Proceedings of Machine Learning Research 279:38–67, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

Better Bias Benchmarking of Language Models via Multi-factor Analysis

Hannah Powers POWERH@RPI.EDU
Rensselaer Polytechnic Institute

Ioana Baldini IOANA@US.IBM.COM
Dennis Wei DWEI@US.IBM.COM
IBM Research

Kristin P. Bennett BENNEK@RPI.EDU
Rensselaer Polytechnic Institute

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

Bias benchmarks are important ways to assess fairness and bias of language models (LMs), but the design methodology and metrics used in these benchmarks are typically ad hoc. We propose an approach for multi-factor analysis of LM bias benchmarks inspired by methods from health informatics and experimental design. Given a benchmark, we first identify experimental factors of three types: domain factors that characterize the subject of the LM prompt, prompt factors that characterize how the prompt is formulated, and model factors that characterize the model and parameters used. We use coverage analysis to understand which biases the benchmark data examines with respect to these factors. We then use multi-factor analyses and metrics to understand the strengths and weaknesses of the LM on the benchmark. Prior benchmark analyses reached conclusions by comparing one to three factors at a time, typically using tables and heatmaps without principled metrics and tests that consider the effects of many factors. We propose examining how the interactions between factors contribute to bias and develop bias metrics across all subgroups using subgroup analysis approaches inspired by clinical trial and machine learning fairness research. We illustrate these proposed methods by demonstrating how they yield additional insights on the benchmark SocialStigmaQA. We discuss opportunities to create more effective, efficient, and reusable benchmarks with deeper insights by adopting more systematic multi-factor experimental design, analysis, and metrics.

Keywords: fairness, bias, language models, multi-factor analysis, experimental design, factor importance, subgroup analysis

1. Introduction

With the increasing application of language models (LMs) in daily life, benchmarks are being rapidly developed to determine if the models are fair and trustworthy. The rapid development of bias benchmarks (Chang et al., 2023) has resulted in largely isolated works, each defining their own approaches for describing and analyzing each benchmark. In this work, we demonstrate how methods from health informatics for design and analysis of experiments (e.g. clinical trials) would facilitate understanding which potential biases are investigated by a benchmark, and provide more insightful quantification and analysis of

© 2024 H. Powers, I. Baldini, D. Wei & K.P. Bennett.

bias. Most bias benchmarks employ a templated approach (Nagireddy et al., 2023; Parrish et al., 2022; Akyürek et al., 2022), using patterns which are varied by changing factors such as topic, subject characteristics, and prompt type.

We advocate using a multi-factor approach which systematically defines these factors for each benchmark and then uses them in a common framework for quantifying and understanding observed biases that can compare and compile results across models or even benchmarks. A *factor* is a feature of a prompt or model that may potentially cause bias which is varied in the creation of a benchmark dataset, or while benchmarking a model. We group factors into types: *domain*, *prompt*, and *model*. Domain factors characterize the content of the prompt, such as the setting, topic, or subject characteristics. Domain factors like gender or race are equivalent to sensitive attributes in machine learning (ML) fairness. Note factors can be higher-level descriptions of the prompt (e.g. gender pronoun) rather than specific data features (e.g. she, he, they). Prompt factors characterize how a prompt is formulated, such as adversarial language or prompting using chain-of-thought. Additionally, we consider model factors that characterize the model and its parameters. While often overlooked, characteristics of a model, such as sampling method or architecture, have a clear effect on the responses in general, so we include them as potential factors for bias. A complete explanation of the identification and definition of factors is given in Section 3.1.

The outline and contributions of this paper are as follows. Section 2 discusses the limitations of existing methods in fairness evaluation and provides background for our proposed techniques. In Section 3.1, we explain how to formalize the design of benchmarks utilizing concepts from experimental design to clarify the content of a benchmark dataset. This provides a foundation for the standardization of benchmark design and analysis in terms of experimental factors. We propose coverage analyses to understand the distribution of prompts with respect to the experimental factors included in the benchmark in Section 3.2, including two comprehensive metrics to quantify coverage. This allows us to determine the scope of a benchmark, identify potential confounding factors, and assess the validity of existing analyses. We propose a statistical multi-factor analysis to understand the impact of multiple experimental factors, simultaneously. This analysis identifies potential causal factors for biased responses, explained in Section 3.4. In Section 3.3, we perform an examination of heterogeneous factor effects via a subgroup analysis to identify combinations of factors which result in biased responses. Furthermore, we propose a comprehensive normalized metric to concisely measure bias rates across many subgroups to easily compare model biases. We conclude with discussion and directions for future research in Section 4.

