Controlling Distributional Bias in Multi-Round LLM Generation via KL-Optimized Fine-Tuning

Yanbei Jiang[1] **Amr Keleg**[2] **Ryandito Diandaru**[2] **Jey Han Lau**[1]
Lea Frermann[1] **Biaoyan Fang**[3] **Fajri Koto**[2]

[1]The University of Melbourne [2]MBZUAI [3]Oracle

Correspondence: yanbeij@student.unimelb.edu.au

Abstract

While the real world is inherently stochastic, Large Language Models (LLMs) are predominantly evaluated on single-round inference against fixed ground truths. In this work, we shift the lens to *distribution alignment*: assessing whether LLMs, when prompted repeatedly, can generate outputs that adhere to a desired target distribution, e.g. reflecting real-world statistics or a uniform distribution. We formulate distribution alignment using the attributes of gender, race, and sentiment within occupational contexts. Our empirical analysis reveals that off-the-shelf LLMs and standard alignment techniques, including prompt engineering and Direct Preference Optimization, fail to reliably control output distributions. To bridge this gap, we propose a novel fine-tuning framework that couples Steering Token Calibration with Semantic Alignment. We introduce a hybrid objective function combining Kullback-Leibler divergence to anchor the probability mass of latent steering tokens and Kahneman-Tversky Optimization to bind these tokens to semantically consistent responses. Experiments across six diverse datasets demonstrate that our approach significantly outperforms baselines, achieving precise distributional control in attribute generation tasks.[1]

1 Introduction

Although generative Large Language Models (LLMs) are inherently probabilistic, existing evaluation methods typically focus on a single-round generation (De-Arteaga et al., 2019; Soundararajan and Delany, 2024; Pan et al., 2025). This narrow view becomes problematic in real-world applications, where repeated prompting can lead to distributional properties in LLM output that deviates from the intended statistics. We refer to this phenomenon as *distribution bias* (Shrestha and Srinivasan, 2025), defined as the difference between the observed output distribution and the target distribution. Distribution bias matters because it influences how LLMs represent social attributes such as gender, race, and sentiment, with implications for fairness (Gallegos et al., 2024), personalization (Salemi et al., 2024), and trustworthiness (Litschko et al., 2023).

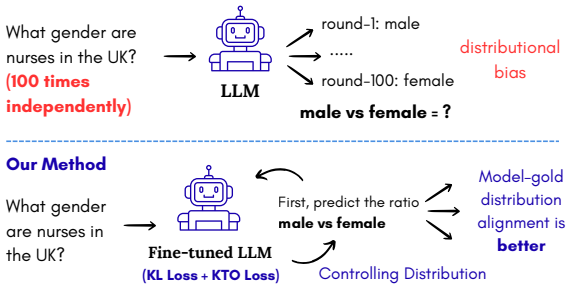


Figure 1: Distributional Bias in Multi-round LLM Generation

While Shrestha and Srinivasan (2025) examine distributional bias at the word level probability, our work focuses on multi-round generation. Consider the example in Figure 1: if we independently prompt an LLM 100 times with "*Write a story about a nurse in the UK!*", the aggregate gender distribution may reflect parity (equally many men and women nurses) or reflect and thereby reinforce stereotypes (predominantly women nurses).[2] This observation highlights a fundamental limitation: LLMs lack distributional understanding and controllability in repeated generations.

This question arises in two scenarios, reflecting different desiderata. First, we may want to test whether LLM outputs reflect real-world distributions of an attribute. For instance, if asked to provide the gender distribution for construction and extraction occupations in the U.S. in 2024, the expected ratio is approximately 96% male and 4%

[1]Code and data are available at https://github.com/YanbeiJiang/Distribution-Debias.

[2]We acknowledge that treating gender as binary unduly simplifies the concept, but adopt this simplification for experimental clarity. Our method extends to > 2 classes.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17634–17649

female.[3] Second, we may wish to enforce a user-specified distribution, e.g., where educational texts should represent all genders equally, as is often the case in the context of *debiasing* (Rooein et al., 2025).

This paper investigates two core questions: (i) *What distributional biases do LLMs exhibit when prompted repeatedly, relative to real-world statistics or a user-specified distribution?* (ii) *Can we control attribute distributions in multi-round generation to match a desired target distribution?* We formulate this challenge as distributional bias control in multi-round LLM generations repeating the same instruction in a number of independent trials.

Our contributions can be summarized as follows: (i) We show that distributional biases exist in multi-round generation of off-the-shelf LLMs, both relative to real-world statistics and user-specified uniform distributions. (ii) We create distribution-aware preference datasets focusing on occupations in the UK and US, and attributes such as gender, race, and sentiment across two settings: attribute generation and story generation. (iii) We introduce a novel method to control output distributions in multi-round generation by fine-tuning LLMs with a hybrid objective combining distribution calibration (KL divergence) and semantic consistency (KTO), steering LLM outputs toward the desired target distribution. Through extensive experiments, we demonstrate that our approach significantly outperforms baselines.

2 Related Work

Control Generation Control generation focuses on guiding and constraining the LLM output on certain conditions, e.g., styles (De Langis et al., 2024; Toshevska and Gievska, 2025; Miura et al., 2025) and attributions (Liang et al., 2024; Lorandi and Belz, 2023; Pang et al., 2025). It has been implemented in various ways, including (1) prompt-based generation (Jie et al., 2024; Suzgun et al., 2022; Liu et al., 2024; Jiang et al., 2025a), (2) classifier-guided generation (Konen et al., 2024; Mai et al., 2023), and (3) reinforcement learning–based control generation (Shulev and Sima'an, 2024; Deng et al., 2022). Distribution alignment is in line with control generation in restricting LLM outputs on certain attributes. However, control generation targets on generating the desired output per generation while distribution alignment focuses on controlling LLMs' generation over several and independently repeated outputs.

Debiasing Distribution alignment is closely related to debiasing in LLMs, which aims to ensure LLMs generate fair responses/distribution to each protected attribute group (e.g., gender (Dinan et al., 2020; Fan and Gardent, 2022; Soundararajan and Delany, 2024) or ethnicity (Narayanan Venkit et al., 2023; Fang et al., 2024)). Recent work explores various debiasing methods, including (1) prompt-based debiasing (Wan and Chang, 2025; Bansal et al., 2022; Huang et al., 2024); (2) pretraining-based debiasing (Zakizadeh and Pilehvar, 2025; Gira et al., 2022; Shrestha and Srinivasan, 2025); (3) reinforcement learning/preference–based debiasing (Xia et al., 2024; Fan et al., 2025; Jiang et al., 2025b). Those methods are tested on carefully curated datasets with desired attribute ratios and evaluated based on the output result over a few runs for each data point. Whether debiasing methods can be applied to guide LLMs to align desired distribution on multi-round generation remains an open question.

LLM Output Consistency Research on LLM consistency investigates why the same model can produce different or even contradictory outputs under minor—or no—changes to input or conditions (Wu et al., 2025; Vázquez et al., 2024; Yang et al., 2024). Increasing decoding hyperparameters such as temperature or top-p sampling often leads to more diverse outputs (Peeperkorn et al., 2024). While related, distribution alignment differs from consistency: it seeks to match the overall attribute distribution across repeated generations while permitting individual variation. Both approaches address variability in generation and are therefore complementary. Our work builds on these insights by analyzing default temperature and top-p settings in multi-round generation and evaluating their impact across different configurations.

3 Methods

We introduce *distribution alignment*, a framework designed to control attribute distributions in multi-round LLM generation. The objective is to align the output distribution with a specified target distribution for sensitive attributes (e.g., gender, race, sentiment) while preserving semantic quality. We propose a two-stage optimization strategy: (1) constructing a distribution-aware preference dataset,

[3] https://www.bls.gov/cps/cpsaat39.htm (29 Dec, 2025).

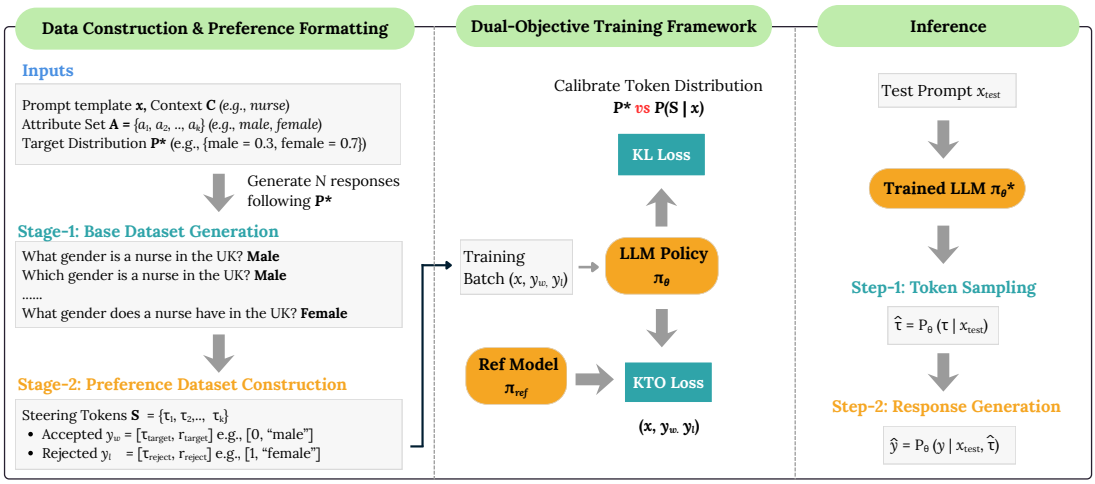


Figure 2: Mitigating Distributional Bias with KTO and KL Loss

and (2) fine-tuning the model using a hybrid objective that combines Kullback-Leibler (KL) divergence for distribution calibration and Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) for semantic consistency.

