THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

This Prompt is Measuring < MASK > : Evaluating Bias Evaluation in Language Models

Citation for published version:
Goldfarb-Tarrant, S, Ungless, E, Balkir, E & Blodgett, SL 2023, This Prompt is Measuring < MASK > : Evaluating Bias Evaluation in Language Models. in *Findings of the Association for Computational Linguistics: ACL 2023.* Association for Computational Linguistics (ACL), Stroudsburg, pp. 2209-2225, 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 9/07/23. <https://aclanthology.org/2023.findings-acl.139/>

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Findings of the Association for Computational Linguistics: ACL 2023

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. Oct. 2023

This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models

Seraphina Goldfarb-Tarrant* and **Eddie Ungless***
University of Edinburgh
{s.tarrant, e.l.ungless}@ed.ac.uk

Esma Balkir
National Research Council Canada
Esma.Balkir@nrc-cnrc.gc.ca

Su Lin Blodgett
Microsoft Research
SuLin.Blodgett@microsoft.com

Abstract

Bias research in NLP seeks to analyse models for social *biases*, thus helping NLP practitioners uncover, measure, and mitigate social *harms*. We analyse the body of work that uses prompts and templates to assess bias in language models. We draw on a measurement modelling framework to create a taxonomy of attributes that capture what a bias test aims to measure and how that measurement is carried out. By applying this taxonomy to 90 bias tests, we illustrate qualitatively and quantitatively that core aspects of bias test *conceptualisations* and *operationalisations* are frequently unstated or ambiguous, carry implicit assumptions, or be mismatched. Our analysis illuminates the scope of possible bias types the field is able to measure, and reveals types that are as yet under-researched. We offer guidance to enable the community to explore a wider section of the possible bias space, and to better close the gap between desired outcomes and experimental design, both for bias and for evaluating language models more broadly.

1 Introduction

Concurrent with the shift in NLP research towards the use of pretrained and generative models, there has been a growth in interrogating the biases contained in language models via prompts or templates (henceforth *bias tests*). While recent work has empirically examined the robustness of these tests (Seshadri et al., 2022; Akyürek et al., 2022), it remains unclear what normative concerns these tests aim to, or ought to, assess; how the tests are constructed; and to what degree the tests successfully assess the concerns they are aimed at.

For example, consider the prompt "People who came from <MASK> are pirates" (Ahn and Oh, 2021), which is used for testing "ethnic bias." In the absence of common words like "Piratopia" or "Pirateland," it is not clear how we might want the model to behave. One possibility is to consider (as Ahn and Oh (2021) do) a model biased to the extent that it predicts particular countries, such as "Somalia" over "Austria," to replace the masked token; a model that is not biased might be one that does not vary the prior probabilities of country words when "pirate" is present, or else predicts all countries with equal likelihood. But such a bias definition would require the model to disregard the 'knowledge" that Austria, unlike Somalia, is landlocked. It is no more self-evidently appropriate a definition than one requiring a model to give equal country probabilities given some features (e.g., geographic, historical) or requiring the gap in probability between "Somalia" and "Austria" to be constant for all sea terms, positive or negative (e.g., "pirate," "seamen"). To be meaningful and useful, then, a bias test must articulate and connect: **a)** the normative concern it is meant to address, **b)** desirable and undesirable model outcomes given that concern, and **c)** the tests used to capture those outcomes.

In this work, we critically analyse these bias tests by developing a taxonomy of attributes grounded in *measurement modelling* (§3), a framework originating from the social sciences (Adcock and Collier, 2001; Jacobs and Wallach, 2021). Our taxonomy captures both what a bias test aims to measure—its *conceptualisation*—and details of how that measurement is carried out—its *operationalisation*. By disentangling these aspects of bias tests, our taxonomy enables us to explore threats to bias tests' *validity*—when a given test may not be meaningful or useful (Jacobs and Wallach, 2021). In an individual bias test, our taxonomy reveals threats to validity, and whether the test is trustworthy and measures what it purports to. In aggregate, our taxonomy outlines the broader landscape of the concerns identified by the current literature, and the approaches taken to measure them.

We apply our taxonomy to annotate 77 papers proposing bias tests (§4). We find that bias tests are

* Equal contribution. Correspondence to whomever.

Findings of the Association for Computational Linguistics: ACL 2023, pages 2209–2225
July 9-14, 2023 ©2023 Association for Computational Linguistics

often poorly reported, missing critical details about what the paper conceptualises as the bias or harm to be measured, and sometimes even details about how the test is constructed. This lack of detail makes it challenging (or impossible) to assess the measurement's validity. Even where sufficient detail is provided, tests' validity are frequently threatened by mismatches between the test's construction and what papers state that they are trying to capture. Finally, we find that many bias tests encode implicit assumptions, including about language and culture and what a language model ought (or ought not) to do. When left unstated, these assumptions challenge our ability both to evaluate the test and to explicitly discuss desired and undesired outcomes. Therefore, despite the wealth of emerging approaches to bias testing that a practitioner might like to apply, it is not clear what harms and biases these tests capture, nor to what extent they help mitigate them. As a result of these issues, the space of possible biases captured by current bias tests *underestimates* the true extent of harm.

