Assessing Gender Bias of Pretrained Bangla Language Models in STEM and SHAPE Fields

Noor Mairukh Khan Arnob[1,3], **Saiyara Mahmud**[1,3], **Azmine Toushik Wasi**[1,2†]

[1]STEM Team, HerWILL Inc., [2]Shahjalal University of Science and Technology,
[3]University of Asia Pacific
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Gender bias continues to shape societal perceptions across both STEM (Science, Technology, Engineering, and Mathematics) and SHAPE (Social Sciences, Humanities, and the Arts for People and the Economy) domains. While existing studies have explored such biases in English language models, similar analyses in Bangla—spoken by over 240 million people—remain scarce. In this work, we investigate gender-profession associations in Bangla language models. We introduce *Pokkhopat*, a curated dataset of gendered terms and profession-related words across STEM and SHAPE disciplines. Using a suite of embedding-based bias detection methods—including WEAT, ECT, RND, RIPA, and cosine similarity visualizations—we evaluate 11 Bangla language models. Our findings show that several widely-used open-source Bangla NLP models (e.g., `sagorsarker/bangla-bert-base`) exhibit significant gender bias, underscoring the need for more inclusive and bias-aware development in low-resource languages like Bangla. We also find that many STEM and SHAPE-related words are absent from these models' vocabularies, complicating bias detection and possibly amplifying existing biases. This emphasizes the importance of incorporating more diverse and comprehensive training data to mitigate such biases moving forward. Code available at `https://github.com/HerWILL-Inc/ACL-2025/`.

1 Introduction

Textual representations play a powerful role in reinforcing gender biases, particularly in how professional roles are described and associated with gender. In both STEM and SHAPE (Social Sciences, Humanities, and the Arts for People and the Economy) (Black, 2020) domains, written content often reflects implicit assumptions—depicting roles like "receptionist" as female-coded and "scientist" as male-coded (Eckert and McConnell-Ginet, 2013). Such patterns are not merely descriptive but normative; they help entrench gendered expectations about who belongs in which fields. These biases contribute to the marginalization of SHAPE disciplines and those who pursue them, often women, by diminishing their public visibility and perceived value. Over time, consistent exposure to gendered language in text influences how individuals internalize societal roles and professional aspirations (European Commission, 2012). Recognizing and addressing gender bias in text is therefore essential to creating better representations across disciplines.

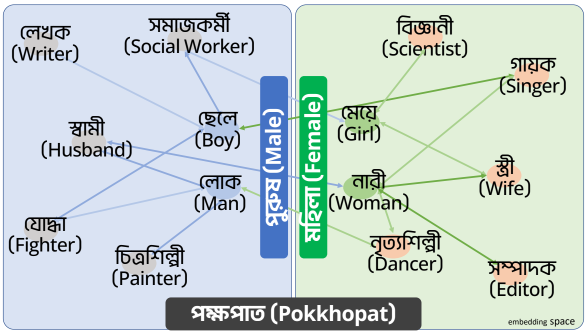


Figure 1: Assessing Gender Bias of Pretrained Bangla Language Models (PBLMs)

Studies exploring the biases of English language models (Nadeem et al., 2021) do not explicitly assess gender bias in SHAPE fields. For example, Therapist, Educationalist, Economist, Lobbyist, Archaeologist, Journalist, Actor, Dancer, Cartoonist etc. are absent in the StereoSet Dataset used in the study by Nadeem et al. (2021). Although gender bias detection of NLP systems is a well-studied task for the English language (Sun et al., 2019), it remains largely unexplored for the low-resource language, Bangla (often referred as *'Bengali'*). While several open-weight Bangla language models are available in public repositories, their gender bias remains largely unexplored. This leaves room for

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 268–281
August 1, 2025 ©2025 Association for Computational Linguistics

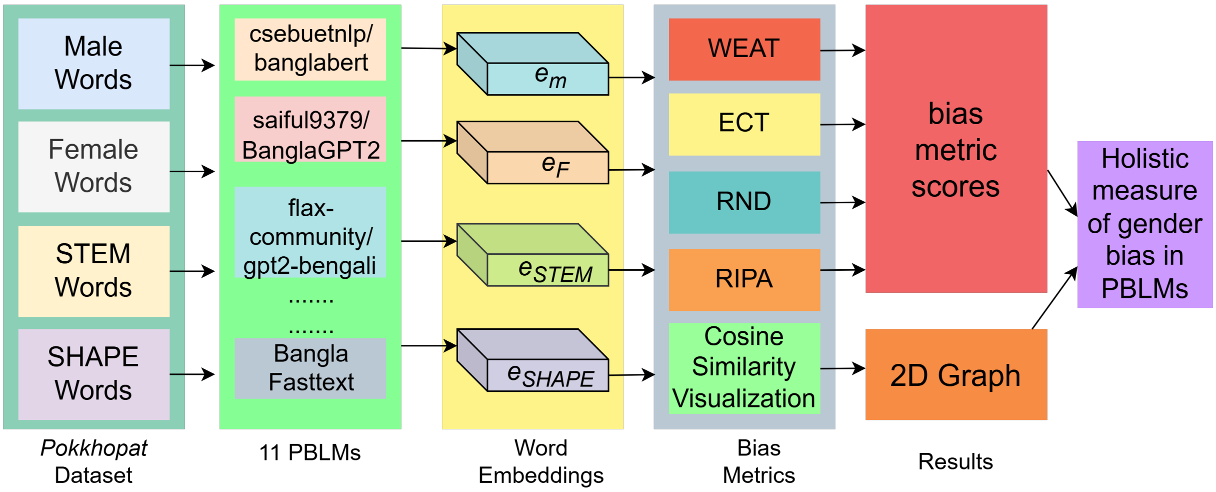


Figure 2: Overall methodology of our paper, including dataset details, models, and evaluation metrics.

these models to be deployed while leaving the risk of exhibiting gender bias in real-world scenarios; where over 131 million Bangla-speaking internet users (Dhaka Tribune, 2023) can experience and be influenced by gender bias.

In Bangla, there is no grammatical gender; instead, the gender system relies on semantics (Mukherjee, 2018). Gender distinctions in Bangla are indicated by specific lexical choices reflecting the gender of the entity. For example, in terms of grammatical gender in Bangla language, "সে" (ʃe) is a pronoun and it can refer to both "she" or "he". "সন্তান" (ʃɔnt̪ˈan) signifies "child" and can represent either a son or a daughter. In terms of Semantic gender, Bangla has separate words for both the genders. For example: "পুরুষ" (puːruːʃ) refers to "man" and "মহিলা" (mo-hi-la) means "woman", “শিক্ষক" (ʃik^{h}ɔk) specifies "male teacher" and "শিক্ষিকা" (ʃik^{h}ika) translates to "female teacher". Although the “শিক্ষক" and "শিক্ষিকা" nouns are lexically similar, their usage does not affect sentence structure, verb conjugation, and adjective agreement; resulting in the absence of grammatical gender in Bangla. Due to the absence of grammatical gender in Bangla, it is difficult to analyze implicit gender bias in STEM and SHAPE fields.

