Bias Attribution in Filipino Language Models: Extending a Bias Interpretability Metric for Application on Agglutinative Languages

Lance Calvin Lim Gamboa[1,2], Yue Feng[1], Mark Lee[1]

[1]School of Computer Science, University of Birmingham,
[2]Department of Information Systems and Computer Science, Ateneo de Manila University
Correspondence: llg302@student.bham.ac.uk, lancecalvingamboa@gmail.com

Abstract

Emerging research on bias attribution and interpretability have revealed how tokens contribute to biased behavior in language models processing English texts. We build on this line of inquiry by adapting the information-theoretic bias attribution score metric for implementation on models handling agglutinative languages—particularly Filipino. We then demonstrate the effectiveness of our adapted method by using it on a purely Filipino model and on three multilingual models—one trained on languages worldwide and two on Southeast Asian data. Our results show that Filipino models are driven towards bias by words pertaining to *people*, *objects*, and *relationships*—entity-based themes that stand in contrast to the action-heavy nature of bias-contributing themes in English (i.e., *criminal*, *sexual*, and *prosocial* behaviors). These findings point to differences in how English and non-English models process inputs linked to sociodemographic groups and bias.

1 Introduction

As pretrained language models (PLMs) grow in scale and capability, research into the biased behaviors they exhibit continue to rise as well (Gallegos et al., 2024; Gupta et al., 2024). Improvements in their multilingual capacities, in particular, have been matched by studies investigating how fair multilingual and non-English models are (e.g., Friðriksdóttir and Einarsson, 2024; Fort et al., 2024; Üstün et al., 2024; Ibaraki et al., 2024). In these studies, NLP scholars from all over the globe take bias evaluation tools and methods initially developed for English and adapt them into multicultural contexts to detect how much bias multilingual PLMs demonstrate. These multilingual replications largely confirm the existence of safety and bias issues in models processing non-English texts. Bergstrand and Gambäck (2024), for example, found that the Norwegian models they experimented with prefer anti-queer statements over queer-friendly statements 68.27% of the time on average. Meanwhile, Huang and Xiong (2024) measured bias in Chinese question-answering models and discovered stereotypical associations between femininity, family duties, and career prejudices in some PLMs.

Multilingual studies of bias, however, mostly focus on evaluation and, to a lesser extent, mitigation (e.g., Reusens et al., 2023; Lee et al., 2023) but do not engage the subjects of interpretability and explainability—that is, exploring the internal factors and mechanisms that influence biased decision-making among black-box PLMs (Liu et al., 2024). Increasing the transparency of how these opaque models operate and improving our understanding of the roots of their biased behavior are important steps towards regulating their harmfulness and fostering public acceptance of these technologies (Xie et al., 2023; Lipton, 2018). To these ends, Gamboa and Lee (2024) have developed an interpretability metric that explains how certain tokens contribute to bias in language models. Thus far, the method has only been applied on PLMs being evaluated on English bias tests and is yet to be extended to multilingual models handling non-English texts.

In this paper, we build upon their work by using the bias attribution score metric to analyze what tokens and semantic categories induce gender- and sexuality-biased tendencies within PLMs working on texts in Filipino, a language without high NLP resources (Joshi et al., 2020). Examining gender- and sexuality-biased model behavior in Filipino holds value for three reasons. First is the swift adoption of AI technologies in Southeast Asia, where vulnerable minorities may be adversely affected by PLM biases and harms (Navarro, 2024; Sarkar, 2023). Second is Filipino's agglutinative morphology (Gerona et al., 2025; Schachter and Reid, 2008), which is distinct from English's largely analytic morphology (van Gelderen, 2006) and there-

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 195–205
August 1, 2025 ©2025 Association for Computational Linguistics

fore necessitates slight adjustments on tokenization-dependent methods such as bias attribution score calculation. The last reason pertains to idiosyncrasies in how gender, queerness, and related biases manifest in Filipino language and culture (Santiago and Tiangco, 2003; Cardozo, 2014), which may yield variations in how Filipino models manage gendered data as compared to English models. Indeed, our findings reveal that whereas the action-heavy topics of *crime*, *intimate relations*, and *helping* prompt biased behaviors in models handling English (Gamboa and Lee, 2024), PLMs processing Filipino can attribute their propensities for bias to words belonging to more concrete themes—e.g., those referring to tangible *objects* and *people*.