3.1 Preliminaries and Problem Formulation

Let $\mathcal{M}$ denote an LLM taking a prompt $x \in \mathcal{X}$ to generate a response $y \in \mathcal{Y}$. We define a domain of *contexts* $\mathcal{C}$ involved in the prompt, which represents the specific scenarios serving as the basis for distribution evaluation. In this work, we instantiate $\mathcal{C}$ as the domain of **occupations** (e.g., doctor, nurse), though the framework applies to other domains.

We focus on a sensitive attribute set $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ (e.g., {Male, Female} or {Positive, Negative, Neutral}) associated with a specific context $c \in \mathcal{C}$. Given a prompt x conditioned on context c, let $\mathbb{P}^*(a_i|c)$ represent the *target distribution* over attributes $a_i \in \mathcal{A}$. This target distribution is defined by a user and may for instance reflect real-world statistics (e.g., labor census data) or a neutral uniform distribution.

Our objective is to train a parameterized policy π_θ such that, for a given prompt x, the empirical distribution of the generated attributes over N independent sampling trials approximates $\mathbb{P}^*(a_i|x)$.

3.2 Optimization Framework

Inspired by Group Distributional Preference Optimization (GDPO) (Yao et al., 2025), our training objective combines **(1)** calibration loss to align steering token probabilities with the target distribution and **(2)** preference loss to match generated text to the steering token. The summary of our method is presented in Figure 2.

Steering Token To facilitate explicit control over the attribute distribution, we introduce a set of Steering Tokens, denoted as $\mathcal{S} = \{\tau_1, \tau_2, \ldots, \tau_k\}$, where each token τ_i uniquely corresponds to an attribute a_i. These tokens are added to the model's vocabulary and serve as precursors to the response.

KL Calibration The first objective ensures that the model's probability of generating a specific steering token matches the target distribution $\mathbb{P}^*$. We apply a KL Divergence loss on the logits of the steering tokens immediately following the prompt x. Let $P_\theta(\tau_i|x)$ be the probability assigned by the model to the steering token τ at the generation step. We minimize:

$$\mathcal{L}_{KL}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[D_{KL}\left(\mathbb{P}^*(\cdot|x) \parallel P_\theta(\cdot|x)\right)\right] \quad (1)$$

where the domain of the distribution is restricted to the set of steering tokens $\mathcal{S}$. This forces the model's "intent" to match the target demographic constraints.

Semantic Alignment via KTO While the KL objective calibrates the probability of the steering token τ_i, it does not guarantee that the subsequent response r semantically adheres to the attribute implied by τ. To enforce this consistency (e.g., ensuring text following a "Positive" token is indeed positive), we employ KTO loss (Ethayarajh et al., 2024). Unlike traditional preference optimization which strictly compares pairs, KTO defines a value function for individual examples based on whether they are desirable or undesirable. In our context, we treat the target-aligned sequence $y_w = [\tau_{target}; r_{target}]$ as *desirable* and the rejected sequence $y_l = [\tau_{reject}; r_{reject}]$ as *undesirable*, where $[;]$ is concatenation. Note that both

sequences explicitly condition on their respective steering tokens.

We first define the implicit log-likelihood ratio $R_\theta(x, y)$ and the reference point z_0 as the KL divergence between the policy π_θ and reference model (initial base model) π_{ref} under the ideal distribution:

$$\begin{aligned} R_\theta(x, y) &= \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\ z_0 &= \text{KL}(\pi_\theta(y'|x) \| \pi_{\text{ref}}(y'|x)) \end{aligned} \tag{2}$$

The KTO value function $v(x, y)$ is then defined specifically for our desirable (y_w) and undesirable (y_l) cases:

$$v(x, y) = \begin{cases} \lambda_D \sigma(\beta(R_\theta(x, y) - z_0)) & \text{if } y = y_w \\ \lambda_U \sigma(\beta(z_0 - R_\theta(x, y))) & \text{if } y = y_l \end{cases} \tag{3}$$

where σ is the sigmoid function, and λ_D, λ_U are weighting hyperparameters for desirable and undesirable outputs, respectively. β controls how far π_θ drifts from π_{ref}. The final semantic alignment loss is the expectation over our constructed dataset $\mathcal{D}_{pref}$:

$$\begin{aligned} \mathcal{L}_{\text{KTO}}(\theta) = \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_{\text{pref}}} \Big[&(\lambda_D - v(x, y_w)) \\ &+ (\lambda_U - v(x, y_l)) \Big] \end{aligned} \tag{4}$$

By minimizing this objective, the model maximizes the value of responses that are semantically consistent with their prefix steering tokens while suppressing inconsistent generations.

Final Objective The final training objective is a sum of the calibration and alignment losses:

$$\mathcal{L}_{total} = \mathcal{L}_{KTO} + \mathcal{L}_{KL} \tag{5}$$

Intuitively, $\mathcal{L}_{KL}$ acts as a macro-controller adjusting the distribution of the steering tokens, while $\mathcal{L}_{KTO}$ ensures the micro-level generation is logically consistent with the chosen token.

3.3 Data Construction

We first synthesize a base dataset that adheres to the target statistical distribution, and subsequently restructure it into a preference-based format to facilitate our contrastive optimization objectives.

Base Dataset Construction For each instruction x (an occupation-specific prompt), we generate a base set of N responses, denoted as $\mathcal{D}_{base} = \{(x, r_i, a_i)\}_{i=1}^{N}$, where r_i is the textual response and $a_i \in \mathcal{A}$ is the corresponding attribute.

To embed the target distribution $\mathbb{P}^*$ into the training data, we explicitly control the generation frequency such that the count N_k of responses exhibiting attribute a_k satisfies:

$$N_k = \text{round}\left(N \cdot \mathbb{P}^*(a_k|x)\right) \tag{6}$$

For instance, given a target distribution of {Male: 0.99, Female: 0.01} and $N = 100$, $\mathcal{D}_{base}$ will contain 99 responses with the Male attribute and 1 with the Female attribute.

Transformation to Preference Pairs To apply our hybrid loss function, we transform $\mathcal{D}_{base}$ into a preference dataset $\mathcal{D}_{pref} = \{(x, y_w, y_l)\}$. For each sample (x, r, a) from the base dataset, we construct a preferred response y_w and a rejected response y_l as follows:

1. **Accepted Response (y_w):** We define the accepted sequence by prepending the steering token τ_a corresponding to the instance's attribute a to the original response content r:

$$y_w = [\tau_a; r] \tag{7}$$

2. **Rejected Response (y_l):** We construct a rejected sequence to provide a contrastive signal. We first sample a negative attribute a_{neg} from the set of remaining attributes $\mathcal{A} \setminus \{a\}$. To prevent introducing secondary biases, we enforce a uniform distribution for this sampling:

$$a_{neg} \sim \text{Uniform}(\mathcal{A} \setminus \{a\}) \tag{8}$$

The rejected response is then formed using the steering token for the negative attribute and a corresponding generated content r_{neg}:

$$y_l = [\tau_{a_{neg}}; r_{neg}] \tag{9}$$

This construction ensures that y_w strictly follows the target distribution (e.g., 99% Male), with the rejection signal y_l providing balanced contrastive examples across the remaining attribute space.

3.4 Inference and Sampling Strategy

During inference, we employ a two-step generation process to explicitly manipulate the output distribution.

1. **Token Sampling:** Given a test prompt x, we first compute the logits for the set of steering tokens $\mathcal{S}$. We convert these logits to probabilities and sample a token $\hat{\tau}$:

$$\hat{\tau} \sim P_\theta(\tau|x), \quad \tau \in \mathcal{S} \tag{10}$$

The samples should approximate the target distribution $\mathbb{P}^*$ after training to minimize $\mathcal{L}_{KL}$.

