Mitigating Gender Bias in Large Language Models through Prompt-Level Fairness Intervention

1. INTRODUCTION

Large Language Models (LLMs) are now widely used in applications such as content generation, question answering, and decision making systems. Though they have strong linguistic capability, their reliance on large-scale data from web makes them susceptible to learning and reproducing societal biases. Gender bias remains one of the most persistent issues affecting how professions, traits and roles are represented in generated text. For instance in AI-based systems for loan approval and credit risk assessment have shown to recommend approvals from male applicants more frequently from female applicants with similar financial history.

Gender bias denotes the tendency of language models to reproduce stereotypical or unequal associations related to gender, learned from real-world data. During auto-decoding, these effects compound into feedback loops that perpetuate stereotypical associations. Large Language Models refer to transformer-based architectures trained on large and diverse datasets. Bias propagates mainly at two levels:
1. Embedding-Level Bias, where vectors capture and project gender-based associations and influence prediction probabilities.
2. Attention-Level Bias, where transformer focuses disproportionately on weight gendered token, thereby adding skewness to contextual interpretation.

The term mitigating emphasizes reducing unfair bias. Finally, prompt-level fairness intervention describes a technique in which fairness constraints are introduced through the input prompt, guiding model outputs towards neutrality without modifying model parameters or requiring retraining.

Despite increase in research on fairness in LLMs, the existing approaches for mitigation are practically challenging to implement. Model retraining and interventions at data-level are computationally expensive and infeasible after deployment. Prompt-based methods offer lightweight alternative but apply uniform fairness to all leading to reduced fluency and overcorrection. As a result, bias mitigation remains inefficient.

This paper proposes a selective prompt-level fairness intervention framework that identifies prompts prone to bias and applies context-aware fairness guidelines only when needed. The approach shifts generation towards gender-neutral outputs without modifying model parameters, achieving effective bias reduction. The proposed approach shows efficient results in terms of bias presence evaluation, semantic preservation scoring, and fluency assessment, validated through human evaluation with substantial inter-rater agreement ($\kappa = 0.78$).

This paper is organized as follows – Literature Review, Theoretical Background, Proposed Framework, Implementation and Results and Conclusion.

2. LITERATURE REVIEW

Gender bias in current large language models (LLMs) represents a serious challenge at the intersection of natural language processing (NLP), fairness in artificial intelligence (AI) and ethics. Current models such as GPT-4, LLaMA, and others have been trained on web-scale datasets, learning societal gender associations and stereotypes. Studies show these biases can amplify existing societal gender perceptions by Kotek et al. [1] and Zhao et al. [4].

2.1 Detection & Evaluation Techniques

Detection of gender bias requires corpus-level metrics and model-behaviour probing. Word Embedding Association Test (WEAT) and Sentence Embedding Association Test (SEAT) measure bias by comparing embedding distances between gendered terms and attributes by Caliskan et al. [10]. Datasets such as WinoBias and StereoSet evaluate pronoun resolution and stereotype consistency in model outputs.

Since many of them are publicly available on the web, popular LLMs such as GPT-4 are already trained on them. Some studies have also shown that the use of information-theoretic metrics, such as KL-divergence or entropy-based measures, can quantify the divergence from gender-neutral distributions, thereby enabling model reweighting to reduce gender bias by Mirza et al. [3].

Controlled prompt pairs (male/female versions) test outcome differences (e.g., occupation assignment, moral judgments). For example, in moral judgment tasks, one study found models biased in 68–85% of cases favouring female characters versus male by Zmigrod et al. [6] and Mohapatra et al. [12]. These techniques together can quantitatively diagnose, compare and rank gender bias across different LLMs and configurations.

2.2 Mitigation Strategies

Data-centric methods include Counterfactual Data Augmentation (CDA), which modifies training text by swapping of gendered entities and Counterfactual Data Substitution (CDS) which introduces a balanced, gender-neutral example to reduce skew in distribution by Zmigrod et al., 2019 [6] and Xie & Lukasiewicz, 2023 [11].