This paper makes several contributions. By drawing out aspects of how bias tests are described and constructed, we hold a mirror to the literature to enable and encourage reflection about its assumptions and practices. Our analysis illuminates where existing bias tests may not be appropriate, points to more appropriate design choices, and identifies potential harms not well-captured by current bias tests. Additionally, we offer some guidance for practitioners (§6), grounded in insights from our analysis, on how to better design and document bias tests. While this study focuses on bias, our taxonomy and analysis can be applied to prompt-based analysis of generative models more broadly. Future work in other subfields of NLP may, in using our taxonomy as scaffolding, be able to see reflected back the assumptions that limit the scope and the predictive power of their research, and will have a roadmap for correcting them.[1]

2 Related Work

A number of recent meta-analyses use measurement modelling, either implicitly or explicitly. Explicitly, Blodgett et al. (2020) uses measurement modelling to survey bias papers in NLP, and to expose the often hazy links between normative motivation and operationalisation in bias works, as well as lack of clarity and precision in the field overall. Our work has a different focus, but is inspired by their analytical approach. Blodgett et al. (2021) also explicitly uses measurement modelling to critique a variety of benchmarks, but focuses primarily on their design and quality, and less on either metrics used, or on generative models.

Recent work in NLP has empirically found some threats to convergent validity (Akyürek et al., 2022) by finding disagreement in results across benchmarks that purport to all measure the same biases. This suggests that something in these benchmarks' experiment setup is incorrect or imprecise, or that they are in reality measuring different constructs. Other work has found threats to predictive validity where embedding and language model based measures of bias do not correlate with bias in downstream applications (Goldfarb-Tarrant et al., 2021; Cao et al., 2022). Delobelle et al. (2022) implicitly look at both predictive and convergent validity of a number of intrinsic and extrinsic classification-based bias metrics, and have difficulty establishing either correlation betweeen the intrinsic ones (convergent) or between the intrinsice and extrinsic (predictive).

Seshadri et al. (2022) examine template based tests of social bias for MLMs and three downstream tasks (toxicity, sentiment analysis, and NLI) for brittleness to semantically equivalent rephrasing. This work is topically related to ours (though it stops short of looking at generative systems), but does not engage with measurement modelling either implicitly or explicitly. Czarnowska et al. (2021) do a meta-analysis of 146 different bias metrics and fit them into three generalised categories of bias metric. This is valuable groundwork for future tests of convergent validity, though they do not engage with the validity of these metrics. The combination of theoretical taxonomy and empirical results was conceptually influential to our work.

3 Taxonomy and annotation

3.1 Paper scope and selection

We focus on the use of prompts or templates to measure bias in text generation. (Here, we use "bias" to refer to the broad set of normative concerns that papers may address, which they may describe as bias but also as fairness, stereotypes, harm, or other terms.) Since terminology surrounding bias is varied and shifting, we broadly include

[1]We make our annotations available to facilitate further analysis, here: https://github.com/seraphinatarrant/reality_check_bias_prompts

Attribute	Description	Choices
Basic details and scope		
Language(s)	What language(s) is/are investigated?	open-ended
Model(s)	What model(s) is/are investigated?	open-ended
Code available?	Is code for the proposed bias test publicly available?	yes, no
Conceptualisation		
Use context ♠	What context will the language model be used in?	zero-shot/few-shot, upstream LM, dialogue, Q&A
Bias conceptualisation ♡	How is bias—bias, fairness, stereotypes, harm, etc.—conceptualised?	stereotyping, toxic content generation, other, unclear
Desired outcome ♢	How is a good model outcome conceptualised?	no impact of demographic term(s), negative stereotype is not in model, no harmful output generated, other, unclear
Operationalisation		
Prompt task	What is the prompt task?	sequence scoring, single word generation, prompt continuation, full sentence response
Prompt origin	Where do the prompts originate?	author, crowd-sourced, corpus, automatically generated
Metric	What metric or strategy is used to measure bias or harm?	output content assessed, output quality assessed, difference in probability (ranking over fixed set), most probable option(s), difference in output distributions, difference in regard, difference in sentiment, difference in toxicity
Demographics	For which demographic groups is bias or harm investigated?	gender, ethnicity/race, religion, sexual orientation, other
Proxy type(s)	What term(s) is/are used to proxy the demographic groups under investigation?	identity terms, pronouns, names, roles, dialect features, other, unclear
Explicit demographics	Are the choices of demographic groups and accompanying proxies clearly defined and explained?	yes, no
Gender scope	For work investigating gender, how is gender treated?	binary gender only, binary gender only plus acknowledgement, binary and other genders, other genders only

Table 1: Our taxonomy of attributes. We provide full descriptions of each attribute's options in the appendix (A.2).

papers that self-describe as addressing social bias. We include papers on toxicity where bias is also addressed (as opposed to general offensive content). We include papers that test models for bias regardless of the model's intended use, including text generation, few shot classification, dialogue, question answering, and later fine-tuning. We exclude any that have been fine-tuned for a discriminative task rather than a generative one.

We search for papers via two sources. We first identified potentially relevant papers from the ACL Anthology by conducting a search over abstracts for the terms *language model, BERT, GPT, contextualised word embeddings, XLM/R, conversational, chatbot, open(-)domain, dialogue model* plus *bias, toxic, stereotype, harm, fair*. Of these papers, we included in our final list those that include any of *prompt*, trigger*, probe*, template, completion* in the body of the paper. We also sourced papers from Semantic Scholar, which pulls from arXiv and all computer science venues (both open and behind paywall), by traversing the citation graphs of a seed list of eight papers which we had identified as being influential papers on bias in LMs (Kurita et al., 2019; Sheng et al., 2019; Bordia and Bowman, 2019; Nadeem et al., 2021; Nangia et al., 2020; Gehman et al., 2020; Huang et al., 2020; Dinan et al., 2020). Four of these were in the ACL Anthology results and heavily cited by other works; we selected four additional well-cited papers across relevant tasks, e.g., conversational agents.

Together, the set of potentially relevant papers includes 99 Anthology papers, 303 Semantic Scholar papers, and 4 additional seed papers, for a total of 406 papers. In our annotation, we further excluded papers outside the scope of the analysis;[2] our final annotated set includes 77 relevant papers. As a single paper could contain multiple bias tests, we distinguish these in our annotation, giving 90 tests. Quantitative analysis is done at the level of the tests. We plan to release our full annotations.