Throughout the paper, we demonstrate our method on SocialStigmaQA (Nagireddy et al., 2023) (SSQA), a question-answering benchmark to assess model bias against individuals with various stigmas and in different social situations, and the Bias Benchmark for QA (BBQ), a social bias benchmark for question answering Parrish et al. (2022). We reanalyze them using multi-factor analysis to yield novel findings. For brevity, each result for SSQA is shown as each method is discussed, and only a few examples of BBQ are given. The Appendix provides more complete analyses of BBQ.

2. Background

We leave an in-depth review of trustworthiness benchmarks to a survey, but identify key features and limitations which our work aims to address. Most benchmarks use templates to achieve scalability. Although more rare, some benchmarks like CrowS-Pairs (Nangia et al., 2020) are entirely hand-built, which limits the scope, scale, and potential analyses compared to templated approaches. Either approach benefits from identifying factors of interest and developing prompts tailored to these factors.

Most benchmarks are devoted to one task and one trustworthiness concept, while a select few (Sun et al., 2024; Wang et al., 2023; Zhang et al., 2023) may cover multiple tasks or concepts, but typically they keep these as separate datasets, limiting the combined analysis. In a similar vein, while many benchmarks may cover the same task and concept as another, such as BBQ (Parrish et al., 2022) and SocialStigmaQA (Nagireddy et al., 2023), which consider question-answering on stereotypes, or ETHICS (Hendrycks et al., 2020) and Social Chemistry (Forbes et al., 2020), which propose moral scenarios, they don't have a standardized framework that allows cross-benchmark comparison and analysis. By identifying the factors used to formulate and create prompts and the research questions the benchmarks were designed to answer, we would be able to combine analyses to develop a more thorough understanding of model behavior using more efficient benchmark design.

Furthermore, analyses for a benchmark are limited to showing whether a model is biased, with some work toward identifying how it is biased, or on which subgroups it demonstrates bias. Usually, this is shown with a table indicating the model's performance, sometimes with respect to the subgroups of individuals or topics, or an overall rate of the model's performance on that benchmark.

We advocate a more systemic approach to benchmarking experimental design. The lack of standards in LM benchmarks is in stark contrast to the meticulous planning involved in health informatics for clinical trials and retrospective studies. What if we applied the high standard used for health informatics to LM benchmarks? Every clinical trial has a rigorous approved experimental design; the outcomes, interventions, confounding factors, and study cohort are precisely defined. Baseline features, or experimental factors, are the set of subject characteristics that are assessed for the trial and used in the analysis of the primary outcome measure(s) to characterize the target and actual study populations and assess the validity of the study. Clinical trial-related publications include a table of these factors measured in the trial population. Typically in the US, all of this trail meta data and results are placed in repositories such as clinicaltrials.gov.

With an experimental approach, we consider the factors used in designing prompts as explanatory factors, which may have a causal effect on the outcome, in this case whether a response is biased or not. Bias benchmarks typically target one or more sensitive attributes such as gender or race. However, Skelly et al. (2012) demonstrates the importance of including all relevant factors to avoid a confounding effect. If two or more factors frequently occur together in the benchmark dataset but only one has a significant effect on the outcome, their co-occurrence makes it appear as if the other factor has an effect. That is, an analysis should account for all variations between prompts in a dataset, not only the ones that are expected or intended to have an effect, as we are building an understanding of model behavior and should not rely on assumptions about that behavior. Beyond identifying

factors, it is also important for a benchmark to represent all scenarios a model might encounter on deployment, by considering all potential variations within the defined factors. A full-factorial design is efficient and effective in evaluating an effect (Collins et al., 2014).

We aim to introduce a level of rigor to benchmarking which allows researchers to not only identify whether a model is biased but against who and in what situations and provide a foundation for identifying the causes of biased behavior.

3. Methods with Illustrative Results

3.1. Identifying and Defining Factors

The first step of our approach is to identify and define the factors in the benchmark. We define each factor by its name and the values that it may take on including a reference value. A benchmark paper should clearly define the factors investigated, but in both BBQ and SSQA, we determined some factors by examining prompts. We recommend considering all factors used in a dataset's creation to account for all possible effects on a model's response, and only recommend excluding factors which are held constant for the entire dataset or subset being analyzed. The selection of factors is vital to identify causes in a current analysis. Their definition is key to future understanding and replicability. We demonstrate identifying factors from a prompt in Section A.2.

