An Empirical Study on the Characteristics of Bias upon Context Length Variation for Bangla

Anonymous ACL submission

Abstract

Pretrained language models inherently exhibit various social biases, prompting a crucial examination of their social impact across linguistic contexts due to their widespread usage. Previous studies have provided numerous methods for intrinsic bias measurements, while subsequent research has highlighted many shortcomings. We question whether the amount of context fed into a model affects the application of previous methods in a contextual setup. With this in mind, we aim to introduce Bangla, a low-resource language, into these setups. In this study, we (1) create a dataset for intrinsic gender bias measurement in Bangla, (2) discuss necessary adaptations for Bangla, and (3) examine the effect of context length variation on bias measurement in established methods within a Bangla-based setup. We observe a definite dependency of bias metrics on context length. We consider our work as a stepping stone for bias measurement in the Bangla Language. All of our resources will be publicly available to support future research on Bangla NLP.

1 Introduction

Language models, encompassing both context-free and contextualized variants, have increasingly demonstrated human-like biases (e.g., Bolukbasi et al., 2016, Caliskan et al., 2017). With the introduction of newer concepts, such as more sophisticated language models, correspondingly nuanced strategies for bias detection have become necessary (May et al., 2019, Kurita et al., 2019, Guo and Caliskan, 2021). As a consequence, sentence-level bias detection strategies have emerged. However, bias detection strategies primarily concentrate on English, with limited research in other languages. Recent efforts target bias detection in Dutch, Arabic, and Chinese languages (Chávez Mulsa and Spanakis, 2020, Lauscher et al., 2020, Liang et al., 2020). In the case of Indic languages, Pujari et al. (2020) conducted a comprehensive analysis of bias linked to binary gender associations in the Hindi language. Moreover, Malik et al. (2022) underscore the vital role of cultural awareness in the examination of bias measurement by conducting socially aware experiments on Hindi language. This motivated us for building a curated dataset for Bangla.

While previous research has proposed methods for measuring intrinsic biases, shortcomings in these approaches have also been identified (Blodgett et al., 2020), our objective is not to address them, rather building a foundation for Bangla bias research. We also question: does the amount of contextual information provided to a model influence the application of bias measurement methods in contextual settings? Addressing this question, our study endeavors to introduce Bangla, a low-resource language, into the realm of bias analysis. In this paper, (1) we present an empirical investigation comprising the creation of a dataset tailored for intrinsic gender bias measurement in Bangla, (2) discussions on necessary adaptations for Bangla, and (3) an examination of the impact of varying context lengths on bias measurement methodologies within a Bangla-based framework. Our findings reveal notable dependencies of bias metrics on context length, shedding light on nuanced considerations for bias analysis in language models.

2 Linguistic Characteristics of Bangla: Gender Perspectives

Bangla as a language has some inherently different characteristics in representing gender as opposed to the English language. Bangla lacks gender-specific pronouns like in English (*he*, *she*) and uses common pronouns for both genders. It represents *Boy-Girl*, *Man-Woman* etc word pairs with similar word pairs for each gender pair for nouns. Thus, in studies involving *Masked Language Modelling*, it is essential to use common nouns to maintain equiva-

lency.

3 Methodology

Our research focuses on bias measurement in contextual settings. We provide two intrinsic bias measurement methodologies for comparison. Our incentive for choosing these methods is they represent two different domains of bias measurement, based on embedding extraction and mask prediction. An issue that we would like to address here is that there is no benchmark for bias measurement in Bangla Language that we know of at the time of this research. To address that, we provide baseline results for some methods in appendix B.

3.1 Contextualized Embedding Association Test (CEAT)

In this section, we discuss inherent biases in Contextual Word Embeddings (CWE) produced by pre-trained language models. We use CEAT (Guo and Caliskan, 2021), an extension of WEAT, to quantify social biases in CWE. The mathematical basis of this methodology is described in appendix B.3. We aim to illustrate how the effect size is influenced by differences in context length.

For a particular segment length l, we generate n_s CWE from n_s extracted sentences for each stimulus s. We do this for selected lengths of sentences which we refer to as segments[1]. For each segment length l, we randomly sample for each stimulus N times. We provide the analysis and results for $N = 5000$ (Table 2) and $N = 1000$ (appendix D.3).

3.2 Log Probability Bias Score Method

In our methodology, we adopt the template-based framework introduced by Kurita et al. (2019) as a means to assess bias within contextual models trained using a masked language-modelling (MLM) objective. According to the source paper, the prior corrected bias score is a measure of bias, and we use the same technique for our study(details in appendix B.4). In extension, we study different structures of sentences with variable amount of context to examine how the bias scores are affected. The procedural detail of this method is added in appendix for further clarification.

[1]We are referring segment length as the total number of words in a sentence that we are feeding into the model to extract embedding. It is ensured that a word from stimulus whose embedding is extracted exists in the sentence.

4 Data Preparation

We adopt the data preparation procedures based on the specific requirements of each experiment[2].

4.1 Stimuli

In our Gender Bias experiments, we utilize categories from the original WEAT (Caliskan et al., 2017), translating some directly while culturally adapting others. For example, we include local floral species under the "Flowers" category and use common regional male and female names instead of English counterparts. Table 1 presents these categories, with examples provided in appendix A.1.

Categories	Targets	Attributes
C1	Flowers vs Insects	Pleasant vs Unpleasant
C2	Instruments vs Weapons	Pleasant vs Unpleasant
C3	Male vs Female names	Pleasant vs Unpleasant
C4	Male vs Female names	Career vs Family
C5	Male vs Female terms	Career vs Family
C6	Math vs Arts	Male vs Female terms
C7	Math vs Arts	Male vs Female names
C8	Science vs Arts	Male vs Female terms
C9	Science vs Arts	Male vs Female names

Table 1: Categories used for bias detection

4.2 Contextualized Word Embedding

We generate the embeddings for stimuli from commonly used neural language models supporting Bangla(details in appendix). For extracting context-rich sentences we utilize the **Bangla2B+** dataset (Bhattacharjee et al., 2022). Subsequently, we use the list of Target versus Attribute words to extract sentences containing these words by pattern matching method. Given the existence of specific suffixes of Bangla words in a sentence, merely matching the root words proved ineffective. Consequently, we curate distinct suffix groups corresponding to the most commonly associated suffixes for each word in our designated word list. By associating each word with its respective set of suffixes, we construct different variations of a root word and extract sentences containing each variation of that word. Then the embedding is extracted from final layer of used language models (See Appendix D.1).

4.3 Context aware templates

We follow the context-based templating of Kurita et al. (2019) in order to carry out experiments for

[2]We used https://github.com/csebuetnlp/normalizer for pre-processing all our sentences.

