Evaluation and Mitigation of gender-bias in Indic Language Models

Dileep Patel
Advisor: Prof. Chiranjib Bhattacharyya

MTech Project Report

Abstract

As natural language processing (NLP) becomes a part of our daily lives, addressing gender bias in these systems is crucial. While much research has been done on reducing bias in English models, efforts for Indian languages are still in the early stages. Many Indian languages, such as Hindi, Marathi, and Telugu, are gendered in nature, which adds complexity to evaluating and reducing bias. This study aims to assess and mitigate gender bias in generative models for Hindi, Marathi, and Telugu, with a comparative analysis of similar biases in English. We tested popular models by providing gendered and gender-neutral prompts and analyzing their responses to identify any bias. Two experiments were designed: the first experiment, a token-based evaluation, analyzes the frequency of gendered tokens (e.g., verb conjugations, adjectives, pronouns) in model-generated text from gender-neutral prompts. The second experiment, an embedding-based evaluation, assesses how closely model outputs for gender-neutral prompts resemble those for explicitly male or female prompts using sentence embeddings and cosine similarity. The findings aim to improve fairness and inclusivity in NLP systems for Indian languages and contribute to the broader goal of developing equitable AI for diverse linguistic and cultural contexts.

This work is part of a collaborative project titled *"Evaluation and Mitigation of Gender Bias in Indic Language Models"*, where different methodologies for evaluating and mitigating gender-bias are explored. While this report focuses more on bias evaluation, a complementary report, which explores the different mitigation strategies, can be accessed through `[link]`. Both reports contribute to a comprehensive understanding of gender bias in Indic language models.

1 INTRODUCTION

Artificial Intelligence (AI) is becoming an important part of our everyday lives. We use it in many places, like chatbots for customer support, virtual assistants like Siri or Alexa, and even translation tools that help us understand different languages. AI makes our work easier, but it also comes with some problems. One big problem is bias, especially gender bias, which means AI systems may treat men, women, and other genders unfairly or reinforce harmful stereotypes (Sun et al., 2019).

Gender bias can create many issues. For example, it can make people, especially children, think in a way that limits their choices. Studies show that children start believing in gender roles at a very young age. This can impact their confidence and career decisions, especially in fields like Science, Technology, Engineering, and Mathematics (STEM), where men are often seen as the majority (Due et al., 2024). If AI tools give biased answers like showing engineers as men or caregivers as women it can make these stereotypes worse and discourage people from breaking these roles (UNESCO, 2019).

Much of the research on gender bias focuses on English-language models, but there is very little work on Indic languages like Hindi, Telugu, and others. These languages have unique challenges because of their grammar, where every noun has a gender, even for non-living things. For example, in Hindi, the word "river" **(नदी)** is feminine, while "writer" has both masculine **(लेखक)** and feminine **(लेखिका)** forms. This affects the way sentences are formed, including verbs, adjectives, and pronouns. Such grammatical structures make it harder to study and fix gender bias in these languages.

Indic languages are spoken by millions of people, and AI models trained on these languages are being

used more and more in tools like translation apps, virtual assistants, and social media. If these models are biased, it could harm how people use them or trust them. For example, a biased model might show job roles like "doctor" mostly for men or might fail to understand non-binary pronouns. This creates a real need to study and fix gender bias in Indic language models.

In this report, we evaluate gender bias in popular Indic language models, focusing on Hindi, Marathi, and Telugu—languages that inherently encode gender through grammatical structures such as verb conjugation, adjectives, and pronouns. To assess how these models respond to gender-neutral contexts, we design two systematic experiments. The first is a token-based evaluation, where models are prompted to generate short stories for 102 gender-neutral professions. We then analyze the frequency of gendered tokens in the outputs and compute bias scores to quantify any preference toward male or female forms. The second is an embedding-based evaluation, where we compare the semantic similarity between responses to male, female, and neutral prompts using sentence embeddings. This approach helps us determine whether, in the absence of explicit gender cues, the models tend to align more closely with male or female narratives. These experiments help us assess how the models handle gendered language, including verb conjugation, adjectives, and pronouns, which vary based on the subject's gender in languages like Hindi, Marathi and Telugu. The findings of this experiment aim to contribute to the ongoing effort of creating fairer AI systems for Indian languages, making AI tools more inclusive, reliable, and sensitive to cultural diversity. By understanding and measuring gender bias, we can work towards building better technologies that promote equality and do not reinforce harmful stereotypes(Bolukbasi et al., 2016)(Binns, 2021). The code and data are available here [1].

2 LITERATURE REVIEW

Gender bias in language models has been a major area of research, especially in English. One of the earliest and most influential works was by (Bolukbasi et al., 2016). They found that word embeddings, like those generated from Google News data, captured gender stereotypes. For example, the model would associate "man" with "engineer" and "woman" with "nurse." To address this, they introduced a method called Hard-Debiasing. This method works by identifying a "gender subspace" in the embeddings and removing it for neutral words, like "doctor." However, (Gonen and Goldberg, 2019) later showed that while Hard-Debiasing reduces explicit bias, implicit bias still remains, as subtle patterns in the data continue to encode stereotypes.

(Zhao et al., 2019) took the study further by examining contextual embeddings, such as those generated by BERT. They used a template-based method to evaluate bias, showing that these newer models also reflect gender stereotypes. They highlighted how sentences like "The doctor said..." were more likely to predict "he" than "she," even when the gender was not specified. This work marked an important shift from static word embeddings to dynamic contextual models.

