Gender Bias in Indic LLMs

Asad Ali

January 10, 2025

Mid-term MTech Project Report

Abstract

Gender bias involves unfair treatment or discrimination against people based on their gender. This bias can manifest in texts through the use of gender-specific language that might incorrectly assign certain traits to individuals solely on the basis of gender. Such language usage contributes to reinforcing stereotypes, which in turn perpetuate existing gender inequalities throughout society [1]. As natural language processing (NLP) technologies become a part of our daily lives, addressing gender bias in these systems becomes essential. Extensive research for analyzing and mitigating gender bias has been done for English language, efforts for Indic languages are still at an early stage. This project focuses on evaluating and reducing gender bias in generative models designed for Indian languages. We'll analyze the word embeddings and word likelihood given the context in Indic LLMs. The findings of this study will help in improving the fairness and inclusivity of NLP systems for Indic languages, contributing to the larger goal of equitable AI development 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"[1], where different methodologies for evaluating gender-bias are explored. While this report focuses on the text-generation-based evaluation, a complementary report, which analyzes the model generations for obtaining a bias score, can be accessed through github. Both reports contribute to a comprehensive understanding of gender bias in Indic language models.

1 INTRODUCTION

A Large Language Models(LLMs) are a class of sophisticated AI models trained on massive textual datasets(like books, articles, websites), capable of understanding and generating human language. With the rising demands of AI integration with industries and businesses to enhance productivity, many powerful LLMs like ChatGPT, Gemini, LLAMA, Claude, etc. are continously trained by their respective organizations and rolled out to meet their demands. However, there is a large section of individuals who cannot benefit from these technologies directly simply because most of these LLMs work in English. LLMs in local languages are crucial to promote cultural inclusivity, improve access to information, and ensure that AI benefits all communities equally. In Indian context, many Indic LLMs are released like OpenHathi, Airavata, Sarvam-1, Krutrim AI, etc.

However, content generated by large language models (LLMs) is not without its challenges, particularly when it comes to issues of harm and bias. Even with extensive and diverse training datasets, English language generation has been found to exhibit various biases, including gender bias. For instance, Bolukbasi et al. (2016)[2] demonstrated that word embeddings trained on large corpora often encode and amplify societal stereotypes, such as associating professions with specific genders. Many researches[2][3][4][5] have been performed for English languages for analysis and mitigation of gender bias but there is very little work on Indic languages. 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

[1]Link to project: https://github.com/dileepp982/Gender_Bias

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'll analyze the initial word embeddings using two prominent analysis methods, i.e. Word Embedding Association Test(WEAT)[3] and Relative Norm Difference(RND)[6]. Thereafter, we'll analyze the log-likelihood scores of generating stereotypical or anti-stereotypical sentences using All Unmasked Likelihood(AUL)[4] method for encoder based models mBERT, mURIL, XLM-RoBERTa,etc. and Conditional Log-Likelihood(CLL)[7] method for decoder based models like OpenHathi, Airavata, Sarvam-1,etc.

This study is important because it highlights how AI systems can impact fairness and equality in diverse societies like India. By addressing gender bias, we can make AI tools more inclusive, reliable, and trusted by everyone. It also helps in building better technologies that do not reinforce stereotypes but instead support a society where everyone has equal opportunities.[2]

2 Literature Survey

Gender bias in Large Language Models (LLMs) has been extensively studied, and research has demonstrated how these models reinforce and magnify the gender biases present in society. The paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"[2] is the first comprehensive study on gender bias in word embeddings. It reveals that embeddings like Word2Vec capture and amplify societal stereotypes, such as associating men with professions and women with domestic roles. The authors proposed neutralize and equalize method to reduce this bias by identifying and neutralizing a gender subspace while preserving semantic relationships, paving the way for fairness in AI systems. Thereafter many other metrics emerged to analyze gender bias in static word embeddings like Word Embedding Association Test(WEAT)[3] and Relative Norm Difference(RND)[6]. The WinoBias dataset[8] highlights biases in job-related terms and advocates for debiasing techniques like gender-swapping and unbiased word vectors, which effectively reduce bias without affecting accuracy. These works were extended for specifically studying BERT-like models, proposing new ways to quantify bias and explored mitigation strategies tai- lored for these advanced models[9][10] These works 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. It's especially useful in downstream tasks like sentiment analysis or question answering. After ChatGPT gained popularity, a study[11] demonstrated that large models like GPT often associate professions like "doctor" with men and "nurse" with women. Another study[12] found similar patterns, with male- associated tasks described as technical and female- associated tasks described as administrative.

