菲律宾语言模型中的偏见归因：扩展偏见可解释性指标以应用于黏着语

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

关于偏见归因和可解释性的新研究表明，在处理英语文本的语言模型中，标记如何导致偏颇行为。我们通过调整信息论偏见归因分数指标，将其应用于处理黏着语的模型，尤其是菲律宾语模型，以继续这条研究路径。接着，我们通过将其应用于一个纯菲律宾语模型和三个多语言模型来展示我们调整方法的有效性——其中一个模型是训练于全球语言数据上，另两个则是东南亚数据。我么的结果表明，菲律宾语模型往往因与人物、物体和关系有关的词汇而产生偏见——这些实体为主题的词汇对比于英语中的偏见贡献主题（比如犯罪、性和社会行为）更多地偏向于动作。这些发现指出英语和非英语模型在处理与社会人口群体和偏见相关的输入时存在差异。

1 介绍

随着预训练语言模型（PLMs）在规模和能力上的增长，研究它们所表现出的偏见行为也在不断增加 (Gallegos et al., 2024; Gupta et al., 2024) 。尤其是，它们在多语言能力上的改进，与调查多语言和非英语模型公平性如何的研究相匹配（例如，Friðriksdóttir and Einarsson, 2024 ； Fort et al., 2024 ； Üstün et al., 2024 ； Ibaraki et al., 2024 ）。在这些研究中，来自全球各地的自然语言处理研究者最初开发的英文偏见评估工具和方法被调整到多文化背景中，以检测多语言预训练语言模型表现出多少偏见。这些多语言仿制研究在很大程度上证实了处理非英语文本的模型中存在安全和偏见问题。例如，Bergstrand and Gambäck (2024) 发现他们实验的挪威模型在平均 68.27 % 的时间里，更倾向于反酷儿的声明而非支持酷儿的声明。同时，Huang and Xiong (2024) 在测量中文问答模型的偏见时，发现了某些预训练语言模型中性别、家庭职责和职业偏见之间的刻板关联。

然而，多语言偏见研究大多集中在评估和较少程度的缓解（例如，Reusens et al., 2023 ； Lee et al., 2023 ），但并未涉及可解释性和可解释性主题——即探索影响黑箱 PLM (Liu et al., 2024) 偏见决策的内部因素和机制。增加这些不透明模型运作的透明度，并提高我们对其偏见行为根源的理解，是规范其危害性和促进公众接受这些技术的重要步骤 (Xie et al., 2023; Lipton, 2018) 。为此，Gamboa and Lee (2024) 开发了一种可解释性指标，用于解释某些令牌如何助长语言模型中的偏见。到目前为止，该方法仅应用于评估英语偏见测试的 PLM，尚未扩展到处理非英语文本的多语言模型。

在本文中，我们基于他们的工作，使用偏倚归因分数度量来分析在处理菲律宾语（一个缺乏高资源自然语言处理工具的语言）文本的预训练语言模型（PLMs）中，哪些词汇和语义类别引发了性别和性取向偏见的倾向。研究菲律宾语中性别和性取向偏见的模型行为具有三方面的价值。首先，东南亚对人工智能技术的快速采用，其中弱势少数群体可能会受到 PLM 偏见和伤害的不利影响。其次，菲律宾语的黏着形态结构与主要分析性形态结构的英语不同，因此需要对诸如偏倚归因分数计算等依赖于分词的方法进行些许调整。最后一个原因涉及性别、性别酷儿及相关偏见在菲律宾语言和文化中的表现的特殊性，这可能导致菲律宾语模型在处理性别数据时与英语模型的管理方式有所不同。确实，我们的研究发现，当处理英语的模型在涉及犯罪、亲密关系和帮助这些强调行为的主题时表现出偏见行为，而处理菲律宾语的 PLMs 则将其偏见倾向归因于属于更具体主题的词汇——例如，那些指代有形物体和人群的词汇。

我们的贡献有三方面：

- 我们是首个利用并调整可解释性指标来研究在处理非英语文本的多语言模型中，单个标记如何导致偏差行为的研究者。
- 我们调整了偏见归因评分度量的方法，最初该方法仅用于英语，现在用于像菲律宾语这样的黏着语。[1]

