Robust Bias Detection in MLMs and its Application to Human Trait Ratings

Ingroj Shrestha
University of Iowa
ingroj-shrestha
@uiowa.edu

Louis Tay
Purdue University
stay@purdue.edu

Padmini Srinivasan
University of Iowa
padmini-srinivasan
@uiowa.edu

Abstract

There has been significant prior work using templates to study bias against demographic attributes in MLMs. However, these have limitations: they overlook random variability of templates and target concepts analyzed, assume equality amongst templates, and overlook bias quantification. Addressing these, we propose a systematic statistical approach[1] to assess bias in MLMs, using mixed models to account for random effects, pseudo-perplexity weights for sentences derived from templates and quantify bias using statistical effect sizes. Replicating prior studies, we match on bias scores in magnitude and direction with small to medium effect sizes. Next, we explore the novel problem of gender bias in the context of *personality* and *character* traits, across seven MLMs (base and large). We find that MLMs vary; ALBERT is unbiased for binary gender but the most biased for non-binary *neo*, while RoBERTa-large is the most biased for binary gender but shows small to no bias for *neo*. There is some alignment of MLM bias and findings in psychology (human perspective) - in *agreeableness* with RoBERTa-large and *emotional stability* with BERT-large. There is general agreement for the remaining 3 personality dimensions: both sides observe at most small differences across gender. For character traits, human studies on gender bias are limited thus comparisons are not feasible.

1 Introduction

Pre-trained Masked Language Models (MLMs) (e.g., BERT (Devlin et al., 2019)) are valuable but show systematic biases favoring certain demographics. E.g., these indicate men as more likely to be engineers, and women as being more emotional (Gallegos et al., 2024; Lee, 2018; Parikh et al., 2019; Booth et al., 2021b). These biases amplify societal marginalization and discrimination in automated decision-making and diminish trust in AI systems (Solaiman et al., 2023). Our research contributes to the active stream on bias detection in MLMs. Our **first goal** is to propose methods correcting limitations commonly exhibited by MLM bias detection approaches that hinder robust inferences. Our **second goal** is to use our methods to gauge if MLMs are biased across gender in the novel domain of perceptions of character and personality. Such bias would undoubtedly make it risky to use MLMs in socially critical contexts such as hiring and promotion decisions.

1.1 Methodological limitations

MLM bias detection typically begins with sentence templates from which parallel sets of sentences are derived, one for each demographic group considered. In essence, templates specify a universe of sentences used to 'probe' an MLM to gauge how it associates a demographic group and a concept representing a domain such as *employment*; this assessment yields a score. Templates are essential since it is infeasible to consider all possible relevant sentences related to a target concept.

With few exceptions, template selection and the derivation of probe sentences are done manually. Template sets also tend to be inherited from one paper to the next. A key problem is that it is typical to treat all templates used as equal. However, there could be variations across templates that cloud the detection of MLM bias. We handle this problem using a *mixed effects model* where template variations are handled as random effects. Using parallel logic, we also handle variations across different domain words as random effects (e.g., variations across words for different jobs). Again, such variations should not confound bias assessment.

A second problem is that probe sentences derived from the same template can vary greatly. For example, both *She is a considerate person* and *She*

[1] Our code and data are available at https://github.com/IngrojShrestha/robust_mlm_bias_detection_human_trait_ratings

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4857–4873
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

is a concerned person derive from the template *[gendered-word] is a/an [trait-word] person*. They have pseudo-perplexity scores of 1.8 and 13.8, respectively, when evaluated using BERT-large (uncased version). Given this difference treating these as equivalent for bias detection is also risky. It makes sense to weigh sentence bias estimate by commonality (estimated as pseudo-perplexity).

The next problem is one of making statistically robust inferences. The minimal approach is to see if the association score difference across demographic groups is statistically significant or not. However, in some cases, even this is absent, relying only on the raw difference in association scores to assess bias (Limisiewicz and Mareček, 2022; Guo et al., 2022; Kaneko and Bollegala, 2021). We consider it important to go beyond significance and consider *effect size*. Effect size tells us how much of an observed difference in association scores is explained by the demographic variable of interest. There may be a sizable and significant difference, but if the effect size is small, then the bias is also small. This can happen if other factors unaccounted for in the study are responsible for score differences. Our first goal is to advocate for a bias detection methodology that does not have these limitations.

1.2 Character and personality perceptions

MLMs have become deeply entrenched in different societal contexts. In psychology, MLMs are being used to estimate human attributes like personality and character from social media texts (Park et al., 2015; Liou et al., 2023; Pang et al., 2020). In organizational settings, they are used in hiring to infer individual attributes from language in job applications (Thompson et al., 2023), interviews (Hickman et al., 2022), surveys (Speer et al., 2023), video interviews and resumes (Booth et al., 2021a; Gagandeep et al., 2023). Clearly, biases in these MLM applications would jeopardize the integrity of outcomes and perpetuate stereotypes. While several studies in psychology study bias (or at least differences) when humans rate males and females at least on personality traits, MLMs have not been assessed in the same context. Thus, our second goal is to assess MLMs for gender biases in character and personality ratings.

We draw on two major psychological frameworks specifying key human traits. One known as *human virtue* or *character traits* (Peterson and Seligman, 2004) specifies key positive traits of individuals. Another specifies *personality traits*, which are enduring descriptive characteristics of individuals. We use lexical approaches based on adjectives describing people as outlined by John et al. (1988). From lexical studies, there are four key character dimensions (*empathy*, *order*, *resourceful*, *serenity*) (Cawley III et al., 2000) and five personality dimensions (*extroversion*, *agreeableness*, *conscientiousness*, *emotional stability*, *openness*) (Goldberg, 1992). Our second goal is to assess MLMs for gender biases along these nine trait dimensions. In summary:

1. We propose a better bias detection methodology for MLMs achieved with a mixed effect model accommodating fixed and random effects. We also weigh probe sentences and estimate effect size.
2. We assess seven MLMs for gender bias in the novel domain of character and personality traits. Gender bias detection is critical for the societal contexts in which MLMs are used.

We first describe our bias detection method. Then we present results from two replication studies followed by our main results on MLM bias in human trait perception with additional analysis. We then present related works and conclusions ending with limitations and an ethics statement.

2 Methodology

We follow the standard template-based approach to estimate bias in MLMs (Gallegos et al., 2024; Delobelle et al., 2022; Stanczak and Augenstein, 2021). A template is a sentence structure with two variables representing a *demographic* attribute word (A) and a *domain* target word (T), along with other words. Here, the attribute is gender, and the target is human traits (character/personality). Templates are used to derive probe sentences (S_1, S_2,..., S_n). Bias is assessed by analyzing the MLM estimates of the association between attribute and target words in probe sentences.

Measuring association: We follow the approach of Kurita et al., 2019. Briefly, we mask the attribute A in a probe sentence ($S^{(A)}_{\text{masked}}$), provide it as input to the MLM, and obtain the likelihood[2] of the attribute (p_A), i.e., $p_A = p_{\text{MLM}}([\text{MASK}] = A|S^{(A)}_{\text{masked}}; \theta)$, where θ represents the MLM's parameters. However, since the likelihood of predict-

[2] While recognizing that this is actually a pseudo-likelihood (Salazar et al., 2020), we use the common approach of using it as a proxy for likelihood.

ing different attribute values could differ even in the absence of a target, we also compute the 'implicit prior bias' across attribute values. To do this, we mask both attribute and target and then obtain the likelihood of the attribute value ($S^{(A,T)}_{\text{masked}}$). We refer to this as $p_{\text{prior}} = p_{\text{MLM}}([\text{MASK}] = A | S^{(A,T)}_{\text{masked}}; \theta)$. Association score (association$_{\text{score}}$) is $\log\left(\frac{p_A}{p_{\text{prior}}}\right)$. Where the MLM splits attribute word into multiple tokens, we take the product of the likelihood of sub-tokens, as commonly practiced (Shahriar and Barbosa, 2024; Ahn and Oh, 2021).

Masking example:

S_i: "The lady is known for her empathy."

$S^{(A)}_{i,\text{masked}}$: "The [MASK] is known for [MASK] empathy."

$S^{(A,T)}_{i,\text{masked}}$: "The [MASK] is known for [MASK] [MASK]."