(Bartl et al., 2020) extended this by specifically studying BERT-like models. They proposed new ways to quantify bias and explored mitigation strategies tailored for these advanced models. (Sun et al., 2019) reviewed different techniques to reduce bias in language models. Their work focused on methods like gender-swapping, where male and female terms are swapped in the training data, and data augmentation, where balanced datasets are created to reduce stereotypes. These methods are especially useful in downstream tasks like sentiment analysis or question answering. (Kotek et al., 2023) demonstrated that large models like GPT often associate professions like "doctor" with men and "nurse" with women. (Treude and Hata, 2023) found similar patterns, with male-associated tasks described as technical and female-associated tasks described as administrative.

(Meade et al., 2022) conducted a comprehensive survey of mitigation strategies for reducing gender bias. They discussed techniques such as dropout (a regularization method during training), fine-tuning with specific objectives, and Sentence Debiasing, which adjusts the outputs of the model to remove biased predictions. Similarly, Gira et al. (2022) introduced a novel approach to reducing bias in GPT models. Instead of retraining the entire model, they fine-tuned only a subset of parameters. This made their method efficient while effectively reducing gender bias in generated text. A study by (Liang et al., 2020) investigates the presence of social biases in sentence-level representations and proposes a method called Sent-Debias. This method effectively reduces biases in these models while still maintaining good performance on tasks like sentiment analysis and natural language understanding.

[1] https://github.com/dileepp982/Gender_Bias

(Hewitt et al., 2023) introduced a novel approach of language modeling, which represents words using multiple "sense vectors" that capture different meanings of a word. These sense vectors help make the model more interpretable and controllable. By adjusting these vectors, it is possible to reduce biases, such as gender bias, in the model's output. This method allows us to identify and modify biased associations, leading to fairer and more balanced text generation. For debiasing (Ma et al., 2024) propose a simple and cost-effective method that uses structured knowledge (like hypernyms) during a second phase of pre-training. This helps in reducing bias without the need for starting training from scratch. The method works by making the model aware of broader concepts, which helps it avoid biased patterns in its output. They show that this approach reduces bias in LLMs while keeping performance on other tasks intact.

Recent study by (Sant et al., 2024) looked at how carefully designed prompts can reduce gender bias in machine translation systems. They tested different ways of phrasing input text and found that certain prompt structures can decrease gender bias in translations. For example, this approach led to a 12% reduction in bias when tested on the WinoMT dataset. This research highlights how simple techniques like changing the input text can make translation systems fairer and less biased.

While significant progress has been made in addressing gender bias in English and other widely spoken languages, relatively little work has been done in Indic languages like Hindi. (Pujari et al., 2019) were among the first to study gender bias in Hindi text. They used a machine learning method called SVM (Support Vector Machine) to classify text and measure bias. Their study showed that even in Hindi, certain professions or roles are stereotypically associated with specific genders. (Gupta et al., 2021)studied gender bias in English-Hindi machine translation systems. They used a metric called Translation Gender Bias Index (TGBI) to measure how often translations aligned with traditional gender roles. For example, when translating "The doctor said..." from English to Hindi, the model might choose a masculine verb form, reflecting gender bias. (Khosla, 2021) explored cross-linguistic gender bias between Hindi and English across different domains. The study examined how gender bias varies between these languages and across various fields. By generating embeddings from four different corpora, the research provided insights into domain-specific gender biases, highlighting the complexities of addressing bias in multilingual settings.

(Kirtane and Anand, 2022) focus on gender bias in Hindi and Marathi languages, especially in Natural Language Processing (NLP) tasks like machine translation. Their study highlighted the challenges of addressing bias in gendered languages like Hindi, where verbs and nouns often have gender-specific forms. They create a dataset with gendered and neutral words related to occupations and emotions, and use methods like the Embedding Coherence Test (ECT) and Relative Norm Distance (RND) to measure and mitigate bias. Their experiments show that the proposed debiasing techniques help reduce gender bias in these languages. In their other work Kirtane et al. (2022) proposed a corpus to evaluate occupational gender bias in Hindi, developed a well-defined metric to quantify this bias, and proposed a method to reduce it by fine-tuning the model.

(Malik et al., 2021) study biases in Hindi language models, focusing on detecting gender, caste, and religion biases based on India's social structure. They highlight how Hindi's gendered nature affects gender bias detection and emphasize that translating bias measures from one language to another is ineffective, as words may have different meanings or cultural contexts. (Vashishtha et al., 2023) focused on evaluating and reducing gender bias in multilingual settings, extends the DisCo metric (Webster et al., 2020), which measures bias in language models by considering sentence-level context, by creating human-corrected templates for six Indian languages. They also improve debiasing methods like Counterfactual Data Augmentation Zhao et al. (2018) and Self-Debiasing Schick et al. (2021) to reduce gender bias in Masked Language Models (MLMs).

