SUST Journal of Science and Technology

journal homepage: https://journals.sust.edu

SUST JOURNAL OF SCIENCE AND TECHNOLOGY

Evaluating Bias in Low Resource MLMs: A Comprehensive Study on Measuring Gender Bias in BERT for Bangla

Ahsan Habib[1], Sadia Tasnim Meem[1], Sadia Islam Hridi[1]

Shahjalal University of Science and Technology, Sylhet – 3114, Bangladesh

ARTICLE INFO

Keywords:

Large Language Models
Low resource LLMs
Bias measurement
Language specific metric
BERT
Bangla LLMs.

ABSTRACT

This paper investigates gender–profession associations in Bengali BERTfamily models using a controlled masked language model (MLM) probing framework. We construct a systematic evaluation corpus of 64,800 sentences using a full factorial design over gender-denoting person words, profession terms, and sentence templates. Using prior-normalized log-probability and sigmoid-based scoring, we measure how profession context influences the likelihood of gendered terms. Across three models (mBERT, BUET BanglaBERT, and Sagor Sarker BanglaBERT) we observe consistent variations in association strength, with a general tendency toward stronger alignment with male-denoting words, alongside notable exceptions across profession categories. These findings reflect learned distributional patterns rather than explicit rules.We further examine the applicability of an Englishoriginated bias measurement framework in Bengali, highlighting challenges in interpretation due to linguistic and data differences. Overall, this work provides a controlled methodology and empirical analysis for studying genderprofession associations in low-resource language models.

1. Introduction

Large Language Models (LLMs), particularly transformer-based models such as BERT [1], have become central to modern natural language processing. Alongside their success, however, concerns have emerged regarding their tendency to learn and reproduce societal biases present in training data. This issue is especially pronounced in low-resource languages like Bengali, where limited and potentially imbalanced corpora can amplify such biases.

In this work, we examine gender bias in Bengali BERT-family models through the lens of the Masked Language Modeling (MLM) objective [2]. Rather than relying on static representations, we probe the model's predictions directly by measuring how strongly profession-denoting words associate with male and female contexts. To quantify this, we compute prior-normalized log-probabilities of target tokens and apply a sigmoid transformation to obtain comparable bias scores.

To ensure that our findings are not artifacts of prompt design, we construct a fully controlled evaluation corpus using a strict factorial setup, varying person words, professions, and sentence templates. We also introduce a with/without names ablation to test whether gender bias depends on explicit gender cues or persists within the semantic representation of profession words themselves. Finally, we evaluate this behavior across multiple models, including multilingual and Bengali-specific BERT variants, to better understand how training data and model design influence bias.

Key Contributions and Insights

Objective: Systematically measure gender–profession bias in Bengali BERT models using MLM probing.

* Corresponding author.
E-mail addresses: ahabib-iict@sust.edu (A. Habib)

https://doi.org/10.63512/sustjst.2025.1002
Received 5 January 2026; Received in revised form 22 April 2026; Accepted 11 May 2026

Contributions:

1. A 64,800-sentence fully factorial Bengali evaluation corpus
2. A log-probability framework for bias quantification
3. A with/without names ablation to isolate the source of bias
4. Comparative analysis across multilingual and Bengali-specific BERT models

Novelty: Unlike most prior work that relies on static embeddings, we use MLM-based probing to study gender–profession associations and explicitly test whether bias is context-triggered or embedded in profession semantics.