Note that we also mask gendered pronouns (e.g., 'her') to prevent leakage of gender information. For such cases, while computing p_A and p_{prior}, we consider the likelihood of the attribute word in the first [MASK] position only (e.g. see dataset description of Bartl et al. (2020) paper).

2.1 Templates

Templates provide the skeletal structure for probe sentences. Clearly, one can only consider a sample of all possible templates (Limisiewicz and Mareček, 2022; Ahn and Oh, 2021; Bartl et al., 2020). The dominant approach has been manual template design (Doughman et al., 2023; Felkner et al., 2023; Mei et al., 2023; Delobelle et al., 2022) with a few exceptions such as Guo et al. (2022); Shin et al. (2020); Liang et al. (2020).

We select templates using a semi-automatic process designed to capture common expressions of human traits using Wikipedia[3] and GPT-4 (see Appendix A.3 for an overview). We have two template types. *Direct* explicitly include the word *personality*, and *Indirect* do not. The idea is to see if the word *personality* guides the model more effectively. E.g., *clean* may then be more easily perceived as representing a personality trait instead of its more common meaning of physical cleanliness.

Our 6 templates (Table 1) align with the common practice of using 2-5 templates (Steed et al., 2022; Limisiewicz and Mareček, 2022; Bartl et al., 2020; Qian et al., 2019). However, unlike previous research, we do not assume that all templates are equal for bias detection.

Attributes and targets: Attributes are 94 pairs of gender-denoting words adapted from Kaneko and Bollegala (2021) and listed in Table 8. Targets are character trait words (Cawley III et al., 2000) listed in Table 9 and personality trait words (Goldberg, 1992) listed in Table 10. Tables are in the Appendix A.1 and A.2. Again, we handle variations across target words as random effects in our model.

Sentence generation from templates: We also pay particular attention to selecting the appropriate article (limited to *a/an*) and determiner (*the*) or pronoun (limited to *my, your, our, their*) to form wholesome sentences. To avoid arbitrariness, we do this by estimating the psuedo-perplexity of a sentence. Using the MLM for which we are conducting bias detection, we select the candidate (e.g., pronoun) with the least perplexity for each gender. When selections differ across genders, e.g., *my father* and *your mother*, we add the alternatives *your father* and *my mother* for balance.

Across templates, the average sentence count ranges from 1,447 to 4,119 for character traits and 1,437 to 1,757 for personality traits, with a small coefficient of variation of 3.5% to 10.8%. Since sentence selection is MLM specific, the numbers can vary within each template and trait dimension. Thus, we provide averages to ensure a consistent overview in Table 13 in Appendix.

2.2 Linear mixed model configuration

In our mixed effect model (Baayen et al., 2008), represented below, *gender* (values: *male* or *female*) is a fixed effect (predictor) and association score (association$_{\text{score}}$) the response variable. Unlike prior research, we account for variability across templates and trait words as random effects represented using standard notation (1|random_effect). These make the statistical estimates more generalizable, which is a critical feature of our methodology. We use the lme4 (Bates et al., 2015) package in R to fit the mixed models. This package incorporates and estimates the influence of both fixed and random effects in a statistically robust manner.

Besides structure, sentences derived from templates can also differ in their popularity. Thus, we weigh each sentence using pseudo-perplexity (Salazar et al., 2020). Specifically, we give higher weights to sentences with lower pseudo-perplexity

[3] Our approach is somewhat analogous to research on bias in auto-regressive language models where Wikipedia sentences are truncated semi-automatically to create prompts (Lucy and Bamman, 2021; Dhamala et al., 2021).

Category	TID	Templates
Indirect	t$_1$	[DET/PRONOUN] [attribute] is [ARTICLE] [target] person. e.g. my father is a cautious person.
	t$_2$	[DET/PRONOUN] [attribute] is [target]. e.g. my mother is cautious.
Direct	t$_3$	[DET/PRONOUN] [attribute] possesses [ARTICLE] [target] personality. e.g. my father possesses a cautious personality.
	t$_4$	[DET/PRONOUN] [attribute] is known for [PRONOUN] [target] personality. e.g. she is known for her cautious personality.
	t$_5$	People admire [DET/PRONOUN] [attribute] because of [PRONOUN] [target] personality. e.g. people admire him because of his cautious personality.
	t$_6$	[DET/PRONOUN] [attribute]'s [target] personality is valued at [PRONOUN] work. e.g. the woman's cautious personality is valued at her work.

Table 1: Templates (TID: template id, attribute: gendered-words, target: character trait words/personality trait words (above examples use character trait words), Determiner (DET): the, PRONOUN: my, your, our, their)

calculated using the same MLMs being analyzed for bias. This weight is introduced during model fitting, adjusting residual variance rather than directly modifying the association score. The overall linear mixed-effect model is as follows:

model$_{\text{lme}}$: association$_{\text{score}}$ $\sim$ gender + (1 $|$ template) + (1 $|$ trait_words)

where, weight = 1 / (sentence pseudo-perplexity)

Pseudo-perplexity is computed by masking one word at a time in the sentence and obtaining the likelihood of the original word. It is the exponential of sum of logs of losses in predicting original words. Following common notation we refer to this as (an estimate of) perplexity.

2.3 Bias assessment

Bias score: This score is given by the coefficient of the *gender* variable in the model. It represents the difference in association scores between genders across targets and templates. A positive (negative) bias score refers to bias against females (males).

We make robust conclusions as follows. First, we test significance of bias score (95%, using Welch's t-test (Welch, 1947)). While somewhat common practice (e.g., Köksal et al. (2023); Bartl et al. (2020)), some bias papers do not test significance (e.g., Guo et al. (2022); Kaneko and Bollegala (2022); Ahn and Oh (2021)). We also consider effect size, a step rarely taken in the bias literature (e.g., while Dayanık et al. (2022); Bartl et al. (2020); Kurita et al. (2019) measure effect size, Kim et al. (2023); Guo et al. (2022); Kaneko and Bollegala (2022); Limisiewicz and Mareček (2022) do not). Effect size measures the magnitude of differences *while accounting for variability within each gender group*. In contrast raw score differences disregard within-group variability. Thus, we consider bias score, its significance *and* effect size.

Effect size measurement: We choose R^2 over Pearson correlation as our measure because R^2 accounts for relationships involving random effects, unlike Pearson correlation *r*. Since our main goal is to analyze bias across gender, we focus on R^2 for the *gender* attribute only. Assuming significance at 95%, the higher the R^2 the more important gender is in explaining differences in association scores. We follow R^2 interpretation guidelines provided by Cohen (1988): very small: [0, 0.01), small: [0.01, 0.09), medium: [0.09, 0.25), large: [0.25, 0.64), very large: [0.64, 1.0]. To understand the relative magnitude of small effect size, i.e., whether it is closer to medium or very small, we further break down it into three groups ▿: [0.01, 0.03), ▵: [0.03, 0.06), and ▴: [0.06, 0.09). We annotate medium to very large effect as $*$.

R^2 confidence intervals (CI): We conduct 1000 parametric bootstrap iterations. In each, we sample with replacement to create a new dataset of the same size. The resulting 1000 R^2 values are used to estimate the confidence intervals using the partR2 library (Stoffel et al., 2021).

Determination of Bias: As is standard, bias scores that are not significant indicate *neutral* or *unbiased* stance. In addition, we consider significant bias scores but with effect size, $R^2 < 0.01$, as *unbiased*.

2.4 MLMs assessed for bias

We analyze four pre-trained MLMs, both base and large, except for distilbert: bert-base-uncased (bert-large-uncased) (Devlin et al., 2019), roberta-base (roberta-large) (Liu et al., 2019), albert-base-v2 (albert-large-v2) (Lan et al., 2020), distilbert-base-uncased (Sanh et al., 2019). All models used are the uncased versions[4]. Note that both ALBERT models are uncased by default. We provide input text in lower case for all models for consistency. We implement MLMs using Hugging Face on NVIDIA Tesla P100 PCIE (16GB) GPU. Each MLM model took about 3 hours on average per trait.

3 Results

3.1 Prior work replication results

First, we replicate our methods on two prior studies (Bartl et al., 2020; Limisiewicz and Mareček, 2022), that share key methodological features, namely the use of templates, the same association scores to measure bias (Section 2), and focus on gender bias. We explore their targets: *profession* in Bartl et al. (2020) and *profession/non-profession* indicated by *Nouns* in Limisiewicz and Mareček (2022)[5]. We aim to see if we obtain comparable results despite using our analytic methods.