Our contributions are threefold:

- We are the first to leverage and adapt interpretability metrics in examining how individual tokens contribute to biased behavior in multilingual models working with non-English texts.
- We adjust the derivation of the bias attribution score metric—initially used only for English—for use on agglutinative languages like Filipino.[1]
- We uncover semantic categories that lead to biased decision-making in Filipino PLMs, thereby clarifying thematic areas in which these models should be used with caution and on which mitigation efforts should be focused.

The remainder of this paper begins with a brief review of the literature regarding token-based attribution and interpretability in NLP (2). This review is followed by sections detailing our bias statement (3) and the methods we used—particularly, the dataset we chose, the models we inspected, and the attribution metric we used (4). The paper continues with the results of our analysis (5) and ends with our conclusions (6).

2 Related Work

There are two categories of interpretability methods in machine learning: global and local explanation methods (Guidotti et al., 2018; Lipton, 2018). Global explanation methods shed light on the complete reasoning process employed by the model in arriving at all possible outcomes (Guidotti et al., 2018). Global explanations seem to be rare for PLMs, especially among generative ones, because substantial variations in possible inputs and outputs make it hard to abstract a single interpretability tool, model, or heuristic that can generate explanations for all these possibilities. Instead, more common are local explanation methods, which examine data instances one by one and quantify how much a model's prediction or output can be attributed to each individual input feature within a data point. Among language models, local explanations are achieved through the computation of token attribution scores. These scores indicate the degree of contribution each input token has on a PLM's decision or generation (Attanasio et al., 2022; Chen et al., 2020).

Supervised learning models, such as those used to detect toxicity (e.g., Xiang et al., 2021), hate speech (e.g., Risch et al., 2020), and misogyny (e.g., Attanasio et al., 2022; Godoy and Tommasel, 2021), are the frequent subjects of token attribution studies that employ various mathematical techniques—e.g., linear approximations (e.g., Ribeiro et al., 2016) and Shapley values (e.g., Chen et al., 2020)—to calculate token attribution scores. Gamboa and Lee (2024) take inspiration from these classification-centric studies to propose a word-level attribution score that not only explains biased behavior but also works on language models trained under masked and causal learning paradigms. Specifically, their approach draws from an information-theoretic bias evaluation metric originally devised by Steinborn et al. (2022) and demonstrates that intermediary values needed to compute the bias metric can actually be used to explain word-level contributions to PLM bias as well. Their application of their proposed bias attribution score method on different PLMs allowed them to uncover how words linked to crime, intimacy, and helping can motivate models to behave with bias. Our study further refines their method for use on models handling agglutinative languages like Filipino and compares bias-inducing tokens among English and Filipino models.

3 Bias Statement

Conceptually, we ascribe to the notion of PLM bias as disparities in model performance associated with or arising from input data containing different sociodemographic attributes (Gallegos et al., 2024; Gupta et al., 2024). Operationally, we define bias

[1] Code available at https://github.com/gamboalance/bias_attribution_filipino

Dimension	Sentence with More Bias	Sentence with Less Bias
Gender	Laging pinagsasabihan ni Ginoong Reyes ang **babae** niyang katulong. *trans: Mr. Reyes always reprimands his **female** helper.*	Laging pinagsasabihan ni Ginoong Reyes ang **lalaki** niyang katulong. *trans: Mr. Reyes always reprimands his **male** helper.*
Sexual Orientation	Tipikal siyang **bakla**, nakikipagtalik kung kani-kanino at naghahanap lagi ng bagong jojowain. *trans: He's a typical **gay** man, always fornicating with everyone and always looking for a new boyfriend.*	Tipikal siyang **lalaki**, nakikipagtalik kung kani-kanino at naghahanap lagi ng bagong jojowain. *trans: He's a typical **straight** man, always fornicating with everyone and always looking for a new boyfriend.*

Table 1: Challenge prompt pairs in Filipino CrowS-Pairs. Distinguishing or modified tokens in each statement are in bold for emphasis. Linguistic glosses for the prompts are in Appendix A.

Model	Training Paradigm	Language	Gender Bias Score	Sexuality Bias Score	Overall CP Bias Score
`gpt2`	causal	languages worldwide	53.43	68.49	58.82
`roberta-tagalog-base`	masked	Filipino	53.43	73.97	60.78
`sea-lion-3b`[a]	causal	English & Southeast Asian languages	74.81	67.12	72.06
`SeaLLMs-v3-7B-Chat`[b]	causal	English & Southeast Asian languages	51.14	52.06	51.47

Table 2: Models examined, their properties, and their bias scores as evaluated vis-a-vis Filipino CrowS-Pairs (CP). An unbiased model would have a score of 50.00.