Meade et al. (2022)[13] reviewed strategies to reduce gender bias, such as dropout during training, fine-tuning with specific goals, and Sentence Debiasing, which adjusts model outputs to remove bias. Similarly, Gira et al. (2022)[14] proposed fine-tuning only a part of GPT models to reduce bias efficiently. Liang et al. (2020)[15] introduced Sent-Debias to remove biases in sentence-level representations while keeping good performance in tasks like sentiment analysis. Hewitt et al. (2023)[16] introduced "sense vectors" to represent different meanings of a word, making language models more interpretable. By adjusting these vectors, they reduced gender bias in outputs. Ma et al. (2024)[17] used structured knowledge, like hypernyms, in an additional training phase to reduce bias in LLMs without retraining the entire model. Sant et al. (2024)[18] studied how carefully designed prompts can reduce gender bias in machine translation. They found a 12% reduction in bias on the WinoMT dataset using specific input phrasing.

Although progress has been made for English, work on Indic languages like Hindi remains limited. Pujari et al. (2019)[3] studied gender bias in Hindi text using SVM classifiers and found stereotypical associations of professions with genders. Gupta et al. (2021)[19] measured bias in Hindi-English machine translation using the Translation Gender Bias Index (TGBI), showing that translations often reinforce gender roles. Khosla et al. (2021)[20] explored gender bias between Hindi and English in different domains and showed how bias varies across fields. Kirtane et al. (2022)[21] tackled bias in Hindi and Marathi NLP tasks, creating datasets of gendered and neutral words. They used

metrics like the Embedding Coherence Test (ECT) and Relative Norm Distance (RND) to measure and reduce bias. They also developed a metric for occupational gender bias in Hindi and proposed fine-tuning methods to reduce it.

Malik et al. (2021)[22] analyzed biases in Hindi models related to gender, caste, and religion, highlighting how cultural context affects bias detection. Vashishtha et al. (2023)[23] extended the DisCo metric for six Indian languages and improved debiasing methods like Counterfactual Data Augmentation and Self-Debiasing. Khandelwal (2024)[7] studied biases in LLMs like GPT-3.5 using the Indian-BhED dataset, showing significant caste and religion biases. Sahoo et al. (2024)[24] introduced IndiBias, a dataset tailored for Indian social biases, covering gender, caste, religion, and more. They tested ten language models and found biases across multiple groups. Hada et al. (2024)[25] conducted a detailed study on gender bias in Hindi language technology. They combined data analysis and field studies, involving rural women, to understand diverse views on gender bias. Their work emphasized the need for context-specific approaches to reduce bias effectively.

These studies highlight increasing awareness of gender bias in Hindi language models and ongoing efforts to address it. Tackling these issues is crucial for building fair and inclusive AI systems that respect diverse languages and cultures.

3 Methodology

This section outlines the methodology used to analyze gender bias in the outputs of Hindi language models, specifically OpenHathi-7B-Hi-v0.1-Base, ai4bharat/Airavata, sarvamai/sarvam-1, and Llama-3-Nanda-10B-Chat. The analysis involved collecting data, examining the word embeddings of Hindi LLM models using the Relative Norm Difference method and WEAT analysis, and using log likelihood methods like All Unmasked Likelihood(AUL) score and Conditional Log Likelihood(CLL) for encoder and decoder based models respectively.