3.1.1 Bartl et al. (2020)

Overview: The authors study gender bias w.r.t. profession for both English and German. We focus on their results for English. Consistent with their work, we use bert-base-uncased MLM and their templates (their Table 1 *English*). They consider 3 categories with 20 professions in each: *Balanced*, *Female* and *Male*. *Female* (*Male*) refers to professions where females (males) dominate in the real world while in *Balanced* professions both have roughly equal participation, decided using US Department of Labor (2020). Consistent with their work, we compute averages of association scores (Section 2) for each gender across all templates (they call this *Pre*) and then take bias score as male minus female average (Table 4 in their paper), i.e., *m-f*. In our mixed model approach (Section 2.2), professional words from their paper substitute for trait words. While they consider the significance of bias scores and effect size, it is unclear whether the reported effect size relates to Pre-association or Post-association (after fine-tuning). So, we exclude their effect size in our comparison.

Results and Analysis: Replication results are in the first 6 rows of Table 2. Our bias scores are significant throughout. Our model, model$_{\text{lme}}$, scores are close to prior results in magnitude and the same in direction. As before for *Female* (*Male*) professions bias is high and against males (females) while for the *Balanced* professions bias scores of around 0.41 are lowest but still favoring males.

	Method	**Bias score**	**Effect Size** (R^2)
Balanced	Prior work	0.40	
	model$_{\text{lme}}$	0.41	0.13*
Female	Prior work	-1.18	
	model$_{\text{lme}}$	-0.83	0.24*
Male	Prior work	0.99	
	model$_{\text{lme}}$	1.00	0.50*
104	Prior work	0.35	
Nouns	model$_{\text{lme}}$	0.40	0.06$^▲$

Table 2: Prior work replication results. The first six rows refer to Bartl et al. (2020). Last two rows refer to Limisiewicz and Mareček (2022). All bias scores and effect sizes (R^2) are significant at 95% confidence level. Please see Table 3 legend for notation.

While our bias scores and direction match those of Bartl et al. (2020), we provide additional meaningful analysis based on effect sizes. Our model yields larger effect sizes – medium to large : 0.13 (*Balanced*), 0.24 (*Female*) and 0.5 (*Male*). E.g., a sizable 13% of score variations in *Balanced* are explained by gender differences.

3.1.2 Limisiewicz and Mareček (2022)

Overview: We focus on their investigation of gender bias in the context of 104 gender neutral professional and non-professional *Nouns* (e.g., professional: ‘chef’, ‘programmer’, ‘painter’; non professional: ‘victim’, ‘customer’, ‘patient’). We use their templates (their Table 1), bert-base-cased MLM, and their masking strategy involving determiners, pronouns and the nouns of interest and compute bias score as in their paper. We replace trait words with their *Noun* words in our models. Notably, the authors do not consider bias score significance, effect size, or control for random effects.

Results and Analysis: As seen in the last two rows of Table 2 our bias score is close to theirs (0.352, see column MEAN, row 1 in their Table 3) in magnitude and matches direction favoring males. However, while it is statistically significant (95% confidence), our effect size is small, 0.06. Thus, we conclude that the bias found in Limisiewicz and Mareček (2022) is small. This underlines the importance of considering random effects and of incorporating sentence weights.

Summary: In both replications, our bias scores match the original findings in magnitude and direc-

[4] All models except RoBERTa use WordPiece tokenization; thus we use the uncased version. RoBERTa uses Byte-Pair Encoding, supporting both cased and uncased text inherently.

[5] A third work by Kurita et al. (2019), also shares the same methodological features. However, we do not replicate it given its significantly small dataset size.

Traits		Language Models						
		Base				Large		
		BERT	RoBERTa	ALBERT	DistilBERT	BERT	RoBERTa	ALBERT
Character traits	*empathy*	-0.36$^{\triangledown}$	-0.19	-0.19	-0.37$^{\vartriangle}$	0.99$^{\vartriangle}$	-0.70$^{\blacktriangle}$	-0.30
	order	-0.08	0.12	-0.06	-0.07	0.69$^{\triangledown}$	-0.30$^{\triangledown}$	-0.41
	resourceful	-0.30	-0.20	-0.15	-0.22$^{\triangledown}$	0.86$^{\vartriangle}$	-0.74$^{\vartriangle}$	-0.24
	serenity	-0.33	-0.43$^{\triangledown}$	-0.47	-0.36$^{\vartriangle}$	0.47$^{\triangledown}$	-1.08*	-0.35
	Effect Size (R^2)	[1E-3, 1.03E-2]	[1E-3, 1E-2]	[0, 2E-3]	[1E-3, 3.7E-2]	[1E-2, 3.5E-2]	[1E-2, 0.127]	[1E-3, 3E-3]
Personality traits	*extroversion*	-0.28	-0.39$^{\triangledown}$	-0.26	-0.25$^{\triangledown}$	0.71$^{\triangledown}$	-0.86*	-0.38
	agreeableness	-0.16	-0.21	-0.16	-0.05	0.61$^{\triangledown}$	-0.77$^{\blacktriangle}$	-0.36
	conscientiousness	-0.20	0.54$^{\triangledown}$	0.05	-0.23$^{\triangledown}$	0.64$^{\triangledown}$	-0.77$^{\blacktriangle}$	-0.23
	emotional stability	-0.18	-0.08	-0.21	-0.21$^{\triangledown}$	1.01$^{\vartriangle}$	-0.26	-0.56
	openness	-0.27	0.15	-0.20	-0.16	0.35	-0.61$^{\vartriangle}$	0.12
	Effect Size (R^2)	[2E-3, 4E-3]	[0, 2E-2]	[0, 1E-3]	[1E-3, 2E-2]	[9E-3, 4.1E-2]	[5E-3, 0.104]	[1E-3, 4E-3]

Table 3: Results for Seven MLMs using model$_{\text{lme}}$. Values presented in each cell (e.g., -0.36 for *empathy*) represent bias scores. Rows labeled 'Effect Size (R^2)' presents a range of effect sizes across traits for each model. Notably, symbols next to bias scores indicated where each trait falls within effect size range. Notation: Black font: significant (*p-value* < 0.05), blue: marginally significant (*p-value* $\in [0.05, 0.10]$), red: not significant (*p-value* > 0.10). Positive (negative) score: bias against females (males). Effect size (R^2). *: Medium [0.09, 0.25) to very large [0.64, 1]; Small: $\triangledown$:$R^2 \in$[0.01, 0.03) $\vartriangle$: $R^2 \in$[0.03, 0.06), $\blacktriangle$: $R^2 \in$[0.06, 0.09).

tion. But while we find a medium to large effect bias for all three professions in Bartl et al. (2020), the effect size is relatively small (0.13) for the balanced category – likely the most important group in their study. Our Limisiewicz and Mareček (2022) replication indicates small bias, an inference possible because of effect size analysis. By accounting for random effects and sentence pseudo-perplexity and examining effect size, we offer more robust and quantitative estimates of bias than in prior work.

3.2 MLMs and human traits: bias results

3.2.1 Bias across MLMs (binary gender)

Table 3 presents our model$_{\text{lme}}$ results.

(1) Base MLMs: Most scores (29/36) are significant, with 2 more being marginally significant. The range of significant scores (ignoring direction) is [0.07, 0.47] for character and [0.15, 0.54] for personality. To the best of our knowledge, bias scores in the literature are in [0.16, 5.6] (Limisiewicz and Mareček, 2022; Ahn and Oh, 2021; Bartl et al., 2020) thus ours are at the lower end.

Effect sizes are 'at most small' for all base models. DistilBERT exhibits these for 6/9 dimensions; the largest ranging in [0.030, 0.037] are for *empathy* and *serenity*. RoBERTa does the same for 3/9 dimensions. Interestingly, ALBERT stands out as unbiased across all dimensions followed by BERT (its one small effect size, for *empathy*, is actually close to negligible (0.01)). Overall, effect sizes indicate close to no gender bias for both trait sets; those observed almost exclusively favor females.

(2) Large MLMs: Scores [0.23, 1.08] are higher in magnitude compared to base models. Interestingly, RoBERTa favors females, while BERT favors males, perhaps due to differences in training goal and corpus and requires further exploration beyond the scope of this paper.