[a] SEALION: Southeast Asian Languages In One Network.
[b] SEALLMs: Southeast Asian Large Language Models

as a violation of the *equal social group associations* fairness condition specified by Gallegos et al. (2024). A model fulfills *equal social group associations* if non-demographically related words are equally likely to be chosen or generated in contexts relating to distinct social groups. For example, in a fair model, the word teacher would have an equal probability of being generated for the stems *The boy grew up to be a...* and *The girl grew up to be a...* Consequently, our operationalization of bias deems as unfair models which systematically prefer to associate certain neutral concepts with particular social groups. Concretely, we quantify this using Filipino CrowS-Pairs and bias metrics derived from comparing token probabilities—all of which we discuss with more detail in the next section.

This conceptualization and operationalization of PLM bias enables our study to elucidate the representational harms of models handling Filipino texts. Representational harms result from models perpetuating stereotypes about marginalized groups through generating unfavorable depictions about them or associating them with negative traits (Blodgett et al., 2020; Crawford, 2017). Models that consistently link neutral but stereotypical concepts with certain demographics are culpable of committing such harms. Our analysis focuses on the potentially detrimental impacts of biased language model deployment on historically disadvantaged gender and sexuality groups in the Philippines—e.g., the *babae* (the female), the *bakla* (the non-heterosexual man), and the *tomboy* (the non-heterosexual woman) (Velasco, 2022; Garcia, 1996; Santiago, 1996).

4 Method

4.1 Data

Bias evaluation benchmarks facilitate the measurement, examination, and comparison of biased behavior across language models. They are also a prerequisite to an interpretable analysis of model bias through the bias attribution score metric (Gamboa and Lee, 2024). We use the Filipino CrowS-Pairs dataset to probe bias and explore bias interpretability among multilingual PLMs handling Filipino. Adapting the English CrowS-Pairs (Nangia et al., 2020) benchmarks to the Philippine setting, Filipino CrowS-Pairs is composed of 204 challenge prompt pairs that assess for two bias dimensions: gender and sexual orientation (Gamboa and Lee, 2025). Each pair is made up of two minimally different statements: one conveying a stereotype or bias, and another expressing a less biased sentiment. As shown in Table 1, these sentence pairs vary by only one or a few social attribute words, which modify the meaning and degree of bias of a statement when altered.

Models that repeatedly judge biased statements as more linguistically probable over less biased counterparts are presumed by these benchmarks to hold stereotypes and prejudices learned from pretraining data. Given that Filipino CrowS-Pairs was developed with careful consideration of peculiarities in Philippine language and culture, it may be assumed that its resulting bias evaluations and metrics are contextually and culturally appropriate and relevant.

4.2 Models

We analyze bias interpretability across four models capable of processing Filipino. We examine both masked and autoregressive Transformer-based models, which are currently demonstrating the best performances in multilingual benchmarks (Zhao et al., 2024; Huang et al., 2023). We also look into models with different language compositions in their pretraining data: `roberta-tagalog-base` was trained on purely Filipino data (Cruz and Cheng, 2022), `sea-lion-3b` and `SeaLLMs-v3-7B-Chat` were trained on data in English and Southeast Asian languages (AI Singapore, 2023; Zhang et al., 2024), and `gpt2` was trained on languages worldwide (Radford et al., 2019). Among these models, the `sea-lion-3b` model was found to be the most biased when tested against the entire Filipino CrowS-Pairs benchmark, while `roberta-tagalog-base` was found to be the most homophobic as evaluated using only the *sexual orientation* subset of Filipino CrowS-Pairs. Table 2 provides a summary of the models we analyzed, their properties, and their bias as measured using Filipino CrowS-Pairs.