2. Literature Review

A significant amount of work has already been done on developing the metric to asses different kinds of biases in large language models (LLMs) but they are predominantly dedicated for English language. Basta et. al. analyzed contextualized word embeddings adapting gender bias measures for word embedding methods from previous work [3]. Their work finds insights into how bias is projected in contextual word embeddings compared to standard embeddings. Zhao et. al. quantified gender bias exhibitied in ELMo's contextualized word vectors by using WinoBias dataset, a benchmark for evaluating male and female coreference resolution examples [4]. Their contribution include a coreference system relying on ELMo inheriting bias. In a similar approach Nadeem et. al. introduced StereoSet, a comprehensive dataset comprising 16,995 Context Association Tests (CATs) across four domains: gender, profession, race and religion [5]. CAT frameword measures stereotypical bias by analyzing contexts which include stereotypical, anti-stereotypical and unrelated contexts. Aiming to quantify social biases in language models May et. al. developed three novel metrics: Target Bias (TB), Bias Amount (BAmt), and Persona Bias (PB) [6]. Zhang et al. investigates biases in clinical language models, particularly focusing on marginalized populations [7]. The methodology includes pretraining BERT on MIMIC-III medical notes and assessing biases through a fill-in-the-blank approach and log probability scoring. The authors propose best practices for mitigating bias in clinical applications of word embeddings Focusing on their reinforcement of stereotypes, Kotek et al. examines gender biases in large language models (LLMs) [8]. The methodology encompasses a fresh testing paradigm on the WinoBias dataset to assess biased occupational associations. Findings show that LLMs are 3–6 times more likely to match occupations associated the traditional roles related to their genders, as societal impressions rather than actual job statistics.

For the exploitation of underlying MLM (Mask Language Modelling) Kurita et al. proposed an approach of a template-based method to quantify bias in BERT embeddings [9]. The methodology demonstrates that this approach yields more consistent results in capturing social biases compared to traditional cosine similarity methods. Bartl et al. investigates gender bias in BERT models using a methodology that involves creating a template-based corpus in English and German to measure bias through gender denoting sentence templates [10].

In recent bengali specific works, Sadhu et al. [11] present one of the earliest studies applying MLM-based probing with prior-normalized log-probability scoring to Bengali. Their work adapts established techniques to a low-resource setting and demonstrates that measured bias is sensitive to contextual factors such as sentence length. However, their evaluation focuses on general semantic associations (e.g., pleasant vs. unpleasant, career vs. family) rather than domain-specific biases such as profession–gender relationships.

Subsequent work by Sadhu et al. [12] shifts attention toward generative large language models, evaluating gender and religious bias in Bengali using template-based and naturally occurring prompts. Their use of Disparate Impact (DI) on generated outputs highlights bias at the level of model behavior, but does not directly analyze token-level associations within encoder models. In contrast, Arnob et al. [13] examine gender bias in Bengali models using static embedding-based metrics such as WEAT and cosine similarity, focusing on STEM and SHAPE profession categories. While their findings provide useful insights, the absence of sentence-level context limits the ability to capture how bias manifests during actual language understanding. Other studies further expand the scope of analysis. Hosain and Morol [14] investigate how linguistic formality and context length influence bias, extending both embedding-based and probabilistic methods. Meanwhile, Reaj et al. [15] adopt a broader, cross-lingual and community-grounded approach, combining retrieval, translation, and human-informed data collection to highlight the limitations of English-centric bias evaluation frameworks in Bengali.

Despite these advances, a key gap remains. Existing Bengali bias studies either rely on static embeddings, focus on general semantic traits, or evaluate generative outputs, leaving profession–gender associations in encoder-based models relatively underexplored. Moreover, none of these works explicitly examine whether such bias depends on surface-level gender cues or is embedded within the semantic representation of profession words themselves.

This work addresses these gaps by applying MLM-based probing with prior-normalized log-probability scoring to measure gender–profession bias in Bengali BERT-family models, including mBERT and Bengali-specific variants. Unlike prior studies, we introduce a with/without names ablation that isolates the role of explicit gender markers, enabling us to distinguish between context-triggered and representation-level bias. Through a fully controlled factorial evaluation design, our study provides a more fine-grained and systematic analysis of how gender bias manifests in profession-related language modeling tasks.

3. Dataset

Our dataset follows a template-based evaluation framework, where sentences are generated by filling fixed templates with controlled lexical slots. We focused on different gender denoting words and their association with different professions.