3.1 Word Embedding Association Tests

3.1.1 Data Collection for WEAT Analysis

The data collection process for conducting the Word Embedding Association Test (WEAT) involves creation of **Target and Attribute Word Sets:**

- *Target Words:* Two sets of contrasting concepts are identified. For gender bias analysis, target sets might include {man, male, he, him} and {woman, female, she, her}.
- *Attribute Words:* Two sets related to stereotypical associations being studied are selected. For instance, a career vs. family bias analysis may use {engineer, scientist, lawyer, doctor} and {homemaker, nurse, teacher, caregiver}.

This dataset creation method is motivated from method of data collection used in HindiWEAT[22].

3.1.2 Model Embeddings Extraction

For Word Embedding Analysis,we extracted the 4096-dimensional non- contextualized word em- beddings from the hindi LLMs using huggingface hooks. Since these models use subword tokenizers, we averaged the embeddings of subtokens using mean pooling to obtain the word embeddings. This is only possible for open source Huggingface models since word embeddings for closed source models like Krutrim AI, Hanooman.ai isn't released publically yet.

3.1.3 WEAT evaluation

The WEAT metric proposed by Caliskan et. al.(2017)[3] is used to measure biases in word embeddings related to Gender, Caste, Religion, and Occupation etc. Let X and Y be two equal-sized sets of target words' embeddings, and A and B be two sets of attribute words' embeddings. Then the test involves comparing the cosine similarities between these embeddings. A larger WEAT score indicates a larger bias.The formula for calculating WEAT is:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where the quantity, s(w, A, B) is calculated using cosine similarity as follows:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b)$$

The amount of bias in WEAT is analyzed by effect size d calculated as:

$$d = \frac{\text{mean}_{x \in X}\ s(x, A, B)\ -\ \text{mean}_{y \in Y}\ s(y, A, B)}{\text{stddev}_{w \in X \cup Y}\ s(w, A, B)}$$

To compute the test significance (p-value), Let {Xi , Yi} denote all the partitions of X Y into two sets of equal size. So,here 10,000 permutations, P of the combined list is generated. For the i-th list in P, it is split in new pairs of Xi and Yi lists. Then We calculate the p-value using the test statistics as:

$$p = \frac{1}{|P|} \sum_{i \in P} [s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

3.2 Relative Norm difference(RND)

3.2.1 Data Collection for RND Analysis

For the analysis of word embeddings using the Relative Norm Difference metric, dataset includes:

- List of gender neutral professions
- Pair of lists containing gender specific nouns and verbs for male(e.g. पिता, चलता, था) and female(e.g. बेटी, पढ़ती, थी) genders.
- Optionally, we can also include pair of axes like income, caste, etc. for comparing professions in that axes.

3.2.2 Model Embeddings extraction

For each of the word in pair of lists as mentioned in section3.2.1, and the professions we've to obtain the word embeddings for analysis. This can be done in the same way as we did in WEAT analysis.3.1.2

3.2.3 RND evaluation

In RND,the objective is to average the embedding vectors within the target set T, and for every attribute a A, the norm of the difference between the average target and the attribute word is calculated, and subsequently subtracted. The formula for calculating RND is:

$$\sum_{x \in A} (\|\text{avg}(T1) - x\|_2 - \|\text{avg}(T2) - x\|_2)$$

The higher the value of the relative distance from the norm, the more associated the attributes are with the second target group(here female), and vice versa. We have used the visual interactive tool WordBias[6], for exploring biases against different intersectional groups(gender,religion,etc.) encoded in word embeddings.

3.3 Log Likelihood based methods

There are many methods for quantifying gender bias in encoder based model using this method like DisCo[5], CrowS-Pairs Score[26],etc. but its uncommon to find one that works for decoder based models[7].

3.3.1 Dataset Creation

For experiments we've used nearly 100 pairs of text inspired from CrowS pairs method[26]. The format of sentences is:

"MASK जब कड़ी मेहनत MASK हैं तो उन्हें भूख लगती है "

We'll have a pair of stereotypical and non-stereotypical list of words which can be inserted in above sentence to complete it.