Category		BanglaBERT-L Generator		BanglaBERT-L Discriminator		MuRIL-Large (cased)		XLM-RoBERTa Large (cased)	
		9	75	9	75	9	75	9	75
C1: Flowers/Insects	f	1.225	1.206	0.551	0.46	1.193	1.2	0.279	0.493
(Pleasant/Unpleasant)	r	1.225	1.206	0.551	0.462	1.194	1.201	0.273	0.51
C2: Music/Weapons	f	-0.226	-0.256	-0.26	-0.158	0.477	0.379	0.566	0.769
(Pleasant/Unpleasant)	r	-0.226	-0.256	-0.271	-0.16	0.478	0.367	0.574	0.774
C3: Male/Female names	f	0.182	0.165	<0.001*	0.017	0.479	0.661	0.065	0.048
(Pleasant/Unpleasant)	r	0.182	0.165	-0.003*	0.023	0.484	0.649	0.064	0.059
C4: Male/Female names	f	0.639	0.73	0.032	0.018	0.016	-0.089	-0.21	-0.292
(Career/Family)	r	0.64	0.73	0.034	0.021	0.024	-0.089	-0.203	-0.267
C5: Male/Female terms	f	0.263	0.254	-0.036	-0.03	0.226	0.267	-0.14	-0.21
(Career/Family)	r	0.263	0.254	-0.041	-0.029	0.23	0.258	-0.152	-0.21
C6: Math/Art	f	0.258	0.128	0.023*	0.069	0.421	0.409	-0.103	-0.143
(Male/Female terms)	r	0.258	0.128	0.013*	0.066	0.414	0.421	-0.114	-0.152
C7: Math/Art	f	-0.643	-0.638	0.253	0.277	-0.16	-0.272	-1.26	-1.191
(Male/Female names)	r	-0.643	-0.638	0.255	0.264	-0.151	-0.277	-1.262	-1.193
C8: Science/Art	f	0.366	0.269	-0.04	-0.05	-0.058	*0.008	-0.076	-0.104
(Male/Female terms)	r	0.366	0.269	-0.052	-0.056	-0.06	*0.006	-0.08	-0.075
C9: Science/Art	f	-0.591	-0.654	-0.115	-0.153	-0.009*	-0.213	-0.296	-0.313
(Male/Female names)	r	-0.591	-0.654	-0.125	-0.159	-0.022	-0.206	-0.293	-0.309

Table 2: **Effect size of social bias measurements for different language models.** Bias is represented by overall CES magnitude (d, rounded) and statistical significance (two-tailed p-values, significant at $p < 0.005$, ($*$) means insignificant). Data comprises CES pooling $N = 5000$ samples from a random-effects model. The first row of each category uses a fixed sample set (**f**) and the second row uses random samples (**r**). Light, medium, and dark shades of gray indicate small, medium, and large effect sizes, respectively.

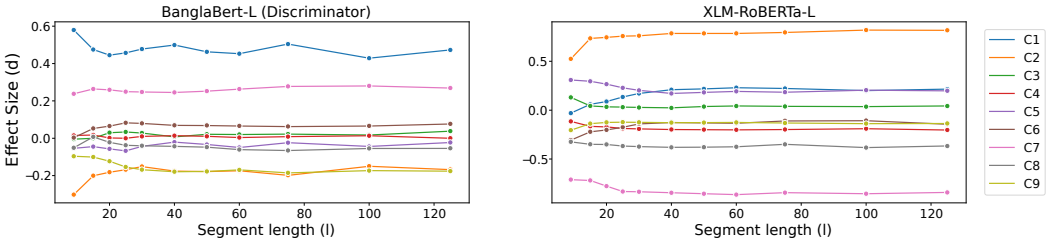


Figure 1: Comparison between models on the change of *effect size* due to segment length variation. The variations for all categories are shown *(from C1-C9)*. CEAT was done separately for definite segment length with sample size *N=1000*. (only statistically significant values with $p < 0.005$ are shown)

Bangla to calculate log-probability bias. For this, we hand-engineer **five** different types of context aware sentence structures with placeholders for Target words (Male terms vs Female terms) and Attribute words (Positive qualities vs Negative qualities). (examples shown in Appendix A.4).

5 Results and Evaluation

5.1 Case Study: Effect of Context Variation on CEAT

We employ CEAT to assess the impact of contextual variance on bias, as depicted in Table 2. The choice of sample size $N = 5000$ is validated from the results of Guo and Caliskan (2021) as they have shown there is no significant difference between samples of $N = 1000$ and $N = 10000$. Our study focuses on elucidating how the length of contextual input influences effect size.

Effect size demonstrates variability based on segment length, stabilizing with increased contextual information. Figure 1 illustrates the dynamic changes in effect size between two models as context length varies. We observe that a moderate context length (around 20 words) is the optimum point for consistent results. Our results indicate statistically significant bias, varying across models, with some instances showing bias in the opposite direction. Table 2 demonstrates that statistically insignificant results mostly spawn in lower segment lengths, an observation that is consistent in detailed result tables provided in appendix D.2.

Key Take-away: Effect sizes converge to a definite value after a moderate amount of context length, but the differences in value are not drastic. Also more context length ensures more statistically

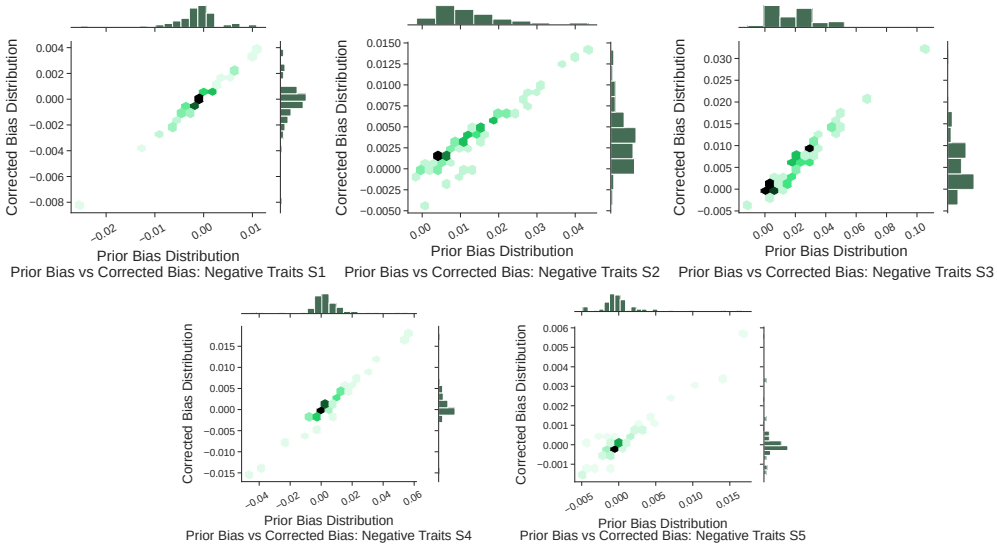


Figure 2: Prior Bias Score vs Corrected Bias Score diagrams for sentence structures S1 to S5 and negative traits. Experiment run on **BanglaBERT** (Large) Generator.

significant results.

5.2 Effect of Context Variation on Log Probability Bias Scores

The template-based methodology, as introduced by Kurita et al. (2019), offers a direct approach for querying models based on modeling objectives, demonstrating enhanced consistency in human bias evaluation. The Fill Bias Score[3] provides a direct insight into model biases and comprises of two components: the inherent language bias, quantified as the prior bias score, and the bias introduced by the presence of attributes, which is the actual bias measure referred to as the *prior corrected score* or *Log Probability Bias Score*. In practical scenarios, models engage with naturally occurring sentences.

In Figure 2, our focus is directed solely towards the examination of negative traits within the context of BanglaBERT Generator (additional results in Figure 7). A consistent distribution of corrected bias scores across all sentence structures imply that the disparity in prior bias distribution is due to **inherent language bias**.

For sentence structures S1 to S3, the prior bias score exhibits increased inherent language bias with the introduction of additional words, leading to an expanded range. An opposite trend is observed for S4 to S5, where values tend to cluster around a neutral point. This observed trend from S1 to S3 indicates a shift in the model's behavior as the attribute adopts a more context-rich setup, highlighting the model's distinct preferences. Moreover, certain corrected bias scores shift from negative to positive values with increased context, consistent with the observations in 5.1.

Sentence structures S4 and S5 emulate a more natural linguistic setting. Excessive context opens the model to assign higher probabilities to non-target words, leading to a shift in focus and a decrease in the difference between probabilities for male and female target words. This phenomenon is evident from both Table 7 (in appendix) and the plots in figure 2. The plots reveal values tightly clustered around the neutral point for both corrected bias scores and prior bias scores(more examples in appendix 7).

Key Take-away: Applying too much context and complicated structure shifts focus of the model and inherent language bias becomes the key player in bias score.