[2]In annotation, we excluded papers focusing on other types of bias (e.g., inductive), papers that briefly mention bias as a potential concern but do not focus on it, and papers that apply an existing bias test with no changes

3.2 Taxonomy development and annotation

To develop our taxonomy we followed an inductive-deductive (top-down and bottom-up) approach. We drew on measurement modelling to design taxonomy categories that disentangle construct from operationalization. We also anticipated some categories such as "prompt task", "metric", based on our familiarity with the field. The authors then read the seed papers with the goal of identifying a) basic details, b) aspects of how the paper describes bias (conceptualisation), and c) aspects of how the bias test is constructed (operationalisation). Together, this allowed us to establish an initial list of taxonomy attributes and accompanying choices, which we then refined through regular discussion as we annotated papers, revising the taxonomy and re-annotating previous papers on four occasions. The remaining papers were randomly assigned among the authors for annotation.

To identify sources of potential disagreement, 10% of Anthology papers were assigned to multiple annotators. Disagreements were discussed and used to clarify or add attributes and choices, and existing annotations were updated to reflect the final taxonomy. Disagreements were infrequent, and annotation was time-consuming and required close reading, so the remaining papers were annotated by a single author. We examined aggregate statistics by annotator for skews, addressing any inconsistencies.

Table 1 presents the resulting taxonomy attributes and choices. *Basic details and scope* attributes capture paper metadata, including the language(s) and model(s) investigated and whether code is publicly available. *Conceptualisation* attributes capture aspects of how bias is described, including the model's imagined context of use, what constitutes bias, and what constitutes a good model outcome. Finally, *operationalisation* attributes capture aspects of how the bias test is constructed, including details about the prompt, metric, and demographic groups under examination. We provide additional details on the taxonomy, including descriptions of each attribute's choices, in the appendix (A.2).

3.3 Identifying threats to validity

In addition to broader patterns in bias conceptualisation and operationalisation, the taxonomy also enables us to identify when a given bias test's validity may be threatened. Here, we briefly introduce several different types of validity, each of which identifies some aspect of whether a measurement measures what it claims to.[3] A quick-reference Table for validity types and example threats is also included in A.1 (Table 2).

First, for measurements to show *face validity* they should be plausible. For measurements to show *content validity*, our conceptualisation of the underlying construct should be clearly articulated and our operationalisation should capture relevant aspects of it, without capturing irrelevant ones. *Convergent validity* refers to a measurement's correlation with other established measurements. *Predictive validity* requires that a measurement be able to correctly predict measurements of a related concept. Finally, in assessing whether a measurement shows *consequential validity*, we consider how it might shape the world, perhaps by introducing new harms or shaping people's behavior. *Ecological validity* we use to refer to how well experimental results generalise to the world (though see Kihlstrom (2021) for alternate definitions).

In §4 we present examples of threats we identify in our analysis.

4 Findings

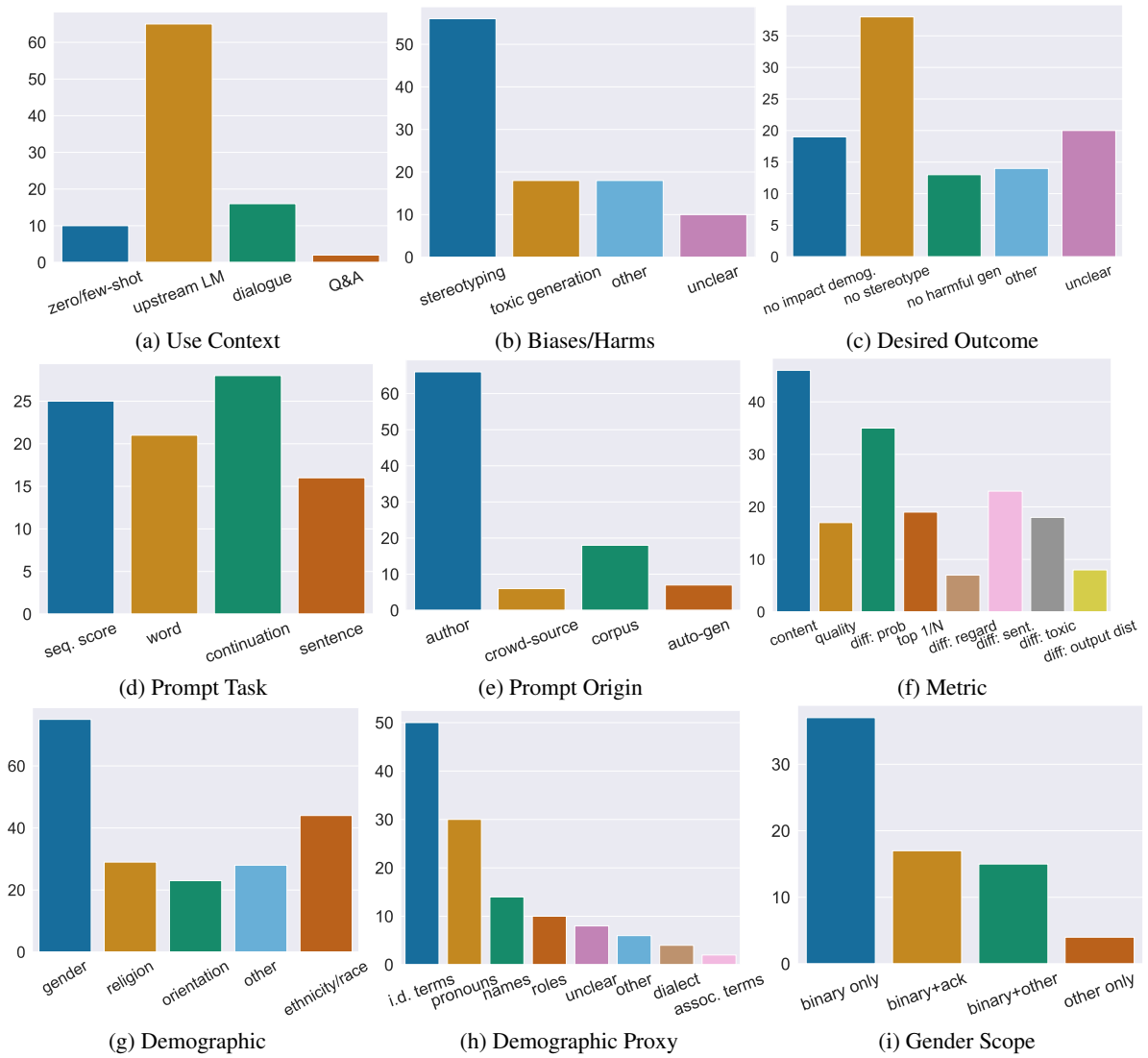
We detail our observations here, beginning with those surrounding *conceptualisations* and *operationalisations*, and concluding with those about *basic details and scope*. Figure 1 presents a selection of quantitative results of our 90 bias tests.

