Analyzing Gender Bias in Hindi Language Models

Arunima Roy, Asad Ali & Dileep Patel
Indian Institute of Science
Bengaluru, KA, India
arunimaroy@iisc.ac.in, asadali@iisc.ac.in, dileeppatel@iisc.ac.in

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 (Menegatti & Rubini, 2017). In this report, we analyze gender biases in Hindi language models to assess their effects on fairness and inclusivity in language technologies. Recent advancements have spotlighted biases in natural language processing, making it essential to evaluate the fairness of responses from large language models (LLMs). While research has predominantly focused on biases in English and related to Western societies, our study specifically examines gender-related biases in Hindi language representations. Newly developed state-of-the-art (SOTA) models like Krutrim AI, OpenHathi, and Airavata have been designed to enhance natural language processing in Indic languages. This project involves a systematic review of the prompts and responses from Krutrim, Airavata, and OpenHathi AI to identify and measure any biases present in their outputs in the Hindi language.

1 Introduction

Artificial Intelligence (AI) language models have greatly changed many fields, such as natural language processing and sociological research. Due to their ability to create text that resembles human writing and handle various natural language tasks, these models have significantly changed how language models are developed. In the context of Indic languages, LLMs can enhance accessibility to technology for Indic language speakers, enabling them to interact with devices and applications more naturally. This can lead to improved communication, content creation, and information access in Indic languages, ultimately promoting digital inclusion and cultural preservation. But, the problem arises if LLMs express biased assumptions about men and women, specifically those aligned with people's perceptions, rather than those grounded in fact. If a system frequently associates certain professions with a specific gender, this creates a representational harm by perpetuating inappropriate stereotypes about what activities men and women can or allowed or expected to perform, for e.g. making that there are less professional females in STEM (McGuire et al., 2021). When such representations are used in downstream NLP applications, there is an additional risk of unequal performance across genders (Gonen & Webster, 2020).

Our project[1] dives into this problem for Hindi and aims to make language models that are more fair and unbiased. This involves a systematic analysis of stereotypical associations between male

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

and female gender and professional occupations in responses from state-of-the-art(SOTA) Hindi LLM models. As part of these efforts, Krutrim AI (OLA, 2023), OpenHathi (Sarvam.ai, 2023) and Airavata (Gala et al., 2024) are recently unveiled SOTA language models, acknowledged as state-of-the-art by the Indic LLM leaderboard curated by Cognitive AI (CognitiveLab, 2024) designed to enhance natural language processing in Indic languages. Krutrim AI became India's first AI unicorn and has Generation capabilities for Hindi, English +8 Indian languages. OpenHathi is series of open Hindi language models and are soon to be launched as gen AI platform. Airavata is an instruction-tuned model for Hindi built by fine tuning OpenHathi with diverse, instruction-tuning Hindi datasets to make it better suited for assistive tasks. These models aim to improve language understanding and generation in Indic languages like Hindi, Bengali, Tamil, and others, addressing the specific linguistic challenges of these languages.

2 Related Work

In word embedding, words with similar meanings are often represented by vectors that are close together. These vectors often encode gender biases, as shown by Bolukbasi et al. (2016) who found stereotypes in embeddings from Google News. The WinoBias dataset by Zhao et al. (2018) 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. Pujari et al. (2020) utilized an SVM-based classifier to assess and debias gender-neutral words by modifying vector properties based on their distance from the decision surface. The Iterative Null-space Projection (INLP) method by (Ravfogel et al., 2020) aims to strip neural representations of biased information through repeated classifier training and projection techniques. (Hewitt et al., 2023) introduced Backpack Language Models, using sense vectors to lessen gender stereotyping in generated content.

