Statistical Bias in Bias Benchmark Design

Hannah Powers
Rensselaer Polytechnic Inst.
powerh@rpi.edu

Ioana Baldini, Dennis Wei
IBM Research
{ioana,dwei}@us.ibm.com

Kristin P. Bennett
Rensselaer Polytechnic Inst.
bennek@rpi.edu

Abstract

Social bias benchmarks lack a consistent framework to standardize practices. We advocate for an experimental approach inspired by health informatics. Current work overlooks statistical biases in the benchmark which cause inaccurate conclusions in the benchmark analysis when unaccounted for. We recommend researchers be aware of the potential for statistical biases during benchmark design and analysis. We demonstrate the importance of formalizing explanatory factors and give examples of the presence of statistical biases and their possible effects with BBQ.

1 Introduction

Although important, existing social bias benchmarks are insular and lack a consistent framework to standardize practices. Most benchmarks are designed to test whether language models (LMs) demonstrate a given subset of social biases. Although well motivated, current work often overlooks statistical biases in the benchmark, such as sampling bias and omitted variable bias. When unaccounted for, these can cause inaccurate conclusions in the benchmark result analysis. Potential problems with LMs may be missed. We propose the use of systematic experimental design (ED) for benchmark design. We call for researchers to identify and correct for statistical biases in both design and analysis. We demonstrate the potential of formally defining factors using the Bias Benchmark for QA (BBQ) [4] and use these factors to identify potential cases of statistical bias in the BBQ dataset.

We advocate considering benchmark creation as a multi-factor problem. Benchmarks should clearly define all factors they investigate or vary, even if they are not the intended focus. They should indicate the variations and combinations used of these factors. For example, a dataset with male and female subjects but only prompts with male subjects vary by ethnicity. For a well-founded analysis, datasets should include a reference value for all factors to which all other variations may be compared. References serve as a control or baseline for each factor during analysis. When investigating social bias, the use of a reference prompt, such as one that includes no social characteristics, gives us an understanding of the LM's baseline response in that scenario. The benchmark should also account for other factors that may influence a response. An analysis should first aim to discover bias and only confirm what has already been discovered, not what is assumed.

We support use of ED inspired by clinical trials (CT). ED for CTs clearly defines the outcomes, interventions, confounding factors and involved study cohorts. They use *experimental factors* to characterize subjects, analyze the results of study populations, and assess the study validity. CTs always include a table analyzing baseline factors [1]. Similarily, benchmarks should define the factors and outcomes of interest to show scope and validity. This includes clarifying the choice and combination of values used and identifying where confounding or other biases may be introduced by the choices made. ED leads to effective use and analysis of factors. The distribution of prompts with respect to factors must be designed and reported to avoid intentional and unintentional bias and ensure sufficient power to discover biases. LM bencmarks rarely perform these steps. The actual coverage of every benchmark should be reported. Benchmarks are intended to identify which types of prompt exhibit bias and potential causes of bias. Ad-hoc benchmark design can introduce statistical biases, i.e. errors in the analysis that result from design or analysis flaws. Many benchmarks only consider

Workshop on Evaluating Evaluations (EvalEval 2024).

Name	Variations
stereotyped group non-stereotyped group	10 ethnicities 16 ethnicites
stereotype	20 stereotypes
gender	male, female, mixed, **neutral**
question polarity	**non-negative**, negative
context ambiguity	**ambig.**, disambig., no context
proper nouns only	true, **false**
prompt format	ARC, RACE, **QONLY**

Table 1: Explanatory factors of BBQ Race/Ethnicity benchmark. Potential reference values given in bold.

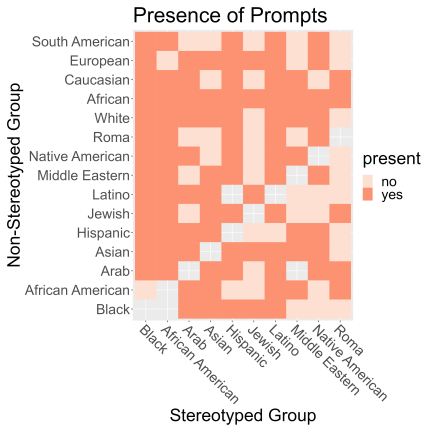

Figure 1: Coverage of BBQ with respect to stereotyped and non-stereotyped groups. Present means a prompt exists within those groups. .

one or two factors, contributing to omitted variable bias in which analyses incorrectly attribute the effect to another factor when excluding a relevant variable [6]. Confounding can come from having an association between the confounding variable and a variable of interest [5]. For example in [3], an anticipated biased response is fixed for a given template, making it challenging to attribute causes.

2 Factors and Biases of BBQ

We demonstrate ED and how to look for statistical biases with BBQ race/ethnicity, a benchmark that assesses racial bias through question-answering templates. We first identify benchmark factors. A factor is defined by its name and potential values, including a reference/control value. The selection of factors is vital for identifying causes of bias. Table 1 contains a list of factors for BBQ including ones missed in initial study. A prompt uses a scenario involving a stereotype with an expected "stereotyped group" and "non-stereotyped group" which are explicitly chosen because of known and referenced biases humans have. Prompt variations are created by using additional question variations.

Inadequate coverage of prompt subgroups can result in mismeasurement and misidentification of LM bias. Figure 1 shows stereotyped and non-stereotyped groups explored by BBQ. Frequently, only groups with suspected bias are compared. Note, Roma is only compared against a few other ethnicities. An analysis may state that including Roma in a prompt has some effect on an LM's response, but maybe the groups which Roma is paired with receive bias in favor of them, resulting in an analysis overstating the effect of Roma's inclusion on a response. Although it is crucial to explore the stereotypes that are considered harmful [2], we must also discover new potential biases. BBQ also has issues with omitted variables. Many prompts are gendered, but gender was not initially considered as a factor. For example, prompts were created to confirm stereotypes around African-American subjects. BBQ's analysis found bias against this group, but failed to account for over half of the nearly two thousand prompts having *male* subjects. Excluding gender from the analysis may have resulted in confounding the bias towards African-American individuals due to gender effects.

3 Conclusion

We show the importance of formalizing explanatory factors and reporting their coverage using the BBQ benchmark. Using ED improves transparency of benchmarks, allowing better understanding of the validity and scope of these studies. Acknowledging and resolving statistical biases in benchmark design and analysis will further improve the validity of claims made on benchmark results, improving performance of and trust in LMs. We must be aware of statistical biases in benchmarking. We encourage research in the standardization of bias benchmark design and reporting, and caution researchers to be mindful of design biases. We advocate for principled design and reporting approaches with a guide for avoiding and accounting for statistical biases.

References

[1] Clinicaltrials.gov. `http://https://clinicaltrials.gov/`.

[2] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *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, Aug. 2021. Association for Computational Linguistics.

[3] M. Nagireddy, L. Chiazor, M. Singh, and I. Baldini. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. *arXiv preprint arXiv:2312.07492*, 2023.

[4] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman. BBQ: A hand-built bias benchmark for question answering. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] A. Skelly, J. Dettori, and E. Brodt. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, 3:9–12, 02 2012.

[6] R. Wilms, E. Mäthner, L. Winnen, and R. Lanwehr. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5:100075, 2021.