The corpus includes 36 person words, consisting of 18

male-denoting and 18 female-denoting expressions, these cover kinship terms (e.g., বাবা/মা, ভাই/বোন), formal honorifics (ভদ্রলোক/ভদ্রমহিলা), common nouns (ছেলে/মেয়ে), colloquial forms (লোকটা/মহিলাটা), and selected proper names (e.g., জামাল/আসমা). The profession set contains 120 occupations, grouped for analysis into 40 male-dominated professions, 40 female-dominated professions, and 40 neutrally/equally dominated professions. This categorization is used strictly for post-hoc analysis. We use 15 sentence templates, resulting in a fully expanded corpus of 64,800 sentences ($36 \times 120 \times 15$).

The central design principle is complete permutation coverage, every person word is paired with every profession across all templates. This ensures that no profession is preferentially associated with any gender marker, eliminating pairing bias and allowing clean attribution of differences in MLM scores to gender substitution alone. Importantly, we deliberately avoid linguistic filtering beyond grammatical correctness, since the objective is not natural sentence modeling but measuring latent model behavior under controlled perturbation of gender context.

To further strengthen causal interpretation, we introduce a with-name variant, containing common bangla gender specific names like আরিফ, জামাল, আসমা, সাহানা, and without-name variant, where common Bangla name-like tokens are removed to test whether pretrained models rely on explicit name-based gender priors or whether bias persists even in their absence. Together, these design choices ensure that observed effects reflect internal learned associations of the model rather than artifacts of prompt design, making this a behavioral probing framework rather than a prompt engineering approach.

4. Methodology

4.1. Masking for data corpus generation

Models like BERT use contextual word embeddings. Replacing type level embedding, they assign every word a different embedding depending on the context. BERT use a masked language modeling, which is a self supervised learning technique. Here the model predicts a missing text which has been "masked" based on the surrounding context. This is introduced to facilitate BERT's bidirectional nature. This intra-structural aspect of BERT allows to query the model to measure the bias of a particular word or a token denoted as `[MASK]`.

Querying the MLM includes creating simple template sentences and then masking them.

We specifically mean gender denoting person words by targets and professions by attributes.

4.2. Bias measurement metric

We rely on two complementary scoring functions based on prior-normalized likelihoods. These metrics are designed to capture how much a profession context shifts the probability of a gendered person word relative to its base frequency in the model.

Log-Probability Association Score

We define the primary bias metric as a log probability ratio:

$$S_{\log} = \log\left(\frac{P(T \mid TM)}{P(T \mid TAM)}\right) \tag{1}$$

We define the gender denoting word to be the Target word T, and profession denoting word to be the Attribute word A.

TM (Target-Masked) is the sentence template where only the person word is masked but profession context is visible, (eg.: `[MASK]` একজন ব্যবস্থাপক). That is, the profession denoting word is not masked. TAM (Target-Attribute-Masked) is the fully masked template, where both the Target and the Attribute is masked (eg: `[MASK]` একজন `[MASK]`).

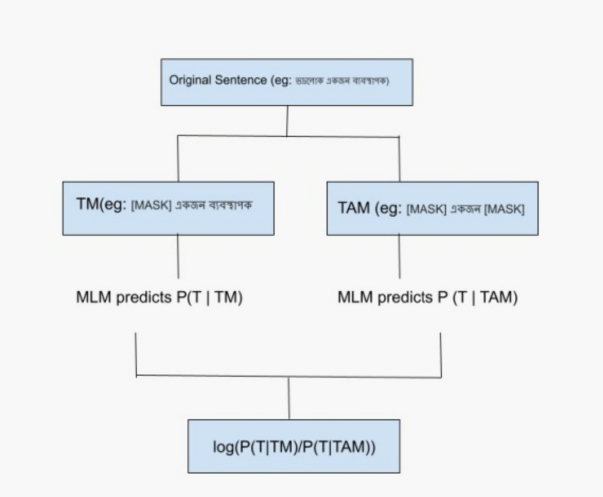

Figure 1. Metric measurement flow

This formulation measures the contextual lift induced by profession information. A higher value indicates that the profession context increases the likelihood of the target person word relative to its prior expectation (the probability without the profession context), while a lower value indicates suppression. This metric is chosen because MLM probabilities are highly skewed and context-sensitive; therefore, raw probabilities are not comparable across tokens. The log-ratio formulation normalizes for token frequency effects and isolates the incremental contribution of profession context to gender prediction.

