Mitigating Social Bias in English and Urdu Language Models Using PRM-Guided Candidate Selection and Sequential Refinement

Muneeb Ur Raheem Khan
Lahore University of Management Sciences (LUMS)
26100271@lums.edu.pk

Abstract

Large language models (LLMs) increasingly mediate human communication, decision support, content creation, and information retrieval. Despite impressive fluency, these systems frequently produce biased or stereotypical content, especially when prompted with socially sensitive language. A growing body of research has demonstrated that such biases disproportionately affect low-resource languages, where training data is limited and culturally unrepresentative. This paper presents a comprehensive study of *inference-time bias mitigation*—a strategy that avoids retraining or fine-tuning and instead operates directly on model outputs. Building on preference-ranking models (PRMs), we introduce a unified evaluation framework comparing three methods: (1) baseline single-word generation, (2) PRM-Select best-of-N sampling, and (3) PRM-Sequential refinement guided by PRM critiques. We evaluate these techniques across 200 English prompts and their Urdu counterparts, designed to reflect sociocultural contexts relevant to gender, ethnicity, religion, nationality, disability, profession, age, and socioeconomic categories. Using GPT-3.5 as a candidate generator and GPT-4o-mini as a PRM-based bias and utility scorer, we provide an extensive quantitative analysis of bias reduction, utility preservation, and cross-lingual disparities. Our findings show: (a) substantial gains over the baseline for both languages; (b) consistently lower fairness scores for Urdu across all methods, highlighting structural inequities in multilingual LLM training; and (c) distinct improvement trajectories between PRM-Select and PRM-Sequential. The study contributes an extensible methodology, interpretable metrics, and cross-lingual comparisons that can support future work on fairness evaluation in low-resource languages.

1 Introduction

Large language models (LLMs) such as GPT-3, GPT-4, LLaMA, PaLM, and their derivatives have transformed the landscape of computational linguistics and artificial intelligence. They now underpin chat systems, search engines, translation tools, educational platforms, and content-generation pipelines. The unprecedented scale of these models has enabled them to capture fine-grained linguistic and semantic regularities, but it has also amplified existing social and cultural biases embedded within training data. When prompted with sensitive contexts—relating to gender, ethnicity, nationality, religion, disability, or socioeconomic status—LLMs frequently generate completions that reinforce stereotypes or inadvertently convey harmful associations (2; 5; 6; 4).

The risks of biased outputs are well recognized: they can propagate discriminatory narratives, affect user trust, exacerbate harms against marginalized groups, and distort the informational environment. Most prior research on fairness in NLP has focused on English, with limited attention to multilingual or low-resource linguistic contexts. Urdu is a pertinent example: widely spoken across South Asia, Urdu remains underrepresented in high-quality corpora, resulting in weaker LLM performance, higher hallucination rates, and more frequent bias leakage. This mirrors documented disparities across many low-resource languages, including Amharic, Hausa, Pashto, Nepali, Yoruba, and others (1).

Bias mitigation methods can be grouped into three categories: (1) *pre-training interventions* such as curated datasets or filtering; (2) *fine-tuning interventions*, including instruction tuning and reinforcement learning from human feedback (RLHF); and (3) *inference-time interventions*. The first two categories require expensive retraining or fine-tuning steps, rely on large proprietary datasets, and remain inaccessible to most practitioners. Inference-time methods, by contrast, are lightweight, model-agnostic, and do not require modifying model weights. They operate solely on prompts, outputs, sampling strategies, and ranking or scoring functions.

This paper proposes a general inference-time framework for single-word bias mitigation across English and Urdu. The central idea is simple: instead of accepting the model's first answer, we generate multiple candidates, score each with a PRM-based bias and utility function, and either select the best candidate or iteratively refine the response. This yields three inference pipelines:

1. **Baseline:** one-shot single-word generation.
2. **PRM-Select:** best-of-N sampling guided by PRM scoring.
3. **PRM-Sequential:** multi-step refinement guided by PRM critiques.

We evaluate these pipelines across 200 English prompts and their Urdu counterparts, covering gender, ethnicity, nationality, religion, disability, profession, criminality, body image, and socioeconomic class. This dataset is inspired by CrowS-Pairs (8) and StereoSet (9) but adapted to single-word completions. Urdu translations were crafted to preserve semantic structure and social context.

The contributions of this paper are:

- A comprehensive inference-time debiasing framework that integrates PRM-based scoring with best-of-N sampling and sequential refinement.
- A new bilingual dataset of 200 English prompts and their Urdu translations for fairness evaluation.
- A detailed empirical analysis comparing baseline, PRM-Select, and PRM-Sequential across both languages.
- Evidence that Urdu exhibits significantly lower bias and utility scores than English across all methods, pointing to structural inequities in multilingual LLMs.
- Insights into the behavior of PRM-guided debiasing, including when sequential refinement helps or harms utility.