[1]代码可在 https://github.com/gamboalance/bias_attribution_filipino 获取

- 我们发现了一些语义类别，这些类别导致菲律宾 PLM 在决策中存在偏见，由此明确了应该谨慎使用这些模型的主题领域，以及应该集中缓解努力的方向。

本文的其余部分首先简要回顾了关于基于 token 的归因和 NLP 中的可解释性的文献（2）。接下来是详细介绍我们的偏见声明（**??**）和我们使用的方法——特别是我们选择的数据集、我们检查的模型以及我们使用的归因度量（3）。本文继续展示我们的分析结果（4），并以我们的结论结束（**??**）。

2 相关工作

在机器学习中，解释性方法分为两类：全局解释方法和局部解释方法 (Guidotti et al., 2018; Lipton, 2018) 。全局解释方法阐明了模型在得到所有可能结果时所使用的完整推理过程 (Guidotti et al., 2018) 。对于大型预训练语言模型（PLMs）来说，全局解释似乎较为罕见，尤其在生成型模型中，因为可能的输入和输出的巨大变化使得很难抽象出一个单一的解释工具、模型或启发式方法来为所有这些可能性提供解释。相反，更常见的是局部解释方法，这些方法逐数据实例进行检查，并量化一个模型的预测或输出在多大程度上可以归因于数据点中每个单独输入特征。在语言模型中，局部解释是通过计算标记属性分数来实现的。这些分数表示每个输入标记对大型预训练语言模型的决策或生成的贡献程度 (Attanasio et al., 2022; Chen et al., 2020) 。

监督学习模型，例如用于检测毒性（例如，Xiang et al., 2021 ），仇恨言论（例如，Risch et al., 2020 ）和厌女症（例如，Attanasio et al., 2022; Godoy and Tommasel, 2021 ）的模型，是令牌归因研究的常见对象，这些研究使用各种数学技术——例如线性近似（例如，Ribeiro et al., 2016 ）和 Shapley 值（例如，Chen et al., 2020 ）——来计算令牌归因分数。Gamboa and Lee (2024) 受到这些以分类为中心的研究的启发，提出了一种词级别的归因分数，该分数不仅解释了偏见行为，也适用于在掩蔽和因果学习范式下训练的语言模型。具体来说，他们的方法借鉴了最初由 Steinborn et al. (2022) 设计的信息论偏见评估指标，并证明计算偏见指标所需的中间值实际上也可以用来解释对 PLM 偏见的词级别贡献。他们在不同的 PLM 上应用他们提出的偏见归因分数方法，使他们能够发现与犯罪、亲密和帮助相关的词语如何促使模型表现出偏见。我们的研究进一步完善了他们的方法，以用于处理像菲律宾语这样粘着语的模型，并比较了英语和菲律宾语模型中导致偏见的令牌。

概念上，我们将 PLM 偏见理解为与或源于包含不同社会人口属性的输入数据相关联的模型性能差异。操作上，我们将偏见定义为违反 Gallegos et al. (2024) 所规定的平等社会群体关联公平条件。一个模型在不同社会群体相关的情境中选择或生成与人口统计无关的词的概率相等，则满足平等社会群体关联。例如，在一个公平模型中，单词“教师”应有相同的概率生成在短语“男孩长大后成为…”和“女孩长大后成为…”中。因此，我们的偏见操作化认为系统性地倾向于将某些中性概念与特定社会群体关联的模型是不公平的。具体而言，我们使用菲律宾 CrowS-Pairs 和从比较词概率中衍生出的偏见指标来量化这一点——关于这些方法，我们将在下一节中更详尽地讨论。

这种对预训练语言模型偏见的概念化和操作化使我们的研究能够阐明处理菲律宾文本的模型在表现上造成的伤害。表现上的伤害源于模型通过生成关于边缘群体的不利描述或将其与负面特质关联而延续刻板印象 (Blodgett et al., 2020; Crawford, 2017) 。模型不断地将中立但刻板的概念与特定的人口群体联系起来，会导致这种伤害。我们的分析侧重于偏见语言模型的部署对菲律宾历史上处于劣势的性别和性取向群体的潜在不利影响——例如，babae（女性）、bakla（非异性恋男性）和 tomboy（非异性恋女性）(Velasco, 2022; Garcia, 1996; Santiago, 1996) 。