ALBERT-large is unbiased. RoBERTa is the most biased with two medium effect sizes (*serenity* and *extroversion*). BERT's bias is intermediate; but effect sizes are at best small. Ranking models by parameters (least to most): ALBERT-base (12M) → ALBERT-large (18M) → DistilBERT (66M) → BERT-base (110M) → RoBERTa-base (125M) → BERT-large (340M) → RoBERTa-large (355M) matches model ranking from least to most biased except for a flip between BERT-base and DistilBERT (but the former is almost completely unbiased, the latter only slightly biased). Possibly the larger architectures capture more complex patterns in training data. We cannot postulate a causal relation between model size and bias. This requires evidence from nuanced, controlled and focused experiments, also beyond this paper's scope.

Same family MLMs also differ in bias. Thus, each model considered for applications should be examined for bias. Across architectures ALBERT is the only one unbiased. Each trait dimension is vulnerable in at least one large MLM; *order*, *emotional stability* and *openness* are least impacted.

3.2.2 Bias: Human vs MLM perspective

Although character traits were proposed in early 2000, there has been little follow-up work in psychology focusing on gender differences using the same lexical framework. Thus, we limit our analysis to personality, where psychology studies on gen-

Traits	**Language Models**																				
	Base												**Large**								
	BERT			**RoBERTa**			**ALBERT**			**DistilBERT**			**BERT**			**RoBERTa**			**ALBERT**		
	M-F	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**	**M-F**	**M-N**	**F-N**
EMP	0.36$^{\triangledown}$	0.71$^{\triangledown}$	0.86	-0.19	1.55$^{\triangledown}$	1.76$^{\triangledown}$	-0.19	8.11*	8.36 *	-0.37$^{\vartriangle}$	0.46$^{\triangledown}$	0.81$^{\triangledown}$	0.99$^{\vartriangle}$	1.46$^{\triangledown}$	0.51	-0.70$^{\blacktriangle}$	0.63$^{\triangledown}$	1.21$^{\triangledown}$	-0.30	3.75$^{\vartriangle}$	3.90 $^{\vartriangle}$
ORD	-0.08	0.89$^{\triangledown}$	0.79	0.12	1.00$^{\triangledown}$	0.94$^{\triangledown}$	-0.06	8.18*	8.31*	-0.07	0.71$^{\vartriangle}$	0.74$^{\triangledown}$	0.69$^{\triangledown}$	0.80$^{\triangledown}$	0.15	-0.30$^{\triangledown}$	0.38	0.61	-0.41	4.12$^{\vartriangle}$	4.34$^{\vartriangle}$
RES	-0.30	0.76$^{\triangledown}$	0.82	-0.20	0.78	0.84	-0.15	7.80*	7.99$^{\blacktriangle}$	-0.22$^{\triangledown}$	0.47$^{\triangledown}$	0.62$^{\triangledown}$	0.86$^{\vartriangle}$	0.38	-0.50	-0.74$^{\vartriangle}$	0.44	1.00$^{\triangledown}$	-0.24	4.00$^{\vartriangle}$	4.09$^{\vartriangle}$
SRN	-0.33	1.31$^{\triangledown}$	1.51$^{\triangledown}$	-0.43$^{\triangledown}$	1.70$^{\vartriangle}$	2.00$^{\vartriangle}$	-0.47	8.03*	8.70*	-0.36$^{\vartriangle}$	0.14	0.43	0.47$^{\triangledown}$	1.38$^{\triangledown}$	0.88	-1.08*	1.51$^{\blacktriangle}$	2.53*	-0.35	3.49$^{\vartriangle}$	3.71$^{\vartriangle}$
EXT	-0.28	0.26	0.40	-0.39$^{\triangledown}$	0.65	1.04$^{\triangledown}$	-0.26	8.26*	8.64*	-0.25$^{\triangledown}$	0.45$^{\triangledown}$	0.68$^{\triangledown}$	0.71$^{\triangledown}$	0.13	-0.38	-0.86*	-0.02	0.81$^{\triangledown}$	-0.38	3.86$^{\vartriangle}$	4.05$^{\vartriangle}$
AGRE	-0.16	1.44*	1.35$^{\vartriangle}$	-0.21	0.58	0.66	-0.16	7.40$^{\blacktriangle}$	7.64$^{\blacktriangle}$	-0.05	0.57$^{\triangledown}$	0.57$^{\triangledown}$	0.61$^{\triangledown}$	0.95$^{\triangledown}$	0.35	-0.77$^{\blacktriangle}$	1.09$^{\vartriangle}$	1.69$^{\vartriangle}$	-0.36	3.11$^{\vartriangle}$	3.40$^{\triangledown}$
CON	-0.20	0.74$^{\triangledown}$	0.67	0.54$^{\triangledown}$	1.33$^{\triangledown}$	0.63	0.05	8.52*	8.43*	-0.23$^{\triangledown}$	0.81$^{\vartriangle}$	1.00$^{\vartriangle}$	0.64$^{\triangledown}$	0.82 $^{\triangledown}$	0.20	-0.77$^{\blacktriangle}$	0.77	1.42$^{\triangledown}$	-0.23	3.91$^{\vartriangle}$	4.18$^{\vartriangle}$
EMS	-0.18	0.61	0.61	-0.08	0.75	0.69	-0.21	7.59*	7.93*	-0.21$^{\triangledown}$	0.21	0.36	1.01$^{\vartriangle}$	0.29	-0.57	-0.26	0.56	0.63	-0.56	2.50$^{\triangledown}$	2.98$^{\triangledown}$
OPN	-0.27	0.63	0.69	0.15	0.41	0.25	-0.20	7.42$^{\blacktriangle}$	7.73$^{\blacktriangle}$	-0.16	0.21	0.35	0.35	0.38	-0.14	-0.61$^{\vartriangle}$	-0.19	0.38	0.12	3.94*	3.56*

Table 4: Results for Seven MLMs using model$_{\text{lme}}$ (**including non-binary *neo* pronoun**). See Table 3 for reported values descriptions and the notation used. M: Males, F: Females, N: Neo-pronouns. M-F result is identical to Table 3. *EMP*: *empathy*, *ORD*: *order*, *RES*: *resourceful*, *SRN*: *serenity*, *EXT*: *extroversion*, *AGR*: *agreeableness*, *CON*: *conscientiousness*, *EMS*: *emotional stability*, *OPN*: *openness*

der differences[6] are available using self-reported questionnaires and Big Five traits.

Hartmann and Ertl (2023); Ock et al. (2020); Russo and Stol (2020); Weisberg et al. (2011); Lippa (2010); Chapman et al. (2007) find that females rate higher on *agreeableness* and the negative trait of *neuroticism*. (Note that *neuroticism* is the opposite of *emotional stability* included in our study.) Our MLM results for RoBERTa-large align on *agreeableness*. However, with BERT-large, the bias direction for this dimension is opposite, favoring males. On *emotional stability* MLMs are largely neutral excepting BERT-large which also favors males as in psychology. But the MLM bias is only small while it is medium sized in psychology. In the three remaining dimensions, several MLMs exhibit small gender bias. This is in general agreement with psychology, which also finds small to little difference. Overall, MLMs vary considerably in bias score, significance and direction, whereas differences in psychological studies are small.

3.2.3 Bias across MLMs (non-binary gender)

While most bias studies in the literature focus on binary genders, there is growing interest in non-binary gender bias (Urchs et al., 2024; Ovalle et al., 2023; Nozza et al., 2022; Dev et al., 2021). In line with this, we extend our analysis with the mixed effect model to non-binary gender bias by including neo-pronouns as attribute words. These are from Hossain et al. (2023) and include *co*, *vi*, *xe*, *cy*, and *ze*. We analyze pairwise gender bias (Table 4).

Bias considering male and female genders alone remains consistent across models compared to our previous Table 3 results. For neo-pronouns, we consistently find mostly small or no bias across MLMs when comparing neo-pronouns to male and and to female genders. We observe a notable exception with ALBERT models: while we do not find gender bias between males and females, we observe small to medium bias against *neo* compared with males/females. Specifically in ALBERT-large, with the exception of *openness* where we find medium bias, there are only small biases against *neo*. In ALBERT-base, we find small to medium-sized biases against *neo*. While larger MLMs exhibit more bias for binary gender (Section 3.2.1), the opposite is the case for non-binary *neo*.