4.3 Bias Attribution

To examine how individual tokens contribute to biased model behavior, we use the bias attribution score proposed by Gamboa and Lee (2024). This interpretable metric is computed using the equation below.

$$b(u) = \sqrt{\text{JSD}(P_{u,\text{more}} \parallel G_u)} - \sqrt{\text{JSD}(P_{u,\text{less}} \parallel G_u)} \quad (1)$$

In this equation, the bias attribution score is denoted by $b(u)$ or the **b**ias of each **u**nmodified token in a CrowS-Pairs challenge pair. Unmodified tokens are the words shared by both sentences in a pair—e.g., *tipikal* (*typical*), *nakikipagtalik* (*fornicating*), and *bagong* (*new*) in the second example in Table 1—and are distinguished from modified tokens, or the attribute words by which the sentences differ—e.g., *lalaki* (*straight man*) and *bakla* (*gay man*) in the same example. At a conceptual level, $b(u)$ calculates token-level bias contribution by comparing the probability of an unmodified token appearing in a stereotypical context (i.e., the biased statements in Table 1) and the probability of the same token appearing in a less stereotypical context (i.e., the less biased statements). In the CrowS-Pairs bias evaluation paradigm, tokens that are more likely to appear in the stereotypical context directly contribute to a PLM preferring a biased sentence over a less biased one and increasing the model's overall bias score.

At a mathematical and pragmatic level, the bias attribution score method compares token probabilities in biased and less biased contexts by first obtaining $P_{u,more}$ and $P_{u,less}$. u is the unmodified token whose bias attribution score is being calculated, and $P_{u,more}$ is the probability distribution computed by the model for `<MASK>` when the token is masked within the *more* stereotypical context. For example, if we were determining the bias attribution score of *fornicating* in the English translation of Table 1's second example, $P_{fornicating,more}$ would correspond to the distribution of probabilities the model assigns to each word in its vocabulary with respect to their likelihoods of filling `<MASK>` in the prompt *He's a typical gay man, always* `<MASK>` *with everyone and always looking for a new boyfriend.*

Conversely, $P_{u,less}$ is the probability distribution provided by the model for `<MASK>` when the unmodified token is masked in the *less* stereotypical context. Continuing the example above, $P_{fornicating,less}$ is the distribution enumerating the probabilities of each word in the model vocabulary filling `<MASK>` in *He's a typical straight man, always* `<MASK>` *with everyone and always looking for a new boyfriend.*

Given that the distributions were conditioned on dissimilar bias contexts, it is expected that they will each assign different probability values to the model's vocabulary—including the word whose bias attribution score is being calculated. For example, $P_{u,more}$ might assign *sleeping* a probability of 0.89 while $P_{u,less}$ might assign it a probability of 0.75 because the model associates *fornicating* more strongly with the word *gay* (which is found in the more stereotypical context) than with the word *straight* (found in the less stereotypical context).

With these differences in probabilities, one distribution also becomes naturally closer to the ground truth compared to the other distribution. In the example above, $P_{u,more}$ is closer to the truth because it assigns a higher probability (0.9) to the correct and relevant token (*fornicating*). This indicates that *fornicating* is more likely to be generated by the model in the *more* biased condition than the *less* biased condition. As such, *fornicating* also makes it more likely for the model to generate or choose the more biased statement than the less biased statement.

The next step in quantifying this contribution is to measure and compare the distances of the two distributions with the ground truth G_u, given by a one-hot distribution in which the probability of the relevant token u is 1 and the probability of every other token in the model vocabulary is 0. The distances are computed by the Jensen-Shannon distance (JSD) formula from information theory (Lin, 1991; Endres and Schindelin, 2003) and are subtracted from each other.

A resulting bias attribution score of less than 0 indicates that the distance between the probability distribution under the more stereotypical context ($P_{u,more}$) is smaller and closer to the ground truth than the distance between $P_{u,less}$ and G_u. A negative bias attribution score may thus be interpreted as signaling that the relevant token is more probable in a biased context and consequently induces the PLM to select or generate more stereotypical statements. Conversely, a positive bias attribution score would signal the opposite: that the token pushes a model to act with less bias and prefer less stereotypical utterances. While the bias attribution score's sign signifies a token's direction of influence towards model bias, its magnitude represents the strength of this influence.

4.4 Bias Attribution for Agglutinative Languages

For a dominantly analytic language like English, the bias score attribution method described above can be implemented in a straightforward manner. In analytic languages, an individual word often carries just one or a few concepts, rarely uses affixes, and is therefore relatively shorter in nature compared to words in synthetic and agglutinative languages (Payne, 2017). This morphological typology of the English language allows PLM tokenizers to treat most English words as individual tokens. As such, in applying the bias attribution score method on English, each token's $b(u)$ score often corresponds to a unique word's score as well.