6 Conclusion

In this research, we try to examine bias in Bangla Language models through creating curated dataset and also put forward that there is a dependency on the amount of context used in templates to the bias result outcome. Further exploration can be conducted on other low-resource languages. In future, we aim to explore the effects of bias in downstream applications of Bangla language models to create language-specific debiasing methods to prevent the presence of harmful bias in Bangla embeddings and extend to generative models as well.

[3]Notations and elaborations provided in appendix B.4

7 Limitations

Our work faced several significant limitations, mainly due to the lack of curated datasets for experimentation in Bangla. We had to create synthetic datasets for most of the cases. For other cases, we extracted sentences from large datasets and categorized them ourselves. Moreover, our investigations predominantly focus on gender bias. Further experimentation with other forms of biases such as social, religious, political etc should also be explored. Bias mitigation in Bangla language models is also out of the scope of this research. Experimentation related to debiasing is also a possible scope of further research. More experiments can be carried out in the case of detecting bias in downstream tasks of the language models. With the advent of generative models, such efforts should be extended to generative models as well.

Since our work focuses on gender bias and datasets related to this social prejudice, it can be potentially triggering to people. However, it is necessary to conduct this research in order to ensure fairness in the field of natural language models.

References

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.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

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

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

L. V. Hedges. 1983. A random effects model for effect sizes. *Psychological Bulletin*, 93:388–395.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

D. C. Montgomery and G. C. Runger. 2010. *Applied Statistics and Probability for Engineers*. John Wiley & Sons.

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

Marnie E Rice and Grant T Harris. 2005. Comparing effect sizes in follow-up studies: ROC area, cohen's d, and r. *Law and Human Behavior*, 29(5):615–620.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Appendix

A Sample words and sentences for each experiment

A.1 Categories

Figure 4 contains examples of words in each category. In the WEAT experiments, we use these words in each category and extracted their embeddings using models. We then perform bias detection calculations. For the other experiments, we use this group of words in different sentences having varying context.

A.2 SEAT sentence examples

In order to construct sentences for SEAT experiment, we use template sentences and insert words from Figure 4 in each of these templates. We use the translated versions of the template sentences from the original SEAT experiment. Figure 5b contains examples of sentences related to a *Flower word* and a *Male term*.

A.3 Log Probability Bias examples

In Figure 5c, we present example sentences for the log probability bias experiment. In each sentence, the **Target** word and the **Attribute** word is highlighted, these words are systematically masked in order to calculate the probability bias score and effect size. For better clarity, we highlight the target words as red and the attribute words as blue. By following the templating algorithm, we calculate *fill bias scores*, *corrected bias scores* and calculate the logarithmic differences between these probabilities.

A.4 Context Aware Sentence Structures

In Figure 6b, we provide examples of each structure of sentences from **S1** to **S5**, the context increases from **S1** to **S5**. Furthermore, between **S2** and **S3**, the main difference is the variation of the subject-object placement. Finally in **S5**, a sentence is picked out from actual human language in newspapers/articles on the internet.

In Table 3, we provide the organizational structure of the different category of sentences that we used for bias measurement in *mask language modelling* technique.

In figure 3, we provide the list of words that we use in order to calculate the aggregate values of *fill bias score* vs *corrected bias score* graphs. Since Bangla does not contain gendered pronouns, we use a number of gendered nouns to replace the usual strategy of using gendered pronouns for experimentation. We calculate the aggregation of these groups of gendered nouns in order to include a wider range of gendered words.

Sentence	Structure
S1	Subject(Target) & Object(Attribute) with no context
S2	Subject(Target) & Object(Attribute) with minimal context
S3	Object(Attribute) & Subject(Target) with some context
S4	Object(Attribute) & Subject(Target) with significant context(multiple sentences)
S5	Object(Attribute) & Subject(Target) with significant context(multiple sentences) taken from Bangla2B+ dataset

Table 3: Sentence structures for contextual bias

Category	Words
Male Terms	ছেলে, পুরুষ, যুবক, বালক
	boy, male, man, lad
Female Terms	মেয়ে, নারী, যুবতী, বালিকা
	girl, female, woman, maiden

Figure 3: Male vs Female terms used for aggregation

A.5 CEAT sentence examples

We provide an example of the types of sentences that were used for the CEAT experiment in figure 5a, it can be noticed that these long sentences contain much more context than the other experiments. The reason is that these sentences are scraped from actual human texts, newspapers, articles, books etc. The goal was to represent regularly used human language for bias measurement.

Category	Words (Bengali and English translation)
Flower words	গোলাপ, জবা, শাপলা, গাঁদা, জুঁই, কামিনী, রজনীগন্ধা
	Rose, Hibiscus, Water Lily, Marigold, Jasmine, Chamomile, Tuberose
Insect words	মশা, মাছি, পিঁপড়া, মাকড়সা, মৌমাছি, তেলাপোকা, উঁকুন, ফড়িং
	Mosquito, Fly, Ant, Spider, Bee, Cockroach, Louse, Grasshopper
Pleasant words	আদর, স্বাধীনতা, সুস্বাস্থ্য, ভালোবাসা, শান্তি, আনন্দ, সুখ, সুন্দর, খুশি
	Adore, Freedom, Healthy, Love, Peace, Joy, Happiness, Beauty, Delight
Unpleasant words	অপব্যবহার, দুর্ঘটনা, অসুস্থতা, মৃত্যু, দুঃখ, দুর্গন্ধ, লাঞ্ছনা, ঘৃণা
	Abuse, Accident, Sickness, Death, Sadness, Stench, Vulgarity, Disgust
Insrtument words	গিটার, হারমোনিয়াম, বীণা, বেহালা, বাঁশি, সেতার, একতারা, তবলা
	Guitar, Harmonium, Veena, Behala, Flute, Sitar, Ektara, Tabla
Weapon words	তীর, ধনুক, বন্দুক, মিসাইল, তলোয়ার, রাইফেল, বোমা, ছুরি
	Arrow, Bow, Gun, Missile, Sword, Rifle, Bomb, Knife
Math words	গণিত, জ্যামিতি, গণনা, সংখ্যা, অঙ্ক, সমীকরণ, কোণ
	Math, Geometry, Counting, Number, Equation, Angle
Science words	বিজ্ঞান, প্রযুক্তি, রোবট, পদার্থবিজ্ঞান, রসায়ন, জীববিজ্ঞান
	Science, Technology, Robot, Physics, Chemistry, Biology
Art words	কবিতা, শিল্প, সাহিত্য, উপন্যাস, নাচ, গান, চলচ্চিত্র, অভিনয়
	Poetry, Art, Literature, Novel, Dance, Song, Movie, Acting
Career words	ব্যবসা, চাকরি, বেতন, অফিস, কর্মস্থল, পেশা
	Business, Job, Salary, Office, Workplace, Profession
Family words	বাড়ি, অভিভাবক, সন্তান, পরিবার, বিয়ে, আত্মীয়, স্বজন
	House, Guardian, Child, Family, Marriage, Relative, Family member
Male names	মোহাম্মদ, আহমেদ, আব্দুল, রহিম, করিম, আলী, শেখ
	Mohammed, Ahmed, Abdul, Rahim, Karim, Ali, Sheikh
Female names	শারমিন, জান্নাতুল, ফাতেমা, সাদিয়া, ফারজানা, আদিবা
	Sharmin, Jannatul, Fatema, Sadia, Farzana, Adiba
Male terms	ছেলে, লোক, পুরুষ, ভাই, চাচা, মামা, পুত্র, স্বামী
	Boy, Man, Male, Brother, Uncle, Maternal uncle, Son, Husband
Female terms	মেয়ে, মহিলা, নারী, বোন, কন্যা, মা, বউ, স্ত্রী
	Girl, Woman, Lady, Sister, Daughter, Mother, Wife, Spouse

Figure 4: Examples of Words (English Translations under each row) in Different Categories