For analyzing gender bias in model responses, a study (Qian et al., 2019) investigated bias alleviation in text generation by direct modification of the loss function, represented as Co-occurrence bias. A word(for example profession) is considered to be biased towards a certain gender if it occurs more frequently with words of that gender. That loss function attempts to equalize the probabilities of male and female words in the output. For analyzing gender bias in Hindi-English Machine Translation models, word embedding based metrics like Word Embedding Assiciation Test(WEAT) (Caliskan et al., 2017), Relative Norm Distance(RND), RNSB, ECT and TGBI (Gupta et al., 2021) can be used. In this project, we'll analyse current SOTA Hindi LLM models by utilizing the generations of these models for calculating Co-Occurance Bias. Using this we can quantify association of a profession with specific gender (Qian et al., 2019). Also we'll use Relative Norm Difference, i.e. RND (Gupta et al., 2021) as a metric to quantify bias in model embeddings of open models.

3 Methodology

This section outlines our approach to analyzing gender bias within the outputs of three Hindi language models: Krutrim, Airavata, and OpenHathi. Our analysis not only involved collecting data and generating outputs from these models but also incorporated a dual-method evaluation using word embeddings analysis with Relative Norm Difference (RND) and the Word Embedding Association Test (WEAT).

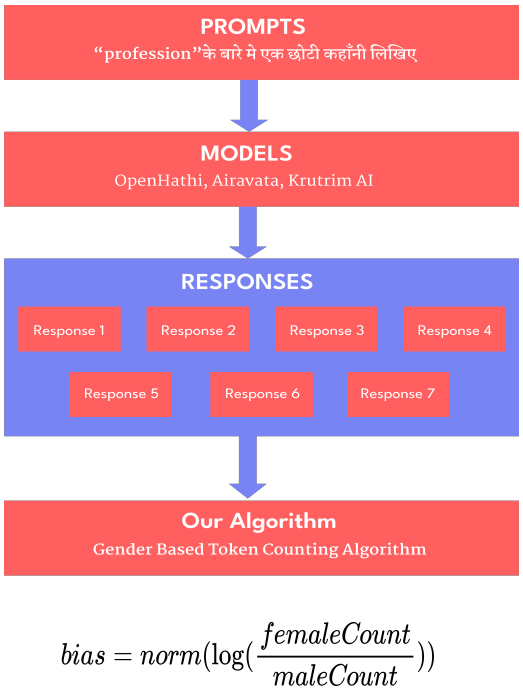
3.1 Data Collection

We categorized over 120 professions, commonly recognized in the Indian context, into three groups: typically male-associated, female-associated, and gender-neutral. The dataset creation was tailored for each model: Krutrim required a custom Python scraper for data extraction due to the lack of an API, whereas Airavata and OpenHathi could be accessed directly via the Hugging Face.

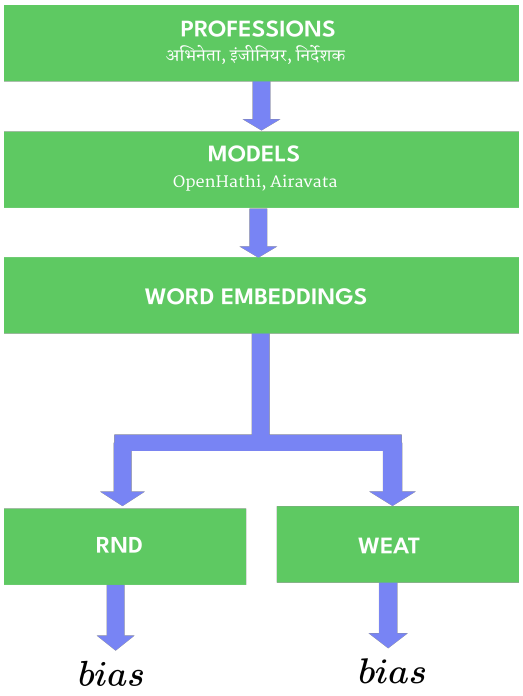
Additionally, we employed the Hindi WEAT dataset and evaluation methodology (Malik et al., 2022) to analyze bias in the word embeddings of the Airavata and OpenHathi models. We also utilized the WordBias dataset for Relative Norm Difference analysis to further investigate biases within these models.

3.2 Model Response Evaluation

Professions from each category were prompted to the models to generate short stories. These prompts were structured as: "profession_name के बारे में एक छोटी कहानी लिखिए " with "profession_name" being a placeholder for each profession. For diversity, we collected seven responses for each prompt.