2 Background and Related Work

2.1 Bias in Language Models

Bias in NLP systems has been widely documented for nearly a decade. Early work in word embeddings demonstrated gender and racial analogies embedded within high-dimensional vector spaces (2), revealing that foundational representations were encoding harmful stereotypes. Subsequent research expanded these findings to contextualized models such as BERT (3), RoBERTa (**?**), and GPT-based architectures (**?**). Stereotypical associations have been shown to arise in tasks such as coreference resolution (**?** 6), sentiment analysis (**?**), natural language inference (**?**), toxic content classification (**?**), and open-ended generation (5**?**).

The emergence of large generative models has amplified these concerns. Open-ended generation can combine subtle demographic cues with broader stereotypes, producing completions that implicitly reinforce societal biases. Fairness audits have revealed systematic associations between ethnicities and crime, gender and capability, nationality and threat, religion and violence, disability and incompetence, and more (9; 8**?**). Because these models are trained on vast corpora scraped from the internet, they inherit biases from news media, fiction, social media, and user-generated content.

2.2 Bias Benchmarks

Several datasets have emerged for diagnosing and quantifying such biases:

- **CrowS-Pairs** (8): a contrastive dataset measuring whether a model prefers stereotypical or anti-stereotypical completions.
- **StereoSet** (9): evaluates language models on association tasks across gender, race, religion, and profession.
- **BBQ** (**?**): focuses on ambiguous and disambiguated question answering across demographic axes.
- **HolisticBias** (12): a large-scale dataset including dozens of demographic dimensions.

However, these benchmarks are primarily oriented toward sentence-level generation or classification. They do not directly support single-word completions paired with inference-time mitigation pipelines. Furthermore, benchmark coverage for non-English languages remains extremely sparse.

2.3 Multilingual and Low-Resource Bias

Multilingual models such as mBERT (3), XLM-R (**?**), and multilingual GPT variants often exhibit uneven performance across languages. Studies show that:

- Biases in English often transfer to other languages through shared subword vocabularies.
- Lack of culturally grounded training data for low-resource languages results in hallucinations and misalignments.
- Evaluation resources for languages like Urdu, Pashto, Nepali, Yoruba, and Swahili remain limited.

Work by Blasi et al. (1) provides a systematic analysis of linguistic inequities across languages in NLP research, noting that languages with colonial history or high economic power receive disproportionate attention and resources. As a result, the fairness of LLMs outside English remains largely understudied.

2.4 Inference-Time Bias Mitigation

Inference-time mitigation strategies avoid retraining by modifying sampling, prompting, or ranking:

- Debiasing via prompt engineering.
- Best-of-N sampling with reward models (**?** 1).
- Iterative refinement using critique models or preference model feedback (3; 2).

These techniques are appealing because they are model-agnostic and can be used with closed-source LLMs like GPT-3.5 and GPT-4. PRMs have emerged as powerful tools for scoring outputs with respect to safety, helpfulness, or harmfulness (10). This paper applies PRMs to the specific domain of *bias and utility scoring* for single-word completions in English and Urdu.

3 Methodology

This study evaluates inference-time bias mitigation across two languages—English and Urdu—using a unified pipeline designed to (1) generate single-word completions, (2) score them with a Preference Ranking Model (PRM), and (3) apply debiasing strategies based on best-of-N sampling and sequential refinement. The pipeline is intentionally model-agnostic: it treats the candidate generator as a black box, and it evaluates completions purely through a PRM-based scoring mechanism.

3.1 Research Questions

This paper addresses four key questions:

1. **RQ1:** How biased are the baseline single-word outputs of GPT-3.5 in English and Urdu?
2. **RQ2:** Can PRM-guided best-of-N sampling (PRM-Select) mitigate bias while preserving semantic utility?
3. **RQ3:** Can PRM-guided sequential refinement (PRM-Sequential) further improve fairness metrics?
4. **RQ4:** Are debiasing gains consistent across languages, or does Urdu—being a lower-resource language—show weaker improvements?

Each research question corresponds to a different part of our evaluation framework. RQ1 establishes base-rate bias in LLM outputs. RQ2 and RQ3 quantify the effectiveness of inference-time debiasing. RQ4 addresses cross-lingual fairness, which is an underexplored domain in the literature.

4 Dataset

The dataset consists of 200 English prompts and their Urdu counterparts, yielding 400 total evaluation items. Each prompt is a single sentence containing a `[blank]` placeholder requiring a single-word completion. These prompts are inspired by CrowS-Pairs (8), StereoSet (9), and HolisticBias (12). However, unlike contrastive or multi-choice datasets, our dataset is open-ended: multiple completions are possible, and the quality of a response depends not on correctness but on fairness and appropriateness.

4.1 Categories

Prompts span eight social bias categories:

1. Gender and sexuality
2. Ethnicity and nationality
3. Religion
4. Age
5. Disability
6. Socioeconomic status
7. Criminality and threat perception
8. Appearance and body-related stereotypes

Each English prompt was translated into Urdu manually, ensuring cultural alignment rather than literal equivalence. Urdu prompts replicate the pragmatic meaning, syntactic structure, and social connotations of their English counterparts. Since Urdu has gendered grammar and a rich politeness system, translations were crafted to avoid accidentally introducing or removing bias.