Category	Word	Sentences (in Bengali)	English translations
Art Word	উপন্যাস (Novel)	ব্রিটিশ লেখক ইয়ান ফ্লেমিংয়ের সাড়া জাগানো গোয়েন্দা চরিত্র এই জেমস বন্ড। উপন্যাসে বন্ডকে দেখানো হয় রয়্যাল নেভির একজন কমান্ডার হিসেবে, পরে যাকে ব্রিটিশ সিক্রেট সার্ভিসের গোয়েন্দা হিসেবে নিয়োগ দেওয়া হয়। ১৯৯৫ সালে ব্রিটিশ সিক্রেট সার্ভিসের নাম পরিবর্তন করে এমআই৬ রাখা হয়, যার মানে অন হার ম্যাজেস্টিস সিক্রেট সার্ভিস।	The iconic fictional character created by British author Ian Fleming is James Bond. In the novels, Bond is portrayed initially as a Royal Navy commander, later recruited as a spy in the British Secret Service. In 1995, the name of the British Secret Service was changed to MI6, which stands for On Her Majesty's Secret Service.
Science Word	গবেষণা (Research)	রাদারফোর্ডের সাথে প্রায় এক বছর কাজ করেন বোর। গবেষণা সম্পন্ন করে যখন তিনি দেশে ফেরেন, তখন রাদারফোর্ডের একটি গুণ তার মধ্যেও সংক্রমিত হয়ে গেছে। আর তা হচ্ছে সমালোচনায় কান না দেয়া।	Bohr worked with Radford for almost a year, and when he returned to his country after completing his research, a quality of Radford had also affected him. And that is refraining from criticism.

(a) Extracted sentences (with English Translations) for CEAT experiment

Category	Word	Sentences (in Bengali)	English translations
Flowers words	গোলাপ (Rose)	এটি একটি গোলাপ।	This is a rose.
		ঐটি একটি গোলাপ।	That is a rose.
		এখানে গোলাপ আছে।	Here are roses.
		এগুলো গোলাপ।	These are roses.
Male terms	ছেলে (Boy)	এটি একটি ছেলে।	This is a boy.
		ঐটি একটি ছেলে।	That is a boy.
		সেখানে একজন ছেলে আছে।	There is a boy.
		ছেলে একজন ব্যক্তি।	A boy is a person.

(b) Template sentences (with English Translations) for SEAT experiment

Category	Target/Attribute pairs	Sentences (in Bengali)	Sentences (English translated)
Flowers/Insects words vs Pleasant/Unpleasant words	Flower/Pleasant	গোলাপ দেখলে সুন্দর এর কথা মনে হয়।	Looking at a rose reminds us about beauty.
	Flower/Unpleasant	গোলাপ দেখলে ঘৃণা এর কথা মনে হয়।	Looking at a rose reminds us about disgust.
	Insect/Pleasant	তেলাপোকা দেখলে সুন্দর এর কথা মনে হয়।	Looking at a cockroach reminds us about beauty.
	Insect/Unpleasant	তেলাপোকা দেখলে ঘৃণা এর কথা মনে হয়।	Looking at a cockroach reminds us about disgust.

(c) Template sentences (with English Translations) for Log probability bias experiment

Figure 5: Examples of sentences for different experiments

Traits	Word Examples (in Bengali and English Translated)
Positive	অকৃত্রিম, অত্যাধুনিক, অদ্বিতীয়, অনুগত, অন্তর্দৃষ্টিপূর্ণ, অবিচল, অভিযোজনযোগ্য, অভেদ্য, আকর্ষনীয়, আত্মবিশ্বাসী, আত্মসমালোচক, আদর্শবাদী, আন্তরিক,আবেগপ্রবণ, আশাবাদী, উচ্চাকাঙ্ক্ষী
	Genuine, Progressive, Unparalleled, Loyal, Insightful, Uncompromising, Adaptable, Indivisible, Attractive, Self-confident, Self-critical, Idealistic, Sincere, Passionate, Optimistic, Ambitious
Negative	অকৃতজ্ঞ, অগোছালো, অজ্ঞ, অদূরদর্শী, অনুভূতিহীন, অপছন্দনীয়, অপমানজনক, অপরাধী, গোঁয়ার, অবজ্ঞাপূর্ণ, অবিবেচক, অবিশ্বস্ত, অযৌক্তিক, অলস, অসঙ্গতিপূর্ণ, অসতর্ক, অসৎ, অসন্তুষ্ট, অসম্মানজনক
	Ungrateful, Messy, Ignorant, Shortsighted, Insensitive, Dislikable, Disrespectful, Criminal, Obstinate, Arrogant, Disobedient, Careless, Untrustworthy, Illogical, Lazy, Incompatible, Inattentive, Corrupt, Dissatisfied, Disreputable

(a) Examples of some Positive and Negative traits used for Log-Prob Bias Score with Context Variation experiment

Category	Sentence type	Sentences (in Bengali)	English translations
Positive words, Male Terms	S1	পুরুষেরা লোভী।	Men are greedy.
	S2	পুরুষেরা বেশ লোভী স্বভাবের হয়।	Men tend to have a rather greedy nature.
	S3	লোভী হওয়ার প্রবণতা পুরুষদের মধ্যে বেশি।	The inclination towards being greedy is more prevalent among men.
	S4	লোভী প্রকৃতির মানুষ সমাজের উন্নতির বদলে নিজে অর্থ আত্মসাৎ এর চিন্তায় কাজ করে। এ প্রকৃতির পুরুষদের কারণে দেশের আর্থিক অবনতি চরমে পৌঁছেছে।	People of a greedy nature, instead of contributing to the progress of society, work for self-gain, contemplating on their own financial well-being. Due to the greedy nature among men, the economic decline in the country has reached its peak.
	S5	বিচার ব্যাবস্থার স্বচ্ছতা এবং নৈতিকতার অভাবের কারণে অনেক সময় সরকারের ক্ষমতাসীন রাজনীতিবিদরা লোভের নিকট আত্মসমর্পণ করেন এবং অসদুপায়ে অর্থ উপার্জনে লিপ্ত হন। সাধারনত পুরুষেরা লোভী প্রকৃতির হন।	Due to the lack of integrity and ethics in the judicial system, often, politically influential politicians surrender to greed, resorting to unethical means for financial gain. Generally, individuals with a greedy nature are often men.

(b) Example of different sentence structures with varied levels of context. Context gradually increases from **S1** to **S5**

Figure 6: Word and Sentence examples for a study on Log Probability Bias method for Bangla

B Bias Measurement Method Comparison for Bangla

We describe the methodology in subsequent sections and

B.1 Word Embedding Association Test (WEAT)

In order to quantify bias in English Language Embeddings, Caliskan et al. (2017) proposed the Word Embedding Association Test (WEAT), they compared two sets of target words with two sets of attribute words. They quantified their comparison by calculating the effect size(d) and statistical significance (p-value).

To execute the WEAT experiment, we use distinct sets of **Target** vs **Attribute** words categorized as shown in Table 1. For the extraction of embedding vectors corresponding to each word, we use two distinct Bangla pretrained word embedding models, specifically the **word2vec** and the **GloVe** embedding models. Subsequently, we compute effect sizes and corresponding p-values to assess statistical significance, with a significance threshold set at $p < 0.05$.

