Analyzing Gender Bias in Hindi Language Models

Progress Review:

- Targeted three state of the art hindi language models (Krutrim AI, OpenHathi base version, Airavata).
- **Dataset Collection for quantifying biases:**
  - Categorized over 120 professions into male-associated, female-associated, and gender-neutral categories.
  - Developed a custom Python scraper for data extraction from Krutrim due to the lack of an API.
  - Accessed and prepared data from Airavata and OpenHathi via Hugging Face.
- **Model Response Evaluation:**
  - Generated and analyzed model responses to assess gender bias.
  - Calculated bias metrics and validated the results through manual labeling.
- **Word Embedding Evaluation:**
  - Extracted 4096-dimensional word embeddings from Airavata and OpenHathi.
  - Applied mean pooling for subword tokenizations to create representative word embeddings.
  - Conducted Relative Norm Difference (RND) and Word Embedding Association Test (WEAT) analyses to measure inherent biases.
- **Methodological Adaptations:**
  - Adapted WEAT and RND from English studies to Hindi, considering linguistic and cultural contexts.
  - Adjusted our analysis to consider how Hindi uses gender in verbs and pronouns, since Hindi's grammar reflects gender more extensively than English.

Introduction

Language models, particularly in the context of natural language processing, have become integral in various applications, from education to healthcare. However, as these models become more pervasive, there is a growing concern about the inherent biases they may carry, particularly gender bias.

Impact of Gender Bias on Society

Gender bias in language models can have far-reaching effects on society. By reinforcing harmful stereotypes, these models can perpetuate unequal treatment and opportunities across various domains. For instance, biased outputs can influence public perception, reinforce discriminatory practices in hiring, and even affect the quality of education and healthcare that individuals receive. Such biases not only perpetuate existing social imbalances but also hinder progress towards a more equitable society.