2. **Response Generation:** We append the sampled token to the prompt and generate the subsequent response autoregressively:

$$\hat{r} = \text{Generate}(x \oplus \hat{\tau}) \tag{11}$$

4 Experimental Setup

Datasets To validate our method, we use the context of US and UK occupations ($\mathcal{C}$), covering a total of 39 distinct roles. We primarily focus on **Gender** ($\mathcal{A}$ = {"Male", "Female"}), constructing two distinct reference distributions: (1) a real-world setting derived from census statistics datasets (Statistics, 2023, 2025), and (2) an uniform setting for all occupations with the notion that an LLM should not prioritize any gender in this context. To demonstrate generalization, we extend our evaluation to **Race** and **Sentiment** within the same occupational contexts. We define the Race attribute set as $\mathcal{A}$ ={"White", "Black or African American", "American Indian or Alaska Native", "Asian", "Native Hawaiian or Other Pacific Islander"}. For Sentiment, we define $\mathcal{A}$ = {"Positive", "Negative", "Neutral"}. Here, for Race and Sentiment, we only use the uniform reference distribution as we have no real-world stats for these attributes.

For all datasets, the prompts ask the model to generate a response for a specific occupation given a target attribute group. To ensure robustness and prevent overfitting to specific phrasing, we utilize GPT-5.1 (OpenAI, 2025) to rephrase the instructions, generating N=100 distinct prompts per occupation for the training, validation and testing sets respectively. The prompt examples for all settings and the dataset size are provided in Appendix A.1.

Evaluation Metrics Our primary metric is the Mean Absolute Error (MAE), which quantifies the deviation of the model's generated distribution from the target distribution. For a given occupation, let $\mathbb{P}^*(a_i)$ be the target probability of attribute a_i, and $\hat{P}(a_i)$ be the empirical probability observed in the model's outputs (calculated over $N = 100$ runs). The MAE is defined as:

$$\text{MAE} = \frac{1}{|\mathcal{A}|} \sum_{a_i \in \mathcal{A}} |\hat{P}(a_i) - \mathbb{P}^*(a_i)| \tag{12}$$

We report the average MAE across all occupations in the test set. Lower MAE indicates better alignment with the target distribution. Note that since some models fail to follow the required output format, we have filtered out those cases. The distribution shown represents only the valid responses, meaning the sample size N is not always 100.

Baselines We compare our approach against five baseline methods comprising both prompting strategies and supervised alignments. First, we evaluate **Zero-shot**, which utilizes standard prompting without any distribution constraints. Second, we test **PE-Explicit**, a prompt engineering method where the target distribution is explicitly specified in the instruction (e.g., "Generate responses such that 90% are Male..."), and **PE-Implicit**, which vaguely instructs the model to "follow real-world statistics" (applied only to real-world distribution tasks). For supervised baselines, we include **Instruction Fine-Tuning (IFT)**, which fine-tunes the model using the standard Supervised Fine-Tuning loss on the $\mathcal{D}_{base}$ training data. Finally, we compare against **Direct Preference Optimization** (**DPO;** Rafailov et al., 2023), a preference alignment baseline trained on our $\mathcal{D}_{pref}$ dataset but excluding our proposed steering token calibration mechanism.

Implementation Details We evaluate models from the Qwen (Team et al., 2024) and Llama (Touvron et al., 2023) families, each instantiated at two parameter scales. All models and baselines are evaluated using the vLLM (Kwon et al., 2023) framework. We use standard sampling parameters with temperature T=1.0, top-p=1.0, and top-k= $-$ 1. For training (IFT, DPO, and Ours), we set the batch size to 4, total epochs to 5, and learning rate to 1e $-$ 6. We use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning with r=8 and α=32. For the DPO and our method, we adopt a two-stage training curriculum: we first perform IFT activation training for 2 epochs to warm up the model's basic instruction-following capabilities, followed by 3 epochs of training to refine distribution alignment. For the KTO loss, λ_D and λ_U are set to 1.0 and β is set to 0.1.

Model	Method	Gender (UK)		Gender (US)		Race	Sentiment	Avg.
		Real	Even	Real	Even	Even	Even	
Qwen2.5-7B-Instruct	Zero-shot	0.132	0.308	0.130	0.306	0.319	0.405	0.267
	PE-Explicit	0.209	0.331	0.231	0.409	0.240	0.444	0.311
	PE-Implicit	0.260	-	0.195	-	-	-	0.228
	IFT	0.144	0.080	0.247	**0.051**	0.123	0.162	0.135
	DPO	0.193	0.500	0.177	0.500	0.264	0.364	0.333
	Ours	**0.093**	**0.046**	**0.086**	0.061	**0.111**	**0.114**	**0.085**
Qwen2.5-1.5B-Instruct	Zero-shot	0.176	0.252	0.159	0.300	0.255	0.287	0.238
	PE-Explicit	0.127	0.324	**0.151**	0.220	0.216	0.443	0.247
	PE-Implicit	0.355	-	0.178	-	-	-	0.267
	IFT	0.122	0.077	0.153	0.080	0.078	0.099	0.102
	DPO	0.215	0.500	0.242	0.500	0.175	0.270	0.317
	Ours	**0.084**	**0.048**	0.158	**0.054**	**0.072**	**0.075**	**0.082**
Llama-3.1-8B-Instruct	Zero-shot	0.146	0.342	0.131	0.356	0.196	**0.159**	0.222
	PE-Explicit	0.220	0.203	0.229	0.237	0.177	0.435	0.250
	PE-Implicit	0.284	-	0.212	-	-	-	0.248
	IFT	0.129	**0.052**	0.147	**0.049**	0.172	0.236	0.131
	DPO	0.178	0.500	0.147	0.500	0.227	0.394	0.324
	Ours	**0.114**	0.076	**0.108**	0.091	**0.108**	0.199	**0.116**
Llama-3.2-1B-Instruct	Zero-shot	0.330	0.320	0.305	0.269	0.185	0.269	0.280
	PE-Explicit	0.390	0.368	0.310	0.338	0.168	0.341	0.319
	PE-Implicit	0.377	-	0.325	-	-	-	0.351
	IFT	0.119	**0.072**	0.237	0.086	0.105	0.231	0.142
	DPO	0.147	0.500	0.176	0.500	0.203	0.314	0.307
	Ours	**0.099**	0.101	**0.068**	**0.075**	**0.093**	**0.182**	**0.103**

Table 1: Main results measured by MAE across four models and six datasets (lower is better). The best performance is highlighted in bold. "Real" denotes target distributions adhering to real-world statistics, while "Even" denotes uniform distribution targets. "Avg." denotes average MAE over six datasets. Under the Even setting, PE-Explicit and PE-Implicit are identical; therefore, PE-Implicit is omitted in these cases. See Appendix A.2 for additional statistical analyses, including confidence intervals and standard deviations across occupations.

5 Results

5.1 Main Results

Table 1 presents the main results across four LLMs and six dataset settings. Firstly, the zero-shot results indicate that LLMs neither adhere to real-world statistics nor to uniform attribute distributions, and this is consistent across model sizes. Secondly, approaches relying solely on prompt engineering prove inadequate for this task: both PE-Explicit and PE-Implicit generally fail to improve distribution alignment. Among the supervised baselines, IFT emerges as the second-best approach, significantly reducing MAE compared to zero-shot inference. More strikingly, DPO, a standard method for alignment, fails catastrophically in the distribution debiasing context. This confirms that standard preference optimization techniques are mode-seeking and force the model to converge on a single best response, which is at odds with the goal of aligning an entire distribution. Our approach consistently achieves the lowest MAE overall (Avg.), and across nearly all individual settings. For instance, on the Qwen2.5-7B model, we reduce the average MAE from 0.135 (IFT) to 0.085,

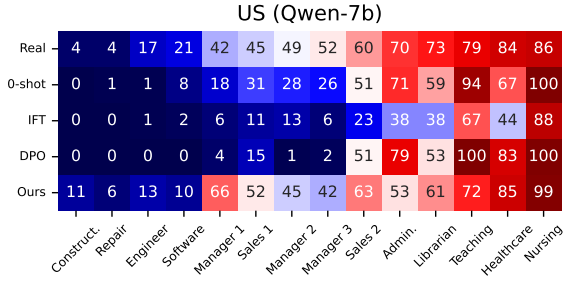


Figure 3: The representation of females in [0, 100] for the 14 considered occupations in the US. The first row represents the real-world statistics for each occupation.

representing a 37% relative improvement. Similar trends are observed on Llama-3.2-1B, where our method outperforms IFT by approximately 27%. These results validate that our dual-loss framework offers a robust and stable solution for manipulating model output distributions regardless of the underlying model architecture.