To address the critical gap in evaluating gender bias in Bangla NLP, our primary contribution lies in a comprehensive empirical assessment of 11 pretrained Bangla Language Models (PBLMs) using five established bias evaluation metrics. Despite the growing use of these models, their implicit gender associations—particularly in relation to STEM and SHAPE domains—remain largely unexplored. Given the absence of gendered pronouns in Bangla, we analyze word embeddings to uncover latent biases, hypothesizing that unbiased models would exhibit similar distances between STEM/SHAPE terms and male/female word embeddings. Our findings reveal that several popular models display measurable gender bias, with stronger biases observed in SHAPE-related vocabulary. To assist in this evaluation, we introduce *Pokkhopat*, a curated dataset comprising gender-categorized Bangla words across STEM and SHAPE fields. This resource is designed to support future bias evaluations, especially by enriching the currently underrepresented SHAPE domain in Bangla gender bias research, as illustrated in Figure 1.

These findings have significant implications for the responsible deployment of Bangla NLP systems, particularly in educational, hiring, and content generation tools where gender neutrality is critical. By uncovering these hidden biases, our work not only sets a precedent for fairness audits in low-resource languages but also encourages the development of more equitable and inclusive language technologies for diverse linguistic communities.

2 Bias Statement

This paper examines how gender bias relates to STEM and SHAPE professions by analyzing word clustering in Bangla language embeddings. Although Bangla lacks grammatical gender (Mukherjee, 2018), biases in language embeddings may reflect stereotypes about STEM and SHAPE fields. Some biases are harmless, but others can be damaging. Biased language models can unfairly reinforce gender roles (Fang et al., 2024). For example, if Bangla embeddings cluster engineering-related words with male-associated words, it suggests a bias linking STEM with males. Conversely, if words related to psychotherapy cluster with female-associated words, it may reflect the stereotype that women are more suited for SHAPE roles (Blow et al., 2008). Such biases can limit diversity in education and the workplace (Funk and Parker, 2018). Ideally, Bangla embeddings should avoid reinforcing gender stereotypes in STEM and SHAPE fields.

The dataset used here represents only two genders: male and female, which may harm those identifying outside this binary (Dev et al., 2021), particularly in Bangladesh, where 12,629 identify as "Third Gender" (BBS, 2022). We present this study to encourage future research that is more inclusive of diverse gender identities.

3 Methodology

Our overall methodology is outlined in Figure 2. We assessed the gender bias of 11 PBLMs using the *Pokkhopat* dataset and 5 bias evaluation metrics to obtain a clear picture of how biased PBLMs are.

3.1 *Pokkhopat* Dataset Development

To investigate gender and professional biases in word embeddings, we developed a dataset named *Pokkhopat*. The dataset comprises four curated word lists: **Male**, **Female**, **STEM**, and **SHAPE**, containing Bangla words, alongside English translations.

We followed the `{ “Subject” : { “Predicate” : [ Object ] } }` format followed by (W3C, 2013) to arrange our dataset. The structure of the JSON file of our dataset is shown below.

```
{ “Gender/Profession” :
    { “Language” :
        [ “Words” ]
    }
}
```

An illustrative sample of the dataset is presented in Figure A6. The dataset includes 57 male-specific words, 56 female-specific words, 47 STEM-specific words, and 73 SHAPE-specific words. The average word lengths of Male, Female, STEM, and SHAPE-related words are 5.61, 5.64, 14.79, and 11.96 respectively. In summary, there are 237 bangla words in the dataset containing 2,242 characters. Average characters per word is 9.46. The standard deviation of word length is 5.93, showcasing the dataset's linguistic variability. The Type-Token Ratio (TTR) (Richards, 1987) of our dataset is 0.970, indicating a high lexical diversity. 51.1% of the words in our dataset contain conjunct consonants. The gender-specific words were extracted from existing Bangla linguistic resources and reviewed by native speakers for contextual and cultural relevance.

For the STEM list, we referenced the occupational taxonomy published by the U.S. Bureau of Labor Statistics [1], identifying professions traditionally classified under science, technology, engineering, and mathematics. Similarly, SHAPE (Social Sciences, Humanities, and the Arts for People and the Economy) professions were selected with guidance from an article from the University of Edinburgh [2]. To ensure the diversity of the dataset, we used synonyms such as "পিতা", "বাবা", "আব্বু", "বাপ" for the word "Father". We also included closely related words like "কাকি" (father's younger brother's wife), "কাকিমা" (respected father's younger brother's wife), "চাচি" (paternal uncle's wife), "পিসী" (father's sister), "ফুফু" (mother's sister's husband), "মাসি" (mother's sister), "খালা" (mother's sister), and "মামী" (maternal uncle's wife). This strategy of including synonyms is followed throughout the dataset to make sure that most words related to males, females, STEM, and SHAPE are abundantly represented in our dataset.

Where Bangla lacked direct equivalents for some words (e.g., "Pharmacist", "Physiologist", "Lobbyist"), careful transliterations were used. The dataset was independently validated by two native Bangla speakers to ensure linguistic accuracy and semantic clarity. To improve transparency, accessibility, and reproducibility, the dataset is made publicly available at Mendeley Data [3]. The curated dataset, *Pokkhopat*, forms the foundation for generating and analyzing the word embeddings used in our experiments.

3.2 Evaluation Methodology

As outlined in Figure 2, we passed the male word list, w_m from the *Pokkhopat* dataset through a PBLM such as ‘`csebuetnlp/banglabert`’ to obtain the word embeddings, e_m. Similarly, we

[1] https://www.bls.gov/k12/students/careers/stem-table.htm

[2] https://cahss.ed.ac.uk/research-ke/serch-research-hub/shape

[3] https://data.mendeley.com/datasets/y3x569kk9t/2

Table 1: Bias evaluation of 11 PBMLs on the *Pokkhopat* dataset across 5 bias metrics. Scores indicating the most bias are in bold.

Model	**Cohen's d**	p-**value**	ECT_{STEM}	ECT_{SHAPE}	RND_{STEM}	RND_{SHAPE}	$RIPA_{STEM}$	$RIPA_{SHAPE}$
csebuetnlp/banglabert	-0.1546	0.77	0.9967	0.9881	0.0284	0.0211	-0.0096	-0.0074
saiful9379/Bangla_GPT2	-0.3530	0.95	0.9943	0.9945	0.0107	-0.0027	0.0359	-0.0484
flax-community/gpt2-bengali	0.2448	**0.09**	0.9977	0.9986	-0.0700	-0.0762	0.0161	0.0150
ritog/bangla-gpt2	-0.2010	0.81	0.9942	0.9963	0.0087	0.0001	**0.1476**	**0.1717**
csebuetnlp/banglat5	-0.4221	0.98	0.9749	0.9675	-0.0364	-0.1090	0.0026	0.0039
neuropark/sahajBERT	0.1322	0.27	0.9207	0.9545	-0.0596	-0.0582	0.0096	0.0059
Kowsher/bangla-bert	-0.2071	0.86	0.9816	0.9532	-0.0499	-0.0899	0.0323	0.0386
csebuetnlp/banglishbert	-0.0916	0.64	0.9946	0.9868	0.0487	0.0605	-0.0488	-0.0479
sagorsarker/bangla-bert-base	**-0.8031**	1.00	0.9729	**0.9319**	**0.3566**	**0.2578**	-0.0636	-0.0222
shahidul034/ text_generation_bangla_model	-0.6987	1.00	0.9907	0.9898	0.0585	0.0037	0.0300	0.0522
Bangla Fasttext	-0.0606	0.70	**0.8776**	0.9359	0.0128	0.0064	-0.0003	-0.0001