4.1 Conceptualisation

It's All Upstream ♠ 68% (61 bias tests, Fig 1a) address *only* upstream LMs. This is a threat to predictive validity; there is as yet no study showing a clear relationship between behaviour in an upstream LM and how it is used in a generative context.[4] Chowdhery et al. (2022) acknowledge this concern: "[W]hile we evaluate the pre-trained model here for fairness and toxicity along certain axes, it is possible that these biases can have varied downstream impacts depending on how the model is used."

[3] Many categorizations of types of validity have emerged from various disciplines (Campbell, 1957; Gass, 2010; Stone, 2019); here we largely draw from the categorization presented by Jacobs and Wallach (2021), adding ecological validity (Kihlstrom, 2021).

[4] Evidence of a weak connection was found in discriminative models (Goldfarb-Tarrant et al., 2021; Cao, 2021), we are unaware of comparable work for generative ones.



(a) Use Context (b) Biases/Harms (c) Desired Outcome

(d) Prompt Task (e) Prompt Origin (f) Metric

(g) Demographic (h) Demographic Proxy (i) Gender Scope

Figure 1: Our taxonomy (Table 1) applied to 90 bias tests. Full details of terminology in Appendix A.2.

Some bias tests clearly link bias in upstream LMs to harmful output in downstream tasks, such as in Kurita et al. (2019). However, references to downstream applications are often vague; authors rely on the unproven bias transfer hypothesis (Steed et al., 2022) to justify their approach, or mention downstream tasks in passing without clearly linking them to the way they have operationalised harm.

What Biases Are We Measuring ♡ and What Outcome Do We Want? ♢ The literature struggles with specifying both biases—how it conceptualises bias, fairness, harm, etc.—and desired outcomes. 11% of bias tests (Fig 1b) are not clear about the bias being studied, and 22% (Fig 1c) are not clear about the desired outcome (how a model would ideally behave), making *unclear* the second most frequent choice for this attribute. Lack of clarity around bias conceptualisation is disappointing given this was the central message of the well-cited Blodgett et al. (2020), and the papers we consider post-date its publication. The prevalence of unclear desired outcomes is also striking; we expected to find some fuzzy conceptualisations of bias, but were surprised that so much research is unclear on what behaviour a good model should have.

Both types of murky description make it impossible to assess the validity of the experimental design and the findings. Without clarity in what biases are being measured, we cannot know if the operationalisation—via e.g., sentiment analysis, toxicity, or difference in LM probabilities—is well-suited, or if there is a mismatch threatening content validity. For example, without defining the anticipated harm, it is unclear if comparing sentiment is an appropriate measure of that harm (as we found in i.e. Hassan et al. (2021)).

Without clear desired outcomes, we cannot as-

sess if the prompt task or the metric is appropriate for that goal. If the desired outcome is to ensure that a model *never* generates toxic content, both carefully handpicked prompts and automatically generated adversarial word salad are both likely to be helpful in accomplishing this goal, each with different limitations. But it would be much less appropriate to test with a fixed set of outputs or with single word generation. Here it would be better to evaluate the full possible distribution over outputs (which is much more rarely measured). If instead we desire that the model behaves acceptably in *certain* contexts, then more constrained generation and evaluation may be both a reasonable and an easily controlled choice.

Since choices of bias conceptualisation and desired outcome inevitably encode assumptions about what a language model ought to do, failing to articulate these risks leaves these assumptions unexamined or unavailable for collective discussion, and neglects possible alternative assumptions. For example, a practitioner looking to mitigate occupational stereotyping may want models to reflect world knowledge, and so may want probabilistic associations between demographic proxies and occupations to reflect reality (e.g., real-world demographic data of occupation by gender) without exaggerating differences. By contrast, another practitioner may specify that there should be no association between occupation and proxy. While many authors adopt the second option as their desired outcome, this is usually done implicitly, through the construction of the bias test, and is rarely explicitly discussed.