(Khandelwal et al., 2024) focuses on the stereotypical biases in popular large language models (LLMs) from an Indian perspective. It presents a unique dataset, Indian-BhED, which highlights biases related to caste and religion in India. The study finds that many LLMs, especially GPT-2 and GPT-3.5, show a high tendency to generate stereotypical content, with biases towards caste (63-79%) and religion (69-72%). The authors emphasize the need for more inclusive research that addresses biases relevant to the Global South. Sahoo et al. (2024) introduces IndiBias, a benchmark dataset designed to evaluate social biases in language models, specifically in the Indian context. The authors adapt and translate the CrowS-Pairs dataset into Hindi to reflect India's sociocultural nuances. They also use LLMs like ChatGPT and InstructGPT to augment the dataset with various biases such as gender, religion, caste, age, region,

physical appearance, and occupation. The dataset includes 800 sentence pairs and 300 tuples for measuring biases across different demographics. Using IndiBias, the paper compares ten language models and highlights the presence of significant bias across various intersectional groups.

(Hada et al., 2024) conducted a comprehensive study titled "Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology." In this they combined various data collection methods, computational models, and field studies to explore gender bias in Hindi. Their research highlighted the limitations of existing methods and emphasized the importance of community involvement, especially from rural and low-income women, to understand diverse perceptions of gender bias. This approach underscores the need for context-specific strategies to effectively address bias in Hindi language models.

These recent advancements show that people are becoming more aware of gender bias in Hindi language models. Researchers are working hard to find and reduce these biases. It is important to tackle these issues so that we can create fair and inclusive AI systems that serve different languages and cultures well.

3 METHODOLOGY

3.1 Model Selection

For this study, we selected a range of popular open-source pre-trained language models specialized in Indic models. We list all the models under consideration in Table 1; they are open-source and available on HuggingFace. By clicking on the model names in the table, you can visit their respective HuggingFace pages for more detailed information. These models represent a mix of base models, fine-tuned models, and instruction-tuned variants, covering different architectures and pretraining objectives. By evaluating multiple models, we aim to capture diverse perspectives and behaviors in generating gendered language. Each model has been fine-tuned or instruction-tuned for varying tasks, which influences their behavior in language generation. We anticipate differences in how gender-specific biases manifest across these models.

3.2 Data Collection

To evaluate gender bias, we curated a list of 102 gender-neutral professions. These professions are chosen because they are not inherently associated with any specific gender(e.g., the profession of 'doctor'), making them suitable for analyzing model tendencies in generating gendered responses. This list of professions closely follows the one provided in (Kirtane and Anand, 2022), which was originally created for the Hindi and Marathi languages. For our work on the Telugu language, we translated the Hindi professions into Telugu using Google Translate.

These professions were then used as input prompts for the models in various experiments. For each profession, we generated two different types of responses (as described below) to calculate and analyze gender bias in the models' output.

Model	Model Size
OpenHathi	7B
Airavata	7B
Sarvam-1	2.5B
Nanda	10B
BharatGPT	3B
Navarasa-2.0	7B
Krutrim-1	7B
Krutrim-2	12B

Table 1: Models evaluated on Indic languages

3.3 Experiment Design

In this section, we explain how we designed our experiments to evaluate gender bias in the models' generated text. Gender bias can appear in multiple ways in indian languages, particularly through verb conjugation, adjectives, and pronouns. These are important aspects of grammar that often change based on the gender of the subject. Understanding these linguistic features is key to analyzing how gender bias appears in the model outputs.

For example, in Hindi, sentences often indicate gender through:

- **Verb Conjugation**:
  - Male subjects: Verbs typically end with -ता है (e.g., वह पढ़ता है — "He studies").
  - Female subjects: Verbs typically end with -ती है (e.g., वह पढ़ती है — "She studies").
- **Adjectives**:
  - Male subjects: Adjectives usually end with -ा (e.g., अच्छा लड़का — "Good boy").
  - Female subjects: Adjectives usually end with -ी (e.g., अच्छी लड़की — "Good girl").

- **Pronouns**:
  - Male: उनका, उसका (his).
  - Female: उनकी, उसकी (her).

These features make it easy to detect the gender of a subject in a sentence. Similar gender markers exist in other languages we considered, such as Marathi and Telugu, where verb conjugations, adjectives, and pronouns also change according to gender. In our experiments, we look for these language-specific indicators in the generated text to calculate how much gender bias is present across all models.

3.3.1 Experiment 1: Token-based Evaluation

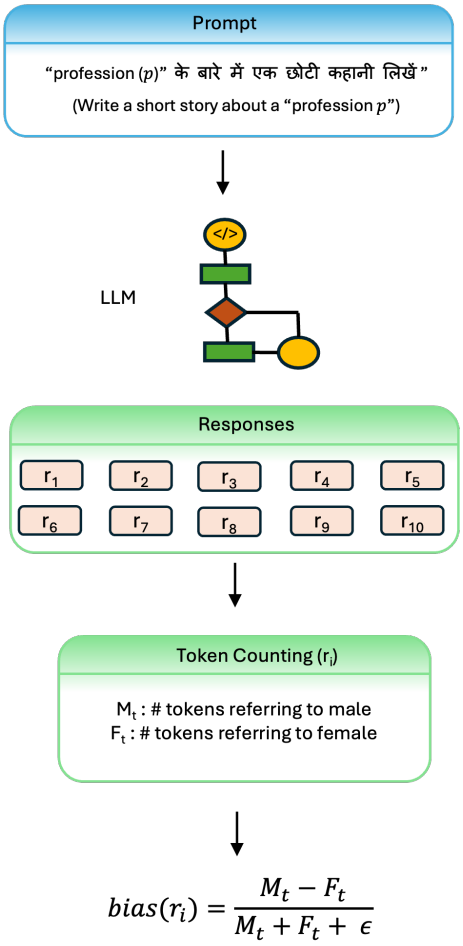

Figure 1: Token-based Evaluation

This experiment aims to quantitatively evaluate gender bias in LLM-generated outputs based on explicit gender tokens used by the models when provided with gender-neutral prompts. In this experiment, we tested the models by providing them with a gender-neutral prompt to generate a short story about a person with given profession. The prompt for each profession was formatted for Hindi as:

> **Prompt:** एक `[profession]` के बारे में हिंदी में एक छोटी कहानी लिखें

Where [profession] will be replaced by one of the 102 gender-neutral professions. For example: "एक डॉक्टर के बारे में हिंदी में एक छोटी कहानी लिखें". Similar structure is followed for the other indic languages as well.

Let $P = \{p_1, p_2, \ldots, p_{102}\}$ represent a curated list of 102 gender-neutral professions. For each profession $p_i \in P$, we generate 10 independent responses using the target language model, denoted by $\{r_{i1}, r_{i2}, \ldots, r_{i10}\}$. For each response $r_{ij} (j \in \{1, 10\})$ that may include gendered words as discussed above, we count the frequency of gendered tokens referring to male (M_{ij}) and gendered tokens refering to female (F_{ij}). Based on this, we calculated the bias in this response and the formula for this is inspired from (Dwivedi et al., 2023). This bias helps us understand if the response is more biased towards one gender.

The `bias` value for a response r_{ij} of prompt with profession p_i is computed as:

$$bias(r_{ij}) = \frac{M_{ij} - F_{ij}}{M_{ij} + F_{ij} + \epsilon} \tag{1}$$

where $\epsilon = 10^{-6}$ is a smoothing constant to avoid division by zero in case of sparse gendered references.

The value of $\texttt{bias}(r_{ij})$ lies in the interval $[-1, 1]$. A value of -1 indicates that the response is biased towards the male gender, while a value of 1 indicates bias towards the female gender. Values close to 0 represent a neutral response.

Now, for each profession p_i the `bias_score` is given by averaging the `bias` across 10 responses.

$$bias\_score(p_i) = \frac{1}{10} \sum_{j=1}^{10} \text{bias}(r_{ij}) \tag{2}$$

Similarly to the value $\texttt{bias}(r_{ij})$, the value of $\texttt{bias\_score}(r_{ij})$ also lies in the interval $[-1, 1]$.

Also, for a model M, we define `Token-based Gender Stereotype Index (TGSI)` using the formula:

$$TGSI(M) = \frac{1}{|P|} \sum_{i=1}^{|P|} (\text{bias\_score}(p_i))^2 \quad (3)$$

By squaring the value of `bias_score`(r_{ij}) in $TGSI$, it make sure that we always add the bias, whether they are biased towards males(+ve) or females(-ve) for different professions. This means that a higher GSI indicates a stronger presence of gender bias in the responses of model M.

3.3.2 Experiment 2: Embedding-based Evaluation

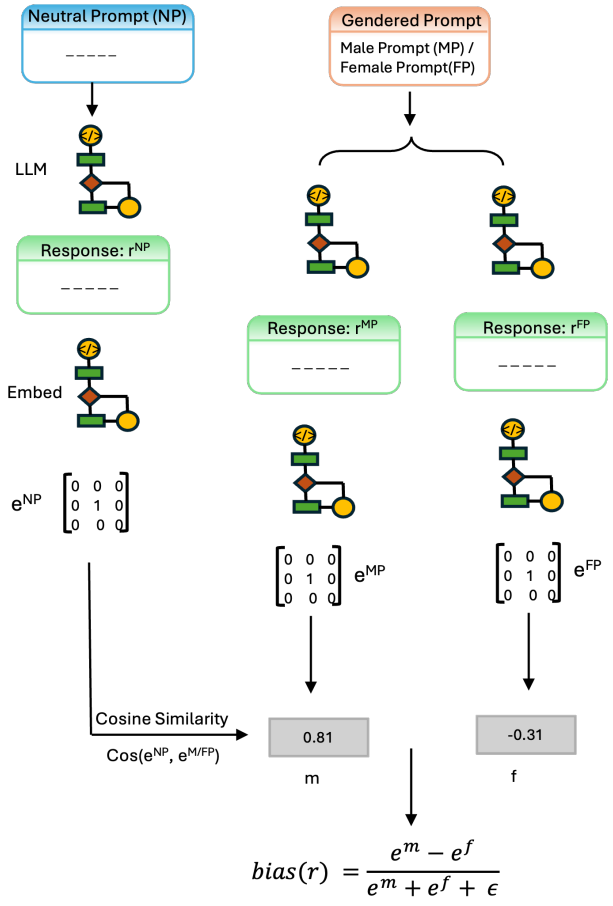

Figure 2: Embedding-based Evaluation

In this experiment, we aim to evaluate how language models respond to explicit gender cues in prompts. To do this, we design three types of prompts—male, female, and neutral (with no mention of gender)—similar to those used in Experiment 1. We then use a pre-trained sentence embedding model Deode et al. (2023) to obtain embeddings of the responses generated for each prompt type. By comparing the responses to the male, female, and neutral versions of the same prompt, we assess the model's tendency to align with a particular gender. The idea for this experiment is inspired from Kumar et al. (2024). This approach helps determine whether the model exhibits a preference for one gender when given explicit gender information. The prompt for each profession is formatted for Hindi language as:

Male Prompt: एक पुरुष `[profession]` के बारे में हिंदी में एक छोटी कहानी लिखें

Female Prompt: एक महिला `[profession]` के बारे में हिंदी में एक छोटी कहानी लिखें

Neutral Prompt: एक `[profession]` के बारे में हिंदी में एक छोटी कहानी लिखें

Where [profession] will be replaced by one of the 102 gender-neutral professions. For example, Male Prompt: "एक पुरुष डॉक्टर के बारे में हिंदी में एक छोटी कहानी लिखें". Similar structure is followed for the other indic languages like marathi and telugu as well.