obtained the word embeddings e_F, e_{STEM}, and e_{SHAPE} from the word lists w_F, w_{STEM}, and w_{SHAPE}. While generating word embeddings, we used the tokenizers recommended by the public repositories of specific models. Since the models are trained on Bangla text corpora, the word embeddings contain contextual information in relation to their meaning and position in sentences in the corpora. We normalized the words using the normalizer recommended by csebuetnlp [4] for getting standardized results. To save time and computational resources, we cached word embeddings to load from local disk. After obtaining the embeddings, we used them to calculate evaluation scores of **WEAT**, **ECT**, **RND**, and **RIPA** using equations outlined in Appendix A.1.1 to A.1.5. The scores obtained from these metrics give us a statistical view of how biased PBLMs are. Furthermore, we plotted the cosine similarity between word embeddings of different word lists to get a visual representation of gender bias in PBLMs. We used 5 different metrics as different metrics can detect various biases in the embedding space of PBLMs with regards to gender and profession. The combination of bias metric scores and the 2D graph give us a holistic view of gender bias in PBLMs.

3.3 Evaluation Metrics

To assess whether the PBLMs exhibit gender bias in the STEM and SHAPE fields, we employ 5 evaluation metrics: **WEAT** (Word Embedding Association Test), **ECT** (Embedding Coherence Test), **Cosine Similarity Visualization**, **RND** (Relative Norm Distance), and **RIPA** (Relational Inner Product Association). Equations for calculating these scores are shown in Appendices A.1.1 through A.1.5. We chose **WEAT** (Caliskan et al., 2017) as it is a widely adopted metric which quantifies implicit bias similar to human implicit bias association test. A WEAT score near 0 implies less bias. The range of WEAT score values is [-1,1]. For example, a WEAT score close to 1 means that the model associates males with STEM professions more than SHAPE professions; whereas a WEAT score close to -1 signifies that the model associates females with STEM more than SHAPE. We compute p-**values** to compute the statistical significance of the WEAT score. The null-hypothesis is that there is no association between gender and profession in the pretrained models' language representations. If $p < 0.05$, we reject this null hypothesis and assert that the model is biased. A higher p-value indicates less gender bias. The **ECT** (Dev and Phillips, 2019) metric was chosen as it can reveal underlying biases in how words are related by examining the overall coherence of the embedding space concerning gender and profession. The value of ECT ranges between -1 and +1, where a value close to +1 indicates less gender bias. For example, an ECT_{STEM} score close to +1 means that STEM word embeddings are equally close to male and female word embeddings. An ECT_{STEM} score closer to -1 indicates that male words are more associated to STEM than female words. Inspired by the figures in Feng et al. (2023), we visualize gender bias in PBLMs by visualizing **Cosine Similarity**. In a Cartesian coordinate system, the x-axis represents the mean cosine similarity between Bangla male-specific and STEM-specific word embeddings, while the y-axis represents the same for female-specific and STEM-specific embeddings in Figure 3. Thus, a point $P(x, y)$ reflects the model's gender bias. The farther P is from the blue identity line ($y = x$), the greater the bias. **RND** (Garg et al., 2018) was also adopted in our study since it complements WEAT by focusing on distance, not association. An RND score close to zero translates to next to no bias. If $RND_{STEM} < 0$, it means the PBLM associates males with STEM professions

[4] https://github.com/csebuetnlp/normalizer

more than females. Similarly, $RND_{SHAPE} > 0$ implies that the PBLM associates females more to SHAPE professions. We used the **RIPA** (Ethayarajh et al., 2019) metric as it uses an aggregated representation of the word relations, which is less likely to be swayed by the nuances of individual word choices. The higher magnitude of the RIPA score indicates higher gender bias. The more close to zero the RIPA score is, the model exhibits lesser gender bias. For example, a negative $RIPA_{STEM}$ score indicates that females are more associated with STEM than men. A $RIPA_{STEM} > 0$ indicates that STEM words are more associated with male words. Similarly, a $RIPA_{SHAPE} < 0$ score means females are more associated with SHAPE than men; which enables the societal construct that women are more suited for the SHAPE professions.

4 Experiments

4.1 Experimental Setup

We evaluated the gender bias of 11 PBLMs in this study. The models we evaluated can be found in Table 3 under Appendix A.3. We selected a mix of popular models such as `csebuetnlp/banglabert` and less-known models like `ritog/bangla-gpt2` to paint a holistic picture of PBLMs. We chose language models of generative architectures (GPT2, T5), sequential architectures (BERT, ELECTRA, ALBERT), and a shallow neural network (Skip-gram) for comparing gender bias across different architectures. The models used in our study range from 18.1055M parameters to 321.577M parameters. Dataset size varies between 250 MB to 40 GB. The models were pre-trained using corpora from various sources, including news websites, wikipedia, social networks, blog sites, etc. Therefore, the models we chose for evaluating are diverse in architecture, number of parameters, and pre-training data source; giving us a comprehensive view of gender bias in PBLMs.

4.2 Implementation Details

Our bias evaluation system was implemented and run on a laptop with `AMD Ryzen 3 4300U` processor (clock speed: 2.7 GHz). We utilized the implementation of AllenNLP (Gardner et al., 2018) to calculate the bias metrics WEAT and ECT. Since p-test requires excessive amount of time to calculate on a single thread, we used the built-in `ThreadpoolExecutor` [5] class of Python to activate $n = 16$ threads for calculating p-values faster. On top of AllenNLP's codebase, we implemented the code for calculating Cosine Similarity, RND and RIPA metric scores based on their equations. We used the normalizer recommended in (Hasan et al., 2020) to normalize Bangla text for standard results. We used the skip-gram version of the Bangla Fasttext model to obtain word embeddings. The word embedding lengths in our study are 768 for ELECTRA, GPT2, T5, ALBERT, and BERT-based models, and 300 for the Skip-gram based model.

5 Experimental Findings

5.1 Evaluation Scores

We had previously identified that gender bias in SHAPE professions is less-studied. Furthermore, the gender bias of PBLMs also remains unchecked. If PBLMs associate females with SHAPE professions, it may enforce societal stereotypes. Therefore, we assessed the gender bias of 11 PBLMs using 5 bias evaluation metrics: WEAT, ECT, Cosine Similarity Visualization, RND, and RIPA to see whether the PBLMs associate specific genders to stereotypical professions. In our findings, we observe that some models alarmingly associate gender with profession by affirming societal stereotypes and also exhibiting bias in contrary to societal notions.