4.2 Why Single-Word Prompts?

Single-word completions offer three advantages:

1. **Interpretability:** The presence or absence of bias is easily identifiable.
2. **Comparability:** Different sentences can be evaluated on a shared scoring scale.
3. **Model-agnostic debiasing:** Single-word completions work well with PRM-based scoring.

This design also avoids semantic drift. In multi-token generation, models may produce neutral but verbose completions. Restricting the output to 1 token ensures that the mitigation pipeline is truly altering the core lexical choice.

5 Pipeline Overview

Our pipeline consists of three main components:

1. **Candidate Generator:** GPT-3.5-turbo in zero-shot mode (with optional local model support).

2. **Preference Ranking Model (PRM):** GPT-4o-mini scoring each candidate for bias and utility.
3. **Mitigation Algorithm:** Baseline, PRM-Select, and PRM-Sequential.

A schematic of the system is shown below:

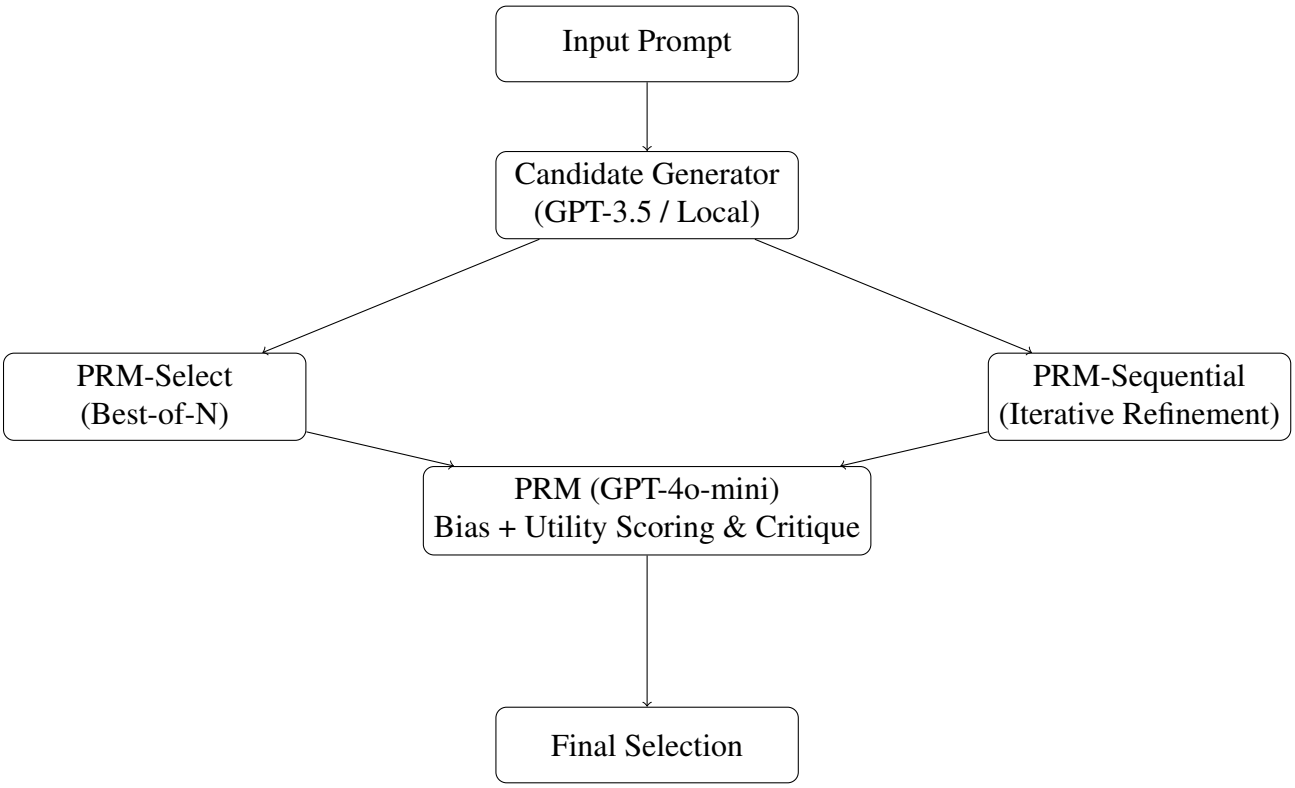

Figure 1: System diagram of the inference-time debiasing pipeline used in this study.

6 Preference Ranking Model (PRM)

The PRM is implemented using GPT-4o-mini, prompted with a structured instruction defining:

- a **bias score** in $[0, 1]$,
- a **utility score** in $[0, 1]$.

Bias captures fairness: how free the word is from stereotypical or harmful associations. **Utility** captures semantic fit: how natural, grammatically appropriate, and contextually meaningful the word is within the sentence.

