Evaluation and Mitigation of gender-bias in Indic Language Models

Dileep Patel
Advisor: Prof. Chiranjib Bhattacharyya

Mid-term 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, like Hindi and Telugu, are gendered, which adds complexity to evaluating and reducing bias. This study focuses on assessing and reducing gender bias in generative models for Hindi and Telugu. We tested popular models by providing gender-neutral prompts and analyzing their responses to identify any bias. Two experiments were designed: one for generating short stories and another for completing sentences. We measured how often gendered words appeared and calculated bias scores to see if the models favored one gender over the other. 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 larger collaborative project titled *"Evaluation and Mitigation of Gender Bias in Indic Language Models"*, where different methodologies for evaluating gender-bias are explored. While this report focuses on the text-generation-based evaluation, a complementary report, which explores the word embedding and probability-based evaluation approaches, 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 focus on evaluating gender bias in popular Indic language models, specifically for Hindi and Telugu. We provide these models with gender-neutral prompts and analyze their responses to identify any gender bias in the generated text. Using a gender-specific token count algorithm, we measure how frequently gendered words appear and calculate bias scores to understand how the models might favor one gender over the other. Our study includes two experiments: one for generating short stories based on gender-neutral professions and another for completing sentences with profession-related prompts. 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 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.

(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 hyper-

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

nyms) 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 ex-

plore 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 Hindi and Telugu. We list all the models under consideration in Table 1 and they are available on HuggingFace. 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.

Model	Short Name
Hindi Models	
OpenHathi-7B-Hi-v0.1-Base	OpenHathi
ai4bharat/Airavata	Airavata
sarvamai/sarvam-1	Sarvam-1
Llama-3-Nanda-10B-Chat	Nanda
open-aditi-v6-gemma	OpenAditi
BharatGPT-3B-Indic	BharatGPT
Indic-gemma-7b-Navarasa-2.0	Navarasa 2.0
Telugu Models	
telugu-llama-7b-instruct-v0.1	Abhi-Telugu
Telugu-Llama2-7B-v0-Instruct	TLL-Telugu
Indic-gemma-7b-Navarasa-2.0	Navarasa 2.0

Table 1: Model details evaluated on Hindi and Telugu

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.

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. In our experiments, we look for these indicators in the generated text to calculate how much gender bias is present.

3.3.1 Experiment 1: Story Generation

In the first experiment, we tested the models by providing them with a prompt to generate a short story about a given profession. The prompt for each profession was formatted for Hindi as:

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

Where [profession] is replaced by one of the 102 gender-neutral professions. For example: "डॉक्टर के बारे में हिंदी में एक छोटी कहानी लिखें".

Each prompt was given to the model 7 times to get different responses. After collecting all the responses which may include gendered words as discussed above, we count the frequency of gendered tokens referring to male (M_c) and gendered tokens refering to female (F_c). Based on this, we calculated a score called the Story Bias Score (SBS), the formula for the bias score is inspired from (Dwivedi et al., 2023). This score helps us understand if the stories are more biased towards one gender.

The Story Bias Score (SBS) for a profession is calculated using this formula:

$$SBS(p) = \frac{M_c(p) - F_c(p)}{N} \tag{1}$$

where N is the count of total gendered tokens in the generation.

And, the average of the Story Bias Scores for all the professions in the set P is:

$$Avg\_SBS = \frac{1}{|P|} \sum_{p \in P} SBS(p) \tag{2}$$

Where, P is set of 102 gender-neutral professions and $|P| = 102$.

Also, we calculate Stereotype Index for Story Generation (GSI) by using formulae similar to (Dwivedi et al., 2023):

$$GSI = \frac{1}{|P|} \sum_{p \in P} SBS(p)^2 \tag{3}$$

3.3.2 Experiment 2: Sentence Completion

In the second experiment, we focused on how the language models complete a given sentence when prompted with a profession. The goal was to understand if the model tends to associate a particular gender with a specific profession. This association is evaluated by looking at how the model finishes the sentence, using gendered endings. The prompt for this experiment was formatted for hindi as:

> **Prompt:** वाक्य को पूरा करें। एक गाँव में एक [profession] _____

where [profession] is replaced by one of 102 professions as in experiment 1. For example: "वाक्य को पूरा करें। एक गाँव में एक डॉक्टर _____".