WEAT Scores. The WEAT scores (Cohen's d) and p-values of PBLMs can be found in Table 1 of Appendix 7. The WEAT score for most of the models in consideration is close to zero, indicating that these models exhibit less bias. Only `sagorsarker/bangla-bert-base` gives a low WEAT score of -0.8031, showing that the model associates females with STEM words; which is contrary to the social stereotype that women are more suited for the SHAPE field. Some of the Cohen's d values closest to 0 are shown by `csebuetnlp/banglabert` and `csebuetnlp/banglishbert`. The dataset used to train these models, as shown in Table 3 under Appendix A.3, is Bangla2B+. The creators of Bangla2B+ tried to eliminate harmful content from the corpus as much as possible (Bhattacharjee et al., 2022), which could have contributed to detecting the lowest bias for models trained on Bangla2B+.

[5] `https://docs.python.org/3/library/concurrent.futures.html`

None of the p-values shown in Table 1 are less than 0.05, therefore we can not reject the null hypothesis. Therefore, no statistically significant bias is revealed by the p-values.

ECT Scores. ECT scores of pretrained Bangla language models can be found in Table 1. All of the ECT_{STEM} and ECT_{SHAPE} values shown in Table 1 are close to 1, meaning that the models exhibit next to no gender bias with regards to gender and profession. The ECT_{STEM} score farthest from 1, is shown by `Bangla Fasttext`, with a score of 0.8776, suggesting that this model is biased against females as it associates STEM-specific words more to male words. The lowest ECT_{SHAPE} score is achieved by `sagorsarker/bangla-bert-base`, which means that this model associates males with SHAPE words, contrary to the societal convention which asserts that males are unsuitable for SHAPE professions.

Cosine Similarity Visualizations. Cosine similarity visualization of male vs female words in STEM is shown in Figure 3. Almost all points for all models fall close to the blue identity line (Ciesielski, 1997) or $y = x$, meaning the STEM roles are represented equally closely to male and female words; with next to no sign of gender bias. The point for `sagorsarker/bangla-bert-base` falls slightly above the identity line, meaning that this model associates females more to STEM professions, compared to males, exhibiting bias against males. The `Kowsher/bangla-bert` model was pre-trained on the largest Bangla corpus (40 GB), as mentioned in Table 3 under Appendix A.3. The large corpus could have contributed to the bias of this model.

A visualization of cosine similarity in light of male and female-specific words in the context of SHAPE is shown in Figure 4. The point corresponding to the `sagorsarker/bangla-bert-base` falls above the identity line, meaning that the model associates females rather than males with SHAPE-specific words. This biased behavior of the model is aligned with the stereotype that females are well-suited for SHAPE professions. The reason for this bias can be rooted in the fact that as Table 3 under Appendix A.3 shows, `sagorsarker/bangla-bert-base` was trained on a corpus collected from various sites from the internet, with potentially biased content. The `Bangla Fasttext` model has 321 Million trainable parameters, as shown in Table 3. Despite having the highest number of trainable parameters, this model exhibits no bias, as shown in Figure 4. This result matches with the results of (Tal et al., 2022), where it is shown that larger models do not always exhibit more bias.

RND Scores. Almost all RND scores for each model shown in Table 1 are close to zero, exhibiting unbiased behavior. However, `sagorsarker/bangla-bert-base` has a higher RND_{SHAPE} score of 0.25, which implies the model associates females more to SHAPE roles. Which means that using this model may enforce the stereotype that women are more suited for SHAPE professions. As mentioned in Table 3 under Appendix A.3, `sagorsarker/bangla-bert-base` adopts the BERT architecture. The biased RND score of `sagorsarker/bangla-bert-base` could have been caused by the bias-inducing components of BERT as identified in (Bhardwaj et al., 2021). The high RND_{STEM} score of `sagorsarker/bangla-bert-base` indicates that the model associates females more to STEM roles as compared to males, exhibiting bias against males.

RIPA Scores. The RIPA scores of the Bangla language models are shown in Table 1. Most of the RIPA scores are close to zero, indicating low gender bias. However, the relatively higher magnitude of $RIPA_{STEM}$ score of `ritog/bangla-gpt2` asserts that this model associates males more with STEM roles, affirming established social stereotypes. The comparatively higher $RIPA_{SHAPE}$ score of `ritog/bangla-gpt2` signifies that that this model exhibits bias in contrary to existing societal norms by indicating that male words are closer to SHAPE words.

5.2 Lost for Words: The Bias We Can not See

Most models in Table 1 do not exhibit statistically significant gender bias. While investigating the reason for such behaviour, we found that if attribute words, such as STEM and SHAPE words, are largely absent from a model's pretraining data (i.e., out-of-vocabulary), detecting gender bias becomes challenging (Chaloner and Maldonado, 2019). The *Pokkhopat* dataset includes many words absent from common pretraining corpora used for PBLMs. Table 2 indicates that only 4.25% of STEM-specific words from the *Pokkhopat* dataset appear in the Bangla2B+ corpus, used to train `csebuetnlp/banglabert`