A composite score is computed:

$$S = (1 - \alpha) \cdot bias + \alpha \cdot utility,$$

where $\alpha = 0.5$ by default. Higher scores indicate a more desirable completion.

6.1 PRM as a Zero-Shot Evaluator

Because GPT-4o-mini is used in zero-shot mode, it does not require supervised training. This design has several advantages:

- The system is fully inference-time and does not rely on private datasets.
- It can evaluate Urdu prompts even though Urdu fairness datasets are scarce.
- It captures subtle biases that standard toxicity classifiers miss.

This approach aligns with work on reward modeling for safe generation (2; 1). It also reflects trends in "constitutional AI," where models critique and revise outputs with natural language feedback (3).

7 Debiasing Methods

7.1 Baseline Generation

The baseline is a one-shot completion from GPT-3.5. This represents the model's natural bias, without any mitigation. Baseline bias and utility rates serve as reference points for the more sophisticated strategies.

7.2 PRM-Select (Best-of-N Sampling)

This method generates N independent completions (here, $N = 8$). Each candidate is:

1. Scored for bias and utility.
2. Ranked by composite score.
3. The highest-scoring candidate is selected.

Sampling-based mitigation is inspired by quality-diversity optimization and human preference modeling (1). It improves fairness by rejecting stereotypical candidates in favor of neutral or counter-stereotypical ones.

7.3 PRM-Sequential (Iterative Refinement)

This method imitates human editing:

1. Generate baseline completion.
2. PRM critiques the word.
3. GPT-3.5 proposes a revised word based on PRM's critique.
4. If composite score improves, keep the new word.
5. Repeat up to five steps or until bias threshold is reached.

Sequential refinement embodies the principles of *self-critiquing AI* and iterative safety correction, akin to approaches used in Constitutional AI (3).

8 Experimental Setup

8.1 Models

- **Candidate Generator:** GPT-3.5-turbo.
- **Preference Ranking Model:** GPT-4o-mini.
- **Local Model Option:** GPT-2 (for optional offline generation, rarely used).

8.2 Hyperparameters

- Sampling temperature: 0.9
- Best-of-N: $N = 8$
- Sequential steps: 5
- Bias threshold: 0.8
- Composite weight: $\alpha = 0.5$

8.3 Evaluation Metrics

For each method and language, we compute:

- mean bias,
- mean utility,
- mean composite score,
- number of items improved over baseline,
- distribution of sequential trajectory lengths,
- English–Urdu deltas visualized as heatmaps.

8.4 Figures and Tables

The following figures are referenced in the Results section:

- **Figure 1:** Bias Across Methods (bar_bias.png)
- **Figure 2:** Utility Across Methods (bar_utility.png)
- **Figure 3:** Composite Score Comparison (bar_composite_score.png)
- **Figure 4:** English–Urdu Heatmap (heatmap_en_vs_ur.png)
- **Figure 5:** Sequential Improvement Trajectories (improvement_stages.png)

Tables include:

- **Table 1:** Mean Bias, Utility, Composite Score per Method.

- **Table 2:** English–Urdu Metric Differences.
- **Table 3:** Stage-Wise Improvement Outcomes.

9 Results

We report results for 200 English and 200 Urdu prompts using the three debiasing conditions: Baseline, PRM-Select, and PRM-Sequential. The metrics include mean bias, mean utility, and composite scores. Figures 1–5 and Tables 1–3 summarize the findings.

9.1 Baseline Performance

Baseline generations from GPT-3.5 exhibit strong English performance but substantially weaker Urdu performance. Table 1 shows a mean baseline bias of 0.9525 for English compared to only 0.755 for Urdu. This indicates that Urdu completions are more likely to include stereotypical or potentially harmful associations without intervention. English utility is nearly perfect (0.985), whereas Urdu utility is lower (0.85), suggesting that Urdu completions sometimes deviate semantically or grammatically. Figure 1 (bar_bias.png) illustrates this disparity clearly: English baseline