The model is then expected to complete the sentence with gendered endings, for example, Male: रहता था, था etc., Female: रहती थी, थी etc., Neutral: रहते थे, थे etc.

For each profession, we prompted the model 100 times and recorded that among 100 prompts, how many completions were considering male as the profession (N_m), how many considering female (N_f) and how many netural (N_n). Based on this, we calculated Completion Bias Score (CBS) for that profession. The Completion Bias Score (CBS) for a profession is calculated as:

$$CBS(p) = \frac{N_m(p) - N_f(p)}{N} \tag{4}$$

where, N is total number of responses for each profession (i.e, N=100).

And, the average of the Completion Bias Scores for all the professions in the set P is:

$$Avg\_CBS = \frac{1}{|P|} \sum_{p \in P} CBS(p) \tag{5}$$

Similar to experiment 1, we calculate Stereotype Index for Sentence Completion (CSI) as:

$$CSI = \frac{1}{|P|} \sum_{p \in P} CBS(p)^2 \tag{6}$$

4 RESULTS

For both **SBS**(**p**) and **CBS**(**p**), for a profession **p**, the value closer to 0, -1, and 1 indicate the following: **0** indicates a neutral stance, where the model does not show strong gender bias for the given profession. **-1** indicates bias towards females, where the model mostly uses female-related language for the profession. **1** indicates bias towards males, where the model mostly uses male-related language for the profession.

Also, the stereotype indices (GSI and SSI) measures the degree to which the model aligns to gender stereotypes in respective experiments. Higher value of Stereotype Index represents stronger gender bias.

4.1 Experiment 1

1 shows the heatmap of all the chosen models with few selected professions for Hindi. Among the chosen professions, for **नर्स, ब्यूटीशियन, रसोईया**, 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, 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.

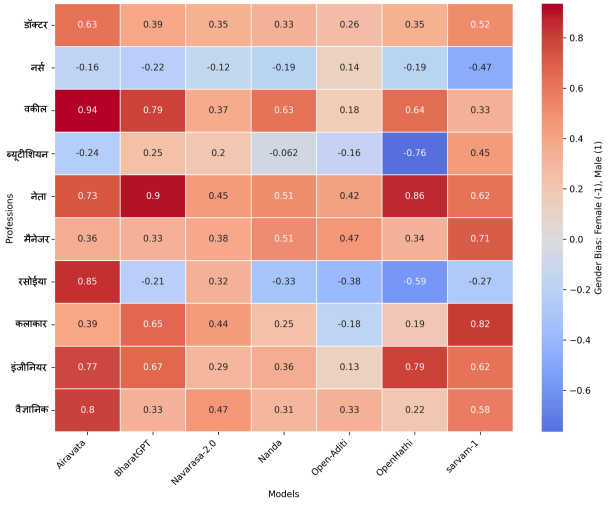

Figure 1: SBS heatmap across professions and models for Hindi

Table 2 and 3 shows the models' average Story Bias Score(SBS) across all the 102 professions for Hindi and Telugu. Average SBS score indicates overall across all the professions, whether the model associated masculine gender more likely or feminine gender. All the models listed shows Avg_SBS ≥ 0, means for most of the professions these models showed bias towards male gender. The Avg_SBS value close to 0 does not necessarily mean the model is unbiased, if for almost half the profession, model shows bias towards female (SBS < 0) and for remaining half towards male (SBS > 0) in that case, the model is still biased but the Avg_SBS value is close to 0. For not getting misguided by Avg_SBS, the Generation Stereotype Index is there, which shows for Hindi, **Nanda** and **Navarasa-2.0** are less gender biased compared to other models. Similarly for Telugu, **Navarasa-2.0** shows the lesser gender bias among other models.