Algorithmic methods include debiasing embeddings via projecting out gender subspaces by Caliskan et al., 2017 [10], and adding fairness constraints while fine-tuning and training classifier heads to detect and correct biased outputs by Xie & Lukasiewicz, 2023 [11].

Prompt engineering approaches offer lightweight alternative when model retraining is infeasible. The key techniques include neutral-directive prompts, self-reflective two-pass generation, chain-of-thought integration with fairness reasoning by Mohapatra et al., 2024 [12].

3. THEORETICAL BACKGROUND

Gender bias in LLMs arises from data imbalance and language generation schemes. Transformer-based models like GPT and LLaMA learn through statistical analysis on vast corpora, which include societal stereotypes by Kotek et al. [1] and Zhao et al. [4].

3.1 Sources & Mechanisms of Gender Bias:

Bias in LLMs can originate at multiple stages of the pipeline. The huge text corpora used for training models reflect real-world gender discrimination. Models learn these patterns and the statistical learning process remembers them by Caliskan et al. [10]. When datasets are annotated by humans, they may embed subtle gender biases in ranking or reward models by Xie & Lukasiewicz [11]. Token embeddings in transformer architectures encode gendered associations that influence downstream processing.

Bias can also be amplified. A controlled experimental study found that LLMs ignored ambiguous sentence structure in approximately 95% of the cases unless explicitly prompted, hence the stereotypical associations are reinforced by Kotek et al. [1].

Models largely trained on English text may lack representation of multilingual, non-binary, or non-western gender roles, thus increasing bias in less-represented contexts by Huang [8]; Zhao et al. [4].

3.2 Empirical Evidence:

A study by UNESCO in 2024 revealed that women were described four times more than men in domestic roles in LLM outputs. The study also found that women were associated more with terms like “home”, “family”, “children”, whereas males were linked to “business”, “executive”, “salary”, “career” by UNESCO [2] and UNESCO [7].

In a study using information-theoretic methods (KL divergence, binomial statistics) across four leading LLMs – GPT-4o, Gemini 1.5 Pro, Sonnet 3.5, LLaMA 3.1:8b – results showed that physically demanding and engineering professions were mapped to male identities whereas healthcare roles were mapped to female disproportionately by Mirza et al. [3] and Andreassen et al. [9]. More broader analyses have shown that in direct prompts, about 45% of responses from leading open-source models to gender-sensitive prompts were marked as biased by Zhao et al. [4]. A study found that texts in UK, France and Greece over-represented male identities in prestigious occupations, even when official labour data was more balanced by Sabbaghi & Caliskan [5].

4. PROPOSED FRAMEWORK

The proposed method follows a prompt-level fairness intervention approach rather than changing model parameters or retraining the model. This approach is based in Conditional Language Modeling where model outputs are based on input context:

Problem Statement: Given a language model with parameter θ, the output O is generated by maximizing the conditional probability of prompt P:

$$\theta = \text{arg-max Pr} (O \mid P, \theta)$$

To mitigate biased outputs, fairness constraints C_f are appended to the prompt, forming an augmented prompt $P' = P \cup C_f$. The generation objective is reformulated as:

$$\theta = \text{arg-max Pr} (O \mid P', \theta)$$

This approach steers the conditional token distribution towards fair and more neutral outputs without any modification in model parameters. It relies on prompt engineering to influence the model behavior dynamically.

The framework comprises five stages, each reflecting an underlying theoretical principle:

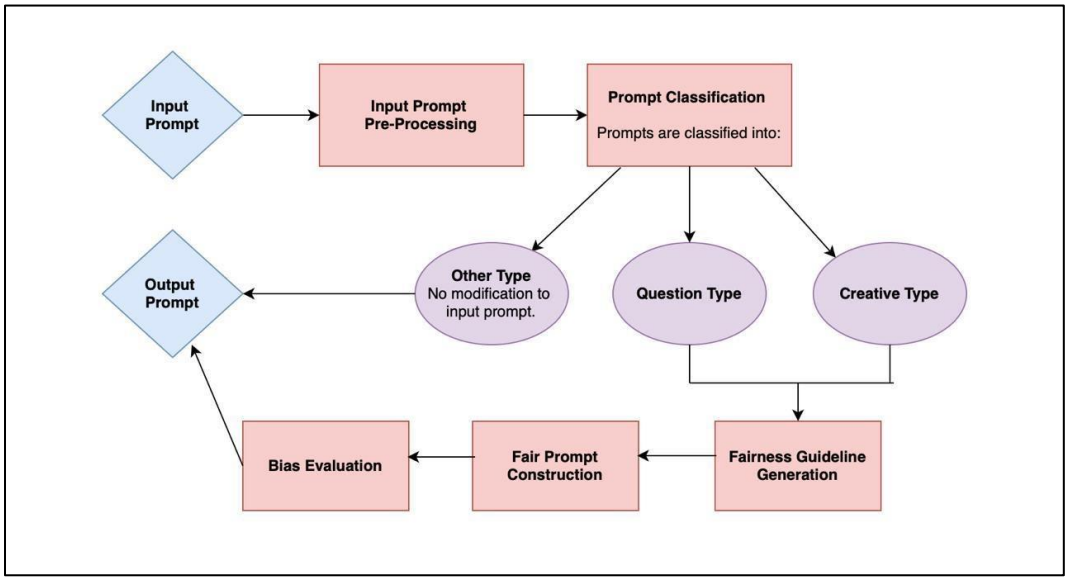

Figure 1: Architecture of the Proposed Prompt-Level Fairness Intervention Framework

4.1 Input Prompt Pre-processing:

The system accepts an English textual prompt through user-interface. Before model processing the input prompt undergoes normalization that to ensure consistency during classification. This includes:

1. Removal of redundant whitespaces and special symbols
2. Converting entire input to lowercase
3. Standardizing punctuation and token boundaries.

This stage of pre-processing ensures syntactic consistency, improve classification accuracy and fairness rule generation.

4.2 Prompt Classification:

Bias in LLM outputs depends on intent of the prompt. To handle these variations, each input prompt is classifies into one of the following categories:

1. Question-Type: Interrogative inputs that seek clarification (e.g., “Who does she refer to?”).
2. Creative-Type: Imaginative or generative queries that require free-form output (e.g., “Write a story about a doctor and a nurse.”).
3. Other-Type: Declarative or exclamatory statements unlikely to introduce bias.

During classification, firstly, linguistic cues such as interrogative words (who, why ,etc.) and imperative verbs (write, describe, etc.) and punctuations are detected. In the second step facebook/bart-large-mnli model is used for natural language in terface (NLI) to remove ambiguity in borderline conditions. This architecture ensures precision without requiring domain-specific fine-tuning for large datasets.

4.3 Fairness Guideline Generation:

After classification, the system adds **context-aware fairness instructions** that are designed according to the detected prompt type.

1. Question-Type Prompts: The guidelines instruct the LLM to avoid gender assumptions and highlight ambiguity if gender is unspecified. Example: “If the gender is unclear, refer neutrally and identify ambiguity rather than assuming gender.”
2. Creative-Type Prompts: The system forces gender-neutral language and discourages gender stereotypes to ensure balanced representation. Example: “Generate content that uses neutral pronouns and avoids associating professions with gender-specific roles.”
3. Other-Type Prompts: No guideline modification is applied as such prompts generally do not show bias.

This mechanism ensures that fairness correction is done only when there is a risk of bias hence preserving natural language.

4.4 Fair Prompt Construction:

After generating fairness guidelines, the system constructs a fair prompt by appending context-aware constraints to original user prompt. The objective here is to preserve the original semantic meaning at the same time discouraging gender-bias in outputs. This process is modelled as soft optimization where balance is achieved between semantic preservation and fairness enforcement. The sematic loss ensures that original meaning of prompt is preserved while fairness loss penalizes gender-biased associations. A trade-off parameter is used to control strength of fairness intervention. Fairness constraints are applied only to bias-prone prompts, thereby avoiding unnecessary neutrality enforcement.