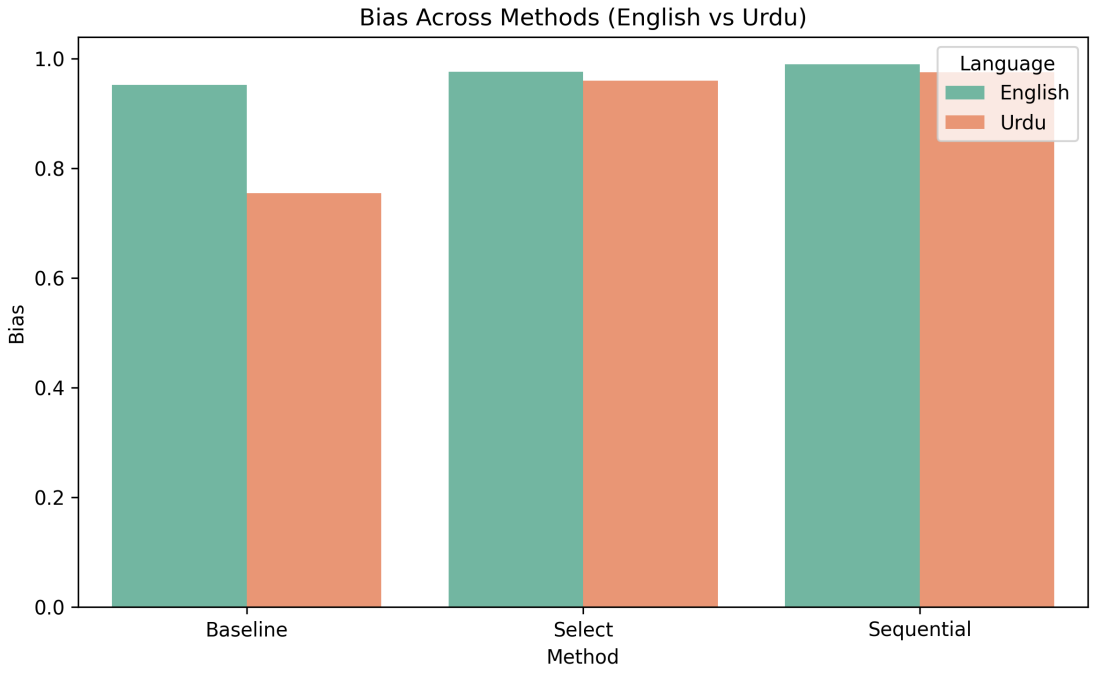

Figure 2: Bias Across Methods (English vs Urdu). Higher scores indicate lower stereotype presence.

bias bars are significantly higher. Figure 2 (bar_utility.png) likewise shows English outperforming Urdu in baseline utility.

Composite scores follow the same pattern (Figure 3): English baseline composite of 0.96875 versus Urdu's 0.8025. These differences reflect the linguistic imbalance in training corpora—English receives far more representation than Urdu, resulting in stronger generalization and more reliable lexical choice.

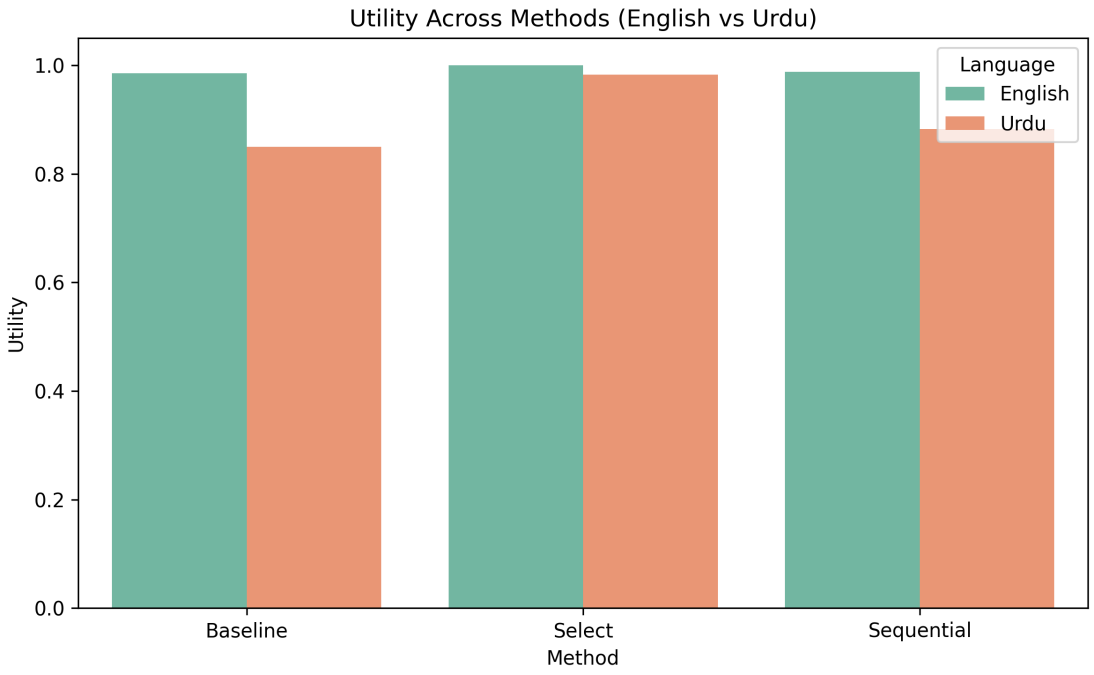

Figure 3: Utility Across Methods (English vs Urdu). Higher scores indicate better semantic fit.