Table 1 further indicates that the models' predictions for the gender linked to each occupation are closer to the real-world distributions than to an even distribution. To further investigate this, we compare real-world female representation in 14

Model & Method		Gen (UK)		Gen (US)		Senti	Avg.
		Real	Even	Real	Even	Even	
Qwen7B	Zero	0.26	0.34	0.26	0.38	0.42	0.33
	Exp	0.26	0.36	0.26	**0.35**	0.33	0.31
	Imp	0.44	-	0.43	-	-	0.44
	IFT	0.27	0.34	**0.25**	0.37	0.42	0.33
	DPO	**0.23**	**0.31**	0.26	0.36	0.42	0.32
	Ours	0.27	0.34	0.26	0.38	**0.32**	**0.31**
Qwen1.5B	Zero	0.27	0.32	0.23	0.32	0.30	0.29
	Exp	0.32	0.34	0.30	0.33	0.26	0.31
	Imp	0.40	-	0.37	-	-	0.39
	IFT	0.27	0.32	0.22	0.31	0.28	0.28
	DPO	**0.26**	**0.30**	**0.21**	**0.28**	0.29	0.27
	Ours	0.28	0.32	0.28	0.31	**0.13**	**0.26**
Llama8B	Zero	0.28	0.36	**0.23**	0.38	0.35	0.32
	Exp	0.32	0.35	0.30	0.35	0.25	0.31
	Imp	0.46	-	0.45	-	-	0.46
	IFT	0.27	0.36	0.25	0.37	0.35	0.32
	DPO	0.28	**0.35**	0.27	**0.35**	0.35	0.32
	Ours	**0.25**	0.36	0.25	0.37	**0.22**	**0.29**
Llama1B	Zero	**0.21**	0.32	0.20	0.33	0.33	0.28
	Exp	0.24	**0.25**	0.27	**0.26**	0.28	0.26
	Imp	0.31	-	0.28	-	-	0.30
	IFT	0.26	0.32	0.20	0.34	0.30	0.28
	DPO	0.26	0.33	**0.20**	0.32	0.31	0.28
	Ours	0.22	0.31	0.26	0.32	**0.19**	**0.26**

Table 2: Story generation results measured by MAE across four models and five datasets. The best performance is highlighted in bold.

occupations (US) with the distribution predicted by the different models (Figure 3). We find that the zero-shot models tend to amplify the skew in real-world distributions, which is ameliorated by our method. For instance, the proportion of Female predictions for *Software developers* is 21% in real world in the US, indicating a representation bias toward Males. However, zero-shot, IFT, and DPO categorically assign less proportions of 8%, 2%, and 0% to Females for this occupation; a sign of bias amplification. More case study analyses are provided in Appendix A.3.3.

5.2 Extension to Story Generation

To evaluate the generalization capability of our framework in more complex, real-world scenarios, we extend our experiments to a story generation task. Instead of explicitly requesting an attribute, we prompt the model to write a "day in the life" story based on a given occupation. Due to the linguistic complexity of narratives where attributes (e.g., gender or sentiment) are often implicit, simple keyword matching is insufficient. We therefore employ a strong LLM judge (Qwen3-30B) to classify the attributes of the generated stories. Table 2 presents the results of this evaluation. We exclude the Race attribute, as incorporating race into storytelling is inappropriate. We observe a divergence in performance across different attributes. In the case of Gender, controlling the distribution in long-form storytelling proves challenging for all methods. Our approach performs comparably to baselines (e.g., IFT and DPO), with no single method achieving dominant debiasing results. This suggests that as generation length increases, the model's internal priors regarding gender-occupation correlations become harder to override. However, our method demonstrates superior performance in **Sentiment** steering. More specifically, our method achieves the lowest average MAE across all four models, highlighting its potential for controlling stylistic attributes even in complex generation tasks.

5.3 Ablation Study

To investigate the individual contributions of our loss components, we conduct an ablation study by training the Qwen models with either the KL divergence loss ($\mathcal{L}_{KL}$) or the KTO loss ($\mathcal{L}_{KTO}$) removed. Table 3 presents the results. We observe that our framework consistently achieves the best performance across the majority of datasets, validating the synergy between token calibration and semantic alignment. Notably, removing $\mathcal{L}_{KL}$ generally leads to the highest MAE, which confirms that the steering token calibration is the foundational mechanism for distribution matching. In comparison, the model without $\mathcal{L}_{KTO}$ ("w/o KTO") performs better than the model without $\mathcal{L}_{KL}$, but still suffers significant degradation compared to the full method. This suggests that while KL sets the statistical target, the KTO preference loss is essential for reinforcing the connection between the token and the response. Furthermore, substituting $\mathcal{L}_{KTO}$ with the standard DPO loss, yields unstable results (e.g., reaching 0.50 MAE in some tasks, implying the model converges to a single attribute). This instability likely stems from DPO's strict pairwise margin maximization, which can be overly aggressive and disrupt the delicate probability calibration established by the KL term. In contrast, KTO decomposes the objective into independent desirable and undesirable value functions. This allows the model to effectively bind the semantic content to the steering token without overriding the distributional constraints.

Model & Method		Gen (UK)		Gen (US)		Race	Senti
		Real	Even	Real	Even	Even	Even
Qwen7B	w/o KL	0.19	0.50	0.28	0.08	0.15	0.36
	w/o KTO	0.12	**0.04**	0.28	0.07	0.14	**0.09**
	swap DPO	0.18	0.26	0.19	0.50	0.12	0.09
	Ours	**0.09**	0.05	**0.09**	**0.06**	**0.11**	0.11
Qwen1.5B	w/o KL	0.34	0.24	0.27	0.10	0.14	0.07
	w/o KTO	0.30	0.05	0.26	0.09	0.19	**0.07**
	swap DPO	0.13	0.50	0.22	0.10	0.14	0.12
	Ours	**0.08**	**0.05**	**0.16**	**0.05**	**0.07**	0.08

Table 3: Ablation study on Qwen models measured by MAE. "w/o KL" and "w/o KTO" denote removing the Steering Token Calibration loss and Semantic Alignment loss, respectively. "swap DPO" denote swapping the KTO loss to DPO loss.

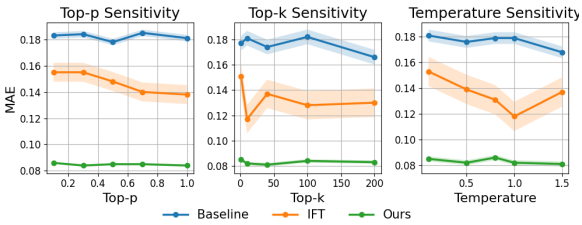
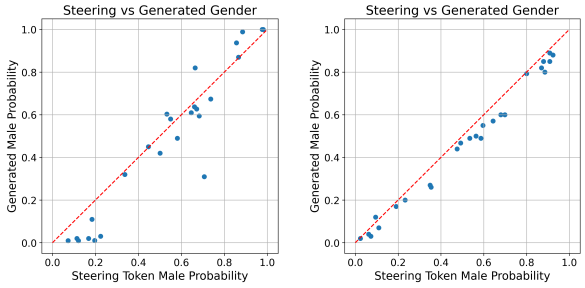


Figure 4: Sensitivity of Top-p, Top-k, and Temperature on Qwen2.5-1.5B with UK Gender Real dataset. Y-axis: MAE; shaded areas: std. dev. over 5 runs.

Model & Method		Gen (UK)		Gen (US)		Avg.
		Real	Even	Real	Even	
Qwen7B	Zero	0.25	**0.11**	0.30	0.26	0.23
	IFT	0.15	0.19	0.12	0.12	0.14
	Ours	**0.09**	0.14	**0.05**	**0.06**	**0.08**
Qwen1.5B	Zero	0.12	0.14	0.29	0.30	0.21
	IFT	0.13	0.24	0.06	**0.04**	0.12
	Ours	**0.09**	**0.10**	**0.06**	0.06	**0.08**

Table 4: Internal alignment measured by MAE based on the Softmax probabilities of gender tokens (Male/Female). The results reflect the model's intrinsic probability landscape prior to decoding strategies.



(a) Qwen2.5-1.5B (b) Qwen2.5-7B

Figure 5: Steering token versus generated Male gender probability using the data with real-world UK distribution. The dashed diagonal represents perfect alignment.

6 Discussion and Analysis

Influence of Generation Parameters To assess the robustness of our framework against decoding hyperparameter variations, we conduct a sensitivity analysis on three key generation parameters: Top-p, Top-k, and Temperature, measuring the MAE across a wide range of values. The results are visualized in Figure 4. As observed, our method (represented by the green line) exhibits remarkable stability and consistently achieves the lowest MAE across all parameter settings. More results are provided in Appendix A.3.1.

Internal Model Logits Behavior To verify whether our training objective successfully reshapes the model's fundamental belief prior rather than merely optimizing for a specific decoding strategy, we analyze the internal logit distributions. Specifically, we extract the probabilities of the attribute tokens (e.g., "Male" and "Female") directly from the model's Softmax layer at the first generation step, before any sampling (e.g., Top-p or Top-k) is applied. Table 4 reports the MAE calculated between these internal probabilities and the target distributions. We observe a consistency between the internal probability landscape and the final sampled outputs presented in previous sections. Ours consistently achieves the lowest MAE across the averaged metrics, significantly outperforming the Zero-shot baseline and IFT. More results are provided in Appendix A.3.2.