(a) Model Response generation & evaluation



(b) Word Embedding Evaluation

Figure 1: Methodology for model bias evaluation

For the evaluation of bias, we assessed gender portrayal in Hindi sentences, focusing on gender-specific words such as verbs and possessive pronouns present in the model responses. We calculated a bias value by comparing the counts of female-specific words to male-specific words, and then normalized it using the tanh function. This normalization allowed us to express bias on a scale

ranging from -1 (indicating male bias) to 1 (indicating female bias). This methodology is applicable in this context due to the nature of our prompts, where we requested the generation of short stories centered around various professions. Consequently, the model constructs narratives focused on characters assumed to hold those professions, thus incorporating gender-specific language. To ensure the reliability of this approach, we validated it by manually labeling approximately 370 responses. In approximately 361 of these responses, the method proved effective.

3.3 Word Embedding Evaluation

For a deeper understanding of inherent biases, we also looked into the non-contextualized word embeddings from the Airavata and OpenHathi models. Due to proprietary restrictions, Krutrim's embeddings were not accessible. Using Python, we extracted 4096-dimensional embeddings, and for subword tokenizations, we applied mean pooling to get representative word embeddings.

Our evaluation utilized the Relative Norm Difference (RND) and Word Embedding Association Test (WEAT) metrics. RND measures the norm of the difference between averaged vectors within a target set and individual attribute words, highlighting gender associations. WEAT assesses biases by comparing cosine similarities between embeddings, quantifying the bias intensity.

By merging these analyses, we crafted a methodology that not only investigates generated text outputs for gender bias but also the underlying word embeddings, offering a holistic view of the biases present in Hindi language models.

3.4 Reflection on Methodological Adaptations from English to Hindi

Much of the bias measurement techniques in NLP have been developed with English data. Applying these methods to Hindi involves both direct adaptations and nuanced modifications to account for linguistic and cultural differences:

WEAT & RND: The use of WEAT and RND is a direct extension from studies conducted on English language models. However, adapting it to Hindi involves selecting culturally and linguistically relevant word sets. This could highlight how biases manifest differently due to the distinct social and cultural contexts that shape word usage and associations in Hindi.

Use of Gendered Verbs and Pronouns: While the approach of analyzing gender through verbs and pronouns is borrowed from English, Hindi's grammatical structure inherently encodes gender more extensively across verbs, adjectives, and even some nouns. This linguistic feature makes gender bias potentially more pervasive and pronounced in Hindi text, thus necessitating adjustments in how bias is quantified.

4 Results

4.1 Model Response Evaluation:

This approach offers a dynamic assessment of bias by analyzing the text generated by the models in response to various prompts. It reflects how models actually perform when used in real-world applications where text generation is required. For some selective professions, the comparison of bias value can be seen in Figure 2

Krutrim Model:

- The bias analysis of the Krutrim model revealed varied results across different professions.
- For gender-neutral professions such as "प्रोग्रामर", "एथलीट", "सिक्योरिटी गार्ड" etc. the model exhibited a significant bias towards masculine gender, and professions such as "नर्स", "ब्यूटीशियन" etc, the model showed a notable bias towards feminine gender.
- Overall, the Krutrim model exhibited a significant inclination towards generating male-biased responses when presented with gender-neutral professions. Conversely, it rarely produced responses considering females for such professions.
- For more detailed view of the results, see link

Airavata Model:

- The bias analysis of the Airavata model revealed similar trends to those observed in the Krutrim model, i.e, with a tendency towards masculine gender biases. However, Airavata demonstrated a notable improvement by generating more responses considering females for gender-neutral professions compared to Krutrim.
- However, Airavata demonstrated unexpected flexibility by producing responses that considered males for traditionally female-specific professions like "नर्तकी" and "गायिका"as well as considering females for traditionally male-specific professions such as "अध्यापक" and "अभिनेता"
- Overall, while the Airavata model showed less bias towards masculine gender in gender-neutral professions compared to Krutrim, it lagged behind Krutrim in accurately linking male and female genders to their respective specific professions.
- For more detailed view of the results, see link