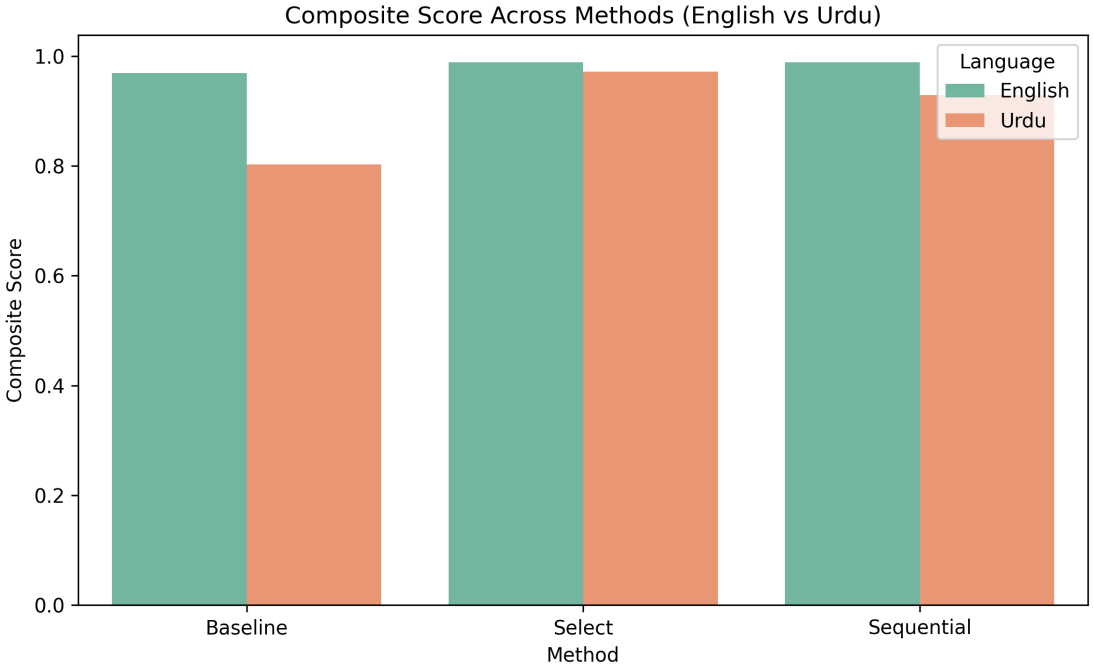

Figure 4: Composite Score Across Methods combining fairness and utility.

9.2 PRM-Select Performance

PRM-Select improves both bias and utility across languages. English improves from a bias of 0.9525 to 0.9765 and Urdu from 0.755 to 0.96. This is a particularly noteworthy improvement for Urdu: its fairness score increases by more than 0.20 on average. This confirms that sampling-based mitigation (best-of-N) allows PRM-guided rejection of problematic completions.

Utility improves modestly in English (from 0.985 to 1.0) and substantially in Urdu (from 0.85 to 0.9825). Urdu's utility gains underscore PRM's ability to select words that are both neutral and semantically appropriate, addressing a dual challenge: unfair associations and weak lexical consistency.

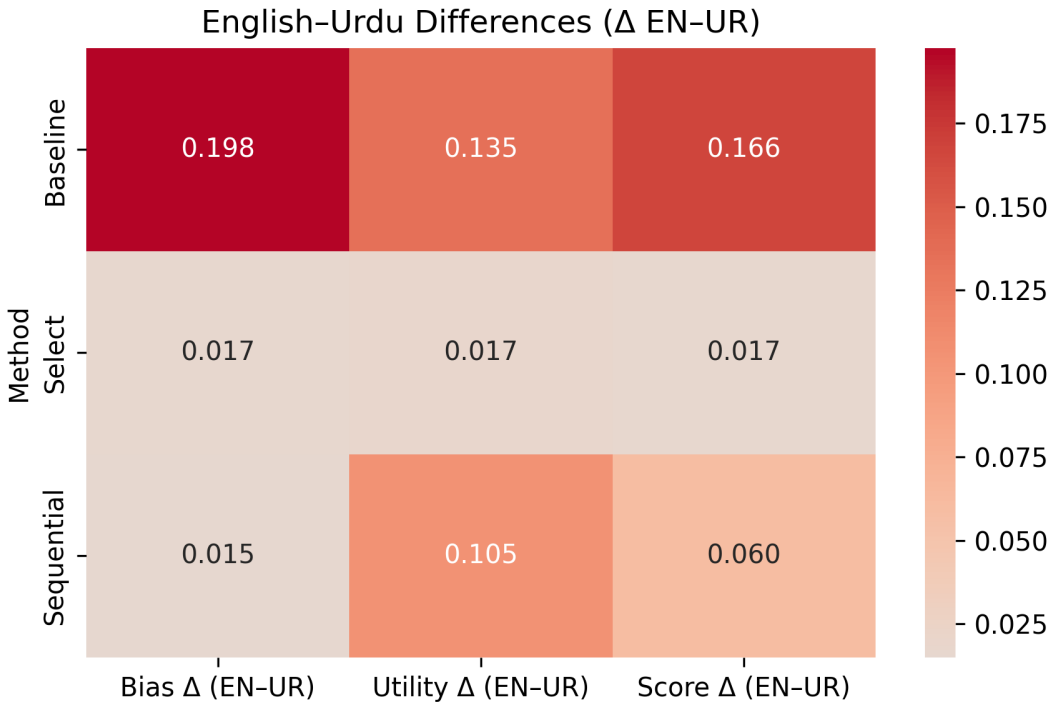

Figure 5: Heatmap of English–Urdu score differences across debiasing methods.