Correlation of Steering Token and Response A core premise of our framework is that the steering token τ, calibrated via $\mathcal{L}_{KL}$, effectively dictates the semantic attribute of the subsequent text response. To verify this link, we analyze the correlation between the model's predicted probability of the "Male" steering token and the actual frequency of "Male" responses generated during inference. From Figure 5, we observe that the data points cluster tightly around the diagonal line, demonstrating a near-linear correlation. This alignment indicates that the model has learned to condition its generation on the steering token: when the model assigns a high probability to the Male token, it almost generates a Male attribute, and vice versa.

Influence on General Capabilities To assess whether our distribution debiasing framework compromises the models' general utility, we evaluate

Model & Method		MMLU	GSM8K	TruthfulQA	IFEval
Qwen7B	Zero	26.91	81.35	59.36	67.62
	IFT	27.57	81.73	57.41	66.31
	DPO	27.57	81.65	58.38	66.31
	Ours	26.87	81.12	60.00	65.23
Qwen1.5B	Zero	28.81	62.32	54.00	42.93
	IFT	27.80	61.64	55.51	39.21
	DPO	28.00	61.87	57.00	41.25
	Ours	27.46	62.33	56.00	42.21

Table 5: Evaluation of model capabilities across general domains using the fine-tuned model from UK Real dataset. Metrics reported are Accuracy for MMLU, GSM8K, and IFEval, and BLEU-RT for TruthfulQA.

Method	UK Real	UK Even	US Real	US Even
Ours (0/1)	0.084	0.048	0.158	0.054
Token (man/woman)	0.088	0.041	0.168	0.047
Token (he/she)	0.078	0.049	0.136	0.050

Table 6: Comparison of methods on UK and US datasets under Real and Even settings using Qwen2.5-1.5B.

them on four diverse benchmarks: MMLU (general knowledge), GSM8K (mathematical reasoning), TruthfulQA (truthfulness), and IFEval (instruction following). Table 5 shows that all evaluated methods incur negligible impact on general capabilities compared to zero-shot settings.

Choice of Steering Tokens In our main experiments (as illustrated in Table 1), we utilized abstract numerical tokens (e.g., "0" and "1") to serve as steering vectors. However, since our framework treats steering tokens as learnable embeddings, their specific surface form should be arbitrary. To empirically verify this, we conducted an ablation study comparing our original numerical tokens against semantically explicit tokens: man/woman and he/she. As shown in the table below, the performance differences across these choices are negligible and fall within the range of standard experimental noise.

Generalized to Unseen Occupations The evaluation is conducted only on occupations seen during training, the improved alignment may simply reflect memorization of distributions for a limited set of occupations. Therefore, we conduct a cross-setting experiment where the model was trained exclusively on the UK dataset and subsequently evaluated on the US dataset. Since the set of occupations differs significantly between these contexts, the US data effectively serves as an unseen test set. As shown in the table below, our method continues

Method	Model	Real	Even
IFT	Qwen2.5-7B	0.158	0.189
	Qwen2.5-1.5B	0.187	0.154
	Llama-3.1-8B	0.109	0.101
	Llama-3.2-1B	0.147	0.085
DPO	Qwen2.5-7B	0.201	0.500
	Qwen2.5-1.5B	0.265	0.500
	Llama-3.1-8B	0.205	0.500
	Llama-3.2-1B	0.178	0.500
Ours	Qwen2.5-7B	0.078	0.050
	Qwen2.5-1.5B	0.095	0.061
	Llama-3.1-8B	0.084	0.134
	Llama-3.2-1B	0.088	0.120

Table 7: Comparison of IFT, DPO, and our method on UK datasets under Real and Even settings across different models.

to significantly outperform baselines on these unseen occupations except Llama model on even distributions, achieving alignment performance consistent with the seen results. This demonstrates that our approach generalizes to new occupations rather than simply overfitting to the training set.

Benefit and Risk of Aligning Model Outputs Aligning model outputs to real-world distributions involves a complex trade-off between practical utility and the risk of perpetuating societal harm. On one hand, real-world distributions contain factual information that is essential for applications where accurate statistics matter, such as medical diagnostics or other safety-critical tasks. On the other hand, directly mirroring real-world distributions can reinforce existing societal inequities and disproportionately disadvantage minority groups, such as in gender–occupation associations. Our paper proposes a method to steer the output distribution toward a desired target distribution, but selecting this target must be application-dependent and requires careful consideration of potential ethical risks.

7 Conclusion

We address the challenge of controlling output distributions in multi-round LLM generation. By integrating Steering Token Calibration with Semantic Alignment, our proposed framework effectively steers model statistics to match target distributions across gender, race, and sentiment. Empirical results confirm that our method significantly outperforms existing alignment techniques, offering a robust solution for probabilistic generation without compromising general capabilities.

Limitations

Despite the efficacy of our framework, several limitations remain. First, while our method excels in controlling explicit attributes in short-form generation, maintaining precise distribution alignment in complex, long-form tasks, such as implicitly steering gender in narrative story generation, remains challenging due to stronger internal priors in pre-trained models. Second, our current evaluation focuses on specific demographic and sentiment attributes within occupational contexts, generalizing this approach to more intersectional attributes requires further investigation. Finally, our reliance on introducing explicit steering tokens necessitates access to the model's vocabulary and weights, potentially limiting applicability in closed-source environments.

References

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Karin De Langis, Ryan Koo, and Dongyeop Kang. 2024. Dynamic multi-reward weighting for multi-style controllable generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6783–6800, Miami, Florida, USA. Association for Computational Linguistics.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Angela Fan and Claire Gardent. 2022. Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

Zhiting Fan, Ruizhe Chen, and Zuozhu Liu. 2025. BiasGuard: A reasoning-enhanced bias detection tool for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9753–9764, Vienna, Austria. Association for Computational Linguistics.

Biaoyan Fang, Ritvik Dinesh, Xiang Dai, and Sarvnaz Karimi. 2024. Born differently makes a difference: Counterfactual study of bias in biography generation from a data-to-text perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 409–424, Bangkok, Thailand. Association for Computational Linguistics.

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

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

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

Dong Huang, Jie M Zhang, Qingwen Bu, Xiaofei Xie, Junjie Chen, and Heming Cui. 2024. Bias testing and mitigation in llm-based code generation. *ACM Transactions on Software Engineering and Methodology*.

Yanbei Jiang, Yihao Ding, Chao Lei, Jiayang Ao, Jey Han Lau, and Krista A Ehinger. 2025a. Beyond perception: Evaluating abstract visual reasoning through multi-stage task. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13–45.

Yanbei Jiang, Chao Lei, Yihao Ding, Krista Ehinger, and Jey Han Lau. 2025b. Propa: Toward process-level optimization in visual reasoning via reinforcement learning. *arXiv preprint arXiv:2511.10279*.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Prompt-based length controlled generation with multiple control types. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1067–1085, Bangkok, Thailand. Association for Computational Linguistics.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. 2024. Controlled text generation for large language model with dynamic attribute graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5797–5814, Bangkok, Thailand. Association for Computational Linguistics.

Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Singapore. Association for Computational Linguistics.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295, Torino, Italia. ELRA and ICCL.

Michela Lorandi and Anya Belz. 2023. How to control sentiment in text generation: A survey of the state-of-the-art in sentiment-control techniques. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 341–353, Toronto, Canada. Association for Computational Linguistics.

Huiyu Mai, Wenhao Jiang, and Zhi-Hong Deng. 2023. Prefix-tuning based unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14847–14856, Singapore. Association for Computational Linguistics.

Takuto Miura, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2025. Style-controlled response generation for dialog systems with intimacy interpretation. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 50–59, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2025. https://openai.com/index/gpt-5-1/.

Birong Pan, Yongqi Li, Weiyu Zhang, Wenpeng Lu, Mayi Xu, Shen Zhou, Yuanyuan Zhu, Ming Zhong, and Tieyun Qian. 2025. A survey on training-free alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4445–4461, Suzhou, China. Association for Computational Linguistics.

Kaiyi Pang, Minhao Bai, Jinshuai Yang, Yue Gao, Minghu Jiang, and Yongfeng Huang. 2025. A plug-and-play method for linguistic alignment in language models. *Knowledge-Based Systems*, page 113597.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased tales: Cultural and topic bias in generating children's stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 52–72, Suzhou, China. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Ingroj Shrestha and Padmini Srinivasan. 2025. LLM bias detection and mitigation through the lens of desired distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1480, Suzhou, China. Association for Computational Linguistics.

Velizar Shulev and Khalil Sima'an. 2024. Continual reinforcement learning for controlled text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pages 3881–3889, Torino, Italia. ELRA and ICCL.

Shweta Soundararajan and Sarah Jane Delany. 2024. Investigating gender bias in large language models through text generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 410–424, Trento. Association for Computational Linguistics.

UK Occupation Statistics. 2023. https://assets.publishing.service.gov.uk/media/5a7f3952ed915d74e622924b/Working_Futures_Headline_Report_final_for_web__PG.pdf.

US Occupation Statistics. 2025. https://www.bls.gov/cps/cpsaat39.htm.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe, editors. 2024. *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*. Association for Computational Linguistics, St Julians, Malta.