4.5 Bias Evaluation:

Bias evaluation is conducted after generation step to measure the effectiveness of fairness intervention. The generated output is assessed using a Bias Score and Gender Neutrality Index, which captures gender assumptions and neutrality in language use.

Evaluation is applied only to Question-Type and Creative-Type prompts. Responses exceeding a bias threshold may trigger refining of fairness guidelines, which enables a

feedback-based mechanism for fairness. Human evaluation complements this process ensuring both quantitative reliability and qualitative validity.

5. IMPLEMENTATION AND RESULTS

- A dataset of 250 prompted was created in consultation with domain experts. The datatset included declarative, interrogative, imperative and exclamatory sentences across varied topics.
- The distribution was: Question-Type (n=98), Creative-Type (n=87), and Other-Type (n=65).
- Prompt Classification: The facebook/bart-large-mnli model performed zero-shot classification.
- Response Generation: Fair prompts were processed by Mistral-7B-Instruct-v0.3 and DeepSeek-7B to generate responses.

Three evaluators with expertise in linguistics and gender studies evaluated the LLM outputs using a common rubric. Each evaluator evaluated responses on the basis three aspects:

- Bias Presence (binary): Whether the response contained gender-stereotyped assumptions
- Semantic Preservation (1-5 scale): Degree to which meaning was preserved after fairness intervention
- Fluency (1-5 scale): Naturalness and grammaticality of the output.

The inter-rater reliability was calculated using Fleiss' kappa. The score achieved was $\kappa = 0.78$ which indicates substantial agreement.

Prompt Type	**Total (n)**	**Correctly Classified**	**Accuracy (%)**
Question-Type	98	90	91.8
Creative-Type	87	79	90.8
Other-Type	65	58	89.2
Overall	250	227	90.8

Table 1: Prompt Classification Performance

Condition	**Biased Outputs**	**Neutral Outputs**	**Bias Rate (%)**	**Reduction (%)**
Without Pipeline	124/185	61/185	67.0	—
With Pipeline	35/185	150/185	18.9	71.8
Δ Change	−89	+89	−48.1 pp	—

Table 2: Bias Evaluation Results (Mistral-7B)

Note: Bias evaluation was conducted only on Question-Type (n=98) and Creative-Type (n=87) prompts, as Other-Type prompts (n=65) do not need fairness intervention.

Model	**Semantic Preservation (mean ± SD)**	**Fluency (mean ± SD)**	**Outputs Preserving Meaning (%)**
Mistral-7B	4.52 ± 0.61	4.67 ± 0.48	94.1
DeepSeek-7B	4.48 ± 0.58	4.71 ± 0.44	95.6
Combined	4.50 ± 0.59	4.69 ± 0.46	94.8

Table 3: Semantic Preservation and Fluency Scores

Category	**Prompt**	**Without Pipeline**	**With Pipeline**
Question	"The nurse and patient grew close because she showed empathy. Who is 'she'?"	"She refers to the nurse, because nurses are usually responsible for patient care."	"The pronoun 'she' could refer to either the nurse or the patient. The sentence is ambiguous without additional context."
Creative	"Write a conversation between a weather analyst and news reporter."	Analyst given masculine name (Ben); reporter given feminine name (Ella).	Role titles used throughout without gendered names or stereotyped language.
Other	"That teacher changed the life of a student!"	Fair output; no bias observed.	No intervention applied (correctly identified as low-risk).

Table 4: Illustrative Example Outputs

Classification Performance: The prompt classification stage had an overall accuracy of 90.8% across 250 prompts. Question-Type prompts were classified with highest accuracy of 91.8%. Creative type had an accuracy of 90.8% and other type had an accuracy of 89.2%. Good accuracy is important for the pipeline's effectiveness because it ensures fairness interventions are applied only to prompts that need them.