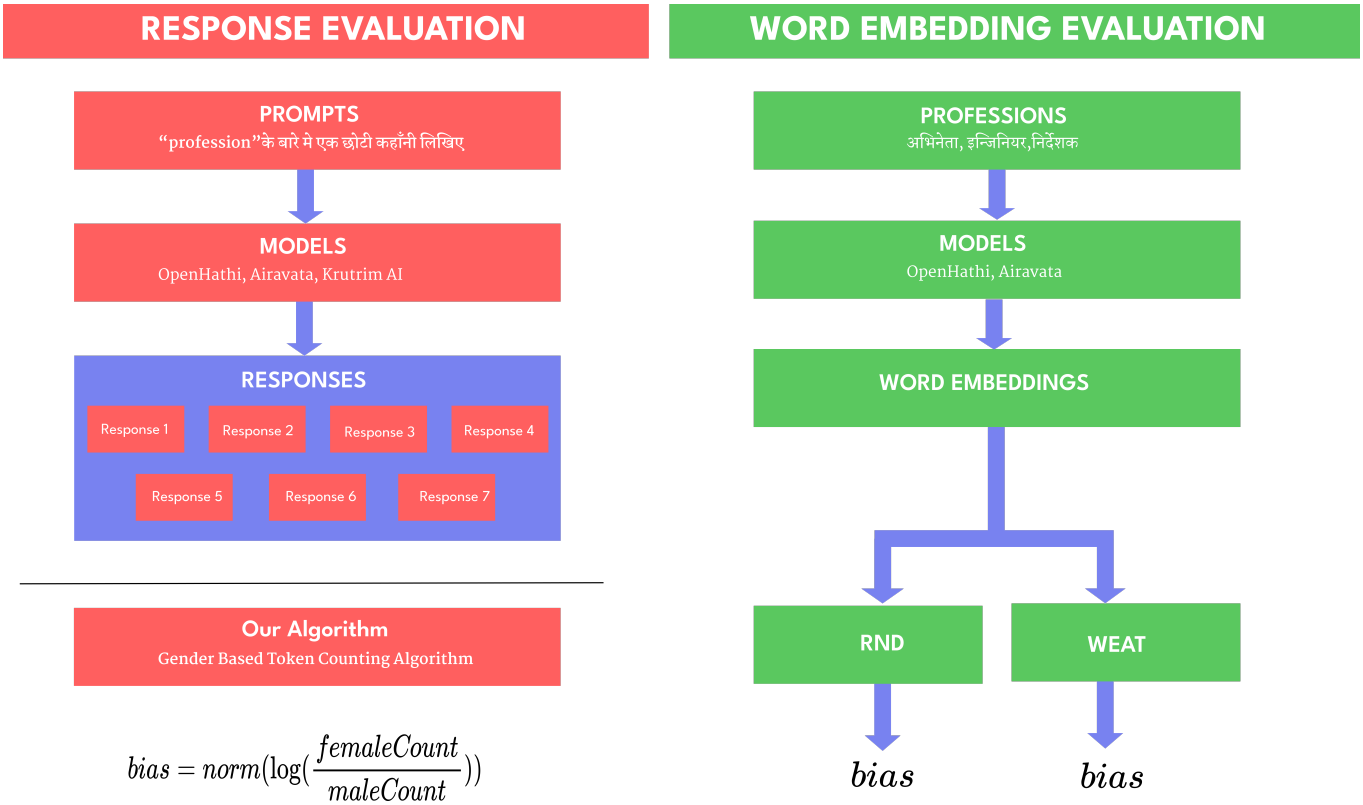
Critical Areas for Gender Debiasing

Debiasing language models is crucial, especially in areas where decisions have a significant impact on people's lives. In education, eliminating bias ensures that learning materials promote equality and inclusivity. In healthcare, unbiased models are essential to provide accurate and equitable medical advice and treatment. Similarly, in legal systems and recruitment processes, ensuring that language models are free from gender bias is vital to prevent discrimination and promote fairness.

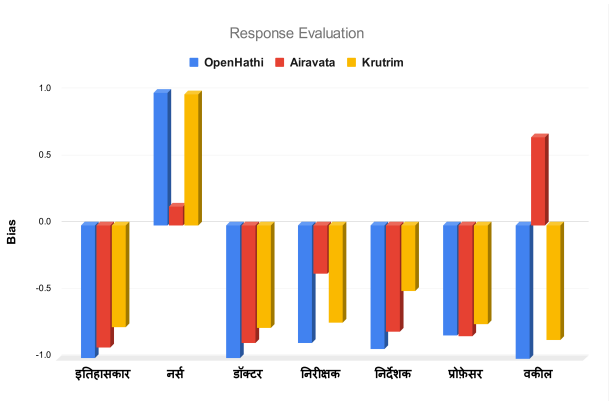
The Role of LLMs for Children

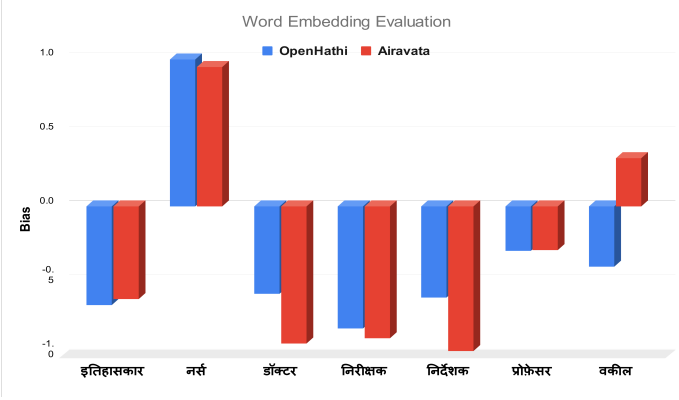
Large Language Models (LLMs) hold immense potential for educational purposes, offering personalized learning experiences and supporting creative endeavors for children. However, it is crucial to develop these models with careful consideration of the content they produce, ensuring that they are free from biases and provide a safe, supportive learning environment. Properly implemented, LLMs can empower children to learn at their own pace while fostering a more inclusive worldview.

Methodology



Results





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", "ब्यूटिशियन" 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.

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.

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.

Potential Research Directions

- Exploration of additional Hindi language models for comparative analysis.
- The current techniques for quantifying bias have limitations, so we need to explore additional methodologies for a more thorough assessment.
- Using the mining techniques for quantifying bias in language models of ACM Best Paper titled: "Akal Badi ya Bias".
- link of the paper:https://facctconference.org/static/papers24/facct24-132.pdf