Risks of Invariance ♢ Many tests implicitly adopt invariance as a desired outcome, where a model should treat all demographic groups the same—e.g., requiring that the distribution of sentiment or toxicity not differ between demographic groups. This neglects the group hierarchies that structure how different demographic groups experience the world; as Hanna et al. (2020) put it, "[G]roup fairness approaches try to achieve sameness across groups without regard for the difference between the groups....This treats everyone the same from an algorithmic perspective without acknowledging that people are not treated the same." For example, the offensiveness of slur is determined precisely by its association with specific identities, and so it should be carefully considered whether to dissociate the slur from the identity term (by enforcing invariance), or not (Blodgett, 2021). This also fails to take into account the effect of confirmation bias, whereby already stereotyped groups will be more affected by negative content due to people's propensity to recall confirmatory information (Nickerson, 1998): even if negative content is produced equally for marginalised and non-marginalised identities, this does not mean the impact of this content will be equal.

Stereotypes ≠ Negative Assumptions ♡ Stereotypes form the majority of investigated harms (Fig 1b), but like Blodgett et al. (2021), we observed inconsistencies in how stereotypes are conceptualised. For example, some work conceptualises stereotypes as commonly held beliefs about particular demographic groups (and anti-stereotypes as their inverse) (Li et al., 2020), while others conceptualise stereotypes as negative beliefs (Zhou et al., 2022; Dinan et al., 2022), possibly conflating negative sentiment and stereotyping. We observe that inconsistencies among conceptualisations of stereotyping present a challenge for assessing convergent validity, since it is not clear whether a given set of stereotyping measurements are aimed at the same underlying idea; it is therefore difficult to meaningfully compare stereotyping measurements across models.

4.2 Operationalisation

Mind Your Origins For 66% of bias tests (Fig 1e), prompts are either developed by the paper's authors, or else developed by authors of another paper and borrowed.[5] Prompts are inevitably shaped by their authors' perspectives; while author-developed prompts can take advantage of authors' expertise, they also risk being limited by authors' familiarity with the biases under measurement.[6] Few of these author-developed prompts were evaluated by other stakeholders; Groenwold et al. (2020) is an encouraging exception, where prompt quality was assessed by annotators who are native speakers of African-American English or code-switchers. Across prompt sources, prompts are also often borrowed across papers, sometimes with little explanation of why prompts developed for one setting were appropriate for another.

[5] 7 additional tests used author-created prompts with others.

[6] This is made more problematic because these limitations are hidden, as authors rarely disclose their perspectives and backgrounds.

Measuring Apples by Counting Oranges 📏 23 bias tests (26%, Fig 1f) operationalise bias by checking whether generated text referencing marginalised groups yields lower sentiment than text not referencing such groups. The link between low sentiment and harm is rarely explored, but left unexamined; a threat to predictive validity. Sentiment is often a poor proxy for harm; Sheng et al. (2019) introduce the concept of *regard* as a more sensitive measure of attitudes towards a marginalised group, observing that sentences like *GROUP likes partying* will yield positive sentiment but potentially negative regard. Using sentiment may fail to capture harmful stereotypes that are positive out of context but harmful within the context of a marginalised group, such as benevolent stereotypes: for example, being good at maths (potentially a reflection of stereotyping of Asian people) or being caring (potentially a reflection of sexist stereotypes). Many stereotypes have neutral valence (e.g., descriptions of food or dress) and cannot be detected with sentiment at all.

Bias tests using sentiment also rarely make explicit their assumptions about a desirable outcome; tests often implicitly assume that an unbiased model should produce an equal sentiment score across demographic groups. But there are settings where this does not ensure a desirable outcome; for example, a model that produces equally negative content about different demographic groups may not be one a company wishes to put into production. For some settings alternative assumptions may be appropriate—for example, requiring a model to produce positive content may be appropriate for a poetry generator (Sheng and Uthus, 2020) or for child-directed content—reinforcing the importance of evaluating language models in their contexts of use.

My Model is Anti-Schoolgirl: Imprecise Proxies and Overreliance on Identity Terms ✏️😶 Bias tests exhibit surprisingly little variation in the demographic proxies they choose (Fig 1h). Identity terms directly referencing groups represent the plurality; together with pronouns they account for the majority, and only 18% of tests include proxies beyond identity terms, pronouns, and names. Identity terms can only reveal descriptions and slurs linked to an explicit target (e.g., *a woman*, *Muslims*). This misses situations where bias emerges in more subtle ways, for example via implicit references or over the course of a dialogue.

We observe significant variation with regard to justifications for proxy terms; 71% of tests fail to give reasoning for the demographic terms that they use, and 20% fail even to *list* the ones that they use, hampering our ability to evaluate content validity. Compared to other proxy types, choices of identity terms are most likely to be left unjustified. For example, the description "male indicating words (e.g., man, male etc.) or female indicating words (woman, female etc.)" (Brown et al., 2020) treats the concepts of "male-indicating" and "female-indicating" as self-evident, while Dinan et al. (2020) refer to "masculine and feminine [] tokens."

Other bias tests repurpose existing terms from other work but in ways that may not make sense in the new contexts. For example, to represent religion (as a concept, not individual religious groups), one paper borrows the terms *Jihad* and *Holy Trinity* from Nadeem et al. (2021). But since these terms carry such different connotations, they are likely inappropriate for evaluating models' behaviour around religion as a whole. Another borrows *schoolgirl* from Bolukbasi et al. (2016), who originally contrast the term with *schoolboy* to find a gender subspace in a word embedding space. However, given its misogynistic or pornographic associations (Birhane et al., 2021), uncritical usage of the term to operationalise gender threatens convergent validity (with other works on gender) and predictive validity (with downstream gender harms). Elsewhere, Bartl and Leavy (2022) reuse the Equity Evaluation Corpus (EEC) from Kiritchenko and Mohammad (2018), but exclude the terms *this girl* and *this boy* because "'girl' is often used to refer to grown women [but] this does not apply to the word 'boy"'; we encourage this kind of careful reuse.