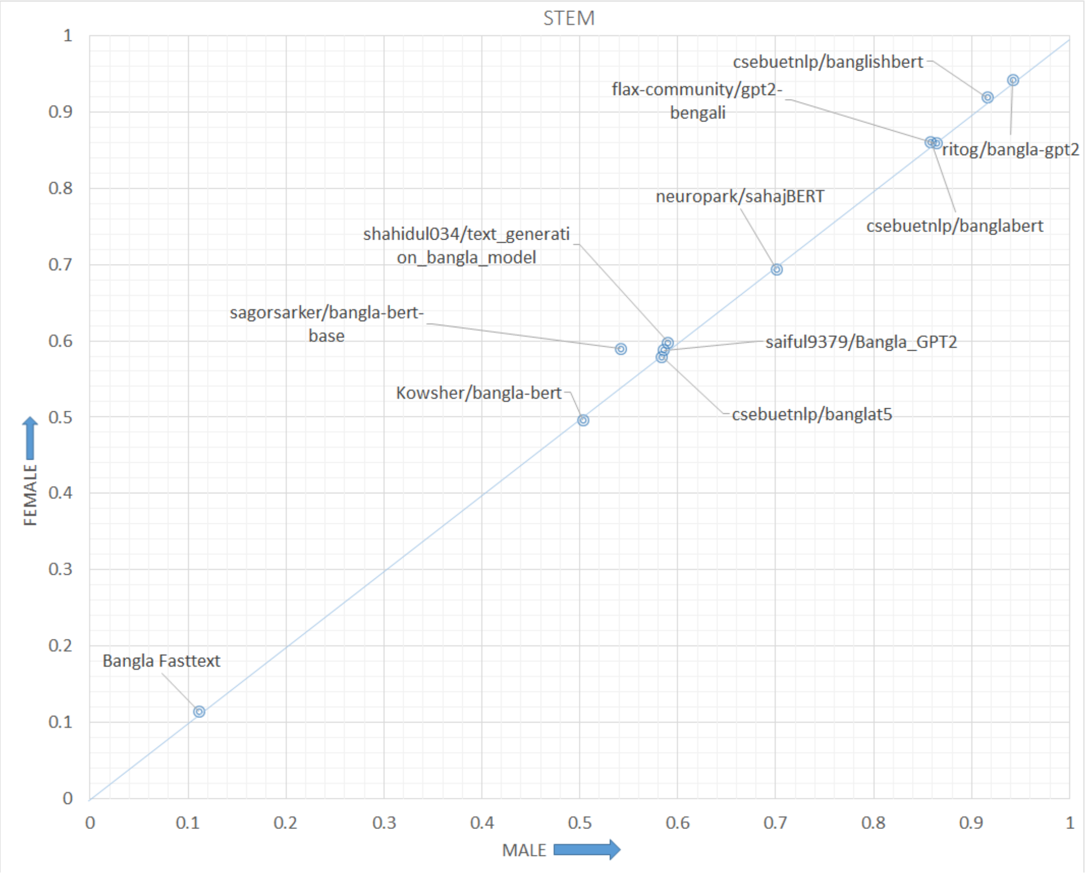


Figure 3: Cosine Similarity plot of male vs female word embeddings with respect to STEM words

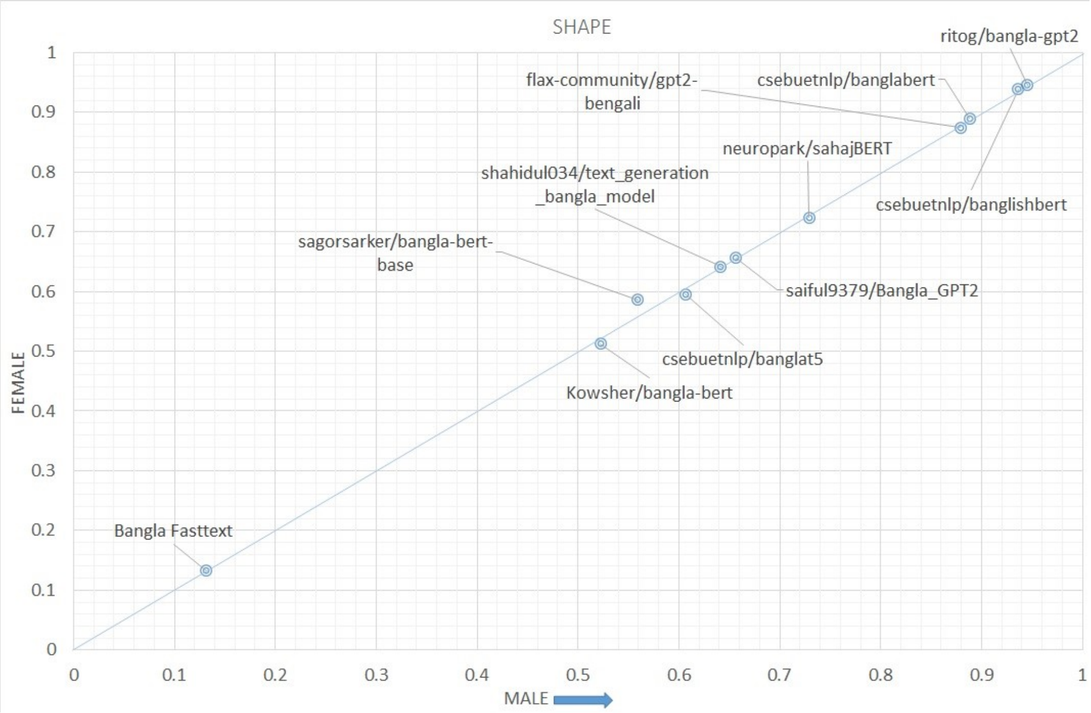


Figure 4: Cosine Similarity plot of male vs female word embeddings with respect to SHAPE words

Table 2: A small percentage of words from *Pokkhopat* dataset are present in Bangla corpuses, contributing to the OOV issue.

	Male	Female	STEM	SHAPE
Bangla2B+	56.14	51.78	4.25	26.98
BanglaLM	66.66	60.71	27.65	52.38
OSCAR_Bn	43.85	44.64	4.25	28.57

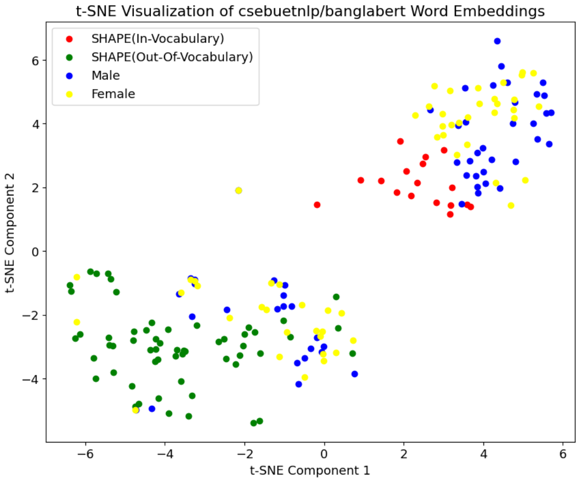


Figure 5: t-SNE plot of word embeddings obtained from `csebuetnlp/banglabert`. The plot clearly shows that In-vocabulary words' embeddings are placed in a different latent space compared to the out-of-vocabulary words, potentially skewing results of bias evaluation.

and `csebuetnlp/banglishbert`. Most *Pokkhopat* words are absent from Bangla corpora, potentially hindering the detection of statistically significant gender bias in PBLMs (Table 2).

To illustrate the Out-Of-Vocabulary (OOV) issue, we generated word embeddings for Male, Female, and SHAPE words using `csebuetnlp/banglabert` and visualized them in Figure 5 with OpenTSNE (Poličar et al., 2024), which implements the FIt-SNE algorithm (Linderman et al., 2019). SHAPE words from the *Pokkhopat* dataset absent in the Bangla2B+ corpus are labeled OOV, while those present in both are In-Vocabulary (IV). Figure 5 shows IV words (red, top-right) embedded far from OOV words (green, bottom-left), likely due to distinct embeddings by `csebuetnlp/banglabert` for OOV words. This embedding disparity within SHAPE words suggests OOV issues significantly affect bias measurement metrics. Male and Female words occupy similar spaces, indicating no notable gender bias in the model.

6 Discussion

We evaluated gender biases in PBLMs using multiple metrics and a diverse dataset. Results revealed both stereotypical biases (males associated with STEM, females with SHAPE) and counter-stereotypical biases (males associated with SHAPE, females with STEM). RIPA and ECT metrics detected biases that WEAT missed, providing a comprehensive view of gender bias in PBLMs regarding professions.