- *Stereotypical*: ['पुरुष','करते']
- *Anti-Stereotypical*: ['महिलाएं','करती']

3.3.2 Bias Computation

We compute the percentage of instances where the model is more likely to generate the stereotypical version of a sentence compared to its anti-stereotypical counterpart. To calculate the difference in likelihoods, we first determine the log-likelihoods of generating a sentence for the encoder and decoder models, accounting for variations in the relative base frequencies of the words being swapped.

1. *Encoder Based Models:* For Encoder Based Multilingual Models like mBERT, mURIL, XLM-RoBERTa, etc. we're using All Unmasked Likelihood(AUL) score[4]. This metric eliminates measurement biases caused by word frequency and input contexts, which were present in DisCo and Crow-S by looking at entire sentence at once instead of one-by-one masking. AUL score for a sentence S is given as:

$$AUL(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \log P(w_i \mid S; \theta)$$

where, |S| is length of sentence, $P(w_i \mid S; \theta)$ is the probability assigned to token w_i during MLM task conditioned on complete sentence S with pretraining parameter θ.

2. *Decoder Based Models* For deoder-based models like OpenHathi, Airavata, Sarvam-1,etc. we are using Conditional Log Likelihood (CLL)[7] method. The CLL score is computed as:

$$CLL(S|w) = logP(S_w|w; \theta) - log(w; \theta)$$

where, w is target stereotypical/antistereotypical word, S_w is the sentence containing target word w, and θ is the pretraining parameter. This metric solves the frequency bias issue of target words pointed out in Crow-S pairs[4].

4 Results

4.1 WEAT scores

Model	WEAT score
Open-Hathi	0.067(0.465)
Airavata	0.110(0.441)
Sarvam-1	-0.793 (0.939)
Nanda	0.301 (0.304)

Table 1: WEAT score for Maths,Arts vs Male,Female

The WEAT scores in Table 1 reveal varying levels of gender bias across the models:

- **Open-Hathi**: Shows a slight positive bias (0.067) associating math with male terms, with high variability (0.465).
- **Airavata**: Exhibits a similar slight positive bias (0.110) but with slightly more consistency (0.441).
- **Sarvam-1**: Demonstrates a strong negative bias (-0.793), associating arts with male terms and math with female terms, with high variability (0.939).
- **Nanda**: Shows the strongest positive bias (0.301) with low variability (0.304), indicating more consistent associations.

These results highlight the presence of stereotypes in the models, with variability suggesting inconsistent biases that require targeted debiasing techniques.

4.2 RND scores

Profession	OH	AV	S	N
इतिहासकार	-0.9968	-0.9168	-0.7642	-0.8926
नर्स	0.9903	0.1407	0.9800	0.7037
डॉक्टर	-0.9988	-0.8824	-0.7696	-0.8836
निरीक्षक	-0.8824	-0.3665	-0.7299	-0.6596
निर्देशक	-0.9287	-0.8000	-0.4966	-0.7418
प्रोफ़ेसर	-0.8270	-0.8304	-0.7402	-0.7992
वकील	-0.9996	0.6559	-0.8592	-0.4010

Table 2: RND Scores for Professions Across Models. OH:OpenHathi, A:Airavata, S: Sarvam-1, N:Nanda

The RND scores in Table 2 reveal consistent male bias (negative scores) for professions like इतिहासकार , डॉक्टर , and निर्देशक across all models, with OpenHathi (OH) showing the strongest bias. Female-associated professions, like नर्स , have positive scores, with Sarvam-1 (S) and OpenHathi (OH) displaying the strongest alignment. Nanda (N) generally balances biases, as seen in निरीक्षक and वकील . Overall, biases align with traditional gender stereotypes.

4.3 AUL scores

AUL scores are valid for encoder based models.