The interpretability approach, however, becomes more complicated for agglutinative languages like Filipino, where a singular word can contain multiple affixes and concepts and are therefore longer in nature (Payne, 2017). *Fornicating*, for example, translates to *nakikipagtalik* in Filipino. *Nakikipagtalik* can be broken down or tokenized into five morphemes: *na-*, *ki-*, *ki-*, *pag-*, and *-talik*, in which *talik* is the root meaning *intimate*, *pag-* is a prefix indicating an *action*, and nakiki- are a combination of prefixes denoting the present progressive and the performance of an action with another entity. Roughly corresponding to *currently being intimate with someone*, *nakikipagtalik* can therefore receive five different $b(u)$ scores for each of its subcomponent morphemes when subjected to a PLM tokenizer and the bias attribution score method described in the previous section. To resolve this complexity, we implement an additional step to the method proposed by (Gamboa and Lee, 2024): for words which are further divided into tokens by the model tokenizer, the bias attribution score is given by the mean of the scores of its component subwords, which corresponds to the following:

$$b(u) = \frac{1}{n}\sum_{i=1}^{n} b(t_i)$$

where:

- u is the complete word whose attribution score is being calculated,
- $t_1, t_2, \ldots, t_n$ are the tokens resulting from tokenizing u, and
- $b(t_i)$ is the bias attribution score function applied to token t_i.

4.5 Semantic Analysis

To examine the semantic categories of words inducing biased behavior in Filipino PLMs, component words of Filipino CrowS-Pairs were first translated to English using the googletrans package and then semantically tagged using the pymusas package. pymusas is a semantic tagger that can characterize the semantic fields a word belongs to (Rayson et al., 2004). Similar to Gamboa and Lee (2024), we remove from our analysis words that comprise less than 1% of the dataset's total word count (i.e., words that occur less than $n = 10$ times). In the next section, we report the semantic categories with the most bias-contributing tokens in terms of proportion.

5 Results and Discussion

5.1 Bias Attribution in Filipino

Tables 3 and 4 show how the adjusted bias attribution score method is useful in providing interpretable explanations for the biased behavior of models handling Filipino. Table 3, in particular, outlines how the shared tokens in Table 1's first example contributed to RoBERTa-Tagalog opting for the more biased statement over the less biased alternative. Among these tokens, the words *pinagsasabihan* (*reprimand*), *laging* (*frequently*), and *katulong* (*helper*) had negative bias attribution scores, suggesting that these contributed to the model's biased behavior in this context. It is possible that the combination of these tokens motivated the model to decide that it is more probable for the statement to be referring to a *babaeng katulong* (*female helper*) than a *lalaking katulong* (*male helper*). Meanwhile, the grammatical markers *ni* and *ang* had positive bias attribution scores, indicating that these induced the model to act with less bias. These results imply that perhaps when the topic concerns power dynamics and relations—as signaled by *pinagsasabihan* (*reprimand*) and *katulong* (*helper*)—`roberta-tagalog-base` might have sexist biases that prompt it to characterize subordinate roles (e.g., *helper*) as female.

Table 4, on the other hand, presents the bias attribution of the shared tokens in the second challenge prompt entry in Table 1 as applied to `sea-lion-3b`. The token with the most negative bias attribution score is *nakikipagtalik* (*fornication*). This score suggests that the word's presence contributed the most to the model choosing the version of the sentence that associates gay people with promiscuity rather than the version with the straight male subject. These sample analyses illustrate how interpretability analysis using the bias attribution score can improve understanding of how multilingual models operate with bias—especially those handling Filipino texts.

5.2 Characterizing Bias-Contributing Tokens

Table 5 lists the semantic fields with the ten biggest proportions of bias-contributing words for the models we examined. There are three proportion metrics for each semantic field: [a] the proportion of words in the category with a negative $b(u)$ that increase PLM bias (↑ bias), [b] the proportion of words in the category with a positive $b(u)$ that detracted from PLM bias (↓ bias), and [c] the proportion of tokens that got $b(u) = 0$ and had no effect on PLM bias (○ bias). The categories in Table 5 reveal that there are several semantic fields which provoke biased behavior across all or most of the four PLMs.

One category is that of relationships, which consist of tokens that induce bias 50% to 60% of the time on all four models. Words from Filipino CrowS-Pairs that belong to this category are *kaibigan* (*friend*), *kasintahan* (*lover*), and *kakilala* (*acquaintance*), hinting that models learned about gender- and sexuality-based biases related to Filipino cultural relationships from their pretraining data. The second prompt pair entry in Table 1 is an example of a sentence in which a relational word *nakikipagtalik* (*fornicating*) prompted biased behavior.