OpenHathi Model:

- OpenHathi was not even able to generate proper sentences, also it was generating sentences as a mixture of Hindi and English sentences.
- After fine-tuning the parameters, we tried to record the responses but as the sentences were not properly generated, the evaluation is not very concrete.
- For bias evaluation, OpenHathi performed relatively similar to Airavata but considered female for male-specific professions such as "अभिनेता", "लेखक" etc.
- Overall, the evaluation of Bias on OpenHathi was mainly similar to Airavata but the generation of OpenHathi sentences was relatively poorer than Airavata.
- For a more detailed view of the results, see link

4.2 Word Embedding Evaluation:

This standard method exposes the bias associations with word embeddings, learned during the training process, which may or may not be evident in the text generated by the models, but are likely to influence it subtly. For some selective professions, the comparison of bias value associated with it can be seen in Figure 2

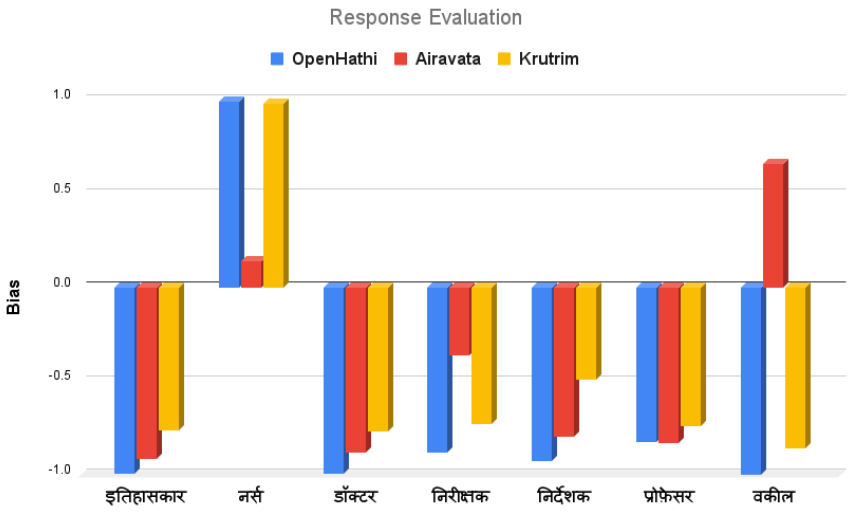


Figure 2: Model Response Evaluation Results

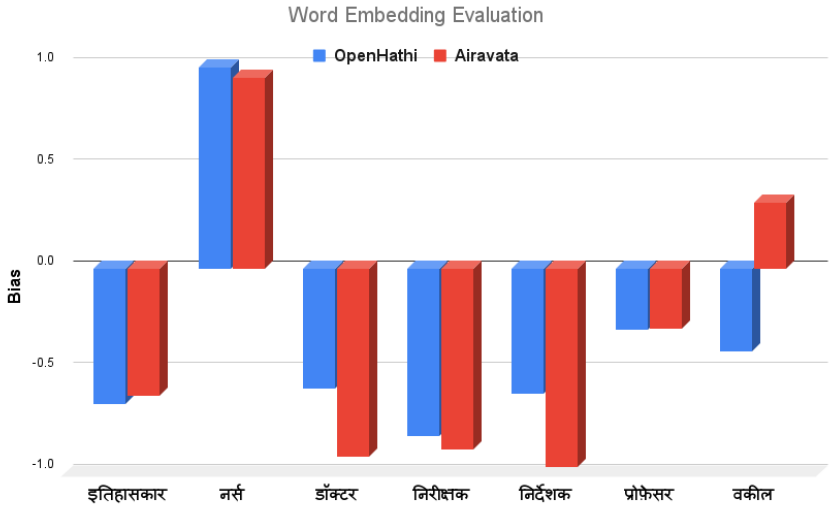


Figure 3: Word Embedding Evaluation Results

In above plots, x-axis represents Gender-Neutral Professions and y-axis represents Bias scores. Higher bias value indicates that female-association words outnumber male-associated ones, suggesting bias towards females in the generation for that profession and vice versa.