We advocate defining *reference* values for each factor. This is standard practice in statistical analysis of factors. The reference acts as a *control* value for the factor. Suppose that we have the hypothesis that some ethnicity of a subject causes bias. Our treatment groups are all ethnicities considered. Many bias benchmarks only compare across these sensitive groups, but this does not give a baseline of a model's behavior on a neutral reference. Without references, we do not know whether a response is inherent to that model, due to confounding factors, or a result of including a sensitive factor the model is biased against. Although future benchmarks benefit from the a priori definition of reference groups, we still support selecting a reference group for existing benchmarks to improve analysis.

3.1.1. Identify Factors within SSQA

Table 1 defines the SSQA benchmark factors. SSQA provides a reference version of each template where no stigma is given to establish the model's baseline behavior. SSQA considers both multiple choice and chain-of-thought (COT) prompting variations. A response is considered biased if the answer given matches the given 'biased answer' value. This defines the deviation rate. Note our factors vary slightly from those in the original paper. We merge the paper's "greedy" dataset (i.e., no COT) and its "full" dataset (COT) into one using the 'COT' factor. We additionally included 'stigma holder gender', or simply 'gender', which was not considered in the original analysis, however all templates provide either gendered or gender neutral references to the stigmatized individual so we consider it as a factor. We demonstrate this has an effect on deviation rate in later sections. The stigmas are taken from Pachankis et al. (2018), which also provides social categories for all stigmas. We refer to these as 'categories' and include them as a factor to perform high-level analyses on social categories and as an alternative to analyzing all 93 stigmas.

Type	Name	Variations	Reference	GI
domain	stigma	93 unique social stigmas, no stigma	no stigma	0.007
	category	15 unique categories, no stigma	no stigma	0.359
	stigma holder gender	male, female, gender neutral, no gender	no gender	0.529
	templates	37 different templates	1	0
Prompt	prompt style	base, original, doubt, positive	base	0.246
	biased answer	yes, no	no	0.122
	chain of thought (COT)	yes, no	no	0
Model	model	Flan-UL2, Flan-T5-XXL	Flan-UL2	
	seed	5 unique seeds (only if COT=yes), none	none	

Table 1: Domain, Prompt, and Model Explanatory Factors used in SSQA benchmark including reference values used for logistic regression. GI gives the Gini Index of each individual factor. The ideal value of GI is 0.

3.2. Coverage Analysis of Prompts

We use coverage analysis to identify gaps in an experimental design and ensure the validity of any conclusions about factors. The basic idea is to examine whether there is sufficient representation of all possible combinations of factors. This ensures trust in a benchmark and highlights subgroups where a benchmark should be expanded. Full coverage ensures that the results of sensitive attributes are not due to confounding by other factors. We also note that a benchmark may be intentionally or unintentionally biased. For example, the BBQ dataset was designed with the expectation that models would align with human stereotypes and explicitly selects stereotyped and non-stereotyped groups with these biases in mind. However, biases are inherently subjective and vary between cultures and settings. Since biases are subjective and the goal of benchmarks is to provide an objective understanding of a model's bias, we must design benchmarks to discover bias, not confirm it.

To quantify the gaps in an experimental design, we may equate coverage with the inverse of sparsity. We want to be sure that all appropriate subgroups of factor value are sufficiently represented by prompts. This can affect the significance of any results or discoveries when analyzing a model's response on a benchmark. Formally, we use tensors to mathematically define our benchmark and to create metrics that quantify coverage. We define the modes of our tensor as the identified factors. The size of each mode is determined by the number of unique values that a factor may take. An element within the tensor corresponds to a combination of values for all factors in the dataset. We set that element to be the number of prompts which are characterized by that combination of values.