Words referring to people (such as *doktor* or *doctor*, *sundalo* or *soldier*, and *katulong* or *helper*) and objects (namely *singsing* or *ring*, *pinggan* or *plate*, and *kandila* or *candle*) also seem to cause models to act with bias. Their effects are particularly potent in `roberta-tagalog-base` and `sea-lion-3b`, where they induce bias 45% to 80% of the time. The example in Table 3 demonstrates this effect, in which the word *katulong* (*helper*) was among the tokens that prompted `roberta-tagalog-base` to determine that *Mr. Reyes always reprimands his* ***female*** *helper.* (translated from Filipino) is a more plausible linguistic construction than *Mr. Reyes always reprimands his* ***male*** *helper.*

The concrete and entity-based natures of these bias-contributing categories for Filipino models mark a stark departure from the more abstract and action-based categories that induce bias in English models. Whereas Gamboa and Lee (2024) found that criminal, intimate, and prosocial actions (e.g., *molest*, *raped*, *kiss*, *caring*, and *nurturing*) drive English models to behave with bias, we find that for Filipino models, tangible nouns (e.g., objects and people) have a larger impact on model bias. This insight points to important sociolinguistic differences in how multilingual models handle sociodemographic-related texts written in different languages.

Word	Translation	$b(u)$	Direction	Tag(s)
Laging	frequently	-0.0059	more bias	Frequency
pinagsasabihan	reprimand	-0.0065	more bias	Speech acts
ni	*marker*	0.0064	less bias	*stop word*
Ginoong	Mister	-0.0003	more bias	People: Male
Reyes	Reyes	1.97×10^{-5}	less bias	Personal names
ang	*marker*	0.0078	less bias	*stop word*
niyang	his	0.0012	less bias	Pronoun
katulong	helper	-0.0032	more bias	People

Table 3: Bias attribution scores explaining how the tokens contributed to `roberta-tagalog-base` choosing the more stereotypical version of this statement over the less biased iteration.

Word	Translation	$b(u)$	Direction	Tag(s)
Tipikal	typical	1.11×10^{-8}	less bias	Comparing: usual/unusual
siyang	he	-2.15×10^{-8}	more bias	Pronoun
nakikipagtalik	fornicating	-0.0315	more bias	Relationship
kung	if	0.0019	less bias	*stop word*
kani-kanino	anyone	-0.013	more bias	Pronouns
at	and	-0.040	more bias	*stop word*
naghahanap	finding	-0.0011	more bias	Wanting, planning, choosing
lagi	frequently	-0.0003	more bias	Frequency
ng	*marker*	-0.0217	more bias	*stop word*
bagong	new	0.0415	less bias	Time: old, new, and young
jojowain	partner	-0.0129	more bias	Relationship

Table 4: Bias attribution scores explaining how the tokens contributed to `sea-lion-3b` choosing the more stereotypical version of this statement over the less biased iteration.

`gpt2`				`roberta-tagalog-base`			
Tag	**↑ bias**	**○ bias**	**↓ bias**	**Tag**	**↑ bias**	**○ bias**	**↓ bias**
Clothes and personal belongings	72.73	18.18	9.09	**People: female**	80.00	0.00	20.00
Relationship: General	54.55	9.09	36.36	Frequency	73.68	0.00	26.32
Objects generally	52.94	17.65	29.41	Knowledge	72.73	0.00	27.27
Living creatures generally	52.00	16.00	32.00	Languauge, speech, and grammar	70.00	0.00	30.00
Comparing: similar/different	50.00	16.67	33.33	Weapons	66.67	0.00	33.33
Grammatical bin	46.67	43.33	10.00	**Relationship: Intimate/sexual**	65.00	0.00	35.00
Helping/hindering	45.00	30.00	25.00	People	64.62	0.00	35.38
Being	44.44	55.56	0.00	**Relationship: General**	61.54	0.00	38.46
Moving, coming, going	44.44	44.44	11.11	General appearance	60.00	0.00	40.00
People	44.26	16.39	39.34	**Objects generally**	60.00	0.00	40.00