Relative Norm Difference(RND):

For analysis of bias in word embeddings using the **Relative Norm Difference** metric, visual interactive tool WordBias (Ghai et al., 2021) was employed. This tool facilitated the exploration of biases encoded in word embeddings against different intersectional groups, i.e. gender, religion, etc. The biases observed for some of the professions are shown in Figure 2. For detailed results see github

Word Embedding Association Test(WEAT):

For the **WEAT** analysis, we utilized the Hindi WEAT dataset (Malik et al., 2022). This dataset is specifically designed for evaluating bias in Hindi word embeddings and contains word pairs representing various gender and profession combinations. Results have been shown in Table-1

Description	**WEAT**	
	OpenHathi	**Airavata**
GenderAttributes		
Maths,Arts vs Male,Female	0.067(0.465)	0.110(0.441)
Science,Arts vs Male,Female	-0.404(0.764)	-0.533(0.832)
Adjectives vs Male,Female	-0.019(0.649)	-0.088(0.499)
Gendered Verbs vs Male,Female	1.838(0.000)	1.783(0.000)
Gendered Adjectives vs Male,Female	0.731(0.080)	0.934(0.033)
Gendered Entities vs Male,Female	0.026(0.619)	-0.053(0.422)
Gendered Titles vs Male,Female	-0.552(0.152)	-0.519(0.557)

Table 1: WEAT scores for gender bias in OpenHathi and Airavata models

4.3 Comparative Analysis of Gender Biases in Hindi and English

When comparing the biases in Hindi LLM's to those in English, several factors come into play:

Magnitude of Bias: Initial observations show that biases in Hindi are qualitatively similar but quantitatively more pronounced in the Krutrim and Airavat Models than English.(for OpenHathi initial generation was a mixture of Hindi and English sentences so the biases remain the same for English). This could be due to several reasons:

- Language Structure: Hindi language structurally encodes gender more explicitly than English, which might lead to more pronounced biases in automated text generation.
- Cultural Contexts: Social roles and gender stereotypes may be more rigidly defined in certain Hindi-speaking contexts, which could be reflected in training datasets derived from regional news, books, and other media.
- Training Data: The volume and variety of data available for training models in English are significantly larger than for Hindi. This disparity can lead to less robust models in Hindi that may not handle nuances of gender as effectively.

Qualitative Differences: While the types of biases (e.g., associating nurses with females and engineers with males) might mirror those found in English, the cultural nuances influencing these associations could differ. For instance, the perception of certain professions might carry different societal implications and stereotypes in Hindi-speaking regions compared to English-speaking ones.

5 Limitations

The evaluation of the Krutrim AI model relies solely on the responses it generates to given prompts, as access to word embeddings is not available since its recent launch and proprietary model. There are limitations to the response evaluation approach. For instance, if the model adds multiple characters to the story for the given prompt: "profession_name के बारे में एक छोटी कहानी लिखिए ", the methodology might yield incorrect results. However, we were careful in framing the prompts to ensure that the model generates stories centered around the main character which the model associates with the given profession. Also, In evaluating word embeddings, a limitation arises when professions are provided for extracting word embeddings that are merely transliterated versions of their English counterparts. In such cases, there may not be direct equivalents in Hindi, leading the model to rely on subword units for generating embeddings. This process can alter the associations and biases associated with the word embeddings, potentially skewing the evaluation results.

Conclusion

This study has provided a comprehensive examination of gender biases in state-of-the-art Hindi language models: Krutrim AI, OpenHathi, and Airavata through a systematic analysis of model responses and word embeddings.

Our findings reveal that despite advancements in NLP technologies, gender biases are still prevalent in models trained on Hindi data. The biases were more pronounced in certain models and varied depending on the nature of the profession involved. For instance, the Krutrim AI model tended to generate responses with a masculine bias for gender-neutral professions, while the Airavata model showed a slight improvement by being less biased towards masculine interpretations in similar contexts. The OpenHathi model, although promising, struggled with coherence in its output, which could mask the true extent of biases present.