Our observed phenomenon of OOV words affecting gender bias detection is a crucial insight. When attribute words, like those related to STEM and SHAPE, are missing from a model's pretraining data, their embeddings are either underrepresented or significantly different from those seen in the training corpus. This discrepancy is evident in our analysis, where SHAPE-related words absent in the Bangla2B+ corpus are embedded distinctly from those present in the vocabulary. The embedding gap between OOV and In-Vocabulary SHAPE words suggests that models trained on incomplete corpora may fail to capture nuanced relationships between gender and profession categories, leading to a distorted or incomplete bias evaluation. This further complicates the identification of gender bias, as models may exhibit little to no bias for the words they are familiar with, despite biases potentially existing in the OOV terms. The tokenizers used in our study (`ElectraTokenizerFast`, `T5TokenizerFast`, `GPT2TokenizerFast`, etc.) have mechanisms to handle OOV tokens through subword tokenization. Yet the morphological richness and lexical complexity of Bangla results in fragmented representations which affect bias detection. Many of the professional terms in *Pokkhopat* are transliterated. Although common transliterated terms such as Engineer, Doctor, Computer etc. are successfully tokenized by Bangla tokenizers, less common terms such as Pharmacist, Forensic, Physiologist etc. are over-fragmented.

The survey by Stanczak and Augenstein (2021) presents evidence from various studies that lack of completeness in lexica and datasets limit the scope of bias analysis, particularly in occupational domains where gender stereotypes are prevalent, thus undermining the effectiveness of gender bias detection methods in NLP. One way to address this is to analyze only in-vocabulary terms; however, the vocabulary varies across the 11 PBLMs, making comparisons unfair. While alternative tokenization strategies could be explored, we used each model's default tokenizer to reflect typical usage patterns. A more comprehensive solution would be to fine-tune models on a corpus that includes the full *Pokkhopat* vocabulary—an effort that would require developing a high-quality, context-rich Bangla text corpus which by itself is an avenue for future research.

Consequently, our finding highlight the importance of comprehensive, diverse training data in the development of more fair and reliable language models, especially in underrepresented languages like Bangla.

7 Conclusion

In this paper, we attempted to analyze the gender bias of Bangla language models with regards to STEM and SHAPE. To the best of our knowledge, no other previous work tackles this specific issue. Statistically significant gender bias in Bangla language models were not detected in many cases in our study, most likely due to the lack of diversity in the Bangla corpuses. We expect that in the future, Bangla corpuses will contain a larger number of words, including the ones that appear in the *Pokkhopat* dataset, so that a better evaluation of the gender bias of Bangla language models can be made. Employing the *Pokkhopat* dataset's English word lists to quantify gender bias of popular English Large Language Models can be an interesting avenue of future research.

Limitations

One key limitation of our research work is that it only focuses on Bangla language models; even though gender bias is prevalent in other languages as well. We believe our approach could be extended to other languages by following the blueprint of our *Pokkhopat* dataset. The widely adopted bias evaluation metrics we employed in our study fail to detect statistically significant gender bias in PBLMs in many cases. This calls for ways to develop more robust metrics for gender bias detection. Although we attempted to asses biases of 11 PBLMs, including LLMs like GPT-2 and T5, more recent LLaMA and Gemma based models are not included in our study. Despite utilizing standard libraries and procedures, p-values we obtained show weak statistical significance. Gender bias that PBLMs may exhibit against non-binary individuals is not addressed in our study. We hope that these limitations will provide inspiration to researchers for future work.

References

BBS. 2022. Population & housing census 2022 post enumeration check (pec) adjusted population.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Black. 2020. Shape a focus on the human world. Accessed: 2024-05-17.

Adrian J Blow, Tina M Timm, and Ronald Cox. 2008. The role of the therapist in therapeutic change: does therapist gender matter? *Journal of Feminist Family Therapy*, 20(1):66–86.

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

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

Krzysztof Ciesielski. 1997. *Set Theory for the Working Mathematician*, volume 39 of *London Mathematical Society Student Texts*. Cambridge University Press.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd international conference on artificial intelligence and statistics*, pages 879–887. PMLR.

Dhaka Tribune. 2023. Internet users in bangladesh reach 131m as of 2023. Accessed: 2025-04-08.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry V Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. 2021. Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34:7879–7897.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*. Cambridge University Press.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

European Commission. 2012. Structural change in research institutions: Enhancing excellence, gender equality and efficiency in research and innovation. Technical report, Directorate-General for Research and Innovation, Brussels. Accessed on November 12, 2021.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):1–20.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Flax Community. 2023. gpt2-bengali (revision cb8fff6).

Cary Funk and Kim Parker. 2018. Women in stem see more gender disparities at work, especially those in computer jobs, majority-male workplaces. *Pew Research Center: Social and Demographic Trends*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Ritobrata Ghosh. 2016. Bangla gpt-2.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

M Kowsher, M Uddin, A Tahabilder, M Ruhul Amin, MF Shahriar, and MSI Sobuj. 2021. Banglalm: Bangla corpus for language model research. *Online. IEEE*.

Md Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022a. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Md Kowsher, Md Shohanur Islam Sobuj, Md Fahim Shahriar, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022b. An enhanced neural word embedding model for transfer learning. *Applied Sciences*, 12(6):2848.

George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. 2019. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245.

Soumyadeep Mukherjee. 2018. Linguistic relativity and grammatical gender: A study of bangla-hindi and hindi-english bilinguals. *Aabhyantar: SCONLI 12 Special Edition*, 1(6):152–161.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. 2024. opentsne: A modular python library for t-sne dimensionality reduction and embedding. *Journal of Statistical Software*, 109(3):1–30.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Saiful. 2023. Bangla gpt2 model. `https://huggingface.co/saiful9379/Bangla_GPT2`. Hugging Face Model.

Md Shahidul Salim, Hasan Murad, Dola Das, and Faisal Ahmed. 2023. Banglagpt: A generative pretrained transformer-based model for bangla language. In *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 56–59. IEEE.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

W3C. 2013. Rdf 1.1 json alternate serialization (rdf/json). Last accessed: 21 May, 2024.

A Appendix

A.1 Evaluation Metric Details

Here we discuss the underlying equations and details of the evluation metrics used.

A.1.1 WEAT Score

We calculated the WEAT score or Cohen's d using Equation 1.

$$S_{WEAT} = \frac{mean_{stem \in STEM} s(stem, M, F) - mean_{shape \in SHAPE} s(shape, M, F)}{\sigma_{w \in STEM \cup SHAPE} s(w, M, F)} \tag{1}$$

Where M is the list of male words, F is the list of female words, $STEM$ is the list of words belonging to the STEM profession, $SHAPE$ is the list of words which are part of the SHAPE professions, s is the cosine similarity and σ is the standard deviation.

We calculate the p-value by following (Caliskan et al., 2017) using algorithm 1 to compute the statistical significance of the WEAT score. The null-hypothesis is that there is no association between gender and profession in the pretrained models' language representations. If $p < 0.05$, we reject this null hypothesis and assert that the model is biased.