3 方法

3.1 数据

偏见评估基准有助于衡量、检查和比较语言模型中的偏见行为。它们也是通过偏见归因分数指标 (Gamboa and Lee, 2024) 对模型偏见进行可解释分析的前提。我们使用菲律宾 CrowS-Pairs 数据集来探测偏见并探索处理菲律宾语的多语言 PLM 中的偏见可解释性。将英语 CrowS-Pairs (Nangia et al., 2020) 基准适应到菲律宾环境，菲律宾 CrowS-Pairs 由 204 对挑战提示组成，这些挑战提示评估两种偏见维度：性别和性取向 (Gamboa and Lee, 2025) 。每对由两个略有不同的陈述组成：一个传递刻板印象或偏见，另一个表达较少偏见的情感。如表格 1 所示，这些句子对仅在一个或几个社会属性词上有变化，这些词的改变会影响句子的偏见程度和意义。

这些基准测试假定，那些反复将有偏见的陈述判断为比无偏见的陈述在语言上更可能的模型，从预训练数据中学习到了刻板印象和偏见。鉴于菲律宾 CrowS-Pairs 的开发充分考虑

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: 菲律宾 CrowS-Pairs 中的 挑战提示对。在每个陈述中进行区分或修改的词汇用粗体强调。提示的语言注释在附录 **??** 中。

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: 研究的模型、它们的特性以及相对于菲律宾 CrowS-Pairs（CP）评估的偏差得分。一个无偏差的模型得分应该是 50.00 。

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

了菲律宾语言和文化的特殊性，可以假设其所得出的偏见评估和指标在上下文和文化上是合适和相关的。

我们分析了四种能够处理菲律宾语的模型的偏差可解释性。我们研究了基于掩码和自回归的 Transformer 模型，这些模型目前在多语言基准测试中表现出最佳性能。我们还查看了在预训练数据中具有不同语言组成的模型：`roberta-tagalog-base` 是用纯菲律宾数据训练的，`sea-lion-3b` 和 `SeaLLMs-v3-7B-Chat` 是用英语和东南亚语言的数据训练的，`gpt2` 则是用全球语言的数据训练的。在这些模型中，`sea-lion-3b` 模型在针对整个菲律宾 CrowS-Pairs 基准测试时被发现偏差最大，而 `roberta-tagalog-base` 在使用菲律宾 CrowS-Pairs 的性取向子集进行评估时被发现为最具恐同倾向的。表格 2 总结了我们分析的模型、它们的属性以及使用菲律宾 CrowS-Pairs 测量的偏差情况。

为了研究单个标记如何导致模型偏差行为，我们使用了由 Gamboa and Lee (2024) 提出的偏差归因分数。这个可解释的度量通过以下方程计算。

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

在这个方程中，偏差归因分数用 $b(u)$ 表示，或者说是 CrowS-Pairs 挑战对中每个未修饰词的偏差。未修饰词是指在一个对中的两个句子中都存在的词——例如在表 1 的第二个例子中，"tipikal"（典型的）、"nakikipagtalik"（淫乱的）和"bagong"（新的）——并与修饰词或使句子不同的属性词区分开来——例如在同个例子中，"lalaki"（直男）和"bakla"（同性恋者）。在概念层面上，$b(u)$ 通过比较一个未修饰词出现在刻板印象语境中的概率（即表 1 中的偏向性陈述）和该词出现在较少刻板印象语境中的概率（即较少偏向性的陈述）来计算词级别的偏差贡献。在 CrowS-Pairs 偏差评估范式中，更可能出现在刻板印象语境中的词直接导致 PLM 更倾向于一个偏向性句子而非较少偏向性的句子，并增加模型的总体偏差分数。