Sigmoid-Transformed Score

To improve interpretability and enable bounded aggregation, we additionally rely on a sigmoid-transformed score:

$$S_{\sigma} = \frac{1}{1 + e^{-S_{\log}}} \tag{2}$$

The sigmoid score is presented as an exploratory variant, not a replacement for the log-probability ratio. Its purpose is to compress unbounded log ratios into $[0, 1]$ for interpretability as a normalized association strength. We do not claim theoretical superiority over the raw log ratio; rather, it serves as a sensitivity check. if both scoring

methods yield consistent bias patterns, the findings are robust to the choice of transformation. We agree that comparative validation against alternative normalization strategies (e.g., min-max scaling, z-score standardization) would strengthen this analysis,

This study focuses on masked language model (MLM) probing, which measures bias through token prediction probabilities rather than embedding geometry. Metrics like WEAT, cosine-based association, relative norm difference, and inner-product scores operate on static or contextualized embeddings — a fundamentally different measurement paradigm. While recent Bangla studies (e.g., Arnob et al., 2025) [13] have applied these embedding-based metrics, they capture where words sit in vector space, not how gendered context causally shifts model predictions. Our log-probability ratio isolates the specific influence of gender context from baseline prediction tendencies, addressing a different layer of bias. Including embedding-based metrics would conflate two distinct measurement approaches; however, we acknowledge that cross-method comparison is a valuable direction for future work.

4.3. Models

Our bias measurement framework relies on masked language model (MLM) probing, which requires models trained with the standard masked token prediction objective and equipped with an MLM classification head to extract token-level probabilities. For this reason we specifically selected BERT based models only. Two of them are mono lingual and one of them are multi lingual model. The mono lingual model used:

1. Bangla BERT (follows the original BERT-base model with 12 layers (transformer blocks), 12 attention heads, and a hidden size of 768): BanglaBERT, developed by the BUET NLP group, is a Bangla-specific language model based on the BERT architecture with 110 million parameters. Trained on a massive 27 GB Bangla text corpus from diverse sources, it leverages the Masked Language Modeling (MLM) objective to learn contextual word representations.
2. Sagor sarker's bangla BERT (Based on BERT-base with 12 layers, 12 attention heads, and 768 hidden units.): Bangla-BERT-Base is a Bengali pretrained language model based on the BERT architecture, trained using the masked language modeling (MLM) objective.

The multilingual model used:

1. mBERT (Multilingual BERT) (Same as BERT-base (12 layers, 768 hidden units, 12 attention heads): Multilingual BERT (mBERT) is a pretrained language model by Google designed for 104 languages, including Bangla. The shared 119,547-token vocabulary was built using a WordPiece tokenizer from multilingual Wikipedia data. Trained using masked language modeling (MLM) and next sentence prediction (NSP) objectives, mBERT supports cross-lingual transfer but may lack depth in individual language representation due to its multilingual focus. Despite this, it is a powerful baseline for multilingual NLP tasks and low-resource languages like Bangla.

We used the hugging face API to examine each of these pretrained models.

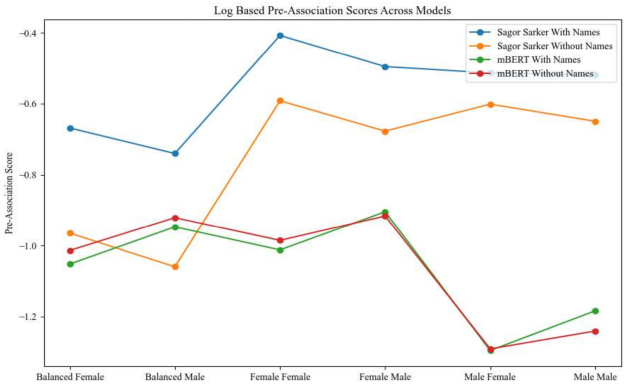

Figure 2. Log based pre-association score.