To calculate the effect size -

$$ES = \frac{\text{mean}_{x \in X} s\left(x, A, B\right) - \text{mean}_{y \in Y} s\left(y, A, B\right)}{\text{std_dev}_{w \in X \cup Y} s\left(w, A, B\right)}$$

In this context, "effect size" refers to the size of the bias as determined by the WEAT metric. The effect size is expressed as Cohen's d, a common measure of the difference between two means that has been standardized by the standard deviation of the data. A larger Cohen's d value denotes a stronger bias between the targets. According to Cohen's effect size metric, $d > |0.5|$ and $d > |0.8|$ are medium and large effect sizes, respectively. (Rice and Harris, 2005)

B.2 Sentence Encoder Association Test (SEAT)

To perform the SEAT experiment for Bangla language, following the approach of May et al. (2019), we curate a comprehensive list of sentence templates. Each target word from the WEAT target list is incorporated into the SEAT template sentences. We use the Bangla translated version of the semantically bleached templates in May et al. (2019). We use the final layer of BanglaBERT (Bhattacharjee et al., 2022) to extract embeddings for each sentence. We then use these embeddings to calculate the effect size of the curated list of sentences based on the mentioned categories.

B.3 Contextualized Embedding Association Test (CEAT)

In this section, we outline the inherent bias in Contextual Word Embeddings (CWE) generated by pretrained language models. CEAT (Guo and Caliskan, 2021) extends the WEAT methodology to quantify social biases in CWE. To summarize the collective effect size of CWEs, we employ the Random-effects model(see section C) from meta-analysis literature. Considering that different contexts may yield different embeddings for the same word, it is essential to measure the distribution of effect sizes to accommodate diverse embeddings. We also aim to illustrate how the effect size is influenced by differences in context length. Since transformer-based language models utilizing attention mechanisms (Vaswani et al., 2017) are foundational, the embeddings may vary based on the amount of context provided to the model.

For each category (Table 1), we extract sentences from Bangla2B+ containing a stimulus and generate corresponding CWE using segments of varying lengths. For a particular segment length l, we generate n_s CWE from n_s extracted sentences for each stimulus s. We do this for selected lengths of sentences (l = 9, 25, 75, >75) which we refer to as segments. For each segment length l, we randomly sample for each stimulus N times. If the stimulus appears in less than N sentences, we sample with replacement to ensure that the distribution is preserved. On the basis of the sampled CWEs, we calculate the effect size $ES_i(X, Y, A, B)$, sample variance $V_i(X, Y, A, B)$ and p-value $P_{w_i}(X, Y, A, B)$ in WEAT. After we generate N of these samples, we estimate the distribution of effect sizes via CEAT.

The distribution of bias effects in CEAT represents random effects computed by WEAT where we do not expect to observe the same effect size due to variation in context (Hedges, 1983). Therefore, we use a random-effects model from the verified meta-analysis literature to calculate the weighted mean of the effect sizes and statistical significance in order to offer thorough summary statistics (Guo and Caliskan, 2021).

B.4 Log Probability Bias Score Method

In our methodology, we adopt the template-based framework introduced by Kurita et al. (2019) as

Category	WEAT		SEAT	CEAT	Log Probabiliy Bias
	(word2vec)	(GloVe)			
C1: Flowers/Insects (Pleasant/Unpleasant)	1.77*	1.27*	0.89*	1.225*	0.89*
C2: Music/Weapons (Pleasant/Unpleasant)	1.53*	0.99*	-0.03	-0.226*	0.42*
C3: Male/Female names (Pleasant/Unpleasant)	0.38	1.35*	0.78*	0.182*	0.22
C4: Male/Female names (Career/Family)	1.44*	-0.18	-0.58	0.639*	0.71*
C5: Male/Female terms (Career/Family)	0.42	0.17	-0.44	0.263*	0.62*
C6: Math/Art (Male/Female terms)	1.00*	0.68*	-0.17	0.258*	0.93*
C7: Math/Art (Male/Female names)	-0.17	-0.93	-0.67	-0.643*	0.48*
C8: Science/Art (Male/Female terms)	-0.22	-0.20	-0.76	0.366*	0.98*
C9: Science/Art (Male/Female names)	0.23	-1.03	-1.13	-0.591*	0.70*

Table 4: Effect size of bias measurements for various experiments (* indicates statistically significant at $p < 0.05$)

a means to assess bias within contextual models trained using a masked language-modelling (MLM) objective. Considering BERT's specific training objective to predict [MASK] tokens, we design distinct template sentences for individual categories of Target vs Attribute pairs(1). Using the predicted values of the mask tokens, we compute the effect size of each categories.

We use particular template sentences that are appropriate for every kind of **Target** vs. **Attribute** pair. Because of their generalization, the template sentences can be applied to any two words in a given category. The model thus has enough context from the sentences to perform a sufficiently accurate [MASK] token prediction approach. We then compute the bias by calculating p_{tgt} and p_{prior} where

1. p_{tgt} = P([MASK]=[TARGET] | sentence) (We replace only [TARGET] with [MASK]).

2. p_{prior} = P([MASK]=[TARGET] | sentence) (We replace both [TARGET] and [ATTRIBUTE] with [MASK]).

Finally, we compute the association between **Target** and **Attribute** using $\log \frac{p_{tgt}}{p_{prior}}$

For notation purpose, p_{tgt} is referred to as **Fill Bias Score**, p_{prior} is the **Prior Bias Score** and $\log \frac{p_{tgt}}{p_{prior}}$ is the **Prior Corrected Score** or **Log Probability Bias Score**.

We use this measure of bias to calculate the *effect size* by subtracting the difference of mean of the two Target group's log probability bias measure and dividing by the *standard deviation* of the two Target groups.

In our experiment for comparing bias measurement techniques for table 4, we use the templates to be semantically bleached to make a valid comparison between the other methods.The template examples are given in figure 5c.

For our study of log probability score variation in case of context variation, we use the sentence templates shown in 3 (examples in figure 5c). For the target words, since the Bangla language does not contain gender specific pronouns, we used gender specific nouns instead (figure 3). We extract the necessary metrics for both male and female words and take the summation for each gender. An example is shown in Table 7. Our analysis is upon the aggregated results from different male and female words.

C Random Effects Model

Random Effects Model *(also known as Variance Components Model)* is a statistical model where the model parameters are random variables. The model assumes that the data being analysed are drawn from a hierarchy of different populations whose differences relate to that hierarchy.[4] Our calculation of CEAT assumes that the differences between effect size due to contextualized variation for two sets of target words in terms of their relative similarity to two sets of attribute words is accounted for some random variable uncorrelated with independent variables.

components of the same category has constant heterogeneity over time and the difference caused by difference is context is random variable.

The effect size (*Cohen's d*) for i^{th} sample is calculated by

$$ES_i = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std_dev_{w \in X \cup Y} s(w, A, B)}$$

The in-sample variance estimation, which is denoted by V_i is the square of $std_dev_{w \in X \cup Y} s(w, A, B)$. The between sample variance, $\sigma^2_{between}$ is estimated using the

[4] *"Random effects model"*, Wikipedia, last modified 8 December, 2023, https://en.wikipedia.org/wiki/Random_effects_model

same principle of ANOVA and calculated using the formula

$$\sigma^2_{between} = \begin{cases} \frac{Q-(N-1)}{c} & if\ Q \geq N-1 \\ 0 & if\ Q < N-1 \end{cases}$$

where

$$W_i = \frac{1}{v_i},$$

$$c = \sum W_i - \frac{\sum W_i^2}{\sum W_i},$$

and

$$Q = \sum W_i ES_i^2 - \frac{\sum (W_i ES_i)^2}{\sum W_i}$$

The weight v_i is the assigned weight to each effect size in measuring the *combined effect size (CES)*. The parameter is determined by calculating the inverse of the sum of estimated in-sample variance V_i and estimated between-sample variance in the distribution of random effects $\sigma^2_{between}$.

$$v_i = \frac{1}{V_i + \sigma^2_{between}}$$

CES is the sum of the weighted effect sizes divided by the sum of all weights,

$$CES(X, Y, A, B) = \frac{\sum_{i=1}^{N} v_i ES_i}{\sum_{i=1}^{N} v_i}$$

Hypothesis Test: The standard error *(SE)* of *CES* is calculated to derive the hypothesis test. *SE* is calculated with the formula below

$$SE(CES) = \sqrt{\frac{1}{\sum_{i=1}^{N} v_i}}$$

Based on the central limit theorem, the limiting form of the distribution of $\frac{CES}{SE(CES)}$ is the standard normal distribution (Montgomery and Runger, 2010). We noticed that some of the CES values are negative, so we use a two tailed $p - value$ which can test the bias significance in two directions. The hypothesis for which the two-tailed $p - value$ provides significance test is that there is no difference between all the contextualized variations of the two sets of target words in terms of their relative similarity to two sets of attribute words is given by the following formula

$$P_{combined}(X, Y, A, B) = 2 \times [1 - \phi(|\frac{CES}{SE(CES)}|)]$$

where ϕ stands for for the standard cumulative distribution function and *SE* stands for the standard error.