在数学和实用层面上，偏差归因评分方法通过首先获得 $P_{u,more}$ 和 $P_{u,less}$ 来比较在偏向和较少偏向的上下文中的标记概率。u 是未修改的标记，其偏差归因评分正在计算中，而 $P_{u,more}$ 是在更刻板的上下文中标记被掩盖时模型为 `<MASK>` 计算的概率分布。例如，如果我们正在确定表 1 的第二个例子中“fornicating”在英语翻译中的偏差归因评分，$P_{fornicating,more}$ 将对应于模型为该词汇中的每个词分配的概率分布，这些概率与其填充提示句“他是一个典型的同性恋者，总是和每个人 `<MASK>` 并总是寻找新男友”的可能性有关。

相反，$P_{u,less}$ 是模型在较不刻板的语境中对未修改的标记被掩盖时提供的 `<MASK>` 的概率分布。继续前面的例子，$P_{fornicating,less}$ 是枚举模型词汇中的每个单词填充 `<MASK>` 的概率

分布，如“他是一个典型的直男，总是和每个人 <MASK> ，总是寻找新的男朋友。”

鉴于这些分布是在不同的偏见背景下进行的条件约束，因此可以预期它们会给模型词汇分配不同的概率值，包括正在计算偏见归因分数的词。例如，$P_{u,more}$ 可能会给“睡觉”分配概率 0.89 ，而 $P_{u,less}$ 可能会给它分配概率 0.75 ，因为模型将“通奸”与“同性恋”这个词（在更具刻板印象的背景中出现）联系得更紧密，而不是与“异性恋”这个词（在较少具有刻板印象的背景中出现）。

由于这些概率上的差异，一个分布比另一个分布自然地更接近真实情况。在上述例子中，$P_{u,more}$ 更接近真实情况，因为它给正确和相关的标记（fornicating）分配了更高的概率（0.9 ）。这表明，在偏向性更强的条件下，模型更有可能生成 fornicating，而不是在偏向性较弱的条件下。因此，fornicating 也使得模型更有可能生成或选择偏向性更强的陈述，而不是偏向性较弱的陈述。

量化这一贡献的下一步是测量并比较两个分布与真实值 G_u 的距离。真实值由一个单热分布给出，其中相关标记 u 的概率为 1 ，而模型词汇表中每个其他标记的概率为 0 。这些距离通过信息理论中的 Jensen-Shannon 距离（JSD）公式 (Lin, 1991; Endres and Schindelin, 2003) 计算，并相互减去。

一个结果性的偏见归因分数小于 0 表示在更典型的情境下的概率分布的距离（$P_{u,more}$ ）比 $P_{u,less}$ 和 G_u 之间的距离更小且更接近于真实值。因此，负的偏见归因分数可以理解为表明相关的词在一个偏见的情境中更有可能出现，进而使得 PLM 选择或生成更多带有刻板印象的陈述。相反，正的偏见归因分数则表明相反的情况：即该词使模型行为更少偏见并倾向于选择不太刻板的表达。虽然偏见归因分数的符号代表了词对模型偏见的影响方向，其大小则表示这种影响的强度。

对于像英语这样的以分析性为主的语言，上述偏置评分归因方法可以以一种简单的方式实现。在分析性语言中，一个单独的词通常只承载一个或几个概念，很少使用词缀，因此相对于综合型和黏着型语言的词来说，其本质上是相对较短的。这种英语的形态学类型学允许 PLM 分词器将大多数英语单词视为单独的词元。因此，在将偏置归因评分方法应用于英语时，每个词元的 $b(u)$ 得分通常也对应于一个独特单词的得分。

然而，对于像菲律宾语这样的黏着语，解释性方法变得更加复杂，因为一个单词可以包含多个词缀和概念，因而自然地更长。例如，fornicating 在菲律宾语中翻译为 nakikipagtalik。Nakikipagtalik 可以分解或标记为五个语素：na-、ki-、ki-、pag- 和 -talik，其中 talik 是表示亲密的词根，pag- 是表示动作的前缀，而 nakiki- 是一组表示现在进行时和与另一实体共同执行动作的前缀。大致对应于“正在与某人亲密”，因此，当使用 PLM 标记器和上一节描述的偏差归因评分方法时，nakikipagtalik 可以获得其各个子语素的五个不同 $b(u)$ 分数。为了解决这一复杂性，我们对 (Gamboa and Lee, 2024) 提出的方法实施了一个附加步骤：对于被模型标记器进一步拆分为词元的单词，偏差归因评分给出其组成子词分数的平均值，如下所示：