The comparative analysis with English language models indicates that the magnitude and impact of these biases can be more significant in Hindi. This is due to the structural and cultural particularities of the Hindi language, which often embeds gender more explicitly in its grammar and vocabulary.

The study underscores the necessity for continuous efforts in developing and refining debiasing techniques, particularly for underrepresented and less-resourced languages like Hindi. These should not only focus on adjusting word embeddings but also on improving the training datasets to reflect a more balanced and inclusive representation of gender. Moreover, future research should explore the intersection of linguistic, social, and cultural factors that contribute to biases, providing deeper insights and more robust solutions to these pervasive issues.

In conclusion, while the analyzed models demonstrate significant potential for supporting Indic language processing, our research highlights the critical need for awareness and proactive measures to ensure that these technologies promote fairness and inclusivity. As language technologies become increasingly integrated into daily life, the imperative to address and mitigate gender biases becomes not just a technical challenge, but a societal responsibility.

Contributions

Arunima Roy:

- Majorly worked on the literature part, reading useful research papers briefly, the related work relevant to this and listing the key findings from the different papers. For ex: listing out different evaluation strategy that can be helpful for gender bias evaluation.
- Played with the prompts of all three models Krutrim AI, OpenHathi and Arivat and found out some useful observations.
- Manually labeling the responses of the different model and tested if the bias methodology used in bias calculation actually works or not. Also, creating the graphs from the evaluation matrices used in the report.

Asad Ali:

- Worked majorly in analysis of OpenHathi and Airvat, the key achievement was taking out the word embeddings of these open source pretrained models.
- Played with these codes, tuned the hyperparameters of OpenHathi and Airvat for better generations.
- From the word embeddings, implementing different bias evaluation methodologies like SEAT, WEAT and RND and Also creating the word bias representation model which is a beautiful way to showcase the bias from the word embeddings.

Dileep Patel:

- Majorly worked on the Krutrim Analysis part, more on finding possible ways to automate the conversation from Krutrim AI model and the key achievement was to writing the code for scrapper using python and recorded responses of possible prompts.
- Listing out the professions, and creating possible prompts that can be used in the model evaluation.
- Implemented the strategy for bias evaluation mentioned in the report.

References

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL `https://www.science.org/doi/abs/10.1126/science.aal4230`.

CognitiveLab. Indic llm leaerboard. 2024. URL `https://www.cognitivelab.in/blog/introducing-indic-llm-leaderboard`.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. Airavata: Introducing hindi instruction-tuned llm, 2024.

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

Hila Gonen and Kellie Webster. Automatically identifying gender issues in machine translation using perturbations. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1991–1995, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.180. URL https://aclanthology.org/2020.findings-emnlp.180.

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

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

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for Hindi language representations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1041–1052, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.76. URL https://aclanthology.org/2022.naacl-main.76.

Luke McGuire, Tina Monzavi, Adam J. Hoffman, Fidelia Law, Matthew J. Irvin, Mark Winterbottom, Adam Hartstone-Rose, Adam Rutland, Karen P. Burns, Laurence Butler, Marc Drews, Grace E. Fields, and Kelly Lynn Mulvey. Science and math interest and gender stereotypes: The role of educator gender in informal science learning sites. *Frontiers in Psychology*, 12, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.503237. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.503237.

Michela Menegatti and Monica Rubini. Gender bias and sexism in language, 09 2017. URL https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-470.

OLA. Krutrim ai. 2023. URL https://olakrutrim.com/.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, pp. 450–456, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450372619. doi: 10.1145/3377713.3377792. URL https://doi.org/10.1145/3377713.3377792.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 223–228, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2031. URL https://aclanthology.org/P19-2031.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020.

Sarvam.ai. Openhathi series: An approach to build bilingual llms frugally. 2023. URL https://www.sarvam.ai/blog/announcing-openhathi-series.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods, 2018.