Model	AUL score
mBERT	45.79
XLMR	60.58
IndicBERT	55.48

Table 3: AUL score for encoder based models

The AUL scores in Table 3 indicate the performance of encoder-based models in avoiding undesirable or biased content, where scores closer to 50 are preferable:

- **mBERT**: With a score of 45.79, mBERT is the closest to the ideal value, indicating a balanced performance in avoiding bias.
- **XLMR**: Scored 60.58, deviating from the ideal range and showing a tendency to avoid bias but potentially at the cost of underrepresentation of certain contexts.
- **IndicBERT**: Scored 55.48, moderately close to the ideal, reflecting relatively balanced behavior in mitigating undesirable content.

These results suggest that mBERT performs best in achieving the desired balance, while XLMR may need fine-tuning to align closer to the optimal range.

4.4 CLL scores

CLL scores are valid for decoder based models.

Model	CLL score
Open-Hathi	58.08
Airavata	66.48
Sarvam-1	49.05
Nanda	51.08

Table 4: CLL score for decoder based models

The CLL scores in Table 4 indicate the performance of decoder-based models, where lower scores are preferable for reducing undesirable content:

- **Open-Hathi**: Scored 58.08, showing moderate performance but with room for improvement in controlling undesirable content.
- **Airavata**: Achieved the highest score (66.48), indicating a relatively higher likelihood of generating undesirable or biased content compared to other models.
- **Sarvam-1**: Recorded the lowest score (49.05), demonstrating the best performance in mitigating undesirable content among the models evaluated.
- **Nanda**: Scored 51.08, performing well and close to Sarvam-1, reflecting strong control over bias and undesirable content.

These results suggest that Sarvam-1 and Nanda are more effective in minimizing biased or undesirable content, while Airavata may benefit from additional fine-tuning or debiasing techniques.

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 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 [16], 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 [27], 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 [28]. This paper explores how bias in large language models (LLMs) arises from internal components like feedforward 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 [29]. 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

[1] Michela Menegatti and Monica Rubini. Gender bias and sexism in language, 09 2017.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016.

[3] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[4] Masahiro Kaneko and Danushka Bollegala. Unmasking the mask – evaluating social biases in masked language models, 2021.

[5] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models, 2021.

[6] Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings, 2021.

[7] Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott 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, September 2024.

[8] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda 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.

[9] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias, 2020.

[10] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review, 2019.

[11] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24. ACM, November 2023.

[12] Christoph Treude and Hideaki Hata. She elicits requirements and he tests: Software engineering gender bias in large language models, 2023.

[13] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In Smaranda Muresan, Preslav Nakov, and Aline 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.

[14] Michael Gira, Ruisu Zhang, and Kangwook Lee. Debiasing pre-trained language models via efficient fine-tuning. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations, 2020.

[16] John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang. Backpack language models, 2023.

[17] Congda Ma, Tianyu Zhao, and Manabu Okumura. Debiasing large language models with structured knowledge. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10274–10287, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[18] Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. The power of prompts: Evaluating and mitigating gender bias in mt with llms, 2024.

[19] Gauri Gupta, Krithika Ramesh, and Sanjay Singh. Evaluating gender bias in hindi-english machine translation, 2021.

[20] Somya Khosla. Investigating cross-linguistic gender bias in hindi-english across domains, 2021.

[21] Neeraja Kirtane and Tanvi Anand. Mitigating gender stereotypes in hindi and marathi, 2022.

[22] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for hindi language representations.

In *North American Chapter of the Association for Computational Linguistics*, 2021.

[23] Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. On evaluating and mitigating gender biases in multilingual settings, 2023.

[24] Nihar Ranjan Sahoo, Pranamya Prashant Kulkarni, Narjis Asad, Arif Ahmad, Tanu Goyal, Aparna Garimella, and Pushpak Bhattacharyya. Indibias: A benchmark dataset to measure social biases in language models for indian context, 2024.

[25] Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. Akal badi ya bias: An exploratory study of gender bias in hindi language technology, 2024.

[26] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[27] 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.

[28] Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation, 2024.

[29] Rameez Qureshi, Naïm Es-Sebbani, Luis Galárraga, Yvette Graham, Miguel Couceiro, and Zied Bouraoui. Refine-lm: Mitigating language model stereotypes via reinforcement learning, 2024.