Fairness Evaluation in Large Language Models: An Analysis of Intrinsic Metrics

Lucas B. Sena [1], Javam C. Machado[1]

Laboratório de Sistemas e Bancos de Dados (LSBD)
DC/UFC – CEP 60440-900 – Fortaleza – CE – Brazil

{lucas.sena,javam.machado}@lsbd.ufc.br

Resumo. *Grandes e médios Modelos de Linguagem (GML e MML) são fundamentais em aplicações de Processamento de Linguagem Natural (PLN), mas ainda apresentam desafios na avaliação de fairness, detecção e mitigação de vieses. Este trabalho compara métricas de avaliação de viés, como SEAT, Adapted WEAT e CEAT, em modelos como BERT e LLaMA-2, e explora a seleção automática de atributos por LLMs. Os resultados mostram que atributos gerados automaticamente potencializam a detecção de vieses e que métricas alternativas à similaridade do cosseno influenciam significativamente os resultados. O CEAT se destaca na captura de vieses interseccionais, contribuindo para o desenvolvimento de métodos mais robustos para avaliação de fairness.*

Abstract. *Large and medium-sized Language Models (LLMs and MLMs) are fundamental in Natural Language Processing (NLP) applications but still present challenges in fairness evaluation, bias detection and mitigation. This work compares bias evaluation metrics such as SEAT, Adapted WEAT, and CEAT on models like BERT and LLaMA-2 and explores the automatic attribute selection by LLMs. The results show that automatically generated attributes enhance bias detection and that alternative metrics to cosine similarity significantly influence the outcomes. CEAT stands out in capturing intersectional biases, contributing to the development of more robust methods for fairness evaluation.*

1. Master's Information

Level	Master's
Admission	02/2024
Qualifying Exam	02/2025
Defense Expected	12/2025

To date, the publication related to this work is [Sena and Machado 2024].

2. Introduction

The advancement of Language Models (LMs) has significantly impacted Natural Language Processing (NLP), enabling progress in tasks such as translation, summarization, classification, and generation. However, their growing use in sensitive domains raises concerns about reproducing and amplifying social biases [Li et al. 2023].

Here, we use the term Language Models (LMs) to refer to both medium-sized models (e.g., BERT) and Large Language Models (LLMs), such as LLaMA-2. This distinction matters because fairness metrics may behave differently depending on model scale [Li et al. 2023].

Studies have identified biases [Caliskan et al. 2017, May et al. 2019, Kurita et al. 2019, Tan and Celis 2019, Li et al. 2023] related to gender, race, and sexual orientation across models, from traditional embeddings (Word2Vec, GloVe) to contextual models (BERT). These biases can subtly influence outcomes and compromise fairness.

We adopt the notion of algorithmic fairness as the absence of systematic and unjustified disparities across social groups [Barocas et al. 2023, Mehrabi et al. 2021]. Fairness evaluation in LMs is crucial to identify and mitigate harmful associations that reinforce discrimination.

Popular metrics such as WEAT and SEAT have been widely used but show limitations, including sensitivity to sentence templates and difficulty addressing intersectionality [May et al. 2019, Tan and Celis 2019].

In database systems, where LMs support information retrieval, query analysis, summarization, and answer generation, bias evaluation becomes especially critical. Inaccurate or unfair outputs can compromise decision-making and the integrity of retrieved data.

Recent efforts include using log-probability metrics [Kurita et al. 2019], bias mitigation adapters [Lauscher et al. 2021], and automatic attribute generation to reduce human bias in word selection.

In this work, we compare fairness evaluation metrics (SEAT, Adapted WEAT, CEAT) in BERT and LLaMA-2 variants and explore the use of automatic attribute generation. Our aim is to highlight metric limitations and suggest paths toward more robust fairness assessments.

3. Theoretical Background

Algorithmic fairness in Language Models (LMs) aims to prevent systematic disadvantages against social groups. This is especially relevant due to the influence of LMs in automated decision-making.

Traditional embeddings, like Word2Vec [Mikolov et al. 2013] and GloVe [Pennington et al. 2014], generate static representations, while contextual models (e.g., BERT [Devlin et al. 2019], GPT-2 [Radford et al. 2019]) produce dynamic embeddings sensitive to context.