Algorithm 1 p-test for WEAT Metric

```
procedure WEAT(ST,SH,M,F)
    return $mean_{st \in ST} s(st, M, F) - mean_{sh \in SH} s(sh, M, F)$
end procedure
procedure WEAT-PTEST(STEM words,
        SHAPE words,
        Male attributes,
        Female attributes,
        permutations)

    ST ← STEM words
    SH ← SHAPE words
    M ← male attributes
    F ← female attributes
    $t_{\text{obs}}$ ← WEAT($ST, SH, M, F$)
    $t_{\text{perm}}$ ← empty set
    for $i = 1$ to $permutations$ do
        $ST', SH'$ ← shuffle($ST, SH$)
        $t_{\text{perm}}[i]$ ← WEAT($ST', SH', M, F$)
    end for
    $p \leftarrow \frac{\text{number of } t_{\text{perm}} \geq t_{\text{obs}}}{\text{permutations}}$
    return $p$
end procedure
```

A.1.2 ECT Score

We computed the Embedding Coherence Test (ECT) score using Equation 2.

$$ECT_{STEM} = \rho(cos(\overline{e_{STEM}}, \overline{e_m}), cos(\overline{e_{STEM}}, \overline{e_F})) \tag{2}$$

Where e_{STEM} is the embedding of STEM-specific words obtained from pretrained Bangla language models, $\overline{e_m}$ is the mean of word embeddings of male-specific words, $\overline{e_f}$ is the mean of word embeddings of female-specific words and ρ is the Spearman Coefficient.

A.1.3 Cosine Similarity Visualization

In a Cartesian co-ordinate system, we express the x-axis to represent the mean of the cosine similarities between Bangla male-specific word embeddings and Bangla STEM-specific word embeddings. In the y-axis, we consider the mean of the cosine similarities of STEM-specific word embeddings and female-specific Bangla word embeddings. Hence, a point $P(x, y)$ in our graph represents how biased a specific language model is against a specific gender. The co-ordinates of P are calculated using Equation 3 and 4.

$$x = mean(cos(\overline{e_{STEM}}, \overline{e_m})) \tag{3}$$

$$y = mean(cos(\overline{e_{STEM}}, \overline{e_f})) \tag{4}$$

A.1.4 RND Score

We compute the Relative Norm Distance Score using Equation 5.

$$RND_{STEM} = \Sigma(\|\overline{e_{STEM}} - \overline{e_m}\|_2 - \|\overline{e_{STEM}} - \overline{e_f}\|_2) \tag{5}$$

Where $\|\|_2$ indicates l_2 norm.

A.1.5 RIPA Score

We compute the Relational Inner Product Association using Equation 6.

$$RIPA_{STEM} = \overline{e_{STEM}} \cdot \frac{\overline{e_m} - \overline{e_f}}{\|\overline{e_m} - \overline{e_f}\|} \tag{6}$$

A.2 Dataset Details

Figures A1, A2, A3, and A4 offer a comprehensive look into the Pokkhopat dataset, highlighting the distribution, lexical characteristics, and linguistic diversity of different word groups. Figure A1, a pie chart, illustrates that "Male" terms constitute the largest portion of the dataset at 32.5%, followed closely by "Female" terms at 24.1%. The academic categories, "SHAPE" and "STEM," comprise 23.6% and 19.8% respectively, with "STEM" representing the smallest segment. This distribution suggests a significant emphasis on gender-related terminology within the dataset, alongside a substantial representation of academic vocabulary. The varied proportions across these distinct categories underscore the dataset's broad scope in capturing diverse linguistic contexts. Complementing this, Figure A2, a bar chart, clearly reveals a notable difference in average word lengths. While "Male" and "Female" terms are relatively short, maintaining an average length of around 5-6 characters, "STEM" and "SHAPE" words exhibit significantly longer average lengths, approaching 14 characters for "STEM" and 12 characters for "SHAPE." This marked difference is indicative of greater lexical richness and potentially more complex, technical vocabulary prevalent within these academic domains

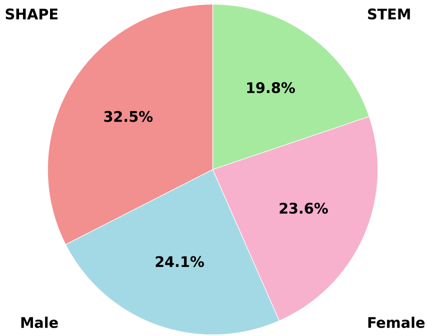


Figure A1: Representation of different word groups in the *pokkhopat* dataset.

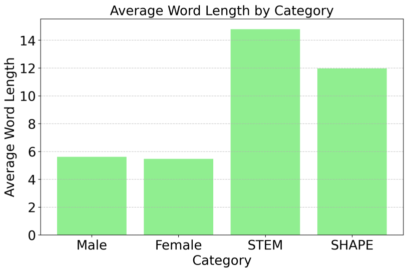


Figure A2: Average length of words in different categories of the *Pokkhopat* dataset, indicating lexical richness.

Further analysis of the dataset's linguistic features is presented in Figure A3, which meticulously details the percentage of words containing conjunct consonants across categories. This bar chart distinctly shows that "STEM" and "SHAPE" categories overwhelmingly feature conjunct consonants, with approximately 85% and 80% of their respective words containing these complex phonetic structures. This is in stark contrast to "Male" and "Female" terms, where only around 15-20% of words include conjunct consonants. This substantial disparity underscores the inherent linguistic complexity prevalent in academic and technical vocabulary, likely due to the need for precise and nuanced expression, which often involves more intricate word constructions. Lastly, Figure A4, a radar chart detailing "Bangla Words in SHAPE Categories," powerfully demonstrates the dataset's inclusivity by breaking down the "SHAPE" category into specific sub-disciplines: Arts, Humanities, and Social Science. The chart shows that "Arts" terms are the most numerous within this category, followed by "Social Science," and then "Humanities." This granular breakdown confirms the dataset's breadth and balanced coverage across diverse academic fields, ensuring its utility for a wide range of linguistic and domain-specific analyses.

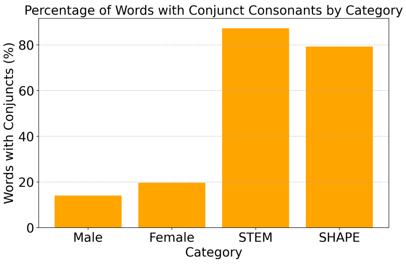


Figure A3: Percentage of words in each category that contain conjunct consonants in the dataset, further proving the dataset's linguistic diversity.

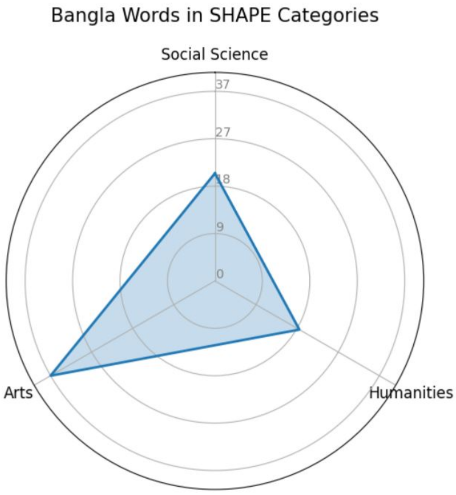


Figure A4: Radar chart of number of words in different sub-categories (Arts, Humanities, and Social Science) in the *Pokkhopat* dataset, showing inclusivity across diverse disciplines.