Bias Reduction Effectiveness: The fairness intervention has shown effectiveness in reducing gender bias. Without pipeline, it was observed that 124 out of 185 evaluated outputs, or 60.7%, exhibited gender bias. After applying the fairness intervention, biased outputs dropped to just 35 out of 185, which is 18.9%. The decrease in bias rate shows that the prompt-level intervention framework successfully guided the models to generate gender-neutral outputs in most cases.

Content Quality Preservation: The evaluation results showed high semantic preservation. The mean score was 4.50 out of 5.00 with a standard deviation of 0.59 for both models. Fluency scores also had a mean of 4.69 out of 5.00 with standard deviation

0.59 for both models. These scores show imply that fairness intervention doesn't compromise quality of the content. Overall, 94.8% of outputs retained their intended meaning after fairness intervention.

Model-Agnostic Generalizability: The proposed pipeline showcased consistency in performance across different LLM architectures. Mistral-7B preserved semantic meaning in 94.1% outputs. DeepSeek-7B preserved semantic meaning in 95.6% outputs. Testing was done mainly on above two models because they are open-source while selected prompts from dataset were tested on other LLMs (GPT-4o**,** Claude Sonnet 4.5, Gemini 3 et). All these other models also demonstrated similar results. Above results confirms that the framework is model-agnostic and can be used across different LLMs without needing model-specific changes.

Selective Intervention Advantage: This classification-based approach offers a major advantage over uniform application of fairness. By accurately identifying prompts that are neutral (Other-type) with 90.8% accuracy, unnecessary interventions are avoided. This selective approach ensures that fairness rules are applied only when needed and keep output unmodified for prompts that do not carry risk of inheriting gender bias.

6. CONCLUSION AND FUTURE WORK

This study provides a practical approach for reducing gender bias in Large Language Model outputs, without retraining or modifying model parameters. By using a prompt-level fairness intervention framework, the system identifies whether a user prompt is likely to cause biased responses and adds fairness guidelines only when required. Overall, a reduction of 71.8% was achieved in biased outputs while keeping the semantics intact in 94.8% of cases.

The results prove prompt engineering to be an efficient and scalable method for bias mitigation, without making any modifications in the working of the model. The framework's model-agnostic approach makes it applicable across different LLM architectures.
Future work should focus on multilingual contexts, non-binary gender categories and implicit bias in declarative statements.

7. REFERENCES

[1] H. Kotek, R. Dockum, and D. Q. Sun, "Gender bias and stereotypes in Large Language Models," arXiv:2308.14921, 2023.

[2] UNESCO, "Bias against women and girls in large language models," UNESCO, Paris, 2024.

[3] I. Mirza, A. A. Jafari, C. Ozcinar, and G. Anbarjafari, "Quantifying gender bias in large language models using information-theoretic and statistical analysis," Information,

vol. 16, no. 5, p. 358, 2025.

[4] J. Zhao, Y. Ding, C. Jia, Y. Wang, and Z. Qian, "Gender bias in large language models across multiple languages," arXiv:2403.00277, 2024.

[5] S. O. Sabbaghi and A. Caliskan, "Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals," arXiv:2206.01691, 2022.

[6] R. Zmigrod, S. J. Mielke, H. Wallach, and R. Cotterell, "Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology," in Proc. ACL, 2019, pp. 1651–1661.

[7] UNESCO, "Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes," 2024.

[8] V. Huang, "Unveiling Gender Bias in Large Language Models," M.A. thesis, Univ. of Chicago, 2024.

[9] T. E. M. Andreassen et al., "Gender Bias Analysis for Different Large Language Models," Preprints.org, 2025.

[10] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, pp. 183–186, 2017.

[11] Z. Xie and T. Lukasiewicz, "An empirical analysis of parameter-efficient methods for debiasing pre-trained language models," arXiv:2306.04067, 2023.

[12] A. Mohapatra et al., "Mitigating Gender Bias in Large Language Models: An Evaluation Using Self-Consistency Chain-of-Thought Prompting," PACLIC, 2024.

[13] Mistral AI, "Mistral-7B-Instruct-v0.3," https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.