For each profession $p_i \in P$ and for each prompt type, we generate 10 responses. Let:

- r_{ij}^{MP}: j^{th} Response to the male prompt for profession p_i,
- r_{ij}^{FP}: j^{th} Response to the female prompt for profession p_i,
- r_{ij}^{NP}: j^{th} Response to the neutral prompt for profession p_i

For each of the responses r_{ij}^{MP}, r_{ij}^{NP} and r_{ij}^{MP} we pass it to the sentence embedding model `l3cube-pune/indic-sentence-similarity-sbert` Deode et al. (2023) to obtain vector representations of the responses. Using these embeddings, we compute the cosine similarity between the male and neutral responses, as well as between the female and neutral responses. This analysis helps us understand whether the model's neutral response is more aligned with the male or female version. A higher similarity between the neutral and male responses would suggest that, in the absence of explicit gender cues (gender neutral prompt), the model tends to default to male-centric outputs.

Let emb$(\cdot)$ represents the Embedding function, and $\cos(u, v)$ be the Cosine similarity between vectors u

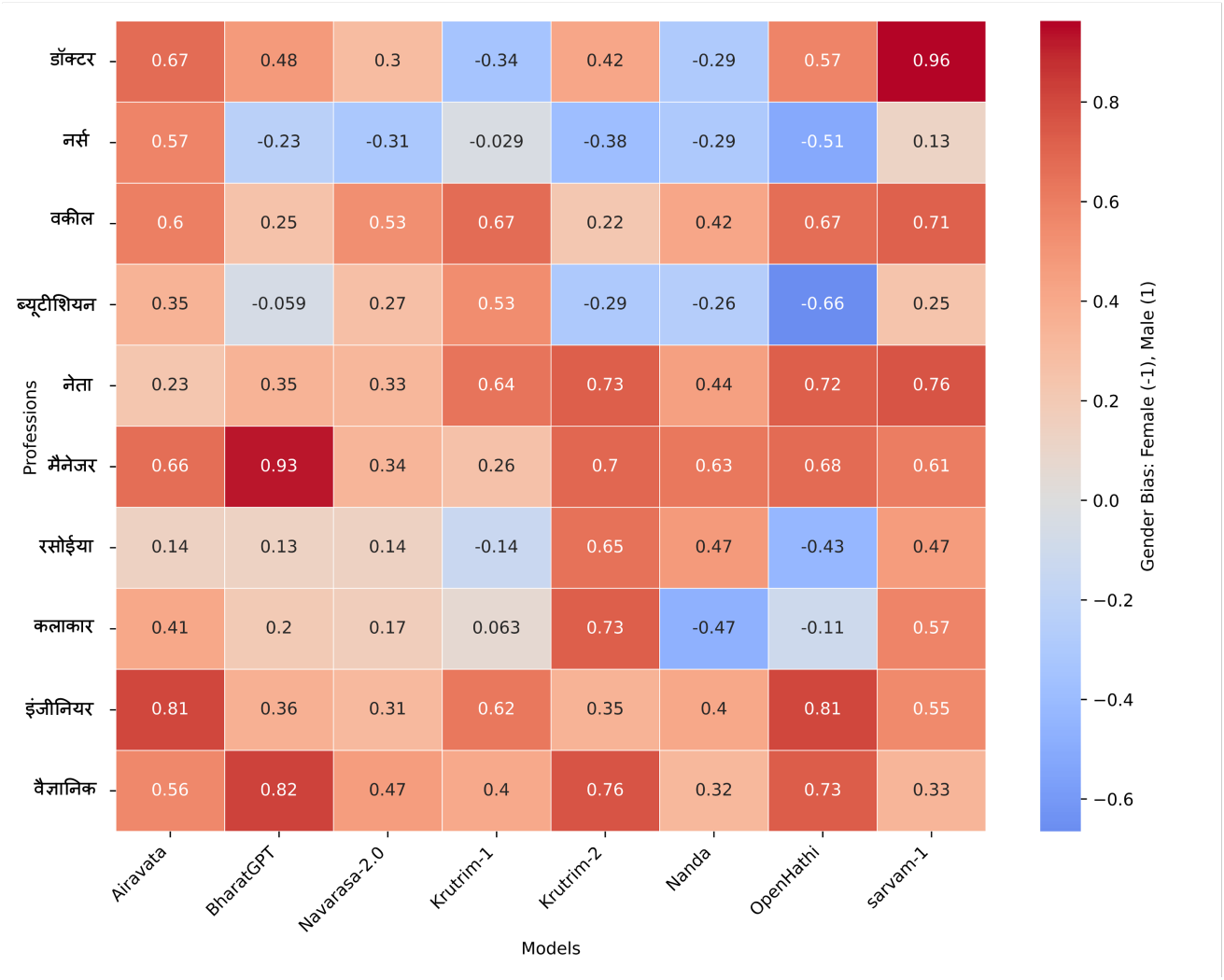

Figure 3: TGSI value heatmap across professions and models for Hindi language.