The MLMs assessed are trained on datasets up to 2019 from sources like Common Crawl, BookCorpus, and Wikipedia. These likely lack adequate representation of non-binary gender patterns (Nozza et al., 2022). Mille et al. (2024) also highlight the underrepresentation of the non-binary groups in Wikipedia. This likely contributes to the bias against *neo*. We recommend carefully controlled experiments for more thorough understanding of the issue.

3.2.4 Additional analyses

We limit analysis to RoBERTa-large (our most biased model for binary gender), except for the analysis of the influence of selected gendered words and the influence of templates, where we analyzed BERT-large (intermediate bias amongst large models).

3.2.4.1 Effect of negative traits

The main experiments are limited to positive trait words. Here, we explore the effect of adding negative traits. We identify a suitable antonym (Appendix Table 11) for each positive character trait (Appendix Table 9) using WordHippo or Merriam-

[6] Note that in computer science inequality in ratings is viewed as bias whereas in psychology differences are observed but not necessarily viewed as bias.

Webster, followed by manual verification. For personality traits, the antonyms (Table 12) are adapted from Goldberg (1992).

	Traits	Positive traits	Positive and Negative traits
Character traits	*empathy*	-0.52 $^{\triangledown}$	-0.08
	order	-0.23	0.10
	resourceful	-0.75 $^{\triangledown}$	-0.40 $^{\triangledown}$
	serenity	-1.00 *	-0.08
Personality traits	*extroversion*	-0.76 $^{\blacktriangle}$	-0.07
	agreeableness	-0.70 $^{\vartriangle}$	-0.01
	conscientiousness	-0.70 $^{\vartriangle}$	-0.24
	emotional stability	-0.17	0.19
	openness	-0.54 $^{\triangledown}$	-0.07

Table 5: Results for RoBERTa-large using t_1 to t_4 templates using model$_{\text{lme}}$. See Table 3 for reported values descriptions and the notation used.

We exclude templates t_5 and t_6 (Table 1) as they are specific to positive traits. To have a consistent interpretation between positive and negative traits in our combined model, we reverse the association for negative traits by multiplying the MLM-derived association$_{\text{score}}$ by -1. Table 5 presents results.

We find small to medium bias favoring females for positive traits, except in *order* and *emotional stability*. When negative traits are included, bias practically disappears except for the *resourceful* trait. This is likely due to the greater prevalence of negative trait sentences (43% to 91% lower perplexity than positive trait sentences in our data) for the bias free dimensions. Consistent with our hypothesis we find that in *resourceful* dimension, where positive trait sentences are more common (53% lower perplexity than negative trait sentences), bias persists albeit with a reduced score. Key to note is that RoBERTa's training corpus is 50% news data from Common Crawl (CC-News), where negative content is more prevalent (Hamborg et al., 2021), which may explain the generally lower perplexity of negative trait sentences compared to positive ones.

	Traits	Fullset	Subset
Character traits	*empathy*	0.99$^{\vartriangle}$	0.39$^{\vartriangle}$
	order	0.69$^{\triangledown}$	0.42$^{\blacktriangle}$
	resourceful	0.86$^{\vartriangle}$	0.39$^{\blacktriangle}$
	serenity	0.47$^{\triangledown}$	0.10
	Effect Size (R^2)	[1E-2,4E-2]	[3E-3,7E-2]
Personality traits	*extroversion*	0.71$^{\triangledown}$	0.11
	agreeableness	0.61$^{\triangledown}$	0.21$^{\triangledown}$
	conscientiousness	0.64$^{\triangledown}$	0.35$^{\vartriangle}$
	emotional stability	1.01$^{\vartriangle}$	0.28$^{\vartriangle}$
	openness	0.35	0.31$^{\vartriangle}$
	Effect Size (R^2)	[9E-3,4.1E-2]	[5E-3,4.9E-2]

Table 6: Full set versus subset of gendered pairs (BERT-large model$_{\text{lme}}$). See Table 3 legend for notation.

3.2.4.2 Influence of selected gendered words

Table 6 compares results using the original *full set* of 94 gendered pairs (used in Table 3) with a *subset* of 7 common gendered words (e.g., daughter-son, girl-boy) (Limisiewicz and Mareček, 2022; Steed et al., 2022; Kurita et al., 2019). Scores fall with the reduced set; a couple cells are no longer significant. Effect sizes stay the same or increase slightly, but all are still at most small. Overall, bias detection is slightly sensitive to gendered words used.

3.2.4.3 Influence of templates

Except for Liu et al. (2021), prior research has largely ignored the impact of templates on bias estimation. Table 7 compares our *direct* and *indirect* templates with BERT-large and also lists their combination ('ALL', same as Table 3 column BERT).

	Traits	Templates		
		Indirect	Direct	ALL
Character traits	*empathy*	2.49$^{\blacktriangle}$	0.07	0.99$^{\vartriangle}$
	order	1.58$^{\vartriangle}$	0.14	0.69$^{\triangledown}$
	resourceful	2.06$^{\blacktriangle}$	0.17	0.86$^{\vartriangle}$
	serenity	1.95$^{\vartriangle}$	-0.32	0.47$^{\triangledown}$
	Effect Size (R^2)	[5E-2,7E-2]	[4E-4,8E-3]	[1E-2,4E-2]
Personality traits	*extroversion*	1.60$^{\vartriangle}$	0.33	0.71$^{\triangledown}$
	agreeableness	1.74$^{\vartriangle}$	0.01	0.61$^{\triangledown}$
	conscientiousness	1.05$^{\triangledown}$	0.32	0.64$^{\triangledown}$
	emotional stability	2.50*	0.28	1.01$^{\vartriangle}$
	openness	1.69$^{\vartriangle}$	-0.22	0.35
	Effect Size (R^2)	[2E-2,1E-1]	[2E-5,1E-2]	[1E-2,5E-2]

Table 7: Comparing templates (BERT-large model$_{\text{lme}}$): see Table 1 for template types and Table 3 for reported values descriptions and the notation used.

Direct templates do not detect bias. Indirect and ALL detect bias, but indirect scores are 1.6-4.8 times higher, suggesting that Direct reduces the capacity of ALL to detect bias. Indirect effect sizes are also larger, with even a medium size effect. Template choice strongly affects bias detection.

3.2.4.4 Influence of pseudo-perplexity

We focus on the *conscientiousness* dimension for RoBERTa-large - the most biased MLM. Pseudo-perplexity of 83% of our 9,066 probe sentences are in [0, 100]. Partitioning these into 20 bins: [0, 5), [5, 10), ..., we find that lower pseudo-perplexity bins have higher sentence density (Figure 1 in Appendix A.4). Bin specific analysis shows significant bias for the first four bins, while this is rare for bins with pseudo-perplexity > 20. Thus, gender bias is visible mainly when MLMs are probed with the most common sentence expressions of traits.

3.2.4.5 Proof-of-concept experiments

We present several proof-of-concept experiments exploring the broader applicability of our methods. Where MLMs are involved, we analyze RoBERTa-large, our most biased model (for binary genders).

Extending our approach to Llama3, an autoregressive generative model: While we focus on MLMs, we present proof-of-concepts for extension of our approach to auto-regressive generative model. We evaluate gender bias, including non-binary gender (*neo* pronouns from Section 3.2.3) in LLama3.1-8B. Method and discussion of results are in Appendix Section A.5.1. We find no significant bias between males and females. However, for all traits, we find medium to large bias against *neo*.

Applying our mixed model to non template bias detection datasets: We present a proof-of-concept for the application of our mixed model to crowd-sourced datasets (non template based). Method and results are detailed in Appendix Section A.5.2. Running our mixed model with the crowd sourced CrowS-Pair sentence set (Nangia et al., 2020), we find — in contrast to their results — no bias in RoBERTa for their 9 bias categories. Unlike us, they do not assess significance or effect size.

Bias mitigation in MLMs: Our focus is on bias detection. For completeness of bias detection and mitigation pipeline, we present a proof-of-concept experiment for bias mitigation in RoBERTa-large, our most biased model. Method and discussion of results are in Appendix Section A.5.3. A standard approach Bartl et al. (2020) successfully mitigates our detected bias.

In future work, we will run more complete tests of our methods in line with these proof-of-concept experiments.