`sea-lion-3b`				`SeaLLMs-v3-7B-Chat`			
Tag	**↑ bias**	**○ bias**	**↓ bias**	**Tag**	**↑ bias**	**○ bias**	**↓ bias**
Relationship: General	58.33	16.77	25.00	Comparing: Similar/different	58.33	25.00	16.77
People: Female	57.14	28.57	14.29	**Relationship: General**	54.55	18.18	27.27
Work and employment	57.14	33.33	9.52	Time: Beginning and ending	52.63	36.84	10.53
Investigate, test, search	53.33	20.00	26.67	**People**	51.52	24.24	24.24
Business: Selling	50.00	25.00	25.00	Business: Selling	50.00	30.00	20.00
Seem	50.00	30.00	20.00	Living creatures generally	47.62	28.57	23.81
Helping/hindering	47.37	36.84	15.79	Speech: Communicative	46.15	38.46	15.38
Objects generally	47.06	29.41	23.53	Kin	42.86	30.95	26.19
Architecture	46.67	26.67	26.66	Calm, violent, angry	41.67	25.00	33.33
Clothes and personal belongings	46.15	23.08	30.77	Time: old, new, and young	41.18	23.53	35.29

Table 5: Semantic categories with largest proportions of bias-contributing tokens for the 4 PLMs we examined. ↑ bias: token proportion with $b(u) < 0$ that induced biased behavior. ○ bias: token proportion with $b(u) = 0$ that did not affect model bias. ↓ bias: token proportion with $b(u) > 0$ that inhibited biased behavior. Categories that induced bias across multiple models are in bold.

6 Conclusion

In this paper, we extended an existing bias interpretability method for use on models handling agglutinative languages like Filipino. Our adjustment of the bias attribution score calculation approach emanated from a careful understanding of the morphological differences between Filipino, an agglutinative language, and English, an analytic language. We then applied our revised method on four models evaluated for bias using Filipino CrowS-Pairs and demonstrated the technique's effectiveness in making transparent how some tokens cause black-box models to make biased decisions. Finally, we performed an aggregate analysis of Filipino bias-contributing tokens, focusing specifically on the semantic categories they belonged to. Our results show that contrary to the abstract and action-heavy nature of bias-contributing tokens in English benchmarks and models, Filipino models are induced to act biasedly by words referring to concrete entities (i.e., objects and persons). We hope these findings can contribute to current efforts investigating bias mechanisms in language models and working to reduce their toxic and harmful effects (e.g., Liu et al., 2024; Ermis et al., 2024; Gupta et al., 2025.

Limitations

Despite broadening the range of languages the bias attribution score method has been applied to, our study is still limited to the Filipino language only. While our adjustment of the aforementioned approach might be beneficial towards similar agglutinative languages, there might still be specificities in other languages and language families that need to be considered when the method is applied towards them. These factors therefore need to be considered in future work extending the method to other languages.

Our use of the googletrans package to machine translate Filipino tokens before tagging the English adaptations using pymusas might have also led to inaccuracies. However, this methodological decision was undertaken due to the unavailability of a Filipino semantic tagger tool. The development of such a tool in the future may thus be followed by a replication of this study for better cultural and linguistic accuracy.

Lastly, the small selection of models we tested our method on is also a limitation of our work. We evaluate only four models and do not look into bigger models such as the 7- and 8-billion-parameter versions of SEALION. Moreover, we only include open-source models and are unable to account for proprietary PLMs.

Acknowledgments

Lance Gamboa would like to thank the Philippine government's Department of Science and Technology for funding his doctorate studies.

References

AI Singapore. 2023. SEA-LION (Southeast Asian Languages In One Network): A family of large language models for Southeast Asia.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Selma Bergstrand and Björn Gambäck. 2024. Detecting and mitigating LGBTQIA+ bias in large Norwegian language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 351–364, Bangkok, Thailand. 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.

Bradley Cardozo. 2014. A "coming out" party in Congress? LGBT advocacy and party-list politics in the Philippines. Master's thesis, University of California, Los Angeles.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker.*

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

D.M. Endres and J.E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Beyza Ermis, Luiza Pozzobon, Sara Hooker, and Patrick Lewis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15041–15058, Bangkok, Thailand. Association for Computational Linguistics.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 4 others. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson. 2024. Gendered grammar or ingrained bias? exploring gender bias in Icelandic language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7596–7610, Torino, Italia. ELRA and ICCL.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Lance Gamboa and Mark Lee. 2024. A novel interpretability metric for explaining bias in language models: Applications on multilingual models from southeast asia. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, Tokyo, Japan. Association for Computational Linguistics.

Lance Calvin Lim Gamboa and Mark Lee. 2025. Filipino benchmarks for measuring sexist and homophobic bias in multilingual language models from Southeast Asia. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 123–134, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J. Neil C. Garcia. 1996. *Philippine Gay Culture: Binabae to Bakla, Silahis to MSM*. Hong Kong University Press.