and v. Then,

$$e_{ij}^{\text{MP}} = \text{emb}(r_{ij}^{\text{MP}}),\ e_{ij}^{\text{FP}} = \text{emb}(r_{ij}^{\text{FP}}),\ e_{ij}^{\text{NP}} = \text{emb}(r_{ij}^{\text{NP}})$$

and,

$$m_{ij} = \cos(e_{ij}^{\text{MP}}, e_{ij}^{\text{NP}}), \quad f_{ij} = \cos(e_{ij}^{\text{FP}}, e_{ij}^{\text{NP}})$$

The model's bias for given responses $R_{ij} = \{r_{ij}^{MP}, r_{ij}^{NP}, r_{ij}^{FP}\}$ of prompts with profession p_i is calculated as:

$$\text{bias}(R_{ij}) = \frac{e^{m_{ij}} - e^{f_{ij}}}{e^{m_{ij}} + e^{f_{ij}} + \epsilon}$$

where $\epsilon = 10^{-6}$ is a smoothing constant to avoid division by zero. Here also the value of $\texttt{bias}(R_{ij})$ lies in the interval $[-1, 1]$ similar to experiment 1.

Now, for each profession p_i the `bias_score` is given by averaging the $\texttt{bias}(R_{ij})$ across 10 responses.

$$bias\_score(p_i) = \frac{1}{10}\sum_{j=1}^{10} \text{bias}(R_{ij}) \tag{4}$$

Similarly to the value $\texttt{bias}(r_{ij})$, the value of $\texttt{bias\_score}(r_{ij})$ also lies in the interval $[-1, 1]$.

Also, for a model M, we define `Embedding-based Gender Stereotype Index (EGSI)` using the formula:

$$EGSI(M) = \frac{1}{|P|}\sum_{i=1}^{|P|} (\text{bias\_score}(p_i))^2 \tag{5}$$

Model	TGSI			
	English	Hindi	Marathi	Telugu
OpenHathi	0.63	0.41	✗	✗
Airavata	**0.41**	0.37	✗	✗
Sarvam-1 †	0.47	0.44	0.21	0.24
Nanda	0.57	**0.18**	✗	✗
BharatGPT †	0.53	0.35	0.23	0.19
Navarasa-2.0	0.62	0.25	0.27	**0.15**
Krutrim-1	0.74	0.26	0.12	0.29
Krutrim-2	0.55	0.20	**0.07**	0.22

Table 2: TGSI value for different languages. †: Bad generation quality.

A higher *EGSI* indicates that the model M exhibits a stronger tendency to favour one gender over another in gender neutral contexts.

4 RESULTS

In this section, we present the findings from our experiments evaluating gender bias in LLM-generated text across English, Hindi, Marathi, and Telugu. We analyze the results from both Token-based Evaluation (Experiment 1) and Embedding-based Evaluation (Experiment 2) to understand how different models exhibit gender bias in their outputs. The stereotype indices *TGSI* and *EGSI* in Experiment 1 and Experiment 2 respectively measures the degree to which the model aligns to gender stereotypes in respective experiments. Higher value of Stereotype Index represents stronger gender bias of the models.

4.1 Experiment 1: Token-based Evaluation

The Token-based Gender Stereotype Index (TGSI) was calculated for each model based on gendered word frequencies in responses to gender-neutral prompts. Figure 3 shows the TGSI value heatmap of all the chosen models with few selected professions for Hindi. Among the chosen professions, for नर्स, ब्यूटीशियन, we see that almost all the models shows strong bias towards the female gender. This indicates that the models during the story generation, are more likely to associate these professions with women. For remaining professions, almost all the models shows bias towards masculine gender. These professions are often percieved as male-dominated, and the models reflect that bias by using more masculine language.

Table 2 shows the Token-based Gender Stereotype Index (TGSI) values for each model and language. TGSI measures how much a model prefers male or female words when it is given a gender-neutral prompt.

For English, most models have high TGSI values (greater than 0.5), Even the best-performing model, Airavata, has a TGSI of 0.41, which is not a big improvement over others. In contrast, for Hindi (lower part of the table), the TGSI values are much lower, around 0.2. This suggests that models like **Navarasa-2.0**, **Krutrim-1**, and **Krutrim-2** are newer and may have been fine-tuned better using carefully selected training data.

For Marathi and Telugu, some models have a ✗ mark, which means they are not trained in these languages. For example, OpenHathi, Airavata, and Nanda are trained only for Hindi and English. Among the models that do support Telugu, **Navarasa-2.0** gives the best results and is also more commonly used

Model	EGSI (x1000)			
	English	Hindi	Marathi	Telugu
OpenHathi	9.7	7.5	✗	✗
Airavata	10.7	7.2	✗	✗
Sarvam-1 †	13.6	3.4	7.1	3.8
Nanda	8.0	5.8	✗	✗
BharatGPT †	12.2	11.1	7.1	9.5
Navarasa-2.0	4.3	**1.1**	**1.1**	**3.4**
Krutrim-1	**2.9**	4.2	3.9	6.0
Krutrim-2	4.5	2.9	3.0	4.9

Table 3: EGSI value for different languages. †Bad generation quality

for Telugu tasks.

