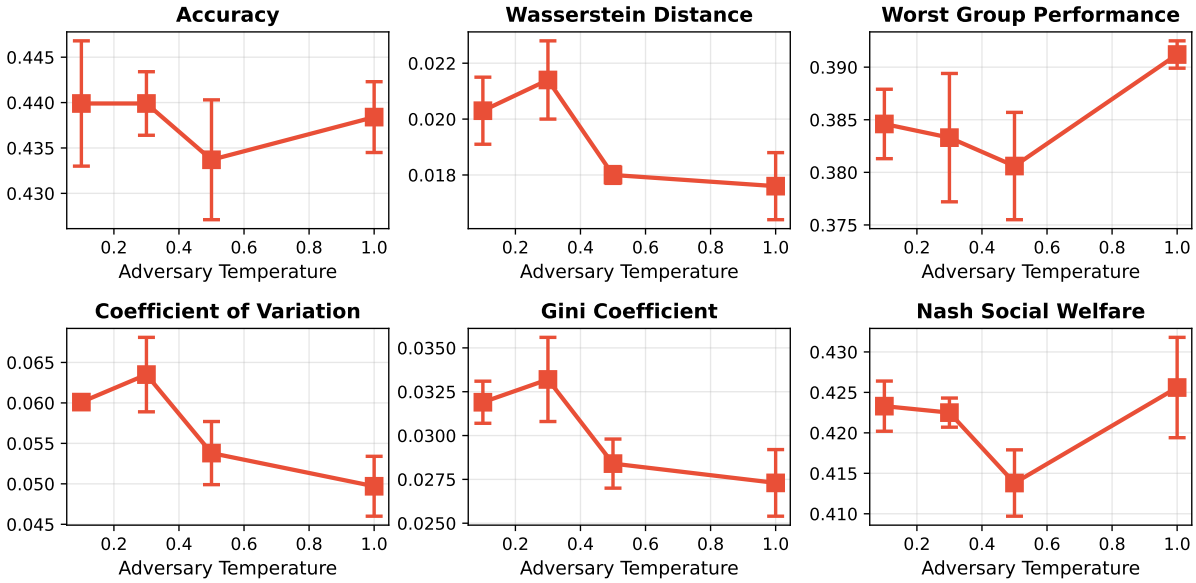


Figure 9: Sensitivity analysis for adversary temperature τ.

Model	Method	Accuracy	Worst Group Perf.	Nash Welfare
Mistral-7B	FairDRO	0.43 $\pm$ 0.0068	0.38 $\pm$ 0.0103	0.42 $\pm$ 0.0084
	ROAD	0.40 $\pm$ 0.0072	0.34 $\pm$ 0.0158	0.38 $\pm$ 0.0109
	SAFO	**0.45** $\pm$ 0.0044	**0.40** $\pm$ 0.0041	**0.43** $\pm$ 0.0067

Table 8: FairDRO and ROAD comparison on the GSS dataset with the Mistral-7B-Instruct-v0.3 backbone. Results are mean $\pm$ std across 3 seeds. **Bold**: best metric value; underline: smallest standard deviation.

late a real survey respondent in a specific temporal context. Each prompt consists of three key components: (1) an explicit instruction describing the simulation objective, (2) a detailed demographic and socioeconomic background of the respondent, and (3) a survey question with predefined answer options. The model is required to select exactly one option, without generating additional explanation, in order to mirror real-world survey behavior.

Table 10 provides an example of the full prompt and corresponding output used in our experiments.

Setting	Method	Accuracy	Worst Group Perf.	Nash Welfare
age×income	ERM	0.44 ± 0.0050	0.07 ± 0.1155	0.28 ± 0.1465
	Group DRO	0.44 ± 0.0081	0.07 ± 0.0714	0.33 ± 0.1506
	FairDRO	0.43 ± 0.0021	0.13 ± 0.0463	0.40 ± 0.0109
	ROAD	0.39 ± 0.0028	0.04 ± 0.0642	0.27 ± 0.0808
	SAFO	**0.45** ± 0.0015	**0.14** ± 0.0100	**0.41** ± 0.0065
race×sex	ERM	0.44 ± 0.0050	0.38 ± 0.0066	0.41 ± 0.0128
	Group DRO	0.44 ± 0.0081	0.40 ± 0.0150	0.42 ± 0.0041
	FairDRO	0.43 ± 0.0021	0.38 ± 0.0260	0.42 ± 0.0101
	ROAD	0.39 ± 0.0028	0.36 ± 0.0027	0.38 ± 0.0037
	SAFO	**0.45** ± 0.0015	**0.41** ± 0.0040	**0.43** ± 0.0031

Table 9: Intersectional fairness evaluation on GSS (LLaMA-3.1-8B). Results are reported as mean ± std across 3 random seeds. **Bold** indicates the best metric value; underline indicates the smallest standard deviation. Optimization directions: Accuracy (↑), Worst Group Perf. (↑), Nash Welfare (↑).

INSTRUCTION:
The survey is conducted in 2024.
You are required to simulate a real survey respondent answering the following question.

== Respondent Background ==
Age: 65+
Sex: Male
Race: White
Marital Status: Divorced
Education: High School
Household Income: $15,000–$19,999
Employment Status: Unemployed, laid off, looking for work
Place of Residence: Medium-size city (50,000–250,000 population)
Region: Mountain

QUERY:
Generally speaking, would you say that people can be trusted or that you can't be too careful in dealing with people?

Options:
A. You usually can't be too careful in dealing with people
B. People can almost always be trusted
C. People can usually be trusted
D. You almost always can't be too careful in dealing with people

ANSWER:
C. People can usually be trusted

Table 10: Example prompt and model output for the social simulation survey task.