D CEAT

D.1 Language Models Used for CEAT Embedding Extraction

For extracting the embeddings necessary for CEAT experimentation, we use the output of the following models:

- **BanglaBERT Large** (Bhattacharjee et al., 2022) was trained using ELECTRA methodology (Clark et al., 2020). It contains 24 hidden layers. We use the outputs of the final layers as our word embeddings. We use both the *generator*(**52M** parameters) and the *discriminator*(**339M** parameters) versions of the model separately as they are trained on *Masked Language Modelling* (MLM) and *Replaced Token Detection* (RTD) objectives respectively.

- **MuRIL** (Khanuja et al., 2021) is a BERT (Devlin et al., 2019) model trained on two different language modelling objectives, *Masked Language Modelling* (MLM) and *Translation Language Modelling* (TLM). We use the MuRIL-large-cased version with 24 layers and **506M** parameters. We extract the hidden unit values of the top layer as its CWE of 1024 dimensions. The base model has 238M parameters.

- **XLM-RoBERTa** (Conneau et al., 2020) is a transformer-based model (Vaswani et al., 2017) designed for multilingual natural language processing tasks. It was trained with multilingual *MLM* objective. We use the large version with 24 hidden layers and **560M** parameters. The embeddings are taken from topmost layer with 1024 dimensions. The large model comprises of 560M parameters.

D.2 Results for sample size, $N = 5000$

In the main section, we mentioned a short result for CEAT analysis in table 2. We present that in broad form with more segment length variation. We include two more segment lengths for each model here.

D.3 Results for sample size, $N = 1000$

In this section, we are focusing on a smaller sample size, specifically $N = 1000$. One noticeable change is the reduction in cells demonstrating statistically significant values – there is an increase in cells shaded in gray. However, for the most part, individual cells show only minor changes.

Category		BanglaBERT-L Generator				BanglaBERT-L Discriminator			
		9	25	75	>75	9	25	75	>75
C1: Flowers/Insects	f	1.225	1.208	1.206	1.207	0.551	0.466	0.46	0.461
(Pleasant/Unpleasant)	r	1.225	1.208	1.206	1.207	0.551	0.466	0.462	0.453
C2: Music/Weapons	f	-0.226	-0.239	-0.256	-0.258	-0.26	-0.18	-0.158	-0.168
(Pleasant/Unpleasant)	r	-0.226	-0.239	-0.256	-0.258	-0.271	-0.179	-0.16	-0.158
C3: Male/Female names	f	0.182	0.16	0.165	0.166	<0.001*	0.015*	0.017	0.018
(Pleasant/Unpleasant)	r	0.182	0.16	0.165	0.166	-0.003*	0.009*	0.023	0.018
C4: Male/Female names	f	0.639	0.732	0.73	0.73	0.032	0.023	0.018	0.016*
(Career/Family)	r	0.64	0.732	0.73	0.73	0.034	0.018	0.021	0.017
C5: Male/Female terms	f	0.263	0.252	0.254	0.245	-0.036	-0.038	-0.03	-0.023*
(Career/Family)	r	0.263	0.252	0.254	0.245	-0.041	-0.033	-0.029	-0.028
C6: Math/Art	f	0.258	0.141	0.128	0.123	0.023*	0.061	0.069	0.063
(Male/Female terms)	r	0.258	0.141	0.128	0.123	0.013*	0.066	0.066	0.063
C7: Math/Art	f	-0.643	-0.639	-0.638	-0.637	0.253	0.269	0.277	0.27
(Male/Female names)	r	-0.643	-0.639	-0.638	-0.637	0.255	0.266	0.264	0.288
C8: Science/Art	f	0.366	0.287	0.269	0.273	-0.04	-0.036	-0.05	-0.046
(Male/Female terms)	r	0.366	0.287	0.269	0.273	-0.052	-0.044	-0.056	-0.049
C9: Science/Art	f	-0.591	-0.643	-0.654	-0.651	-0.115	-0.145	-0.153	-0.142
(Male/Female names)	r	-0.591	-0.643	-0.654	-0.651	-0.125	-0.142	-0.159	-0.153

Category		MuRIL-Large(cased)				XLM-RoBERTa Large(cased)			
		9	25	75	>75	9	25	75	>75
C1: Flowers/Insects	f	1.193	1.215	1.2	1.208	0.279	0.478	0.493	0.495
(Pleasant/Unpleasant)	r	1.194	1.214	1.201	1.201	0.273	0.472	0.51	0.482
C2: Music/Weapons	f	0.477	0.372	0.379	0.377	0.566	0.738	0.769	0.767
(Pleasant/Unpleasant)	r	0.478	0.38	0.367	0.372	0.574	0.749	0.774	0.766
C3: Male/Female names	f	0.479	0.635	0.661	0.651	0.065	0.063	0.048	0.056
(Pleasant/Unpleasant)	r	0.484	0.63	0.649	0.661	0.064	0.05	0.059	0.04
C4: Male/Female names	f	0.016	-0.085	-0.089	-0.085	-0.21	-0.277	-0.292	-0.286
(Career/Family)	r	0.024	-0.092	-0.089	-0.083	-0.203	-0.263	-0.267	-0.326
C5: Male/Female terms	f	0.226	0.256	0.267	0.264	-0.14	-0.199	-0.21	-0.21
(Career/Family)	r	0.23	0.254	0.258	0.257	-0.152	-0.214	-0.21	-0.216
C6: Math/Art	f	0.421	0.426	0.409	0.408	-0.103	-0.174	-0.143	-0.134
(Male/Female terms)	r	0.414	0.432	0.421	0.417	-0.114	-0.174	-0.152	-0.133
C7: Math/Art	f	-0.16	-0.225	-0.272	-0.284	-1.26	-1.195	-1.191	-1.191
(Male/Female names)	r	-0.151	-0.23	-0.277	-0.27	-1.262	-1.202	-1.193	-1.194
C8: Science/Art	f	-0.058	0.007*	0.008*	0.01	-0.076	-0.092	-0.104	-0.104
(Male/Female terms)	r	-0.06	0.013*	0.006*	0.013	-0.08	-0.084	-0.075	-0.106
C9: Science/Art	f	-0.009*	-0.184	-0.213	-0.215	-0.296	-0.311	-0.313	-0.317
(Male/Female names)	r	-0.022	-0.183	-0.206	-0.204	-0.293	-0.307	-0.309	-0.312

Table 5: **Effect size of social bias measurements for different language models.** The bias is reported with overall magnitude of CES (d, with rounded values) and statistical significance with two-tailed p-values (p, significant at $p < 0.005$). The cells shaded with gray are statistically in-significant. The data results from CES pooling $N = 5000$ samples from random-effects model. The first row of each WEAT category uses fixed set of samples for each models, denoted as **f** and the second row uses completely random set of samples denoted as **r**. The light, medium and dark shades of gray are used to indicate small, medium and large effect size respectively.

The key characteristics of the model, as highlighted for the $N = 5000$ sample in Table 2, are still quite evident in Table 6. This suggests that achieving similar results is feasible even with a reduced sample size, especially when faced with resource constraints. Nevertheless, it is crucial to recognize that while the overall trends in the model's behavior remain consistent, there are nuanced alterations in the statistical significance of certain cells.