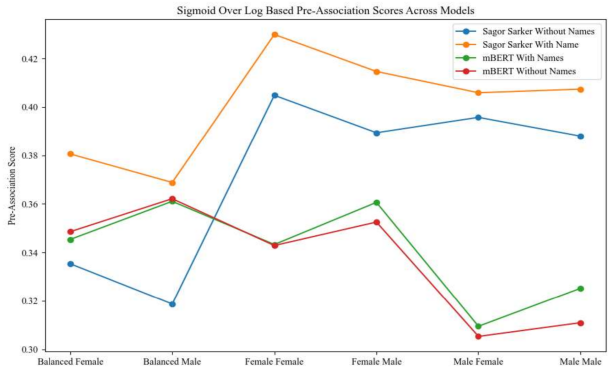

Figure 3. Sigmoid based pre-association score

5. Result Analysis

We applied log scoring on three Bangla language models—mBERT, BanglaBERT by BUET, and BanglaBERT by Sagor Sarker—using two datasets which is shown in table 1. The first dataset contained all gender-denoting words excluding names, while the second included names.

The log scoring produced negative results across all models, with some female words showing strong associations with male terms or balanced professions, which is not meaningful. To address this, we applied a sigmoid scoring method, where the association scores aligned more closely with the expected gender associations.

Male words were, as predicted, more strongly associated with the attributes, even though the scores were positive for all models. Interestingly, the same scoring system did not produce satisfactory results for Bangla, but it did perform well for English. It emphasizes the necessity of creating new assessment metrics and investigating the training set of current Bangla BERT models in order to enhance their performance in tasks pertaining to gender. The figure 2 shows log based pre-association score and the figure 3 shows sigmoid based pre-association score.

Table 1. Preassociation score across models

Model name	Name variant	Gender of Profession	Gender	Pre association sigmoid	Pre association log
mBERT	Without names	balanced	female	0.3485426613	−1.013930731
			male	0.3621372929	−0.9211114627
		female	female	0.3428604917	−0.9854384922
			male	0.3525201868	−0.9164653174
		male	female	0.3052915446	−1.291189707
			male	0.3109551776	−1.240933327
	With names	balanced	female	0.34535423	−1.052416817
			male	0.3610662164	−0.9472101651
		female	female	0.3432783339	−1.012281901
			male	0.3606259946	−0.904504941
		male	female	0.3094550239	−1.296010476
			male	0.3251797598	−1.184121967
Bangla BERT by Sagor Sarker	Without names	balanced	female	0.3353530093	−1.013930731
			male	0.3187089482	−0.9211114627
		female	female	0.4048355216	−0.9854384922
			male	0.3894116663	−0.9164653174
		male	female	0.3957622161	−1.291189707
			male	0.3880238625	−1.240933327
	With names	balanced	female	0.3805831344	−1.052416817
			male	0.3688274694	−0.9472101651
		female	female	0.4298793764	−1.012281901
			male	0.4146573893	−0.904504941
		male	female	0.4058906612	−1.296010476
			male	0.4073835359	−1.184121967
Bangla BERT by BUET NLP	Without names	balanced	female	0.3485426613	−0.9649959402
			male	0.3621372929	−1.060035948
		female	female	0.3428604917	−0.5916077802
			male	0.3525201868	−0.6772515198
		male	female	0.3052915446	−0.6011139274
			male	0.3109551776	−0.649344567
	With names	balanced	female	0.34535423	−0.6683382053
			male	0.3610662164	−0.7401993831
		female	female	0.3432783339	−0.4071133434
			male	0.3606259946	−0.4953420636
		male	female	0.3094550239	−0.513300686
			male	0.3251797598	−0.5186699815

6. Conclusion

This work investigates gender–profession associations in Bengali BERT-family models using a controlled MLM probing framework and a systematically constructed evaluation corpus. Across the three evaluated models, we observe a general tendency toward stronger associations with male-denoting person words across profession categories, with some exceptions where male-, female-, and neutral professions show higher association with female terms. These patterns reflect the distributional behavior learned during pretraining rather than explicit or consistent rules.