Bias can be quantified using several metrics. The *Word Embedding Association Test* (WEAT) [Caliskan et al. 2017] assesses differences in vector similarities. SEAT

[May et al. 2019] adapts WEAT to sentences using neutral templates like "This is a [target word]." CEAT [Tan and Celis 2019] works at the word level, capturing contextual and intersectional biases.

In SEAT, templates such as "This is a nurse." or "This is a lawyer." are applied to targets and attributes. Their neutrality isolates semantic associations but introduces template dependence, since small phrasing changes can affect bias results.

Other approaches use conditional probabilities in masked language models [Kurita et al. 2019], or adapter layers for efficient bias mitigation [Houlsby et al. 2019, Lauscher et al. 2021].

Intersectionality [Crenshaw 2013] is essential, as individuals may face overlapping biases. For instance, a Black woman may experience discrimination distinct from that faced by Black men or white women—an aspect CEAT is designed to capture.

Each metric offers insights into fairness. In SEAT, gender bias is evaluated by comparing associations of "She is a nurse" vs. "He is a doctor" with attributes like "kind" or "assertive." In CEAT, contextual embeddings of words like "nurse" are analyzed for proximity to gender concepts. Adapted WEAT uses log-probabilities from masked sentences like "[MASK] is a nurse" conditioned on target categories.

4. Related Work

Various methods have been proposed to quantify and mitigate biases in contextualized language models such as BERT. These approaches aim to enhance fairness by identifying and reducing biased associations.

The *Sentence Encoder Association Test* (SEAT) [May et al. 2019] adapts WEAT for sentence encoders using cosine similarity on embeddings from simple templates. SEAT has proven effective in detecting biases—such as gender-career associations and stereotypes like the "angry black woman"—but suffers from template sensitivity and reliance on cosine similarity, especially in highly contextualized models.

To address these issues, [Kurita et al. 2019] introduced a metric based on log-probability differences in masked language modeling. This method leverages the model's native training objective, improving stability and interpretability, though it still depends on template phrasing.

Bias mitigation strategies include ADELE (*Adapter-based Debiasing of Language Models*) [Lauscher et al. 2021], which inserts debiasing adapters without altering base model parameters. While promising, residual biases often persist, and standard metrics may not fully reflect impacts on downstream tasks.

The *Contextualized Embedding Association Test* (CEAT) [Tan and Celis 2019] operates at the word level, enabling finer-grained analysis and capturing intersectional biases. It is particularly useful in cases where overlapping social identities (e.g., gender and race) amplify bias.

Our work differs by offering a comparative analysis of multiple intrinsic metrics (SEAT, Adapted WEAT, CEAT) across models of varying scale (BERT and LLaMA-2). Additionally, we evaluate automatic attribute selection, a largely unexplored strategy that, when combined with diverse metrics, provides a more robust and nuanced bias assessment.

5. Methodology

We evaluated fairness using SEAT and alternative metrics—Adapted WEAT, CEAT, Mahalanobis distance, and Pearson correlation—applied to semantically bleached sentence embeddings. The models analyzed were BERT-base and LLaMA-2 (7B and 13B). Target and attribute word lists were sourced both from traditional benchmarks and automatically generated by the *Gemma 3-27B* model.

Mahalanobis distance was included for its ability to account for embedding covariance, potentially revealing directional biases missed by cosine similarity. Pearson correlation, in turn, offers a symmetric and bounded measure of linear association. These alternatives provide complementary views on how bias may manifest in embedding space.

Effect size (d) was computed as the difference between mean similarities normalized by the standard deviation.

Model selection considered both experimental goals and computational efficiency. The *Gemma 27B* model was used to generate diverse and stereotype-rich word lists, requiring only one inference per prompt, which kept resource use low. For the fairness evaluations—which involve intensive embedding extraction and large-scale similarity computations—we used LLaMA-2 7B and 13B. These models offer a good balance between quality and computational cost.

Although newer models like LLaMA-3 exist, their high resource demand makes them less suitable for large-scale intrinsic bias testing. Moreover, our focus is not on benchmarking the latest architectures but on analyzing fairness metrics under consistent and reproducible conditions. LLaMA-2 remains a strong baseline widely adopted for analyzing linguistic representations.

6. Results

We applied fairness metrics to BERT and LLaMA-2 models to assess biased associations and the impact of target word selection methods.