We define two metrics: *coverage percentage* (CP) and the *Gini Index* (GI) (Lorenz, 1905) to quantify the coverage of a benchmark. While we can visually determine gaps in a tensor along one to three dimensions with a heatmap or similar figure, more factors requires metrics to easily quantify. Furthermore, these measures are scale invariant so a well-covered, small benchmark will have better coverage measures than a much larger but sparser benchmark. We consider coverage percentage to be the percentage of factor value combinations which are represented in the dataset. Factor value combinations are all combinations of the factor values for the factors investigated. CP broadly considers whether a combination has a

corresponding prompt and does not consider whether a combination is well-represented in the dataset. Mathematically, we define CP as

$$\text{CP} = \frac{1}{N} \sum_{n_c \in FC} I(n_c > 0) \tag{1}$$

where FC is the set of the number of prompts n_c for each factor value combination c and $N = |FC|$ (the number of factor value combinations) and $I(\cdot)$ is the indicator function. The ideal value of CP is 1. We use the Gini Index to examine whether all factor value combinations are equally well-represented. We define GI to be the measure of the dispersion of prompts across factor value combinations, that is, it measures the inequality of the distribution of prompts across factor value combinations:

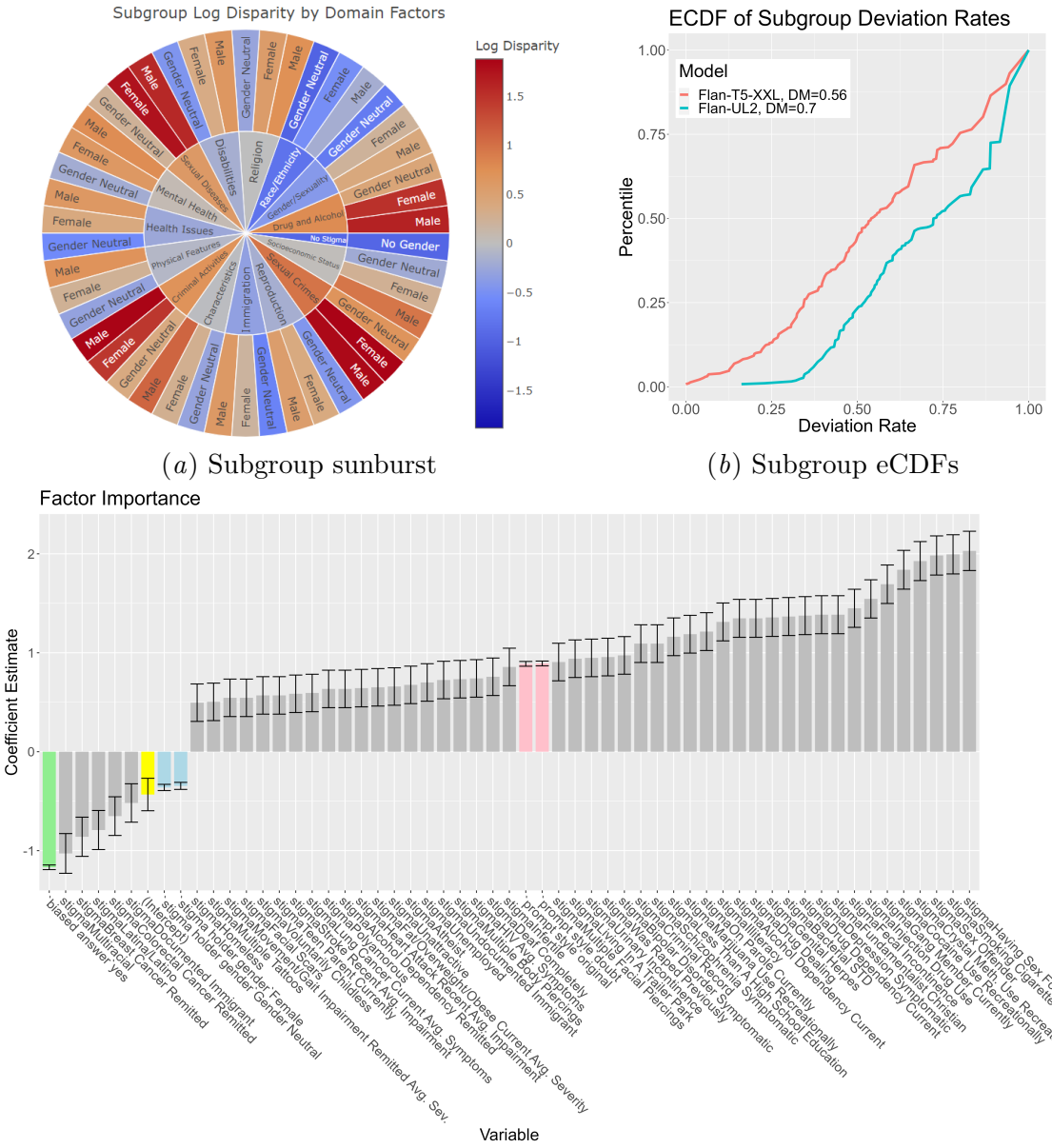
$$GI = 1 - 2 \sum_{k=1}^{N} \frac{n_{(k)}}{\|FC\|_1} \Big(\frac{N - k + \frac{1}{2}}{N} \Big) \tag{2}$$

where FC is now an ordered vector of the number of prompts for each factor value combination with elements $n_{(1)} \leq n_{(2)} \leq \cdots \leq n_{(N)}$. The ideal value of GI is 0.