We also examine the applicability of an English-originated bias measurement framework—based on prior-normalized log-probability and sigmoid scoring—in a Bengali setting. While the framework enables consistent comparison across controlled inputs, its interpretation requires caution due to linguistic and data-related differences.

This study has several limitations. The dataset is template-based and may not capture the full variability of natural Bengali usage. The selection of person words and profession categories is manually curated and not grounded in externally validated sociolinguistic statistics. In particular, due to the absence of a standardized or comprehensive source of workforce data for Bengali contexts, the categorization of professions into male-dominated, female-dominated, and balanced groups was determined by the authors based on general perception, and should therefore be interpreted as an analytical heuristic rather than a definitive classification. Additionally, the analysis is limited to sentence-level MLM predictions.

For future work, extending this framework to more diverse linguistic contexts and exploring bias mitigation strategies, such as data balancing or representation-level adjustments, would be valuable for improving fairness and robustness in Bengali language models.

References

[1] A. Bhattacharjee, T. Hasan, W. U. Ahmad, K. Samin, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding eval-

uation in Bangla," *arXiv preprint arXiv:2101.00204*, 2021. https://arxiv.org/abs/2101.00204.

[2] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. https://arxiv.org/abs/1810.04805.

[3] C. Basta, M. R. Costa-Jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.08783*, 2019. https://arxiv.org/abs/1904.08783.

[4] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.03310*, 2019. https://arxiv.org/abs/1904.03310.

[5] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020. https://arxiv.org/abs/2004.09456.

[6] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," *arXiv preprint arXiv:1903.10561*, 2019. https://arxiv.org/abs/1903.10561.

[7] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: Quantifying biases in clinical contextual word embeddings," in *Proc. ACM Conf. Health, Inference, and Learning*, pp. 110–120, 2020.

[8] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proc. ACM Collective Intelligence Conf.*, pp. 12–24, 2023.

[9] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," *arXiv preprint arXiv:1906.07337*, 2019. https://arxiv.org/abs/1906.07337.

[10] M. Bartl, M. Nissim, and A. Gatt, "Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias," *arXiv preprint arXiv:2010.14534*, 2020. https://arxiv.org/abs/2010.14534.

[11] J. Sadhu, A. Khan, A. Bhattacharjee, and R. Shahriyar, "An empirical study on the characteristics of bias upon context length variation for Bangla," in *Findings of the Assoc. Comput. Linguist.: ACL 2024*, Bangkok, Thailand, pp. 1501–1520, 2024.

[12] J. Sadhu, M. R. Saha, and R. Shahriyar, "Social bias in large language models for Bangla: An empirical study on gender and religious bias," in *Proc. First Workshop on Language Models for Low-Resource Languages*, Abu Dhabi, United Arab Emirates, pp. 204–218, 2025.

[13] N. M. K. Arnob, S. Mahmud, and A. T. Wasi, "Assessing gender bias of pretrained Bangla language models in STEM and SHAPE fields," in *Proc. 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Vienna, Austria, pp. 268–281, 2025.

[14] M. T. Hosain and M. K. Morol, "Intrinsic linguistic bias in formal vs. informal Bengali pragmatics with progressive context inflation," in *Proc. 14th Int. Joint Conf. Natural Language Processing and 4th Conf. Asia-Pacific Chapter of the Assoc. Comput. Linguist.*, Mumbai, India, pp. 384–396, 2025.

[15] M. A. H. Reaj, R. D. Gupta, J. S. Pritha, A. A. Noman, A. Ahmed, G. M. Mohiuddin, and T. H. Liew, "Cross-lingual probing and community-grounded analysis of gender bias in low-resource Bengali," in *Proc. 2025 4th Int. Conf. Smart Cities, Automation & Intelligent Computing Systems (ICONSONICS)*, pp. 37–44, Oct. 2025. https://doi.org/10.1109/ICONSONICS67145.2025.11309238.