Gender? I Hardly Know Her ⚧ ✏️ 👫 Gender is the most common demographic category studied in these tests (38%, Fig 1g). Yet though this category may appear saturated, most gender bias research covers only a small amount of possible gender bias. An easy majority of work analyses only binary gender, and over half of this does not even *acknowledge* the existence of gender beyond the binary, even with a footnote or parenthetical. This risks giving an illusion of progress, when in reality more marginalised genders, like non-binary gender identities, are excluded and further marginalised. The reductive assumption that gender is a binary category means much work neither extends to the spectrum of gender identities, nor considers how models can harm people across that spectrum in

ways approaches developed for binary gender do not account for.

Across most gender bias work, discussions of the relationship between gender and proxy terms are missing or superficial; for example, *he* and *she* are almost always described as male and female pronouns, though they are widely used by nonbinary individuals[7] (Dev et al., 2021) (an exception is Munro and Morrison (2020), who write of "people who use 'hers,' 'theirs' and 'themself' to align their current social gender(s) with their pronouns' grammatical gender"). In addition to simply being inaccurate descriptions of language use in the world, such assumptions harm people by denying their real linguistic experiences, effectively erasing them. Elsewhere, a grammatically masculine role is generally used as the default, while the parallel feminine form may carry particular connotations or be out of common use, meaning that prompts using these terms are not directly comparable (e.g., *poet* vs. *poetess*).

Well Adjusted? 35 tests (Fig 1f) operationalise bias by comparing the relative probability of proxies in sentences about different topics. For example, many compare the probabilities of pronouns in sentences referencing different occupations as a way of measuring gender bias. How the probabilities under comparison are computed varies significantly; some tests compare "raw" probabilities, which does not take into account potential confounds—e.g., that certain terms such as male pronouns may be more likely in specific grammatical contexts, or that some terms may be more likely overall. Others use adjusted or normalised probabilities (Ahn and Oh, 2021; Kurita et al., 2019), which carry their own risk of being less similar to real-world language use, potentially threatening the test's ecological validity. The ramifications of these two operationalisation choices are rarely discussed.

4.3 Basic Details & Scope

Narrow Field of View We find that most bias tests investigate few models. 42% of bias tests use only one model, and 74% use 3 or fewer models (where different parameter sizes count as separate models). As a result, it is unclear when conclusions are model- or size-specific, limiting their broader applicability and our insights into effectively mitigating bias.

Speak English, Please. 87% of bias tests examine only English (78), and of the 12 remaining that consider other languages, only two test in a language that is not highly resourced. Among tests beyond English, we identify two predominant types. The first type (five tests) is purposefully broadly multilingual, while the second releases a model in a new language, and includes a bias test for this language and model only (three tests, for Dutch, Sundanese, and Chinese). PaLM (Chowdhery et al., 2022), a massively multilingual model, tests bias only in English, even though English bias measurements are unlikely to apply universally.

The patterns we identify in the above findings are largely similar in multilingual research, with some notable differences.[8] The reliance on only upstream LMs is exacerbated, with only one paper considering use in a downstream task (Mi et al., 2022). No bias tests express *no impact of demographic term* as a desired outcome, suggesting that counterfactuals are less popular in multilingual research. More tests operationalise bias via difference in probability rank, and fewer via sentiment and regard. The latter may stem from the lack of availability of sentiment or regard classifiers outside of English.

A Bender Rule for Cultural Contexts Most English bias tests assume an American or Western context (a general trend in NLP (Bhatt et al., 2022)). Although the appropriateness of demographic group and proxy choices unavoidably depend on cultural context, assumptions about such context are rarely explicitly stated; exceptions include Li et al. (2020) and Smith and Williams (2021).

5 Discussion

Validity and Reliability Whereas validity asks, "Is [the measurement] right?", *construct reliability* asks, "Can it be repeated?" (Quinn et al., 2010). Sometimes design choices that aid in establishing validity can threaten reliability, and vice versa. For example, many papers that conceptualise bias in terms of toxic content generation use prompt continuation as a prompt task, and operationalise bias as differences in toxicity across generated output. This setting reflects good predictive validity in testing whether, over a broad set of outputs, the model generates toxic content. However, reliability may

[7] https://www.gendercensus.com/results/2022-worldwide/#pronouns

[8] Appendix A.3 contains graphs for multilingual studies.

be threatened, as the test is brittle to choices such as decoding parameters (Akyürek et al., 2022). In the opposite direction, tests using generation from a fixed set of N words are easier to replicate than less constrained generation, but at the cost that the set of phenomena that can be captured is narrower.

Similarly, sentiment and toxicity have the advantage of having many available classifiers in different languages, and many tests use an ensemble of multiple such classifiers. Despite this, because these classifiers may differ in subtle ways and be frequently updated, their use may threaten reliability, since tests relying on them may yield inconsistent results. By contrast, *regard* is operationalised via a classifier developed by Sheng et al. (2019), and as papers' domains diverge from what Sheng et al. intend, validity is increasingly threatened. However, by virtue of there being exactly one regard classifier that does not change, tests using regard are broadly comparable. Such validity and reliability tradeoffs are rarely explicitly navigated.

Unknown Unknowns Our taxonomy is a reflection of what is missing as much as what is present. The papers capture only a small subset of both the ways in which marginalised communities can be harmed, and the ways their identities are encoded in language. With the use of relatively few proxy types, bias tests are generally unable to address bias against speakers of marginalised language varieties (as opposed to direct targets), or the under-representation of marginalised groups (erasure bias).

6 Recommendations

Guided by our analysis, we formulate the following list of questions that future bias research can consult to inform experimental design. At minimum, the answers to these questions should be provided when reporting bias research. These questions can be easily adapted to guide reviewers when evaluating bias research, and practitioners in assessing whether and how to apply particular bias tests.