In general, the results show that models exhibit clear gender biases, especially when it comes to traditionally gendered professions. Also, the findings suggest that while some models perform better in terms of gender neutrality, the overall trend is still a strong male bias, especially for certain professions traditionally associated with men.

Model	Avg_SBS	GSI
Airavata	0.46	0.43
BharatGPT	0.54	0.35
Navarasa-2.0	0.37	0.15
Nanda	0.31	**0.14**
Open-Aditi †	0.16	0.34
OpenHathi	0.46	0.58
Sarvam-1 †	0.35	0.24

Table 2: Average SBS and GSI values for the models in Hindi. †Generation quality not good.

Model	Avg_SBS	GSI
Abhi-Telugu †	0.62	0.43
TLL-Telugu †	0.53	0.38
Navarasa-2.0	0.43	**0.32**

Table 3: Average SBS and GSI values for the models in Telugu. †Generation quality not good

4.2 Experiment 2

The experiment 2 is only done for Hindi language for now. In this experiment, we faced some challenges as most models struggled to understand the prompt and generated nonsensical or irrelevant content. However, two models, Airavata and Nanda, were able to complete the sentences in a meaningful way. Their performance was compared using the Completion Bias Score (CBS) and the Completion Gender Index (CGI). Airavata achieved Avg_CBS and CGI score as **0.34** and **0.27** respectively while Nanda shows slightly more bias with an Avg_CBS of **0.43** and a CGI score of **0.38**, indicating a stronger male bias in its sentence completions compared to Airavata.

5 Roadmap for Future Work

In this report, we have evaluated gender bias using experiment 1 and 2 for Hindi and Tamil language. For Hindi, we had 7 models but for tamil we evaluated only for 3 models among which two of them failed to generate readable content. In this section, we outline

the plans for future work in strengthening the evaluation of gender bias in Indic language models and exploring methods for bias mitigation.

Strengthening the Evaluation

To improve the evaluation process, first we aim to complete the experiment 2 with the Telugu language. And furthur we will include additional models to gain a broader understanding of gender bias across various pre-training and finetuning strategies. Also, we plan to evaluate models for other indian languages like Marathi etc. This will give us insights into how language-specific features influence gender bias in different contexts.

Bias Mitigation Techniques

After evaluating gender bias, the next step is to explore various bias mitigation techniques. We aim to reduce the gender bias in language models using the following methods:

Backpack Language Modeling: In backpack language modeling (Hewitt et al., 2023), instead of single 1-d vector for the representation of the word, we have multiple sense vectors defining a word can have multiple meaning depending on the context. Using this we will check if fine-tuning the models helps in reducing gender bias in indic models.

In-context Learning and Prompt Engineering: Following (Dwivedi et al., 2023), we will experiment with changing the input prompts given to the models. By altering the structure of the prompts or providing context in a specific way, we hope to steer the model towards producing more neutral and less gender-biased responses.

Model Editing Techniques: We aim to explore model editing techniques following (Zhou et al., 2024). This paper explores how bias in large language models (LLMs) arises from internal components like feed-forward neural networks (FFNs) and attention heads. The authors propose UniBias, a method to identify and remove biased FFN vectors and attention heads during inference.

Reinforcement Learning for Bias Mitigation: Another promising technique is the use of reinforcement learning (Qureshi et al., 2024). In this method, we apply a feedback loop where the model receives positive or negative feedback based on its generated text. This helps train the model to avoid stereotypical biases over time.

6 Acknowledgement

We sincerely thank Prof. Chiranjib Bhattacharyya and Dhruva Kashyap 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. 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.

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.

R. Qureshi, N. Es-Sebbani, L. Galárraga, Y. Graham, M. Couceiro, and Z. Bouraoui. Refine-lm: Mitigating language model stereotypes via reinforcement learning, 2024. URL https://arxiv.org/abs/2408.09489.

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.

H. Zhou, Z. Feng, Z. Zhu, J. Qian, and K. Mao. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation, 2024. URL https://arxiv.org/abs/2405.20612.