Overall, the results clearly show that the models still carry gender bias, especially for professions that are traditionally linked to one gender. For most of the 102 professions we tested, the TGSI values are positive (see Figure 3), which means there is a general bias towards male words. This male bias is stronger for jobs that are typically seen as male-dominated.

4.2 Experiment 2: Embedding-based Evaluation

Table 3 displays the Embedding-based Gender Stereotype Index (EGSI) for the same models and languages. Unlike TGSI, EGSI checks how similar the model's neutral response is to male or female responses in embedding space. A higher EGSI means that the model often drifts toward either the male or female direction, even when the prompt is gender-neutral.

Embedding-based evaluation depends on semantic similarity rather than exact tokens. So, models that are trained with balanced data across male and female terms perform better in this kind of evaluation. **Navarasa-2.0** performs consistently well not just for Indic languages but also for English. Other models like Krutrim also show similar performance and come close to the best model. On the other hand, models like BharatGPT and Sarvam-1 have higher EGSI values, which means they show more bias at the semantic level. Also, their generation quality is not very good. This could be because, as smaller models, they have limited capacity to understand and retain complex context and subtle semantic differences. Additionally, their training data may contain more noise or less gender-balanced examples, which affects both their performance and fairness.

Both TGSI and EGSI values show that no model is completely free from gender bias. Some models may appear more neutral in token usage (low TGSI) but still carry implicit bias in semantics (high EGSI), and vice versa. Models fine-tuned with specific focus on Indic languages (like Navarasa-2.0) tend to handle gender better than general-purpose multilingual models.

5 Qualitative Observations of Model Behavior

In addition to the quantitative evaluation of gender bias, we conducted a qualitative analysis of model behavior across several multilingual and fine-tuned language models. This analysis highlights notable issues related to language understanding, generation quality, and unexpected outputs.

- **Nanda Model:** Although prompts were given in English and the model was explicitly instructed to respond in English, it frequently generated responses in Hindi. This suggests a strong bias toward Hindi generation, likely due to the model being pretrained only on English and Hindi.
- **Smaller Models (BharatGPT-3B and Sarvam-1 2B):** These models had difficulty understanding prompts that were longer than a single sentence. The quality of generation was also generally poor, which can be attributed to their smaller size and limited context window.
- **OpenHathi (Base Version):** Regardless of whether the prompt was in Hindi or English, the model often responded with a mixture of Hindi, English, and Hinglish within the same output. This is likely because the model was fine-tuned on separate datasets in English, Hindi, and Hinglish, without sufficient control over the output language.
- **Krutrim-1 and Sarvam-1:** These models occasionally produced Chinese characters or phrases unexpectedly in the middle of responses. This unusual behavior may be due to noise or contamination in the training data, or lack of proper decoding constraints.
- **BharatGPT (Marathi):** When tested on Marathi prompts, BharatGPT showed very poor generation quality. The output contained garbled text and strange special characters, indicating either insufficient training data for Marathi or problems in tokenization for the language.

These observations highlight the challenges faced by current multilingual and fine-tuned models in maintaining language consistency and quality across different prompts. Although these issues are separate from gender bias, they are important to consider when evaluating and deploying these models in real-world multilingual applications.

6 Discussion and Conclusion

This study systematically evaluated gender bias in popular Indic language models through both token-level and embedding-level analysis. Our experiments revealed that despite recent advancements in fine-tuning and instruction tuning, large language models continue to exhibit measurable gender bias, especially in Indian languages like Hindi, Marathi, and Telugu.

Our findings show that most models tend to default to masculine expressions in response to gender-neutral prompts, especially for professions traditionally seen as male-dominated. Models like Navarasa-2.0 and Krutrim-2 showed relatively better balance in both token usage and semantic alignment, likely due to more targeted fine-tuning. However, the embedding-based evaluation revealed that even seemingly neutral responses often aligned more closely with male references, indicating deeper, implicit bias. Additional observations such as language mixing, poor Marathi and Telugu support, or unstable outputs from smaller models further suggest that model size, training data quality, and decoding strategies also influence fairness and usability.

In conclusion, while some Indic models show promise in reducing explicit gender bias, no model was found to be entirely free from stereotype. Token-level neutrality does not always translate to unbiased meaning, and deeper semantic patterns still reflect societal imbalances. For real-world use, especially in linguistically diverse and sensitive domains like India, it is important not only to benchmark bias through such systematic evaluations but also to explore active mitigation strategies such as balanced finetuning, adapter-based debiasing, and domain-specific alignment.

7 Future Work

While this study provides a detailed evaluation of gender bias in current Indic language models, it also opens up several directions for future work. First, the experiments focused on binary gender representation (male and female) due to the linguistic markers available in Hindi, Marathi, and Telugu. Future research should extend this framework to account for non-binary gender expressions, which are increasingly relevant in modern social discourse and require careful linguistic adaptation in Indian languages.

Another key direction is to strengthen the current evaluation methods. While the token-based and embedding-based approaches gave us useful insights, they still have limitations. For example, they rely on detecting explicit gender markers, which works well in gendered languages like Hindi, Marathi, and Telugu. However, many Indian languages such as Bengali or Assamese are largely non-gendered and do not mark gender clearly in grammar. For such languages, our current methods may not be effective. Future work should aim to design new strategies that go beyond

verb endings or pronouns—such as analyzing sentence framing, role assignments, or stereotypical associations present in the generated text.