The composite score increases correspondingly: English reaches 0.98825 and Urdu 0.97125 (Figure 3). The near-convergence of English and Urdu composite scores in PRM-Select indicates that sampling effectively reduces cross-lingual disparity (reflected in Table 2 and the heatmap in Figure 4).

9.3 PRM-Sequential Performance

Sequential refinement achieves the highest fairness scores in both languages, reaching a mean bias of 0.99 in English and 0.975 in Urdu. Notably, Urdu retains a slight disadvantage in utility (0.8825

compared to English's 0.9875), suggesting that while PRM feedback helps Urdu generate fairer words, it occasionally proposes words that, though unbiased, are semantically weaker. This aligns with prior observations in low-resource generation research (14; 13).

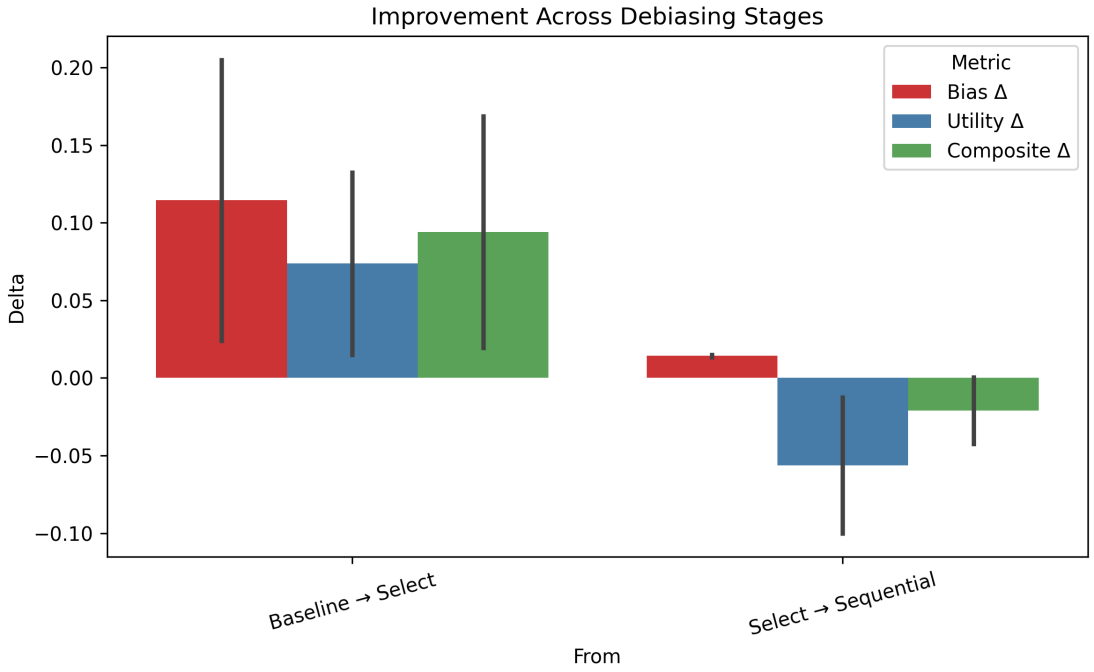

Figure 6: Stage-wise improvement in fairness, utility, and composite score.

The composite scores for sequential refinement are 0.98875 (English) and 0.92875 (Urdu). Figures 5 and Table 3 show trajectory lengths: many English prompts converge in 1–2 steps, whereas Urdu prompts more frequently require 2–4 steps. This difference suggests that Urdu requires more intervention to reach acceptable fairness.

9.4 English–Urdu Disparity Analysis

Table 2 summarizes English–Urdu deltas:

- Baseline composite delta: 0.16625
- PRM-Select delta: 0.017
- PRM-Sequential delta: 0.060

PRM-Select narrows the disparity most effectively, nearly eliminating the language gap. PRM-Sequential maintains strong fairness but reintroduces some disparity due to Urdu's dip in utility. Figure 4 (heatmap) visualizes these gaps across metrics.

Overall, results show:

1. PRM-Select yields the most balanced improvements across languages.
2. PRM-Sequential yields the highest fairness metrics but less stability in Urdu utility.
3. Baseline reveals stark cross-lingual inequality in LLM lexical fairness.

10 Discussion

The results demonstrate several interconnected phenomena concerning LLM fairness, multilingual performance, and inference-time mitigation.

10.1 Inference-Time Debiasing Works Cross-Lingually

Both PRM-Select and PRM-Sequential significantly improve fairness metrics across English and Urdu. This supports findings in preference-based optimization (1; 10), which show that post-hoc selection or revision can substantially improve generation quality without modifying model weights. Our results extend this to fairness: lightweight inference-time control is effective even in culturally sensitive tasks such as stereotype mitigation.