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

其中：

- u 是正在计算其归因分数的完整单词，
- $t_1, t_2, \ldots, t_n$ 是从 u 中分词得到的标记，并且
- $b(t_i)$ 是应用于符号 t_i 的偏倚归因评分函数。

3.2 语义分析

为了检查诱导菲律宾 PLM 中偏见行为的词语的语义类别，首先使用 googletrans 包将菲律宾 CrowS-Pairs 的组成词翻译成英文，然后使用 pymusas 包进行语义标记。pymusas 是一种语义标记工具，可以表征一个词所属的语义领域 (Rayson et al., 2004) 。类似于 Gamboa and Lee (2024) ，我们从分析中去除在数据集中总词数中占比少于% 的词（即，出现次数少于 $n = 10$ 次的词）。在下一节中，我们按比例报告偏见贡献最大的语义类别。

4 结果与讨论

4.1 菲律宾语中的偏见归因

表格 3 和 4 展示了调整后的偏见归因评分方法如何在解释处理菲律宾语言的模型偏见行为中提供可解释的说明。特别是，表格 3 描述了表格 1 的第一个例子中的共享标记如何促使 RoBERTa-Tagalog 选择更偏见的陈述而不是较少偏见的替代选项。在这些标记中，单词 pinagsasabihan（训斥）、laging（经常）和 katulong（帮手）具有负的偏见归因评分，表明这些在这种情况下对模型的偏见行为有贡献。可能是这些标记的组合促使模型决定陈述更可能指的是一个 babaeng katulong（女性帮手）而不是一个 lalaking katulong（男性帮手）。同时，语法标记 ni 和 ang 具有正的偏见归因评分，表

明这些标记促使模型表现出较少的偏见。这些结果意味着，当话题涉及权力动态和关系时—如通过 pinagsasabihan（训斥）和 katulong（帮手）信号化—roberta-tagalog-base 可能存在性别歧视偏见，促使它将从属角色（例如帮手）刻画为女性。

另一方面，表格 4 展示了表格 1 中第二个挑战提示条目中共享标记的偏见归因结果，该条目应用于 sea-lion-3b 。偏见归因得分最为负面的标记是 nakikipagtalik（通奸）。这一得分表明，这个词的存在对模型选择版本的贡献最大，使之将同性恋者与滥交联系在一起，而非选择与直男性为主体的版本。这些样本分析展示了如何使用偏见归因得分的可解释性分析来提高对多语言模型在处理偏见时的操作方式的理解，特别是那些处理菲律宾文本的模型。

表 5 列出了我们研究的模型中，语义字段中对偏见贡献词汇比例最大的十个。每个语义字段有三个比例指标：[a] 在类别中具有负 $b(u)$ 的词的比例，这些词增加了 PLM 偏见（偏见），[b] 在类别中具有正 $b(u)$ 的词的比例，这些词减少了 PLM 偏见（偏见），以及 [c] 获得 $b(u) = 0$ 且对 PLM 偏见没有影响的标记的比例（∘ 偏见）。表 5 中的类别揭示了有几个语义字段在所有或大多数四个 PLM 中引发了偏见行为。

一类是关系类别，其中包含的令牌在四个模型中会导致 50 % 至 60 % 之间的偏差。来自菲律宾 CrowS-Pairs 且属于这一类别的词语有 kaibigan（朋友）、kasintahan（情人）和 kakilala（熟人），暗示模型从其预训练数据中学习到了与菲律宾文化关系相关的性别和性取向偏见。表 1 中的第二个提示对条目是一个例子，其中关系词 nakikipagtalik（通奸）引发了偏颇行为。