Given these observations, we propose that optimizing the sample size could be a promising avenue for further investigation. Determining the optimal sample size, one that ensures reliable results without sacrificing statistical significance, presents an interesting area for future research and in-depth exploration.

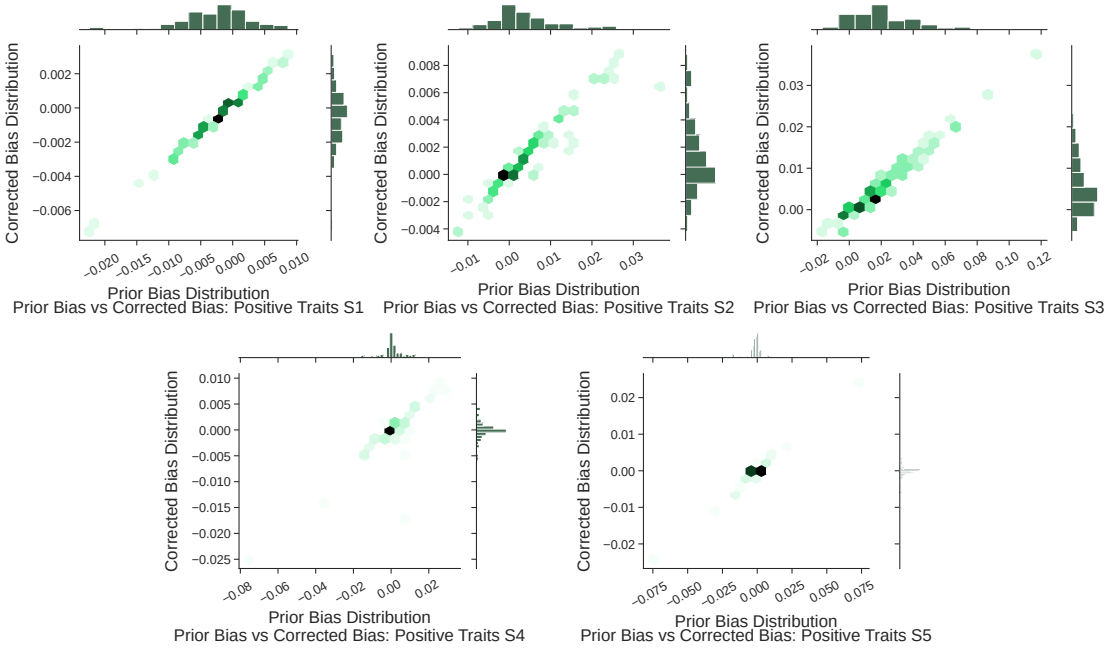
Category		BanglaBERT-Generator				BanglaBERT-Discriminator			
		9	25	75	>75	9	25	75	>75
C1: Flowers/Insects	f	1.226	1.206	1.212	1.22	0.532	0.419	0.461	0.5
(Pleasant/Unpleasant)	r	1.223	1.199	1.212	1.214	0.574	0.462	0.484	0.445
C2: Music/Weapons	f	-0.224	-0.249	-0.245	-0.256	-0.263	-0.159	-0.18	-0.147
(Pleasant/Unpleasant)	r	-0.232	-0.235	-0.254	-0.252	-0.246	-0.192	-0.174	-0.128
C3: Male/Female names	f	0.19	0.149	0.165	0.178	-0.018*	0.026*	0.022*	0.029
(Pleasant/Unpleasant)	r	0.18	0.15	0.166	0.167	-0.002*	0.005*	0.014*	0.028*
C4: Male/Female names	f	0.643	0.734	0.723	0.726	0.019*	0.032	0.015*	0.007*
(Career/Family)	r	0.641	0.751	0.735	0.745	0.016*	0.007*	0.012*	0.017*
C5: Male/Female terms	f	0.259	0.245	0.248	0.243	-0.053*	-0.01*	-0.008*	-0.021*
(Career/Family)	r	0.275	0.262	0.253	0.256	-0.056*	-0.048	0.001*	-0.058
C6: Math/Art	f	0.254	0.176	0.15	0.14	0.023*	0.083	0.054	0.057
(Male/Female terms)	r	0.262*	0.157	0.13	0.125	0.021	0.074	0.075	0.078
C7: Math/Art	f	-0.629	-0.642	-0.639	-0.628	0.252	0.294	0.273	0.295
(Male/Female names)	r	-0.621	-0.63	-0.63	-0.62	0.266	0.282	0.243	0.259
C8: Science/Art	f	0.366	0.306	0.279	0.283	-0.041*	-0.042	-0.055*	-0.046
(Male/Female terms)	r	0.375	0.277	0.282	0.305	-0.031	-0.051	-0.031	-0.082
C9: Science/Art	f	-0.593	-0.635	-0.647	-0.65	-0.124	-0.147	-0.167	-0.159
(Male/Female names)	r	-0.584	-0.635	-0.646	-0.638	-0.117	-0.143	-0.169	-0.163

Category		MuRIL-Large(cased)				XLM-RoBERTa Large(cased)			
		9	25	75	>75	9	25	75	>75
C1: Flowers/Insects	f	1.196	1.218	1.213	1.201	0.275	0.472	0.51	0.482
(Pleasant/Unpleasant)	r	1.192	1.207	1.195	1.201	0.278	0.469	0.499	0.494
C2: Music/Weapons	f	0.478	0.376	0.377	0.358	0.574	0.749	0.774	0.766
(Pleasant/Unpleasant)	r	0.469	0.375	0.382	0.363	0.572	0.733	0.769	0.769
C3: Male/Female names	f	0.48	0.626	0.653	0.648	0.064	0.05	0.059	0.04
(Pleasant/Unpleasant)	r	0.497	0.619	0.655	0.659	0.068	0.044	0.051	0.058
C4: Male/Female names	f	0.015	-0.065	-0.08	-0.093	-0.203*	-0.263	-0.267	-0.326
(Career/Family)	r	0.037	-0.082	-0.076	-0.1	-0.216	-0.279	-0.307	-0.302
C5: Male/Female terms	f	0.205	0.274	0.273	0.245	-0.152	-0.214	-0.21	-0.216
(Career/Family)	r	0.231	0.239	0.244	0.263	-0.158	-0.193	-0.202	-0.206
C6: Math/Art	f	0.412	0.436	0.436	0.386	-0.114	-0.174	-0.152	-0.133
(Male/Female terms)	r	0.411	0.442	0.414	0.419	-0.111	-0.224	-0.139	-0.106
C7: Math/Art	f	-0.141	-0.226	-0.284	-0.288	-1.262	-1.202	-1.193	-1.194
(Male/Female names)	r	-0.131	-0.233	-0.275	-0.276	-1.261	-1.2	-1.187	-1.187
C8: Science/Art	f	-0.068	0.026	0.018	0.014	-0.08	-0.084*	-0.075*	-0.106*
(Male/Female terms)	r	-0.055	0.006*	0.011*	0.011*	-0.076	-0.089	-0.093	-0.102
C9: Science/Art	f	-0.013*	-0.179	-0.231	-0.207	-0.293	-0.307	-0.309	-0.312
(Male/Female names)	r	-0.016*	-0.207	-0.21	-0.205	-0.291	-0.315	-0.309	-0.313