Yixin Wan and Kai-Wei Chang. 2025. White men lead, black women help? benchmarking and mitigating language agency social biases in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9082–9108, Vienna, Austria. Association for Computational Linguistics.

Xiaoyuan Wu, Weiran Lin, Omer Akgul, and Lujo Bauer. 2025. Estimating LLM consistency: A user baseline vs surrogate metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30530–30544, Suzhou, China. Association for Computational Linguistics.

Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. 2024. Aligning as debiasing: Causality-aware alignment via reinforcement learning with interventional feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4684–4695, Mexico City, Mexico. Association for Computational Linguistics.

Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2024. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3343–3353, Bangkok, Thailand. Association for Computational Linguistics.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2025. No preference left behind: Group distributional preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Mahdi Zakizadeh and Mohammad Taher Pilehvar. 2025. Gender encoding patterns in pretrained language model representations. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 489–500, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix

A.1 Datasets

A.1.1 Datasets Overview

We evaluate our framework across six distinct dataset splits derived from three attribute groups (Gender, Race, Sentiment) and two geographical contexts (UK, US). Table 8 details the statistics for each split. For every occupation in our list (25 for UK, 14 for US), we generate $N = 100$ distinct prompts. For Race and Sentiment, we use the US occupations. The datasets are partitioned into Training, Validation, and Test sets to ensure robust evaluation.

Dataset Split	Attribute	Train	Val	Test
UK Occupations (Real)	Gender	2500	300	2500
UK Occupations (Even)	Gender	2500	300	2500
US Occupations (Real)	Gender	1400	300	1400
US Occupations (Even)	Gender	1400	300	1400
US Occupations (Even)	Race	1400	300	1400
US Occupations (Even)	Sentiment	1400	300	1400

Table 8: Statistics of the six dataset splits used in our experiments. "Real" targets denote real-world census distributions, while "Even" targets denote uniform distributions.

A.1.2 Prompts

We assess the model's performance in two distinct generation settings for each of the six dataset splits, resulting in a total of 12 experimental scenarios.

Attribute	Task Setting	Representative Prompt Template
Gender	Attribute Generation	What is the gender of the [OCCUPATION]?
	Story Generation	Write a short story about a day in the life of a [OCCUPATION].
Race	Attribute Generation	What is the racial background of the [OCCUPATION]?
	Story Generation	Describe a scene featuring a [OCCUPATION] at work.
Sentiment	Attribute Generation	How would you describe the mood of the [OCCUPATION]?
	Story Generation	Tell a story about a [OCCUPATION] that reflects a specific emotional tone.

Table 9: Representative prompt templates for the 12 experimental settings (covering 6 dataset splits $\times$ 2 tasks). '[OCCUPATION]' is replaced by the specific role from our US or UK occupation lists. Actual inputs include paraphrased variations of these templates.

(1) **Attribute Generation:** The model is explicitly prompted to identify or assign an attribute to an occupation (e.g., "What is the gender of..."). This setting tests the direct probability distribution of the attribute. (2) **Story Generation:** The model is prompted to write a narrative or describe a scenario involving the occupation (e.g., "Write a story about..."). This setting evaluates whether the debiased distribution persists in open-ended, real-world generation contexts.

Table 9 provides the representative prompt templates used for each attribute group. To ensure linguistic diversity and prevent the model from overfitting to a specific sentence structure, we utilized GPT-5.1 to rephrase these templates into 100 variations for each occupation.

A.2 More Results

A.2.1 Standard Deviation across Occupations

To assess the stability of our method across different contexts, we analyze the variance of the model's performance relative to the specific occupation being prompted. Table 10 reports the MAE accompanied by the standard deviation (SD) calculated across the 25 UK occupations and 14 US occupations.

A.2.2 Confidence Intervals via Bootstrapping

To verify the statistical reliability of our observed metrics and ensure that the performance gains are not artifacts of sampling variance, we calculate the 95% Confidence Intervals (CI) for the MAE scores. We employ non-parametric bootstrapping with 1,000 resamples on the $N = 100$ test responses for each occupation. Table 11 presents the results. Narrow confidence intervals indicate high precision in our estimates. Crucially, we observe that for most dataset splits, the confidence intervals of our method do not overlap with those of the baselines (especially IFT and Zero-shot). This separation confirms that our method's superiority is statistically significant and robust to sampling variations, rather than a result of stochastic generation noise.

A.3 Additional Analysis

A.3.1 Influence of Generation Parameters

Figures 6–10 present the parameter sensitivity analyses on the remaining dataset splits. With the exception of the US Gender Real dataset, all datasets exhibit similar trends: our method (shown by the green curve) demonstrates strong robustness and consistently achieves the lowest MAE across all generation parameter settings.

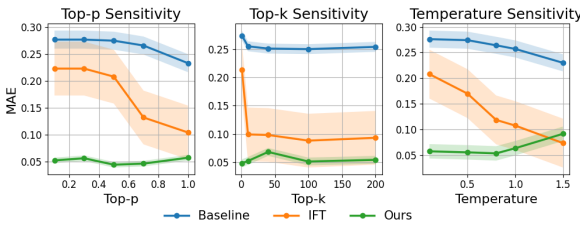


Figure 6: Sensitivity analysis of Top-p, Top-k, and Temperature on the Qwen2.5-1.5B model and the UK Gender Even Datasets.

A.3.2 Internal Model Logits Behavior

Table 12 reports the MAE between the internal probability distributions and the target distributions for two additional models. Our method consistently achieves the lowest MAE on the averaged metrics, substantially outperforming both the zero-shot baseline and IFT.

Model	Method	Gender (UK)		Gender (US)		Race	Sentiment
		Real	Even	Real	Even	Even	Even
Qwen2.5-7B	Zero-shot	$0.132_{\pm 0.109}$	$0.308_{\pm 0.151}$	$0.130_{\pm 0.070}$	$0.306_{\pm 0.163}$	$0.319_{\pm 0.002}$	$0.405_{\pm 0.029}$
	PE-Explicit	$0.209_{\pm 0.133}$	$0.331_{\pm 0.156}$	$0.231_{\pm 0.118}$	$0.409_{\pm 0.126}$	$0.240_{\pm 0.000}$	$0.444_{\pm 0.000}$
	PE-Implicit	$0.260_{\pm 0.196}$	-	$0.195_{\pm 0.135}$	-	-	-
	IFT	$0.144_{\pm 0.098}$	$0.080_{\pm 0.052}$	$0.247_{\pm 0.144}$	$\mathbf{0.051}_{\pm 0.143}$	$0.123_{\pm 0.030}$	$0.162_{\pm 0.046}$
	DPO	$0.193_{\pm 0.139}$	$0.500_{\pm 0.000}$	$0.177_{\pm 0.141}$	$0.500_{\pm 0.000}$	$0.264_{\pm 0.022}$	$0.364_{\pm 0.036}$
	Ours	$\mathbf{0.093}_{\pm 0.092}$	$\mathbf{0.046}_{\pm 0.030}$	$\mathbf{0.086}_{\pm 0.061}$	$0.061_{\pm 0.037}$	$\mathbf{0.111}_{\pm 0.022}$	$\mathbf{0.114}_{\pm 0.013}$
Qwen2.5-1.5B	Zero-shot	$0.176_{\pm 0.118}$	$0.252_{\pm 0.146}$	$0.159_{\pm 0.108}$	$0.300_{\pm 0.128}$	$0.255_{\pm 0.022}$	$0.287_{\pm 0.059}$
	PE-Explicit	$0.127_{\pm 0.108}$	$0.324_{\pm 0.131}$	$\mathbf{0.151}_{\pm 0.096}$	$0.220_{\pm 0.145}$	$0.216_{\pm 0.011}$	$0.443_{\pm 0.003}$
	PE-Implicit	$0.355_{\pm 0.195}$	-	$0.178_{\pm 0.109}$	-	-	-
	IFT	$0.122_{\pm 0.086}$	$0.077_{\pm 0.046}$	$0.153_{\pm 0.124}$	$0.080_{\pm 0.053}$	$0.078_{\pm 0.026}$	$0.099_{\pm 0.049}$
	DPO	$0.215_{\pm 0.146}$	$0.500_{\pm 0.000}$	$0.242_{\pm 0.177}$	$0.500_{\pm 0.000}$	$0.175_{\pm 0.024}$	$0.270_{\pm 0.048}$
	Ours	$\mathbf{0.084}_{\pm 0.081}$	$\mathbf{0.048}_{\pm 0.031}$	$0.158_{\pm 0.113}$	$\mathbf{0.054}_{\pm 0.037}$	$\mathbf{0.072}_{\pm 0.014}$	$\mathbf{0.075}_{\pm 0.032}$
Llama-3.1-8B	Zero-shot	$0.146_{\pm 0.144}$	$0.342_{\pm 0.120}$	$0.131_{\pm 0.081}$	$0.356_{\pm 0.121}$	$0.196_{\pm 0.035}$	$\mathbf{0.159}_{\pm 0.073}$
	PE-Explicit	$0.220_{\pm 0.139}$	$0.203_{\pm 0.122}$	$0.229_{\pm 0.144}$	$0.237_{\pm 0.109}$	$0.177_{\pm 0.046}$	$0.435_{\pm 0.008}$
	PE-Implicit	$0.284_{\pm 0.176}$	-	$0.212_{\pm 0.085}$	-	-	-
	IFT	$0.129_{\pm 0.072}$	$\mathbf{0.052}_{\pm 0.030}$	$0.147_{\pm 0.107}$	$\mathbf{0.049}_{\pm 0.035}$	$0.172_{\pm 0.024}$	$0.236_{\pm 0.039}$
	DPO	$0.178_{\pm 0.133}$	$0.500_{\pm 0.000}$	$0.147_{\pm 0.112}$	$0.500_{\pm 0.000}$	$0.227_{\pm 0.025}$	$0.394_{\pm 0.012}$
	Ours	$\mathbf{0.114}_{\pm 0.087}$	$0.076_{\pm 0.071}$	$\mathbf{0.108}_{\pm 0.062}$	$0.091_{\pm 0.070}$	$\mathbf{0.108}_{\pm 0.022}$	$0.199_{\pm 0.019}$
Llama-3.2-1B	Zero-shot	$0.330_{\pm 0.215}$	$0.320_{\pm 0.081}$	$0.305_{\pm 0.164}$	$0.269_{\pm 0.074}$	$0.185_{\pm 0.023}$	$0.269_{\pm 0.043}$
	PE-Explicit	$0.390_{\pm 0.262}$	$0.368_{\pm 0.046}$	$0.310_{\pm 0.184}$	$0.338_{\pm 0.061}$	$0.168_{\pm 0.017}$	$0.341_{\pm 0.039}$
	PE-Implicit	$0.377_{\pm 0.235}$	-	$0.325_{\pm 0.187}$	-	-	-
	IFT	$0.119_{\pm 0.081}$	$\mathbf{0.072}_{\pm 0.052}$	$0.237_{\pm 0.103}$	$0.086_{\pm 0.055}$	$0.105_{\pm 0.027}$	$0.231_{\pm 0.036}$
	DPO	$0.147_{\pm 0.093}$	$0.500_{\pm 0.000}$	$0.176_{\pm 0.118}$	$0.500_{\pm 0.000}$	$0.203_{\pm 0.025}$	$0.314_{\pm 0.032}$
	Ours	$\mathbf{0.099}_{\pm 0.065}$	$0.101_{\pm 0.079}$	$\mathbf{0.068}_{\pm 0.046}$	$\mathbf{0.075}_{\pm 0.061}$	$\mathbf{0.093}_{\pm 0.022}$	$\mathbf{0.182}_{\pm 0.038}$