指代人（如 doktor 或 doctor，sundalo 或 soldier，以及 katulong 或 helper）和物体（即 singsing 或 ring，pinggan 或 plate，以及 kandila 或 candle）的词语似乎也会导致模型表现出偏见。在 roberta-tagalog-base 和 sea-lion-3b 中，它们的效果尤其明显，这些词语在 45 % 到 80 % 的时间里诱发偏见。表格 3 中的例子展示了这一效果，其中单词 katulong（helper）是促使 roberta-tagalog-base 判断 Reyes 先生总是斥责他的女性帮手的词语之一。（从菲律宾语翻译而来）比起 Reyes 先生总是斥责他的男性帮手，这种语言结构更为合理。

这些为菲律宾模型提供偏倚的分类，因其具体和基于实体的性质，与那些在英语模型中引起偏倚的更为抽象和基于动作的类别形成了鲜明对比。Gamboa and Lee (2024) 发现，犯罪、亲密和亲社会行为（例如，猥亵、强奸、亲吻、关心和养育）驱使英语模型产生偏倚，而对于菲律宾模型，我们发现有形名词（例如，物体和人）对模型偏倚的影响更大。这个见解指出了多语言模型如何处理用不同语言书写的社会人口相关文本时，在社会语言学上的重要差异。

在本文中，我们扩展了一种现有的偏见可解释性方法，以用于处理如菲律宾语等黏着语的模型。我们对偏见归因评分计算方法的调整是基于对菲律宾语（一种黏着语）与英语（一种分析性语言）之间形态差异的深入理解。然后，我们将修订后的方法应用于四个模型，这些模型使用菲律宾语 CrowS-Pairs 进行偏见评估，并展示了该技术在透明化某些标记如何导致黑箱模型做出偏见决策方面的有效性。最后，我们对菲律宾语偏见贡献标记进行了综合分析，特别关注它们所属的语义类别。我们的结果表明，与英语基准和模型中偏见贡献标记的抽象和动作导向特性相反，菲律宾语模型受到指代具体实体（即物体和人物）的词语的影响，从而表现出偏见。我们希望这些发现能为当前调查语言模型中的偏见机制及致力于减少其有害影响（例如，Liu et al., 2024; Ermis et al., 2024; Gupta et al., 2025 ）提供贡献。

尽管偏见归因评分方法的应用扩大了语言范围，但我们的研究仍然仅限于菲律宾语。虽然我们对上述方法的调整可能对类似的黏着语有所帮助，但在应用该方法于其他语言和语言家族时，可能仍需要考虑特定性。因此，这些因素需要在将方法扩展到其他语言的未来工作中加以考虑。

我们在使用 googletrans 软件包将菲律宾语标记进行机器翻译后，再使用 pymusas 对英语改编进行标注，这可能导致了不准确性。然而，由于缺乏菲律宾语语义标注工具，这一方法决定被采取。未来此类工具的开发可能会带来本研究的再现，从而提高文化和语言的准确性。

最后，我们测试过的方法只在少量的模型上应用，这也是我们工作的一个局限性。我们仅评估了四个模型，并未研究诸如 7 亿和 8 亿参数版本的 SEALION 之类的大模型。此外，我们只包括了开源模型，无法涵盖专有的 PLM。

5

致谢 Lance Gamboa 感谢菲律宾政府的科学和技术部为他的博士研究提供资助。

References

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*,

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: 偏差归因分数解释了这些标记如何促使 `roberta-tagalog-base` 选择这段陈述的更刻板版本，而不是偏见较少的版本。

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: 偏差归因分数解释了这些标记如何促使 `sea-lion-3b` 选择该声明更具刻板印象的版本而不是较不偏见的迭代版本。

`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: 对于我们研究的 4 个 PLM，具有最大比例的偏倚贡献标记的语义类别。 偏倚：标记比例为 $b(u) < 0$ ，导致了偏倚行为。○ 偏倚：标记比例为 $b(u) = 0$ ，没有影响模型的偏倚。 偏倚：标记比例为 $b(u) > 0$ ，抑制了偏倚行为。在多个模型中引发偏倚的类别用粗体表示。

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.

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*.

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.

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.

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*