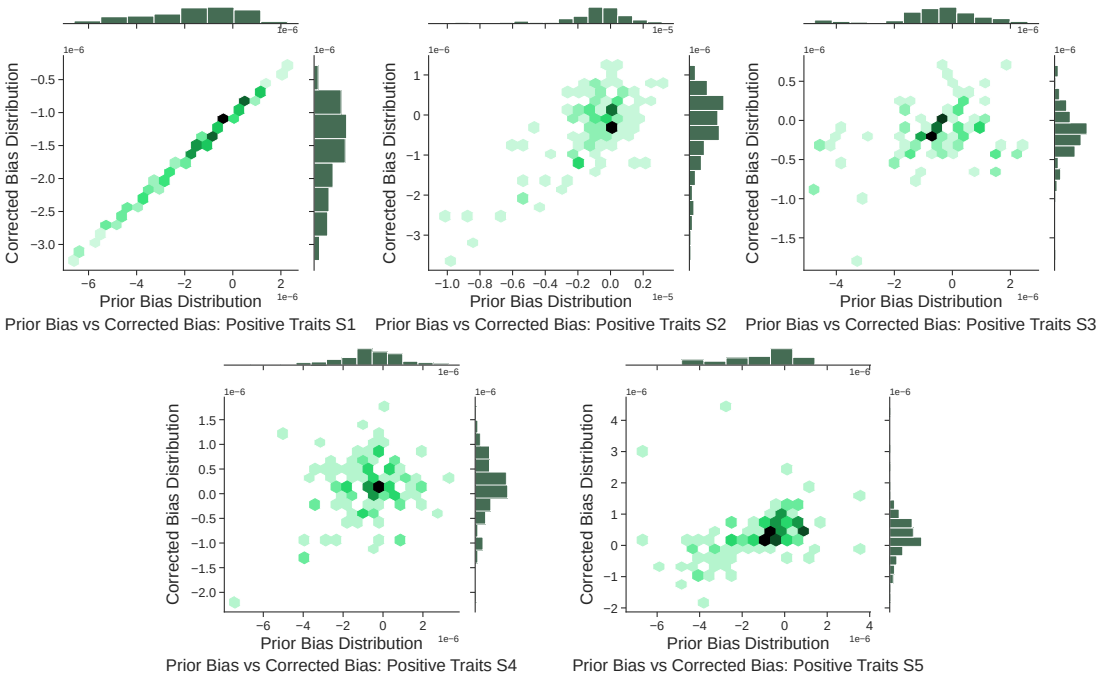
Table 6: **Effect size of social bias measurements for different language models.** The bias is reported with overall magnitude of CES (d, with rounded values) and statistical significance with two-tailed p-values (p, significant at $p < 0.005$). The cells shaded with gray are statistically in-significant. The data results from CES pooling $N = 1000$ samples from random-effects model. The first row of each WEAT category uses fixed set of samples for each models, denoted as **f** and the second row uses completely random set of samples denoted as **r**. The light, medium and dark shades of gray are used to indicate small, medium and large effect size respectively. Compared to pooling with $N = 5000$, more cells with statistically in-significant values are seen.

Model	Score	S1		S2		S3		S4		S5	
		$\bar{x}$ (10^{-3})	σ^2 (10^{-6})	$\bar{x}$ (10^{-3})	σ^2 (10^{-6})	$\bar{x}$ (10^{-3})	σ^2 (10^{-6})	$\bar{x}$ (10^{-3})	σ^2 (10^{-6})	$\bar{x}$ (10^{-3})	σ^2 (10^{-6})
BBGen-L	P_{tgt}	-0.42	2.66	4.02	10.48	6.51	37.35	1.24	24.25	0.15	1.30
	P_{corr}	-0.22	2.66	3.49	12.15	5.58	36.95	0.96	24.71	0.15	1.11
	P_{attr}	0.23	0.71	1.72	102.4	-0.74	28.70	11.79	137.9	10.33	127.9
		$\bar{x}$ (10^{-6})	σ^2 (10^{-6})	$\bar{x}$ (10^{-6})	σ^2 (10^{-6})	$\bar{x}$ (10^{-6})	σ^2 (10^{-6})	$\bar{x}$ (10^{-6})	σ^2 (10^{-6})	$\bar{x}$ (10^{-6})	σ^2 (10^{-6})
MuRIL-L	P_{tgt}	-1.09	$<10^{-10}$	-0.13	$<10^{-10}$	-0.19	$<10^{-12}$	-0.11	$<10^{-12}$	-0.44	$<10^{-12}$
	P_{corr}	-1.87	$<10^{-10}$	0.12	$<10^{-10}$	-0.14	$<10^{-12}$	0.07	$<10^{-12}$	-0.09	$<10^{-12}$
	P_{attr}	1.41	$<10^{-10}$	-0.58	$<10^{-10}$	0.13	$<10^{-12}$	-0.22	$<10^{-12}$	-0.60	$<10^{-10}$

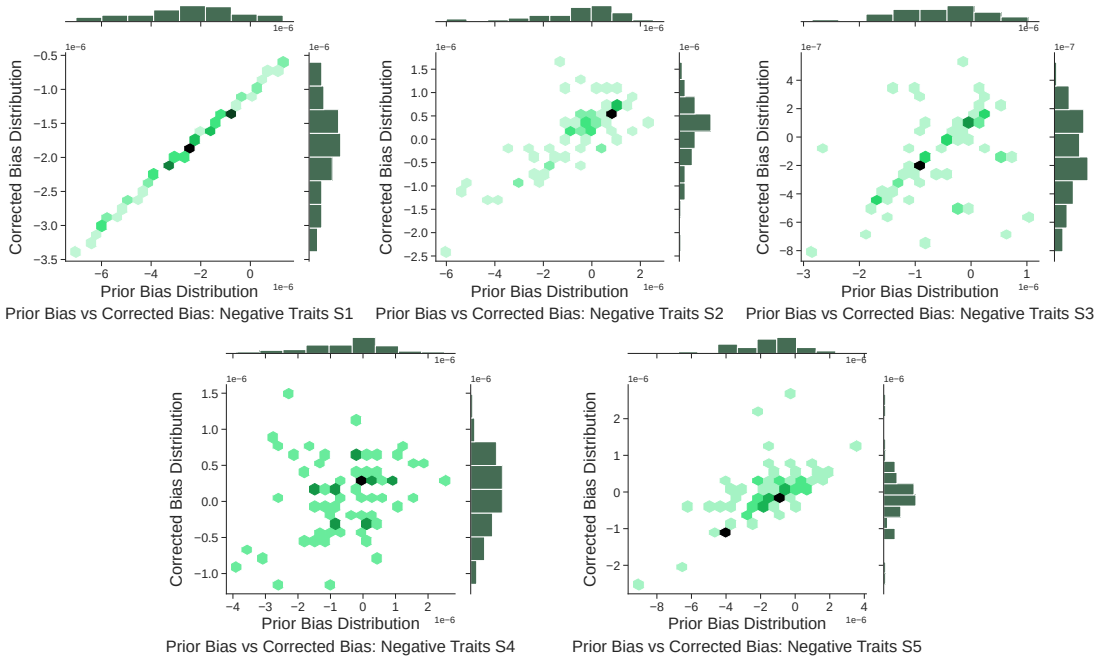
Table 7: Mean($\bar{x}$) and variance (σ^2) for Fill Bias Score (P_{tgt}), Prior Corrected Score (P_{corr}) and Attribute Score (P_{attr}) for two models, **BBGen-L**(*BanglaBERT Large Generator*) and **MuRIL-L**(*MuRIL-Large*) for 5 different sentence structures from S1 to S5 (for negative traits only).P_{attr} is the probability the model assigns to the attribute if the target is given in the sentence.It gives an idea about how much the attribute is associated with the target.



(a) Prior Bias Score vs Corrected Bias Score plots for positive traits in **BanglaBERT** - Large Generator.



(b) Prior Bias Score vs Corrected Bias Score plots for positive traits in **MuRIL** - Large (cased).



(c) Prior Bias Score vs Corrected Bias Score plots for negative traits in **MuRIL** - Large (cased).

Figure 7: A comparison between model behaviors for different sentence structures. Model behaviors are consistent with what was mentioned in the main paper.