Overall, future work will aim at making bias detection in language models more inclusive, scalable, and suitable for the full range of Indian languages and gender identities.

8 Acknowledgement

We sincerely thank Prof. Chiranjib Bhattacharyya for their continuous support, guidance, and constructive feedback during the course of this project. We have used ChatGPT to help improve the language and style of this document. Additionally, we used it to summarize a few documents, but we made sure to verify the accuracy of the summaries provided by ChatGPT.

References

M. Bartl, M. Nissim, and A. Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias, 2020. URL https://arxiv.org/abs/2010.14534.

R. Binns. Fairness in machine learning: Lessons from political philosophy, 2021. URL https://arxiv.org/abs/1712.03586.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. URL https://arxiv.org/abs/1607.06520.

S. Deode, J. Gadre, A. Kajale, A. Joshi, and R. Joshi. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163, Hong Kong, China, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.16/.

S. Due, S. Das, M. Andersen, B. P. López, S. A. Nexø, and L. Clemmensen. Evaluation of large language models: Stem education and gender stereotypes, 2024. URL https://arxiv.org/abs/2406.10133.

S. Dwivedi, S. Ghosh, and S. Dwivedi. Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning. *Rupkatha Journal*, 15(4):10, 2023. doi: 10.21659/rupkatha.v15n4.10. URL https://doi.org/10.21659/rupkatha.v15n4.10.

M. Gira, R. Zhang, and K. Lee. 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, 2022. URL https://aclanthology.org/2022.ltedi-1.8/.

H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *NAACL*, 2019.

G. Gupta, K. Ramesh, and S. Singh. Evaluating gender bias in hindi-english machine translation, 2021. URL https://arxiv.org/abs/2106.08680.

R. Hada, S. Husain, V. Gumma, H. Diddee, A. Yadavalli, A. Seth, N. Kulkarni, U. Gadiraju, A. Vashistha, V. Seshadri, and K. Bali. Akal badi ya bias: An exploratory study of gender bias in hindi language technology, 2024. URL https://arxiv.org/abs/2405.06346.

J. Hewitt, J. Thickstun, C. D. Manning, and P. Liang. Backpack language models, 2023. URL https://arxiv.org/abs/2305.16765.

K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, and S. A. Hale. Indian-bhed: A dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 231–239. ACM, Sept. 2024. doi: 10.1145/3677525.3678666. URL http://dx.doi.org/10.1145/3677525.3678666.

S. Khosla. Investigating cross-linguistic gender bias in hindi-english across domains, 2021. URL https://arxiv.org/abs/2111.11159.

N. Kirtane and T. Anand. Mitigating gender stereotypes in hindi and marathi, 2022. URL https://arxiv.org/abs/2205.05901.

N. Kirtane, V. Manushree, and A. Kane. Efficient gender debiasing of pre-trained indic language models, 2022. URL https://arxiv.org/abs/2209.03661.

H. Kotek, R. Dockum, and D. Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24. ACM, Nov. 2023. doi: 10.1145/3582269.3615599. URL http://dx.doi.org/10.1145/3582269.3615599.

A. Kumar, S. Yunusov, and A. Emami. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 375–392, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.23. URL https://aclanthology.org/2024.acl-long.23/.

P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. Towards debiasing sentence representations, 2020. URL https://arxiv.org/abs/2007.08100.

C. Ma, T. Zhao, and M. Okumura. Debiasing large language models with structured knowledge. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10274–10287, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.612. URL https://aclanthology.org/2024.findings-acl.612/.

V. Malik, S. Dev, A. Nishi, N. Peng, and K.-W. Chang. Socially aware bias measurements for hindi language representations. In *North American Chapter of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:239009591.

N. Meade, E. Poole-Dayan, and S. Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.132. URL https://aclanthology.org/2022.acl-long.132/.

A. K. Pujari, A. Mittal, A. Padhi, A. Jain, M. K. Jadon, and V. Kumar. Debiasing gender biased hindi words with word-embedding. *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:211104718.

N. R. Sahoo, P. P. Kulkarni, N. Asad, A. Ahmad, T. Goyal, A. Garimella, and P. Bhattacharyya. Indibias: A benchmark dataset to measure social biases in language models for indian context, 2024. URL https://arxiv.org/abs/2403.20147.

A. Sant, C. Escolano, A. Mash, F. D. L. Fornaciari, and M. Melero. The power of prompts: Evaluating and mitigating gender bias in mt with llms, 2024. URL https://arxiv.org/abs/2407.18786.

T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021. URL https://arxiv.org/abs/2103.00453.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review, 2019. URL https://arxiv.org/abs/1906.08976.

C. Treude and H. Hata. She elicits requirements and he tests: Software engineering gender bias in large language models, 2023. URL https://arxiv.org/abs/2303.10131.

UNESCO. I'd blush if i could: Closing gender divides in digital skills through education. UNESCO Report, 2019. Accessed: 2025-01-06.

A. Vashishtha, K. Ahuja, and S. Sitaram. On evaluating and mitigating gender biases in multilingual settings, 2023. URL https://arxiv.org/abs/2307.01503.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003/.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings, 2019. URL https://arxiv.org/abs/1904.03310.