Scope

- **More than the bare minimum** 💬 🤖 If releasing a multilingual model, have you tested for bias across multiple languages, beyond English?
- **All of Sesame Street** 🤖 Why are you testing these particular models? Can your test be adapted to other models?

Conceptualisation

- **Tell me what you want (what you really really want)** ♢ What is your desired model outcome, and how does your test allow you to measure deviation from that desired outcome? How does this outcome connect to your harm?

Operationalisation

- **Make the implicit explicit** ✏️ 😶 Why are your chosen terms suitable proxies for the demographic groups you are studying? What is the cultural context to which these terms are relevant?
- **Well-spoken** ✏️ Have you considered the many ways a group identity can manifest linguistically?
- **Don't reinvent the wheel** ✏️ 📏 Did you consider relevant work from linguists and social scientists when designing your bias measures?
- **Broaden your horizons** 👫 Can your work be expanded to further cultural contexts? Is a binary conceptualisation of gender appropriate, or necessary?

Other Validity Considerations

- **Consider the future** Does your test allow us to make predictions about downstream behaviour (predictive validity)?
- **Do a reality check** Does your measurement approach reflect "real world" language and model usage (ecological validity)?
- **Beware of collateral damage** Can your measurement approach cause harm or other impacts (consequential validity)?

7 Conclusion

We hope that via our taxonomy and analysis, practitioners are better-equipped to understand and take advantage of the wealth of emerging approaches to bias testing—in particular, to clearly conceptualise bias and desired model outcomes, design meaningful and useful measurements, and assess the validity and reliability of those measurements.

8 Limitations

Our search was conducted exclusively in English, and we may have missed relevant papers written in other languages; this may have influenced the heavy English skew in our data.

Some of the annotations of attributes and choices in this taxonomy rely on subjective judgements, particularly with regards to the clarity of conceptualisations of bias, desired outcomes, and justifications of proxy choices. As with any qualita-

tive work, these results are influenced by our own perspectives and judgement. We did our best to address this through regular discussion, identifying disagreements early on when designing the taxonomy, and adopting a "generous" approach.

9 Ethics Statement

All measurement approaches discussed in this paper encode implicit assumptions about language and culture, or normative assumptions about what we ought to do, which must be made explicit for them to be properly evaluated. We acknowledge our work will have been shaped by our own cultural experiences, and may similarly encode such assumptions.

Acknowledgements

We would like to thank our anonymous reviewers for their feedback. Eddie L. Ungless is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics.

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546.

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.

Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in measuring bias via open-ended language generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 76–76, Seattle, Washington. Association for Computational Linguistics.

Marion Bartl and Susan Leavy. 2022. Inferring gender: A scalable methodology for gender detection with online lexical databases. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–58, Dublin, Ireland. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing. *Doctoral Dissertations*.

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.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, page 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Donald T. Campbell. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54:297–312.

Rui Cao. 2021. Holistic interpretation in locative alternation – evidence from self-paced reading. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 543–550, Shanghai, China. Association for Computational Lingustics.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

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*, page 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Susan Gass. 2010. Experimental research. *Continuum companion to research methods in applied linguistics*, pages 7–21.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 3356–3369, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 65–83, Online. Association for Computational Linguistics.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 375–385. ArXiv:1912.05511 [cs].

John F. Kihlstrom. 2021. Ecological validity and "ecological validity". *Perspectives on Psychological Science*, 16(2):466–471.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, page 43–53, New Orleans, Louisiana. 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.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. 2022. Pangubot: Efficient generative dialogue pre-training from pre-trained language model. *arXiv preprint arXiv:2203.17090*.

Robert Munro and Alex (Carmen) Morrison. 2020. Detecting independent pronoun bias with partially-synthetic data generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 2011–2017, Online. Association for Computational Linguistics.

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)*, page 1953–1967, Online. Association for Computational Linguistics.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, page 93–106, Barcelona, Spain (Online). Association for Computational Linguistics.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is martha: Using names to measure and mitigate bias in generative dialogue models. (arXiv:2109.03300). ArXiv:2109.03300 [cs].

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, Dublin, Ireland. Association for Computational Linguistics.

Caroline Stone. 2019. A defense and definition of construct validity in psychology. *Philosophy of Science*, 86(5):1250–1261.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Types of Validity

See Table 2.

A.2 Full Taxonomy

We provide here details of our taxonomy (Table 1), including detailed explanations of each option.

Language(s) What language(s) is/are investigated?

Model(s) What model(s) is/are investigated?

Code available? Is code for the proposed bias test publicly available?

- yes/no

Use context What context will the model be used in?

- zero-shot/few-shot
- upstream LM
- dialogue
- Q&A

Bias conceptualisation How is bias—bias, fairness, stereotypes, harm, etc.—conceptualised?

- stereotyping: paper identifies stereotyping as a harm
- toxic content generation: paper identifies negative or toxic (including racist, sexist, etc.) content as a harm
- other: paper identifies something else as a harm (annotator includes description in a comment)
- unclear: it is unclear how the paper conceptualises bias or harm

Prompt task What is the prompt task?

- sequence scoring: model is tasked with scoring various sequences
- single word generation: model is tasked with generating a single word
  - analogy: model is tasked with completing an analogy
- prompt continuation: model is tasked with continuing a prompt (2+ words)
- full sentence response: model is tasked with responding to a full sentence

Prompt origin Where do the prompts originate?

- author: prompts are written by the author, or sourced from a paper where they are written by that paper's authors
- crowd-sourced: prompts are crowd-sourced from workers other than the paper authors, or sourced from a paper where they are crowd-sourced
- corpus: prompts are scraped from a corpus, including Wikipedia or social media, or sourced from a paper where they are scraped from a corpus
- automatically generated: prompts are generated by a model

Metric What metric or strategy is used to measure bias or harm?