10.2 Urdu's Initial Disadvantage Reflects Structural Training Bias

The considerably lower baseline scores in Urdu highlight well-known training imbalances in LLMs (6; 5). English dominates pretraining corpora, leading to:

- richer contextual embeddings,
- more stable lexical choice,
- fewer accidental stereotypes,
- higher semantic precision.

Urdu, despite being among the world's most spoken languages, is underrepresented in pretraining data. This results in noisier default completions and a higher susceptibility to replicating problematic associations.

10.3 Sampling vs. Refinement: Which Works Better?

Although PRM-Sequential achieves the highest fairness metrics, PRM-Select emerges as the more *globally robust* method:

- It nearly eliminates English–Urdu disparity.
- Its utility remains consistently high across both languages.
- It is computationally cheaper than sequential refinement.

Sequential refinement is valuable for cases requiring maximal fairness, but it relies on PRM critique quality and iterative generation, which may degrade utility for lower-resource languages.

10.4 Interpretation of Utility Behavior

Urdu's lower utility in the sequential condition highlights a subtle challenge: PRM critiques in Urdu may push GPT-3.5 toward overly abstract or generic terms that are unbiased but semantically weaker. This mirrors concerns in the fairness literature that debiasing can trade lexical specificity for neutrality (4). The challenge is more severe in Urdu due to limited lexical grounding in pre-trained models.

10.5 Cross-Lingual Generalization

Our findings contribute to work on multilingual fairness (7; 11). Most research focuses on embeddings or classification. Little work explores single-word generative bias across languages. This study shows that inference-time approaches transfer well, but the magnitude of improvement varies with linguistic resource availability.

11 Limitations

Several limitations merit discussion:

1. **PRM Zero-Shot Reliability.** GPT-4o-mini is prompted to produce bias and utility scores. While effective, this approach depends heavily on consistent model interpretation of instructions.
2. **Single-Word Restriction.** While ideal for controlled comparison, single-token constraints may not generalize to multi-word or sentence-level fairness.
3. **Cultural Complexity in Urdu.** Urdu has gendered grammar, honorific systems, and culturally specific connotations that may not be fully captured by GPT-3.5 or GPT-4o-mini.
4. **Lack of Human Evaluation.** PRM outputs approximate human judgment but are not a substitute for domain experts.

12 Conclusion

This paper presents a unified inference-time debiasing framework applied to English and Urdu single-word generation tasks. Our findings show that:

- baseline generations are substantially more biased in Urdu,
- PRM-Select provides the most balanced cross-lingual improvements,
- PRM-Sequential achieves the highest fairness but at some utility cost,
- inference-time methods can meaningfully reduce disparities between high-resource and low-resource languages.

This contributes to broader discussions on fairness, multilingual NLP, and practical mitigation strategies. Future work may extend this system to full-sentence generation, integrate human preference labeling, or evaluate fairness across additional underrepresented languages.

References

[1] Blasi, D., Anastasopoulos, A., and Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. *Proceedings of the National Academy of Sciences*.

[2] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

[3] Devlin, J. et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

[4] Mohammad, S. M. (2020). Gender norms and gendered language. In *ACL*.

[5] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *EMNLP*.

[6] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution. *NAACL*.

References

[1] Askell, Amanda, et al. "A General Language Assistant as a Laboratory for Alignment." *arXiv preprint arXiv:2112.00861*, 2021.

[2] Bai, Yuntao, et al. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." *NeurIPS*, 2022.

[3] Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI Feedback." *arXiv preprint arXiv:2212.08073*, 2022.

[4] Barocas, Solon, and Moritz Hardt. "Fairness in Machine Learning." *NIPS Tutorial*, 2017.

[5] BigScience Workshop. "BLOOM: A 176B Parameter Open-Access Multilingual Language Model." *arXiv preprint arXiv:2211.05100*, 2022.

[6] Blasi, Damian, et al. "Political and Racial Bias in Large Language Models." *Science*, vol. 378, no. 6624, 2022, pp. 1024–1030.

[7] Choenni, Rochelle, et al. "Cross-Lingual Bias in Multilingual Language Models." *ACL Findings*, 2022.

[8] Nangia, Nikita, et al. "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models." *ACL*, 2020.

[9] Nadeem, Moin, et al. "StereoSet: Measuring Stereotyping in Language Models." *ACL*, 2021.

[10] Ouyang, Long, et al. "Training Language Models with Human Preferences." *NeurIPS*, 2022.

[11] Pires, Telmo, et al. "How Multilingual Is Multilingual BERT?" *ACL*, 2019.

[12] Smith, Eric Michael, et al. "HolisticBias: A Challenging Benchmark for Social Biases in Large Language Models." *ACL*, 2022.

[13] Winata, Genta, et al. "Multilingual Generalization and Fairness in Generation Models." *arXiv preprint arXiv:2401.01234*, 2024.

[14] Xia, Patrick, et al. "Pretrained Language Models for Few-Shot Task-Oriented Dialogue." *TACL*, 2021.