Table 10: Distribution alignment performance measured by MAE and its Standard Deviation (SD) across occupations ($Mean_{\pm SD}$). Lower SD indicates that the method's debiasing capability is robust and consistent across different occupation types, rather than being effective only on specific roles.

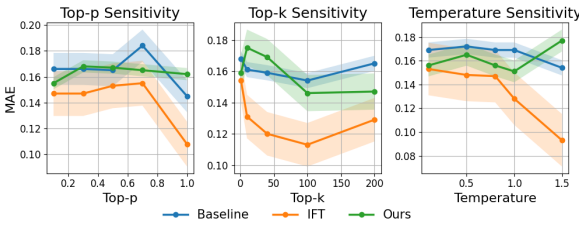


Figure 7: Sensitivity analysis of Top-p, Top-k, and Temperature on the Qwen2.5-1.5B model and the US Gender Real Datasets.

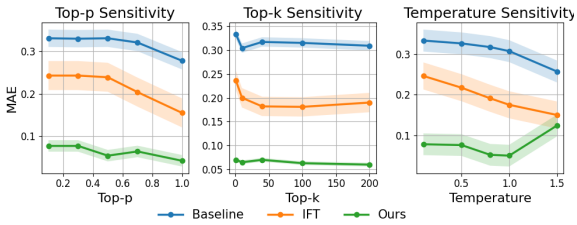


Figure 8: Sensitivity analysis of Top-p, Top-k, and Temperature on the Qwen2.5-1.5B model and the US Gender Even Datasets.

A.3.3 Detailed Representation Percentages for the LLama-8B and Qwen-7B

Figure 11 compares real-world female representation in the different occupations of the UK and the US, with the distribution predicted by the different models. This complements Figure 3.

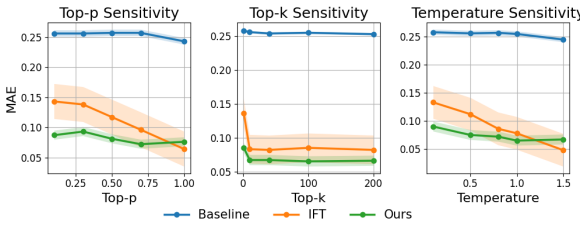


Figure 9: Sensitivity analysis of Top-p, Top-k, and Temperature on the Qwen2.5-1.5B model and the Race Datasets.

Model	Method	Gender (UK)		Gender (US)		Race	Sentiment
		Real	Even	Real	Even	Even	Even
Qwen2.5-7B	Zero-shot	$0.132_{[0.12,0.15]}$	$0.308_{[0.29,031]}$	$0.130_{[0.13,0.15]}$	$0.306_{[0.29,0.32]}$	$0.319_{[0.31,0.32]}$	$0.405_{[0.40,0.42]}$
	PE-Explicit	$0.209_{[0.21,0.22]}$	$0.331_{[0.31,0.33]}$	$0.231_{[0.23,0.24]}$	$0.409_{[0.40,0.41]}$	$0.240_{[0.24,0.25]}$	$0.444_{[0.44,0.44]}$
	PE-Implicit	$0.260_{[0.25,0.27]}$	-	$0.195_{[0.19,0.20]}$	-	-	-
	IFT	$0.144_{[0.13,0.16]}$	$0.080_{[0.06,0.09]}$	$0.247_{[0.24,0.27]}$	$\mathbf{0.051}_{[0.04,0.08]}$	$0.123_{[0.12,0.14]}$	$0.162_{[0.15,0.18]}$
	DPO	$0.193_{[0.17,0.19]}$	$0.500_{[0.50,0.50]}$	$0.177_{[0.16,0.20]}$	$0.500_{[0.50,0.50]}$	$0.264_{[0.26,0.28]}$	$0.364_{[0.35,0.38]}$
	Ours	$\mathbf{0.093}_{[0.09,0.11]}$	$\mathbf{0.046}_{[0.04,0.08]}$	$\mathbf{0.086}_{[0.07,0.11]}$	$0.061_{[0.05,0.09]}$	$\mathbf{0.111}_{[0.11,0.12]}$	$\mathbf{0.114}_{[0.10,0.14]}$
Qwen2.5-1.5B	Zero-shot	$0.176_{[0.16,0.19]}$	$0.252_{[0.25,0.27]}$	$0.159_{[0.15,0.18]}$	$0.300_{[0.28,0.32]}$	$0.255_{[0.24,0.27]}$	$0.287_{[0.28,0.29]}$
	PE-Explicit	$0.127_{[0.11,0.15]}$	$0.324_{[0.32,0.33]}$	$\mathbf{0.151}_{[0.15,0.17]}$	$0.220_{[0.21,0.23]}$	$0.216_{[0.21,0.22]}$	$0.443_{[0.44,0.44]}$
	PE-Implicit	$0.355_{[0.35,0.37]}$	-	$0.178_{[0.17,0.18]}$	-	-	-
	IFT	$0.122_{[0.11,0.14]}$	$0.077_{[0.06,0.09]}$	$0.153_{[0.15,0.16]}$	$0.080_{[0.07,0.10]}$	$0.078_{[0.07,0.09]}$	$0.099_{[0.09,0.11]}$
	DPO	$0.215_{[0.19,0.27]}$	$0.500_{[0.5,0.5]}$	$0.242_{[0.24,0.25]}$	$0.500_{[0.5,0.5]}$	$0.175_{[0.17,0.19]}$	$0.270_{[0.25,0.28]}$
	Ours	$\mathbf{0.084}_{[0.08,0.10]}$	$\mathbf{0.048}_{[0.05,0.08]}$	$0.158_{[0.14,0.18]}$	$\mathbf{0.054}_{[0.05,0.08]}$	$\mathbf{0.072}_{[0.07,0.09]}$	$\mathbf{0.075}_{[0.07,0.10]}$
Llama-3.1-8B	Zero-shot	$0.146_{[0.13,0.16]}$	$0.342_{[0.33,0.35]}$	$0.131_{[0.13,0.14]}$	$0.356_{[0.35,0.37]}$	$0.196_{[0.18,0.21]}$	$\mathbf{0.159}_{[0.15,0.17]}$
	PE-Explicit	$0.220_{[0.22,0.23]}$	$0.203_{[0.20,0.21]}$	$0.229_{[0.22,0.24]}$	$0.237_{[0.23,0.24]}$	$0.177_{[0.17,0.18]}$	$0.435_{[0.44,0.44]}$
	PE-Implicit	$0.284_{[0.28,0.29]}$	-	$0.212_{[0.21,0.23]}$	-	-	-
	IFT	$0.129_{[0.11,0.14]}$	$\mathbf{0.052}_{[0.04,0.06]}$	$0.147_{[0.14,0.16]}$	$\mathbf{0.049}_{[0.04,0.07]}$	$0.172_{[0.15,0.19]}$	$0.236_{[0.21,0.24]}$
	DPO	$0.178_{0.17,0.19]}$	$0.500_{[0.5,0.5]}$	$0.147_{[0.14,0.17]}$	$0.500_{[0.5,0.5]}$	$0.227_{[0.26,0.29]}$	$0.394_{[0.39,0.40]}$
	Ours	$\mathbf{0.114}_{[0.10,0.13]}$	$0.076_{[0.07,0.11]}$	$\mathbf{0.108}_{[0.09,0.12]}$	$0.091_{[0.08,0.12]}$	$\mathbf{0.108}_{[0.10,0.12]}$	$0.199_{[0.19,0.21]}$
Llama-3.2-1B	Zero-shot	$0.330_{[0.31,0.34]}$	$0.320_{[0.31,0.33]}$	$0.305_{[0.30,0.32]}$	$0.269_{[0.25,0.28]}$	$0.185_{[0.18,0.19]}$	$0.269_{[0.26,0.30]}$
	PE-Explicit	$0.390_{[0.38,0.41]}$	$0.368_{[0.36,0.37]}$	$0.310_{[0.31,0.32]}$	$0.338_{[0.31,0.34]}$	$0.168_{[0.16,0.18]}$	$0.341_{[0.33,0.34]}$
	PE-Implicit	$0.377_{[0.37,0.38]}$	-	$0.325_{[0.31,0.33]}$	-	-	-
	IFT	$0.119_{[0.10,0.12]}$	$\mathbf{0.072}_{[0.06,0.08]}$	$0.237_{[0.22,0.24]}$	$0.086_{[0.08,0.09]}$	$0.105_{[0.10,0.14]}$	$0.231_{[0.21,0.23]}$
	DPO	$0.147_{[0.13,0.15]}$	$0.500_{[0.5,0.5]}$	$0.176_{[0.17,0.18]}$	$0.500_{[0.5,0.5]}$	$0.203_{[0.20,0.21]}$	$0.314_{[0.31,0.33]}$
	Ours	$\mathbf{0.099}_{[0.09,0.11]}$	$0.101_{[0.09,0.13]}$	$\mathbf{0.068}_{[0.05,0.09]}$	$\mathbf{0.075}_{[0.07,0.11]}$	$\mathbf{0.093}_{[0.09,0.11]}$	$\mathbf{0.182}_{[0.18,0.20]}$