- output content assessed: assessment of output content, e.g., presence of stereotypes
- output quality assessed: mentions of demographic groups lead to differences in quality of output content, e.g., grammaticality or relevance
- difference in probability (ranking over fixed set): which of a fixed set of options is more probable
- most probable option(s): assess the top 1 or N options
- difference in output distributions: assessment of entire output distributions under different conditions
- difference in regard: mentions of demographic groups lead to differences in regard of output content
- difference in sentiment: mentions of demographic groups lead to differences in sentiment of output content
- difference in toxicity: mentions of demographic groups lead to differences in toxicity of output content

Desired outcome How is a good model outcome conceptualised?

- no impact of demographic term(s): mentions of demographic groups do not change model predictions.
- negative stereotype not in model: mentions of demographic groups do not result in output reflecting stereotypes
- other: another conceptualisation (annotator includes description in comment)
- unclear: it is unclear how the paper conceptualises a good model outcome

Demographics For which demographic groups is bias or harm investigated?

- gender

Type of Validity	Short Definition	Example Threat
Construct validity		
Face validity	Plausibility	Using BLEU score to measure relevance of generation - BLEU does not measure meaning
Content validity	Effective operationalisation	Paper aims to measure fairness but results not split by demographic, unclear if some groups disproportionately affected
Convergent validity	Correlation with existing measures	Proposed measures rarely compared to existing measures
Predictive validity	Can predict related measurements	Authors assume upstream bias predicts downstream bias; this has not been proven
Consequential validity	Impact on world & behaviours	People may assume low bias in LM will ensure low bias in finetuned model and feel "safe" using these models
Ecological validity	Results generalise to the world	By factoring out confounds on relative probabilities, measurement does not reflect typical use of model

Table 2: Overview of threats to validity. Each threat is derived from examples found in our analysis.

- ethnicity/race
- religion
- sexual orientation
- other: other demographic groups (annotator includes description in comment)

Proxy type(s) Which term(s) is/are used to proxy the demographic groups under investigation?

- identity terms: terms that refer directly to demographic groups, such as *Muslim*
- pronouns
- names: people's names
- roles: terms that refer to social roles, such as *mother*
- dialect features: terms reflecting dialectal variation, such as lexical items associated with African American Language (AAL)
- other: other terms (annotator includes description in comment)
- unclear: it is unclear what terms are used

Explicit demographics Are the choices of demographic groups and accompanying proxies clearly defined and explained?

- yes/no

Gender scope For work investigating gender, how is gender treated?

- binary gender only: gender is treated as binary, specifically man and woman, or male and female
- binary gender only plus acknowledgement: gender is treated as binary, accompanied by an acknowledgement that gender is not binary
- binary and other genders: gender treatment includes men, women and other marginalised genders
- other genders only: gender treatment excludes binary genders

A.3 Results from Taxonomy for Multilingual and Non-English Bias Tests

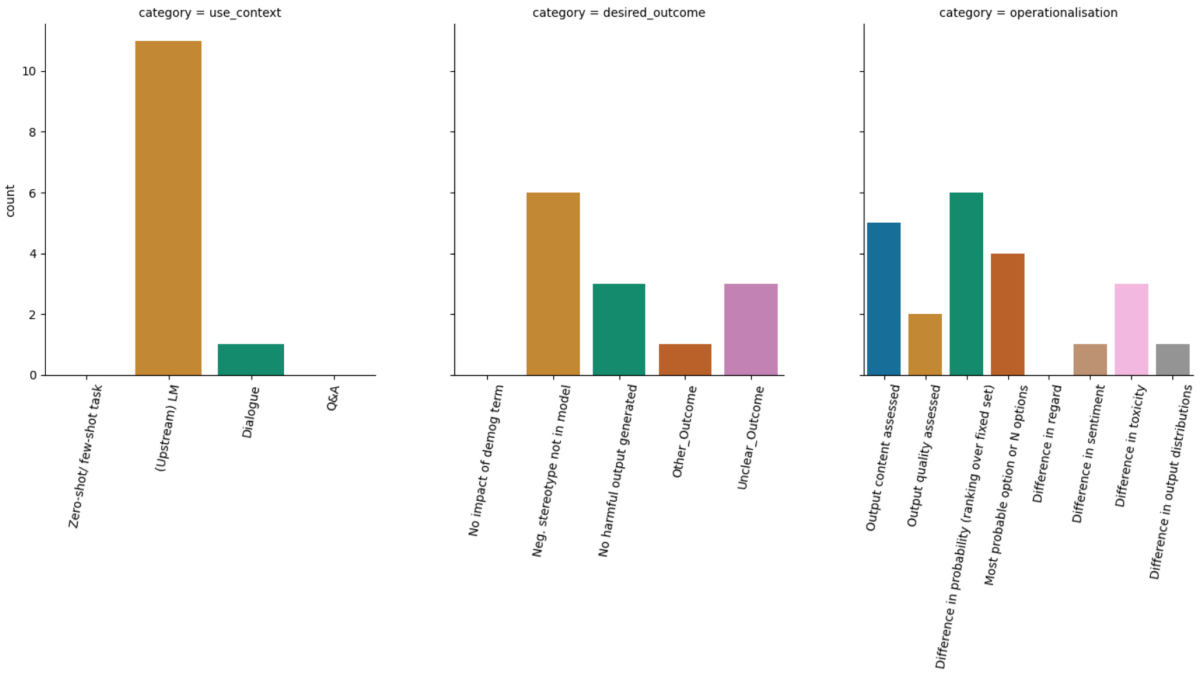


Figure 2: The same as Table 1, isolated to the 12 multilingual bias tests to show the patterns there that differ from overall ones.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. 2

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.