4 Related works

Gender bias studies in MLMs: Profession (Limisiewicz and Mareček, 2022; Bartl et al., 2020) and behavioral concepts (e.g., intelligent) (Guo et al., 2022; Ahn and Oh, 2021) are frequently explored targets in bias studies. Less explored targets include physical appearances (e.g., beautiful) (Kaneko and Bollegala, 2022; Nadeem et al., 2021), stimuli (e.g., career/family), and emotion (Bartl et al., 2020). Exploration of language model biases in human trait perceptions is novel. An exception is Rao et al. (2023) exploring bias in personality perceptions in GPT-4 – they find low gender bias. However, they do not compare with self-reported traits in psychology. The study of biases in *personality* and *character* perceptions is novel.

Gender differences in traits from psychology: Big Five personality traits (Goldberg, 1992), have been studied extensively (Hartmann and Ertl, 2023; Ock et al., 2020; Russo and Stol, 2020; Kuśnierz et al., 2020; Lippa, 2010) using methods of self-reported questionnaires and aggregated through meta-analyses (Lippa, 2010; Feingold, 1994). These, in general, show that females score higher in *agreeableness* and *neuroticism* with small to little difference in the other personality traits. Research on gender differences in character traits from a lexical approach is still lacking despite the framework being proposed in the early 2000s.

MLM bias detection with templates: Prior works average bias scores across templates (Limisiewicz and Mareček, 2022; Bartl et al., 2020), without considering template variability. Some rely solely on magnitude of score differences while others (Steed et al., 2022; Bartl et al., 2020; Kurita et al., 2019) use significance test. Bias quantification using effect size, common in psychology is overlooked.

5 Conclusions

We demonstrate the strength of our proposed mixed model bias detection approach in two replication studies: bias scores match, but our conclusions are stronger. Using our method to assess MLMs for gender bias in trait ratings, we find that larger MLMs tend to show greater bias for binary gender (RoBERTa-large is the most biased), while the opposite for non-binary *neo* (ALBERT-base is the most biased). But almost always any bias detected is small. While choice of target words has little influence choice of template is important. Congruence with observations from psychology in Big 5 traits depends on model and trait. Since MLMs differ in bias it is important to assess them carefully before deploying them in applications critical to society. Within the limits of our research, ALBERT is unbiased for binary gender. However, when considering non-binary gender, no model can be deemed entirely safe.

6 Limitations

Our bias analysis with 3 gender categories, including neo-pronouns (Section 3.2.3), is limited because of challenges faced in the field. In particular, MLM training sets may not adequately represent neo-pronouns. We leave the exploration of bias in other non-binary gender identities for future work. Another limitation is that we do not account for variables such as age and profession, which could influence character/personality ratings, as we focus on a single attribute-target pair. This is left for future work.

Our main study is limited to positive human traits, such as *calm*, and *confident*. As an additional analysis, we include in Section 3.2.4.1 experiments to show that our approach can handle negative traits.

Template design can be quite subjective. We strengthen template quality and representativeness by generating these from a large dataset of sentences. We safeguard quality by favoring popular sentences and sentences of limited size and we constrain these to the present tense. The intent is to favor sentences that are commonly acceptable expressions of human traits with references to the present. As an additional safeguard, we incorporate templates as a random effect in our model. In contrast, the field generally treats all templates as equal for detecting bias. Ensuring template quality has not been emphasized in prior bias studies in MLMs.

Additionally, we focus on template-based bias detection. However, to demonstrate that our method is not limited to this, we present proof-of-concept experiments with crowdsourced CrowS-Pair (Nangia et al., 2020) dataset (appendix A.5.2). We show that our analysis methods can be applied to such approaches.

In order to fully understand gender bias in human traits exhibited by computational models, it is necessary to explore both types of large language models—MLMs and ALMs (e.g., GPT-4 (Achiam et al., 2023) and Llama3 (Dubey et al., 2024)). While our main experiments focus on MLMs, in appendix A.5.1, as a proof-of-concept, we demonstrate the application of our method to Llama3.1-8B ALM.

Finally, our goal is limited to proposing a robust approach for identifying biases in MLMs. We do not mitigate these biases - many papers in the field have this focus. For the reader interested in the complete pipeline we include in Appendix A.5.3 experiments showing successful mitigation of bias using a standard approach in Bartl et al. (2020).

7 Ethical Considerations

Our work proposes a robust approach to bias detection in pre-trained MLMs in the context of human traits. We aim to promote awareness of these biases before using these models, a crucial precursor step for the ethical use of MLMs.

More generally, the intersection of our work with gender differences in human traits in psychology raises the question of whether gender differences in pre-trained MLMs reflect bias or whether they reflect observations of differences between genders. This question is pertinent to the MLM bias detection field as a whole and will require, as a start, in depth meta-analysis in both fields. This larger question is also relevant to the deployment of large language models in applications impacting society.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. 2021a. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 268–277.

Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K

D'Mello. 2021b. Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews. *IEEE Signal Processing Magazine*, 38(6):84–95.

Michael J Cawley III, James E Martin, and John A Johnson. 2000. A virtues approach to personality. *Personality and individual differences*, 28(5):997–1013.

Benjamin P Chapman, Paul R Duberstein, Silvia Sörensen, and Jeffrey M Lyness. 2007. Gender differences in five factor model personality traits in an elderly cohort. *Personality and individual differences*, 43(6):1594–1603.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Erenay Dayanık, Ngoc Thang Vu, and Sebastian Padó. 2022. Analysis of bias in nlp models with regression and effect sizes. In *Northern European Journal of Language Technology, Volume 8*.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Jad Doughman, Shady Shehata, and Fakhri Karray. 2023. Fairgauge: A modularized evaluation of bias in masked language models. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 131–135.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alan Feingold. 1994. Gender differences in personality: a meta-analysis. *Psychological bulletin*, 116(3):429.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Gagandeep, Jaskirat Kaur, Sanket Mathur, Sukhpreet Kaur, Anand Nayyar, Simar Preet Singh, and Sandeep Mathur. 2023. Evaluating and mitigating gender bias in machine learning based resume filtering. *Multimedia Tools and Applications*, pages 1–21.

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*, pages 1–79.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards target-dependent sentiment classification in news articles. In *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16*, pages 156–166. Springer.

Florian G Hartmann and Bernhard Ertl. 2023. Big five personality trait differences between students from different majors aspiring to the teaching profession. *Current Psychology*, 42(14):12070–12086.

Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. 2022. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8):1323.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Michelle Kim, Junghwan Kim, and Kristen Johnson. 2023. Race, gender, and age biases in biomedical masked language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11806–11815, Toronto, Canada. Association for Computational Linguistics.

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747, Singapore. Association for Computational Linguistics.

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.

Cezary Kuśnierz, Aleksandra M Rogowska, and Iuliia Pavlova. 2020. Examining gender differences, personality traits, academic performance, and motivation in ukrainian and polish students of physical education: A cross-cultural study. *International journal of environmental research and public health*, 17(16):5729.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 17–29, Seattle, Washington. Association for Computational Linguistics.

Gloria Liou, Juhi Mittal, Neil KR Sehgal, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2023. The online language of work-personal conflict. *Scientific Reports*, 13(1):21019.

Richard A Lippa. 2010. Gender differences in personality and interests: When, where, and why? *Social and personality psychology compass*, 4(11):1098–1110.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1699–1710.

Simon Mille, Massimiliano Pronesti, Craig Thomson, Michela Lorandi, Sophie Fitzpatrick, Rudali Huidrom, Mohammed Sabry, Amy O'Riordan, and Anja Belz. 2024. Filling gaps in wikipedia: Leveraging data-to-text generation to improve encyclopedic coverage of underrepresented groups. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 16–19.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference* *on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Jisoo Ock, Samuel T McAbee, Evan Mulfinger, and Frederick L Oswald. 2020. The practical effects of measurement invariance: Gender invariance in two big five personality measures. *Assessment*, 27(4):657–674.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Dandan Pang, Johannes C Eichstaedt, Anneke Buffone, Barry Slaff, Willibald Ruch, and Lyle H Ungar. 2020. The language of character strengths: Predicting morally valued traits on social media. *Journal of personality*, 88(2):287–306.

Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. 2019. Addressing bias in artificial intelligence in health care. *Jama*, 322(24):2377–2378.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Christopher Peterson and Martin EP Seligman. 2004. *Character strengths and virtues: A handbook and classification*, volume 1. Oxford university press.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194, Singapore. Association for Computational Linguistics.

Daniel Russo and Klaas-Jan Stol. 2020. Gender differences in personality traits of software engineers. *IEEE Transactions on Software Engineering*, 48(3):819–834.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Arif Shahriar and Denilson Barbosa. 2024. Improving Bengali and Hindi large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8719–8731, Torino, Italia. ELRA and ICCL.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Andrew B Speer, James Perrotta, Andrew P Tenbrink, Lauren J Wegmeyer, Angie Y Delacruz, and Jenna Bowker. 2023. Turning words into numbers: Assessing work attitudes using natural language processing. *Journal of Applied Psychology*, 108(6):1027.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Martin A Stoffel, Shinichi Nakagawa, and Holger Schielzeth. 2021. partr2: partitioning r2 in generalized linear mixed models. *PeerJ*, 9:e11414.

Isaac Thompson, Nick Koenig, Derek L Mracek, and Scott Tonidandel. 2023. Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, 38(3):509–527.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2024. Detecting gender discrimination on actor level using linguistic discourse analysis. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149, Bangkok, Thailand. Association for Computational Linguistics.

Bureau of Labor Statistics US Department of Labor. 2020. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. *Labor Force Stat Curr Popul Surv 2020*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Yanna J Weisberg, Colin G DeYoung, and Jacob B Hirsh. 2011. Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 2:11757.

Bernard L Welch. 1947. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

A Appendix

A.1 Attribute values

Gender	**Gendered words**
female	abbess, actress, airwoman, aunt, ballerina, baroness, barwoman, belle, bellgirl, bride, bride, busgirl, businesswoman, camerawoman, chairwoman, chick, congresswoman, councilwoman, countrywoman, cowgirl, czarina, daughter, diva, duchess, empress, enchantress, female, fiancee, gal, gal, girl, girlfriend, godmother, governess, granddaughter, grandma, grandmother, handywoman, headmistress, heiress, heroine, hostess, housewife, lady, lady, lady, lady, landlady, lass, lass, maam, madam, maid, maiden, maidservant, mama, marchioness, masseuse, mezzo, minx, mistress, mistress, mom, mommy, mother, mum, niece, nun, nun, policewoman, priestess, princess, queen, saleswoman, schoolgirl, seamstress, seamstress, she, sister, sistren, sorceress, spokeswoman, stateswoman, stepdaughter, stepmother, stewardess, strongwoman, suitress, waitress, widow, wife, wife, witch, woman
male	abbot, actor, airman, uncle, ballet_dancer, baron, barman, beau, bellboy, bridegroom, groom, busboy, businessman, cameraman, chairman, dude, congressman, councilman, countryman, cowboy, czar, son, divo, duke, emperor, enchanter, male, fiance, guy, dude, boy, boyfriend, godfather, governor, grandson, grandpa, grandfather, handyman, headmaster, heir, hero, host, househusband, lord, fella, mentleman, gentleman, landlord, lad, chap, sir, sir, manservant, bachelor, manservant, papa, marquis, masseur, baritone, stud, master, paramour, dad, daddy, father, dad, nephew, priest, monk, policeman, priest, prince, king, salesman, schoolboy, tailor, seamster, he, brother, brethren, sorcerer, spokesman, statesman, stepson, stepfather, steward, strongman, suitor, waiter, widower, husband, hubby, wizard, man

Table 8: Attributes: Gendered words. Note that some of the words are redundant, but they are paired with distinct gendered words.

A.2 Trait dimensions (target values)

Character	**Character words**
empathy	affable, charitable, compassionate, concerned, considerate, courteous, empathetic, friendly, gracious, liberal, sensitive, sympathetic, understanding
order	abstinent, austere, careful, cautious, clean, conservative, decent, deliberate, disciplined, earnest, obedient, ordered, scrupulous, self-controlled, self-denying, serious, tidy
resourceful	confident, courageous, independent, intelligent, perseverant, persistent, purposeful, resourceful, sagacious, zealous
serenity	forbearing, forgiving, meek, merciful, patient, peaceful, serene

Table 9: Targets: **Positive character traits** — dimensions and trait words.

Personality	**Personality words**
extroversion	active, adventurous, assertive, bold, energetic, extroverted, talkative
agreeableness	agreeable, cooperative, generous, kind, trustful, unselfish, warm
conscientiousness	conscientious, hardworking, organized, practical, responsible, thorough, thrifty
emotional stability	at ease, calm, contented, not envious, relaxed, stable, unemotional
openness	analytical, creative, curious, imaginative, intelligent, reflective, sophisticated

Table 10: Targets: **Positive personality traits** — dimensions and trait words.

Character	**Character words**
empathy	disagreeable, uncharitable, unfeeling, unconcerned, inconsiderate, discourteous, callous, unfriendly, ungracious, conservative, insensitive, unsympathetic, inconsiderate
order	indulgent, genial, careless, reckless, dirty, liberal, indecent, unmotivated, undisciplined, flippant, disobedient, disorganized, unscrupulous, undisciplined, self-indulgent, frivolous, untidy
resourceful	unsure, cowardly, dependent, stupid, weak, intermittent, aimless, unresourceful, foolish, unenthusiastic
serenity	impatient, unforgiving, assertive, merciless, impatient, disturbed, agitated

Table 11: Targets: **Negative character traits** — dimensions and trait words.

Personality	**Personality words**
extroversion	inactive, unadventurous, unassertive, timid, unenergetic, introverted, silent
agreeableness	disagreeable, uncooperative, stingy, unkind, distrustful, selfish, cold
conscientiousness	negligent, lazy, disorganized, impractical, irresponsible, careless, extravagant
emotional stability	nervous, angry, discontented, envious, tense, unstable, emotional
openness	unanalytical, uncreative, uninquisitive, unimaginative, unintelligent, unreflective, unsophisticated

Table 12: Targets: **Negative personality traits** (Goldberg, 1992) — dimensions and trait words.

A.3 Overview of template selection algorithm

(1) Initially we obtain sentences from the Wikipedia Corpus and Book Corpus (used in BERT pre-training). (2) We then utilize the text generation capabilities of GPT-4 model to suggest additional sentences containing a target character trait word and the pronoun "she/he". (3) We combined all of these sentences. (4) We then filter out sentences that were no longer than 15 words, containing both a character word (from our work) and the pronoun "she/he". (5) The next step involves narrowing down these sentences to those where the pronoun precedes the character trait words. (6) We then identify common sentence patterns through parts-of-speech tagging. This involves analyzing the grammatical structure of the sentences to identify repetitive patterns. (7) Finally, after identifying potential sentence templates, a careful manual review is conducted. The above steps were performed to design *indirect* templates capturing the common expressions of human traits.

To generate *direct templates*, we repeat the process but include the word “personality” in the selection and generation criteria. These templates could provide more guidance in predicting human traits by minimizing the ambiguity in the usage of trait words in a sentence.

Limitations: The character trait word may not be used in the character context in *indirect* templates. This is handled during manual review. Note that this can also be handled by using contextual embedding. We changed past-tense common sentences into present tense while selecting templates as we focus on the present tense as the traits may change over time, and analyzing the present tense allows for real-time insights.

A.4 Influence of pseudo-perplexity

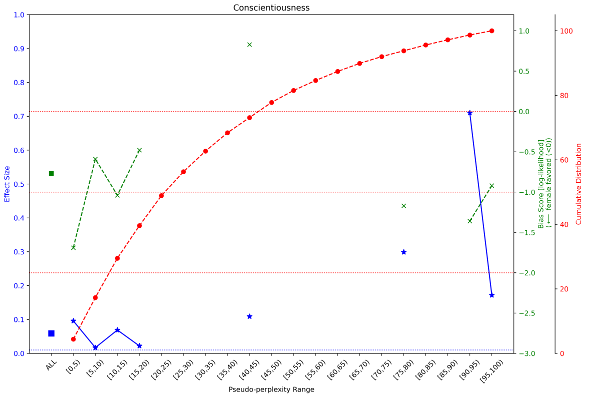

Figure 1: RoBERTa-large (model$_{\text{lme}}$).