Table 11: Distribution alignment performance with 95% Confidence Intervals ($Mean_{\pm CI}$). CIs are calculated via bootstrapping ($k = 1000$) on the test set. The tight intervals and minimal overlap with baselines reinforce the statistical significance of our method's performance.

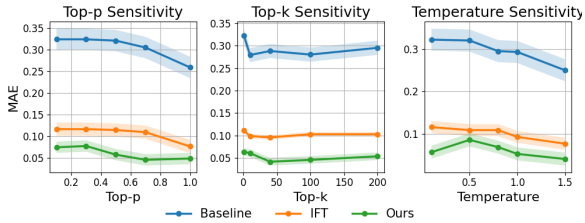


Figure 10: Sensitivity analysis of Top-p, Top-k, and Temperature on the Qwen2.5-1.5B model and the Sentiment Datasets.

Model & Method		Gen (UK)		Gen (US)		Avg.
		Real	Even	Real	Even	
Qwen7B	Zero	0.25	**0.11**	0.30	0.26	0.23
	IFT	0.15	0.19	0.12	0.12	0.14
	Ours	**0.09**	0.14	**0.05**	**0.06**	**0.08**
Qwen1.5B	Zero	0.12	0.14	0.29	0.30	0.21
	IFT	0.13	0.24	0.06	**0.04**	0.12
	Ours	**0.09**	**0.10**	**0.06**	0.06	**0.08**
Llama8B	Zero	0.28	0.30	0.23	0.26	0.27
	IFT	0.18	0.18	0.09	0.08	0.13
	Ours	**0.07**	**0.09**	**0.05**	**0.02**	**0.06**
Llama1B	Zero	0.15	0.10	0.28	0.28	0.20
	IFT	0.12	0.10	**0.05**	0.11	0.10
	Ours	**0.10**	**0.09**	0.10	**0.01**	**0.08**

Table 12: Internal alignment measured by MAE based on the Softmax probabilities of gender tokens (Male/Female). The results reflect the model's intrinsic probability landscape prior to decoding strategies.

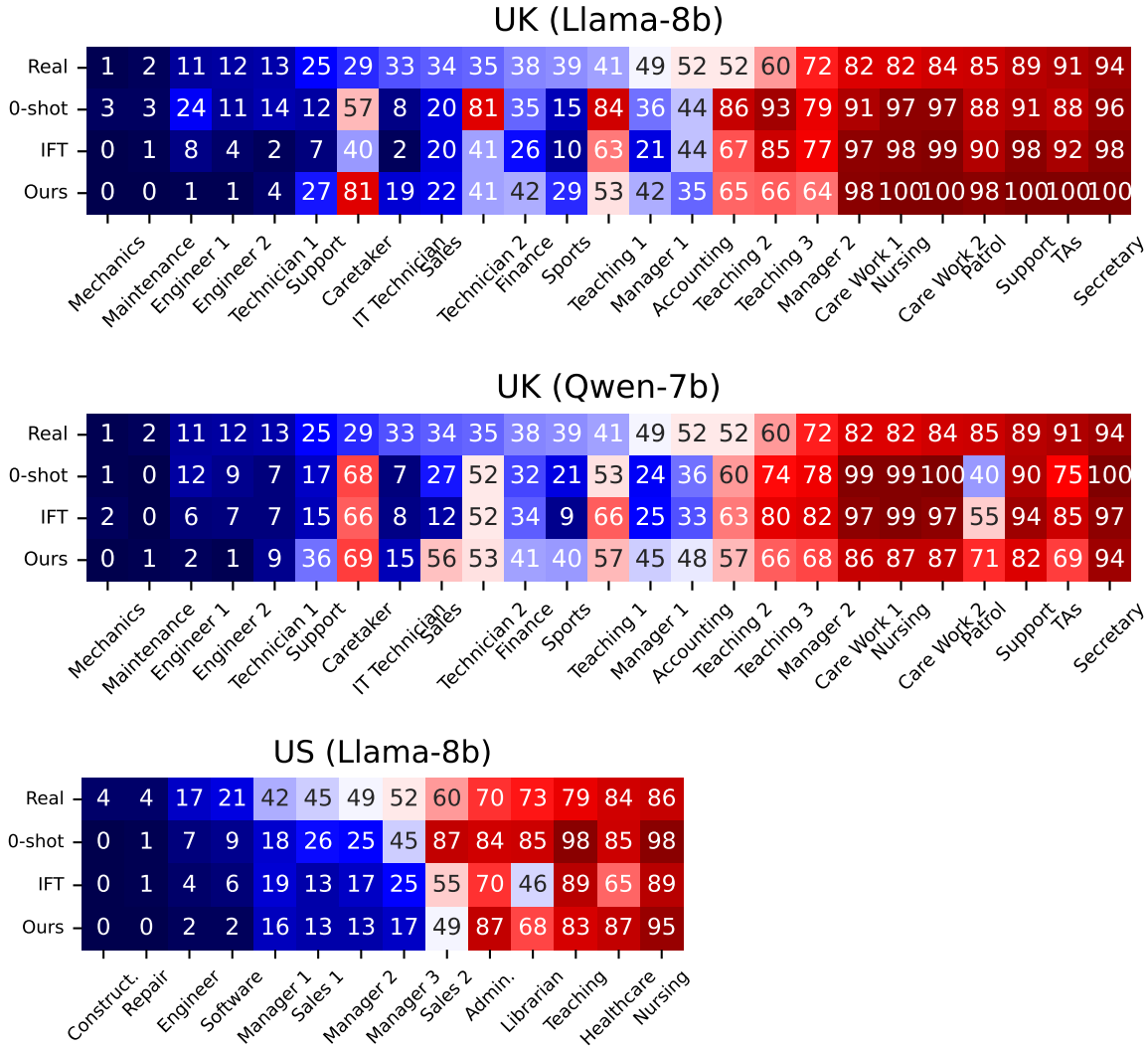


Figure 11: The representation of females in [0, 100] for the 25 occupations in the UK, and the 14 considered occupations in the US. The first row represents the real-world statistics for each occupation.