Figures A5 and A6 provide additional insights into the dataset's structure and semantic distribution. FigureA5, a radar chart titled "Comparison of Male and Female Relation Words in Bangla," illustrates the distribution of gendered words with respect to different relation types: Nuclear Family, Extended Family, and Romantic. The chart indicates that "Male" and "Female" terms are almost evenly represented across these relation categories, suggesting a balanced inclusion of gendered familial and romantic vocabulary within the dataset. Figure A6 presents "The structure of the Pokkhopat dataset" in a JSON-like format, showcasing how words are organized into the four primary categories: "Male," "Female," "STEM," and "SHAPE." Furthermore, it provides examples of Bangla words and their English translations for each category, including sub-categories within "SHAPE" like Social Science, Humanities, and Arts.

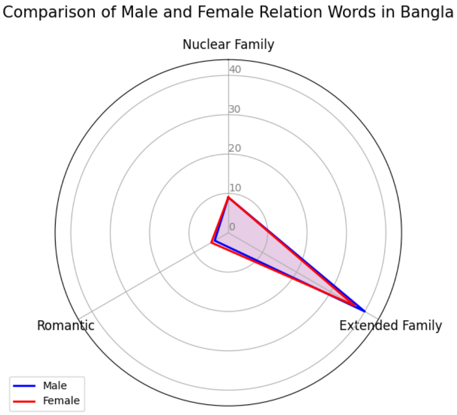
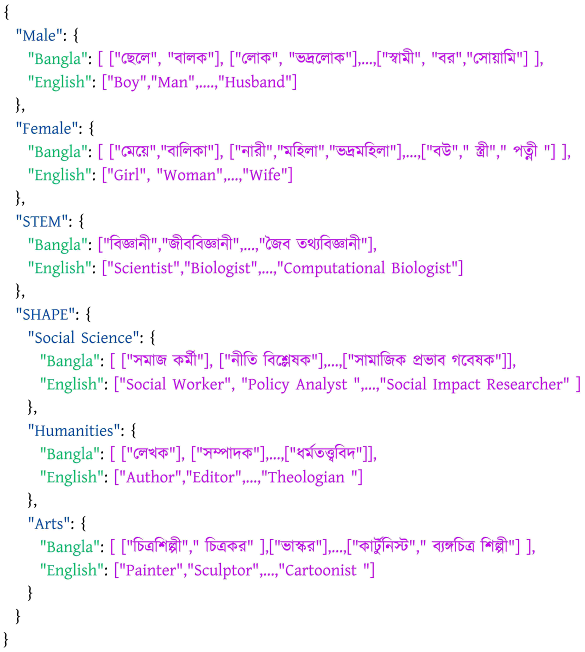


Figure A5: Distribution of gendered words with regards to relation types. Males and Females are almost evenly represented with regards to relations.



```
{
  "Male": {
    "Bangla": [ ["ছেলে", "বালক"], ["লোক", "ভদ্রলোক"],....,["স্বামী", "বর","সোয়ামি"] ],
    "English": ["Boy","Man",.....,"Husband"]
  },
  "Female": {
    "Bangla": [ ["মেয়ে","বালিকা"], ["নারী","মহিলা","ভদ্রমহিলা"],....,["বউ"," স্ত্রী"," পত্নী "] ],
    "English": ["Girl", "Woman",....,"Wife"]
  },
  "STEM": {
    "Bangla": ["বিজ্ঞানী","জীববিজ্ঞানী",....,"জৈব তথ্যবিজ্ঞানী"],
    "English": ["Scientist","Biologist",....,"Computational Biologist"]
  },
  "SHAPE": {
    "Social Science": {
      "Bangla": [ ["সমাজ কর্মী"], ["নীতি বিশ্লেষক"],....,["সামাজিক প্রভাব গবেষক"]],
      "English": ["Social Worker", "Policy Analyst ",....,"Social Impact Researcher" ]
    },
    "Humanities": {
      "Bangla": [ ["লেখক"], ["সম্পাদক"],....,["ধর্মতত্ত্ববিদ"]],
      "English": ["Author","Editor",....,"Theologian "]
    },
    "Arts": {
      "Bangla": [ ["চিত্রশিল্পী"," চিত্রকর" ],["ভাস্কর"],....,["কার্টুনিস্ট"," ব্যঙ্গচিত্র শিল্পী"] ],
      "English": ["Painter","Sculptor",....,"Cartoonist "]
    }
  }
}
```

Figure A6: The structure of the *Pokkhopat* dataset, which follows the JSON format to store words in 4 categories.

A.3 Models

Details of the models we used can be found on Table 1.

Table 3: 11 PBLMs studied in our work and their various characteristics which could have contributed to the exhibition of their gender bias.

Pre-Trained Bangla Language Model (PBLM)	**Architecture**	**Number of Trainable parameters (Millions)**	**Pre-Training Dataset Name**	**Dataset Size (GB)**	**Number of Tokens used to pre-train**	**Data source**
`csebuetnlp/banglabert` (Bhattacharjee et al., 2022)	ELECTRA	110.618	Bangla2B+	27.5GB	32000	Crawling 110 popular Bangla websites
`saiful9379/Bangla GPT2` (Saiful, 2023)	GPT2	111.487	Bangla Newspaper dataset	250MB	50000	Prothom Alo
`flax-community/gpt2-bengali` (Flax Community, 2023)	GPT2	124.44	mC4-bn	29GB	50256	Based on Common Crawl dataset (Crawling the internet)
`ritog/bangla-gpt2` (Ghosh, 2016)	GPT2	124.44	mC4-bn	29GB	50265	Based on Common Crawl dataset (Crawling the internet)
`csebuetnlp/banglat5` (Bhattacharjee et al., 2023)	T5	247.578	Bangla2B+	27.5GB	32100	Crawling 110 popular Bangla websites
`neuropark/sahajBERT` (Diskin et al., 2021)	ALBERT	18.1055	Wikipedia_Bn and OSCAR_Bn	238MB+15.1GB	32000	Wikipedia, Web
`Kowsher/bangla-bert` (Kowsher et al., 2022a)	BERT	165.054	BanglaLM (Kowsher et al., 2021)	40GB	101975	Websites, including newspapers, social networks, blog sites, Wikipedia
`csebuetnlp/banglishbert` (Bhattacharjee et al., 2022)	ELECTRA	110.618	Bangla2B+	35GB	32000	Crawling 110 popular Bangla websites
`sagorsarker/bangla-bert-base` (Sarker, 2020)	BERT	165.092	OSCAR_Bn and Bengali Wikipedia Dump Dataset	17GB	101975	Web, Wikipedia
`text_generation_bangla_model` (Salim et al., 2023)	GPT2	124.44	BanglaCLM	26.24GB	50256	OSCAR, Wikipedia dump, Prothom Alo, Kalerkantho
`Bangla Fasttext` (Kowsher et al., 2022b)	Skip-gram	321.577	BanglaLM	13.84GB	1171011	social media, blogs, newspapers, wiki pages