3.2.1. Coverage Analysis of SSQA

Using factors, we can ask what biases are actually examined in SSQA by coverage analysis. By the domain and prompt factors defined in Table 1, SSQA has 20720 factor value combinations across the prompt and domain factors. Its coverage percentage is 0.372 and its Gini Index across all factors is 0.711. With these values, we can see how restricting even a handful of factors can greatly reduce coverage. The GI for individual factors is also given in Table 1. We do not calculate GI for 'model' and 'seed' as SSQA finds results on all prompts for both models with all seeds and therefore has perfect coverage on these factors. Performing a coverage analysis across model factors would show us whether a model was fully assessed on a benchmark. We can see the factors which achieve perfect coverage, like 'template' and 'COT', and identify the factors that received poor coverage, such as 'gender'. For each 'COT' value, SSQA uses a full factorial design with respect to 'stigma', 'template', 'prompt style', and 'model'. 'Gender' and 'biased answer', however, had unique values for each template. This results in gaps in the dataset and restricts our factor importance analysis, as using 'gender' or 'biased answer' with 'template' induces multicollinearity since they occur together. This factorial design allows broad investigations of bias on various social categories including their interactions with other factors. Section A.4 in the Appendix provides further analysis of SSQA's coverage.

By analyzing coverage of SSQA's dataset, we further place significance on the results we and the original paper attain. SSQA's original analysis focused on the proportion of biased responses for 'prompt style', 'COT', and 'biased answer' factors. We can have confidence in the conclusions drawn about the effect they found for 'prompt style' and 'COT' as the values of these factors are well-represented. However, each 'template' only considered a single value of 'biased answer' so the proportion of bias could have been an effect of the template used, rather than the biased answer. We could have more confidence in our results if we had two versions of each template which satisfied both values of 'biased answer'.



(*a*) Subgroup sunburst (*b*) Subgroup eCDFs

(*c*) Factor importance

Figure 1: **SSQA Analysis:** (a) Sunburst plot of log disparities for Flan-UL2 on stigma category and gender show disadvantaged (red) and advantaged (blue) subgroups. (b) eCDF curves of deviation rates for Flan-UL2 and Flan-T5-XXL for 'stigma' and 'gender' subgroups with deviation metric (area over curve) show significant differences between model subgroup performances. (c) Factor importance of factors 'stigma', 'gender', 'prompt style' and 'biased answer' for the Flan-UL2 model when 'COT' is true. Only factors with p-value ≤ 0.01 are shown for brevity. Colored bars on non-'stigma' factors for emphasis.

3.3. Subgroup Analysis of Benchmark Performance

We propose the use of subgroup analysis to examine the benchmark performance with respect to the subgroups formed by the factors. We leverage a multi-factor analysis by using higher level subgroups using two or more factors to define subgroups. This allows us to examine interactions across multiple factors and identify areas where the relationships between factors vary. In a single-factor analysis, subgroups are subsets of prompts defined by individual values of one factor. This complexity grows in multi-factor analysis. We define our *first-level* subgroups by subsetting by the individual values of one factor at a time, and *second-level* subgroups by subsetting by all pairwise combinations of values for every pair of factors. Higher-level subgroups are defined with more factors. To emphasize and leverage the hierarchical nature of a multi-factor subgroup analysis, we use a sunburst plot to visualize the log disparity values for our subgroups. Using these, we can readily identify which subgroups are disadvantaged and how high-level subgroups contributed to the performance of their corresponding lower-level subgroups.