In the cumulative graphs of Figure 1, data points (sentences) are binned by psuedo-perplexity on the X-axis. The Y-axis represents effect size (left) and bias score (inner right) obtained with the corre-

	Targets	# of target words	Templates					
			t$_1$	t$_2$	t$_3$	t$_4$	t$_5$	t$_6$
Character	*empathy*	13	3067±241	3143±221	2827±233	2882±275	2915±203	2711±103
	order	17	3958±332	4119±313	3676±301	3764±383	3837±231	3489±130
	resourceful	10	2342±211	2447±178	2167±194	2213±239	2236±158	2065 ± 85
	serenity	7	1645±138	1715±145	1506±124	1559±158	1571±116	1447 ± 53
Personality	*extroversion*	7	1619±155	1671±110	1530±151	1543±158	1569±119	1437 ± 57
	agreeableness	7	1653±124	1757±121	1530±123	1554±154	1560 ± 96	1449 ± 59
	conscientiousness	7	1639±135	1692±124	1517±133	1547±152	1580±101	1450 ± 69
	emotional stability	7	1626±133	1730±141	1538±146	1555±155	1593±107	1457 ± 77
	openness	7	1665±123	1711±139	1520±120	1550±152	1608 ± 97	1452 ± 51

Table 13: Mean and Standard deviation ($\mu \pm \sigma$) of the number of sentences for each template in each of character/personality dimensions (includes 94 pairs of gendered words (attributes)) across seven MLMs of variation (σ/μ) ranges from 3.5% to 10.8%. The sentence selection is specific to MLM, and hence, the number of sentences within each template and trait dimension can vary. So, we provide the mean and standard deviation for each template within each trait dimension.

sponding sentence set. Only points with significant bias scores are shown. Effect sizes below 0.01 (blue horizontal line close to the X-axis) have a negligible effect. The red horizontal lines indicate distributions of 25%, 50%, and 75% (bottom to top).

Key to note is that bias score (-0.77) and effect size (0.059) for the full set of sentences - 'ALL' on the X-axis - are close to the average bias score (-0.95) and effect size (0.051) for the first 4 pseudo-perplexity bins.

A.5 Additional proof-of-concept experiments

We limit additional proof-of-concept experiments to RoBERTa-large, our most biased model, except for bias analysis in the auto-regressive language model.

A.5.1 Bias detection in autoregressive language model (ALM)

Our main experiments are on detecting bias in MLMs. Here we show that our approach can be extended to detect bias in autoregressive pre-trained language models, demonstrating this with Llama3.1-8B (Dubey et al., 2024). We analyze pairwise bias for binary gender and non-binary gender using the same neo-pronouns set from Section 3.2.3.

We probe LLama3 with sentences (from our templates) for each gender and trait combination. Similar to work by Hossain et al. (2023), we use sentence loss as a proxy for gender - trait association score. A lower loss indicates a better fit with the model and hence a stronger association between the gendered word and trait word in the sentence. The rest of the bias detection design is as discussed in Section 2.

	Traits	M - F	M - N	F - N
Character traits	*empathy*	0.06	-1.61 *	-1.67 *
	order	0.04	-1.56 *	-1.60 *
	resourceful	0.09	-1.42 *	-1.51 *
	serenity	0.07	-1.64 *	-1.71 *
Personality traits	*extroversion*	0.07	-1.49 *	-1.57 *
	agreeableness	0.06	-1.39 *	-1.44 *
	conscientiousness	0.07	-1.49 *	-1.56 *
	emotional stability	0.05	-1.40 *	-1.45 *
	openness	0.07	-1.64 *	-1.71 *

Table 14: Results for LLama3.1-8B using model$_{\text{lme}}$. M: Males, F: Females, N: Neo-pronouns. See Table 3 for reported values descriptions and the notation used. Negative values indicate larger losses for neo sentences, suggesting weaker association between trait and neo.

Differences between males and females are minimal and there are no biases. However, when comparing males (or females) with neo group, we observe sizable and significant bias scores with medium to large effect sizes (0.15 to 0.26). Hence, there is medium to large bias against neo. The negative differences indicate larger losses for neo sentences, suggesting a weaker association between neo and traits compared to associations for other genders. It has been observed that LLMs generally perform well in tasks with the goal of predicting binary gender while they perform poorly at predicting neo-pronouns (Ovalle et al., 2024; Hossain et al., 2023). This weaker performance in handling neo-pronouns might explain the weaker association between neo and traits compared to binary genders and traits.

A.5.2 MLM bias detection using a crowdsourced dataset without templates

While we focus on template-based design in our main experiments, our work is not limited to bias detection with such datasets. To demonstrate this we present bias analysis on the crowdsourced CrowS-Pairs (Nangia et al., 2020) which does not involve templates.

First, as a sanity check we conduct a replication study using their CrowsPair Score (CPS) metric and achieve a similar value as reported by Kaneko and Bollegala (2022). CPS measures the percentage of stereotypical sentences preferred by an MLM over anti-stereotypical. Additionally we extend the analysis with our approach that focuses on the 'difference' in association scores across these two sentence types. We take stereotype_type as a fixed effect in our model. The association (pseudo-log likelihood PLLScore) is computed using 'score (S)' as in Nangia et al. (2020). To address variations in sentence structure, we grouped sentences by length (short, medium, long) based on the 33rd and 67th percentiles of sentence length, accounting for sentence length variability as a random effect (1|sentence_length_group). The overall model is association$_{\text{score}}$ $\sim$ stereotype_type + (1| sentence_length_group), weight = 1/sentence psuedo-perplexity. We applied our model$_{\text{lme}}$. Bias score is the coefficient of the stereotype_type (i.e., stereotypical PLLScore - anti-stereotypical PLLScore). The rest of the design is the same as in Section 2. Results are in Table 15.

Bias Type	*n*	CPS	Our approach (model$_{\text{lme}}$)
Race/Color	516	64.15	0.59
Gender/Gender identity	262	58.78	0.11
Socioeconomic status/ occupation	172	66.86	0.66
Nationality	159	66.67	1.14
Religion	105	73.33	1.04
Age	87	72.41	1.23
Sexual Orientation	84	64.29	0.88
Physical appearance	63	73.02	1.34
Disability	60	70.00	1.41

Table 15: Results for CrowS-Pair using our approach. CPS: CrowS-Pair Score. n: number of examples.

Stereotypical sentences are preferred over anti-stereotypical ones, but the differences throughout are insignificant, indicating no bias. The problem with CPS is that even minor PLLScore differences contribute to deviations from the 50% ideal and detection of bias. Unfortunately, the magnitude of these differences are not considered. In contrast, our model statistically analyzes the PLLScore differences across sentence types, focusing on both significance and effect size - and we find no bias.

A.5.3 Bias mitigation in MLMs

Our research focus is on bias detection. However, for the reader expecting a complete pipline that includes mitigation - we add this proof-of-concept experiment. We mitigate bias in RoBERTa-large, our most biased model (Section 3.2.1), with a focus on binary gender. We mitigate bias by fine-tuning the model on a gender-swapped GAP corpus (Webster et al., 2018) following Bartl et al. (2020). This process involves dynamic masking during fine-tuning MLM task, following the design described in Liu et al. (2019), to specifically address gender bias.

We tune the model for 3 epochs using AdamW optimizer with a 2e-5 learning rate and a batch size 16. To manage the learning rate adjustment smoothly, we use a polynomial decay scheduler with a linear warm-up phase over the first 500 steps.

	Traits	Before	After
Character traits	*empathy*	-0.70 ▲	-0.31
	order	-0.30 ▽	0.004
	resourceful	-0.74 △	-0.14
	serenity	-1.08 *	-0.39 ▽
Personality traits	*extroversion*	-0.86*	-0.23
	agreeableness	-0.77 ▲	-0.20
	conscientiousness	-0.77 ▲	-0.12
	emotional stability	-0.26	0.03
	openness	-0.61 △	-0.14

Table 16: Bias mitigation performance in RoBERTa-large (model$_{\text{lme}}$). See Table 3 for reported values descriptions and the notation used. **Before** mitigation result is identical to Table 3.

Table 16 presents bias *before* and *after* mitigation in RoBERTa-large. Bias scores reduce by 56% to 98% after mitigation across both sets of traits. There is only one dimension *serenity* that still exhibits some bias - but this has reduced from medium to small. The remaining dimensions have become unbiased as regards gender.