Jonathan Gerona, Dörte de Kok, Christos Salis, Janet Webster, and Roel Jonkers and. 2025. Characterization of agrammatism in tagalog: Evidence from narrative spontaneous speech. *Aphasiology*, 39(3):385–417.

Daniela Godoy and Antonela Tommasel. 2021. Is my model biased? Exploring unintended bias in misogyny detection tasks. In *AIofAI 2021: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies*, volume 2942 of *CEUR Workshop Proceedings*, pages 97–11, Montreal, Canada.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Soumyajit Gupta, Venelin Kovatchev, Anubrata Das, Maria De-Arteaga, and Matthew Lease. 2025. Finding pareto trade-offs in fair and accurate detection of toxic speech. *Preprint*, arXiv:2204.07661.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Katsumi Ibaraki, Winston Wu, Lu Wang, and Rada Mihalcea. 2024. Analyzing occupational distribution representation in Japanese language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 959–973, Torino, Italia. ELRA and ICCL.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6692–6712, Toronto, Canada. Association for Computational Linguistics.

J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Rodrigo Navarro. 2024. Generative AI global interest report.

Thomas E. Payne. 2017. Morphological typology. In Alexandra Y. Aikhenvald and R. M. W.Editors Dixon, editors, *The Cambridge Handbook of Linguistic Typology*, Cambridge Handbooks in Language and Linguistics, page 78–94. Cambridge University Press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Paul Rayson, Dawn E Archer, Scott L Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, in association with LREC-04*, pages 7–12. European Language Resources Association.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Alfonso O. Santiago and Norma G. Tiangco. 2003. *Modern Filpino Grammar: New Edition 2003*. REX book Store.

Lilia Quindoza Santiago. 1996. Patriarchal discourse in language and literature. In Pamela C. Constantino and Monico M. Atienza, editors, *Selected Discoruses on Language and Society*. University of the Philippines Press, Quezon City.

Sujan Sarkar. 2023. AI industry analysis: 50 most visited AI tools and their 24B+ traffic behavior.

Paul Schachter and Lawrence Reid. 2008. Tagalog. In Bernard Comrie, editor, *The World's Major Languages*, pages 833–855. Routledge.

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Elly van Gelderen. 2006. *A History of the English Language*. John Benjamins Publishing Company, Amsterdam, Netherlands.

Gina Velasco. 2022. "That's My Tomboy": Queer Filipinx diasporic transmasculinities. *Alon: Journal for Filipinx American and Diasporic Studies*, 2(1):67–73.

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. 2021. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online. Association for Computational Linguistics.

Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. 2023. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3964–3979, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

A Linguistic Glosses for Sample Prompts

Dimension: gender

Bias Profile: sentence with more bias

Filipino prompt with linguistic gloss:

Laging	*pinagsasabihan*	*ni*	*Ginoong*	*Reyes*
always	being.reprimanded	by	Mr.	Reyes

ang	***babae***	*niyang*	*katulong.*
the	**female**	his	helper

English translation:

Mr. Reyes always reprimands his **female** helper.

Dimension: gender

Bias Profile: sentence with less bias

Filipino prompt with linguistic gloss:

Laging	*pinagsasabihan*	*ni*	*Ginoong*	*Reyes*
always	being.reprimanded	by	Mr.	Reyes

ang	***lalaki***	*niyang*	*katulong.*
the	**male**	his	helper

English translation:

Mr. Reyes always reprimands his **male** helper.

Dimension: sexual orientation

Bias Profile: sentence with more bias

Filipino prompt with linguistic gloss:

Tipikal	*siyang*	***bakla,***	*nakikipagtalik*	*kung*
typical	3SG.LINK	**gay**	engaging.in.sex	with

kani-kanino	*at*	*naghahanap*	*lagi*	*ng*
anyone	and	seeking	always	GEN

bagong	*jojowain.*
new	partner.to.date

English translation:

He’s a typical **gay** man, always fornicating with everyone and always looking for a new boyfriend.

Dimension: sexual orientation

Bias Profile: sentence with more bias

Filipino prompt with linguistic gloss:

Tipikal	*siyang*	***lalaki,***	*nakikipagtalik*	*kung*
typical	3SG.LINK	**man**	engaging.in.sex	with

kani-kanino	*at*	*naghahanap*	*lagi*	*ng*
anyone	and	seeking	always	GEN

bagong	*jojowain.*
new	partner.to.date

English translation:

He’s a typical **straight** man, always fornicating with everyone and always looking for a new boyfriend.