6.1. Comparison of Bias Metrics

We implemented SEAT, Adapted WEAT, and CEAT as proposed in prior work [May et al. 2019, Kurita et al. 2019, Lauscher et al. 2021, Tan and Celis 2019]. Table 1 shows the effect sizes obtained. While SEAT was stronger for LLaMA-2-7B, Adapted WEAT performed best for BERT, and CEAT revealed stronger signals in LLaMA-2-13B.

Table 1. Comparison of alternative metrics for bias evaluation.

Model	SEAT	Adapted WEAT	CEAT
BERT	1.031	**1.574**	0.604
Llama-2-7B	**0.628**	0.511	0.052
Llama-2-13B	0.686	0.647	**1.369**

6.2. Impact of Target Word Generation

We compared traditional word lists with LLM-generated alternatives using SEAT. Automatically generated lists (via *Gemma 27B*) consistently increased bias effect sizes across all models, as shown in Table 2.

Table 2. SEAT results using traditional vs. LLM-generated words.

Model	Traditional	LLM-Generated
BERT	1.031	**1.356**
Llama-2-7B	0.628	**1.093**
Llama-2-13B	0.686	**1.277**

6.3. Alternative Association Metrics

We also evaluated cosine similarity, Mahalanobis distance, and Pearson correlation (Table 3). Cosine and Pearson produced identical results due to embedding normalization, while Mahalanobis yielded an inverse effect, highlighting directional differences.

Table 3. Comparison of similarity metrics.

Metric	Effect Size	Complexity
Cosine	0.2774	$\mathcal{O}(n)$
Mahalanobis	-1.1684	$\mathcal{O}(n^2)$
Pearson	0.2774	$\mathcal{O}(n)$

Discussion of Results

The results demonstrate that CEAT yielded a notably high effect size for LLaMA-2-13B (1.369), suggesting that larger models may encode more complex biases, and that CEAT is more sensitive to intersectional bias detection. The effect size represents the magnitude of association between social concepts — higher values indicate stronger stereotypical associations. The improved values from automatically generated words (e.g., 1.277 for LLaMA-2-13B using LLM words vs. 0.686 with traditional) show that model-generated attributes better capture latent stereotypes. Furthermore, Mahalanobis distance yielded negative effect sizes, which may point to a direction-sensitive representation of bias, unlike cosine or Pearson similarity which are bounded and symmetric. These findings suggest the choice of similarity metric and word list profoundly influences fairness evaluations.

The negative effect size observed with Mahalanobis distance arises from its sensitivity to direction and scale, as it evaluates distances within a covariance-aware space. In this case, the directionality of the embeddings led to a reversed association signal compared to cosine and Pearson. This result illustrates how different similarity metrics can lead to different interpretations of bias, reinforcing the importance of metric selection when conducting fairness evaluations.

7. Conclusion

This work investigated bias evaluation in language models using the *Sentence Encoder Association Test* (SEAT) and alternative metrics, such as the Adapted WEAT and

CEAT. The analysis was conducted on contextual sentence encoders, including BERT and variants of LLaMA-2, considering different strategies for selecting target words, such as traditional lists and lists automatically generated by large language models.

The results indicate that automatic attribute selection contributes to greater bias detection, as evidenced by the increase in *effect size* values compared to conventional lists. Moreover, the comparison between metrics showed that alternatives to cosine similarity, such as Mahalanobis distance and Pearson correlation, significantly influence the results, highlighting the importance of metric choice in fairness analysis.

The CEAT metric proved particularly effective in capturing intersectional bias nuances, demonstrating its usefulness in contexts where multiple social dimensions are present. These findings underscore the need for more refined and adaptive methodologies for algorithmic fairness evaluation in linguistic representations.

As future directions, we propose deepening the analysis of log-probability-based metrics and investigating methods that integrate multiple association metrics for a more comprehensive bias evaluation. Furthermore, future studies may explore the effectiveness of mitigation techniques, such as the use of adapters, in larger-scale models and across different application domains.

This work contributes to advancing the understanding of the limitations and potential of current fairness metrics, providing support for the development of fairer and more equitable NLP systems.

References

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

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

Crenshaw, K. (2013). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Lauscher, A., Lueken, T., and Glavaš, G. (2021). Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*.

Li, Y., Du, M., Song, R., Wang, X., and Wang, Y. (2023). A survey on fairness in large language models. arxiv. doi: 10.48550. *arXiv preprint arXiv.2308.10149*.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sena, L. and Machado, J. (2024). Evaluation of fairness in machine learning models using the uci adult dataset. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 743–749. SBC.

Tan, Y. C. and Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.