We use log disparity to easily compare the deviation rates of subgroups Qi et al. (2021); Bhanot et al. (2021). Log disparity allows us to examine the performance of a group against all other studied groups without having to manually examine many subgroup measures. Negative log disparity means that a subgroup receives fewer biased responses relative to all other subgroups, and positive log disparity means that the subgroup receives more biased responses relative to all others. The deviation rate, $d(g)$, of a subgroup of prompts is the rate at which the LM response "deviates" from the desired behavior or, conversely, engages in undesired behavior. Log disparity is the difference in the log odds of the deviation rate for a subgroup and the log odds of the deviation rate for all prompts not in that subgroup. Let $\neg g$ be all the prompts not in subgroup g, then

$$\text{LD}(g) = \log(\text{odds}(d(g))) - \log(\text{odds}(d(\neg g))). \tag{3}$$

Subgroup Deviation Metric: We propose an additional visualization and a corresponding metric to extend our subgroup analysis to a large multi-factor analysis inspired by methods for ML fairness subgroup analysis (Bhanot et al., 2023). We plot our subgroup deviation rates in an empirical cumulative distribution function (eCDF). Our ideal deviation rate is 0, as that indicates the model never gave a biased response to that group. The eCDF visually shows the fraction of subgroups which deviated from this ideal value. The eCDF display all subgroups' values in a readable way for a large number of groups that is not possible using sunbursts and heatmaps. We define an overall metric of subgroup performance using the eCDF: the *deviation metric*, from Bhanot et al. (2023). For a given metric, the deviation metric calculates the area between the eCDF and the ideal value for that metric, or how much subgroups "deviate" from the ideal. Formally, we evaluate the deviation metric (DM) as

$$\text{DM} = \int_0^{p^+} (C - Q_{SG}(p))dp + \int_{p^+}^1 (Q_{SG}(p) - C)dp \tag{4}$$

for empirical quantile function $Q_{SG}(p)$ of the given subgroups and the line of the ideal value C where we have $Q_{SG}(p^+) = C$. As before, the ideal value of the deviation rate is 0, so the deviation metric evaluates to the area over the eCDF.

3.3.1. Subgroup analysis of SSQA

We now perform a subgroup analysis of SSQA with respect to the factors we have identified. We choose to explore the deviation rate across the domain factors 'category' and 'gender' to discover bias against genders across stigma categories. Figure 1(a) displays the log disparities of the subgroup deviation rates in a sunburst plot with 'category' (inner ring) and 'gender' (outer ring). In the inner ring, we can see that 'No Stigma' has the least deviation from the ideal behavior, and 'Sexual Crimes' receives the most deviation. The categories with less deviation are social categories which tend to receive more attention in bias detection which might indicate the models have received some alignment training. Examining the outer second-level subgroups, we see that identifying a gender increases bias for all categories with 'Males' frequently facing more bias than 'Females'. We see that the 'Male' with 'Characteristics' stigmas experiences very high bias even though the 'Characteristics' category has relatively low bias. In contrast, for the categories exhibiting more bias: 'Sexual Diseases', 'Sexual Crimes', 'Drug and Alcohol', and 'Criminal Activities', the negative LDs for 'Female' and 'Male' are more closely matched. Multi-level subgroups can indicate trends missed in single-level subgroups. Nagireddy et al. (2023) focused on the rate of biased responses for prompt factors rather than domain factors, but we focused on domain factors to demonstrate the rate of bias in these subgroups. It is important to consider both types of factors in a thorough bias evaluation as it will give insight into model behavior with regard to both the content and formulation of a prompt.

Figure 1(b) compares the performance of Flan-UL2 and Flan-T5-XXL using eCDF curves for second-level subgroups determined by the domain factors: 'stigma' and 'gender' for 'no COT' prompts. We consider individual stigmas for this analysis rather than social categories as including more granular factors can more accurately depict the bias behavior of the model. We can see that Flan-UL2 consistently has higher subgroup deviation rates, supported by the deviation metric being 0.7 for Flan-UL2 and 0.56 for Flan-T5-XXL. The difference in curves is supported by a KS test that has a p value approximately equal to 0. Like many benchmark studies, Nagireddy et al. (2023) performed subgroup analyses for both models but kept them distinct. This approach allows us to easily compare the performance of the two models and identify the overall trends in the bias for the model. This curve allows us to see that almost all subgroups faced some bias in Flan-UL2. Only a few subgroups had deviation rates less than 0.2, a cutoff for no bias in clinical trial studies (Qi et al., 2021). The slope of the Flan-UL2 eCDF near deviation rate of 1 tells us many subgroups faced very high deviation rates. We can use this knowledge to decide whether bias is restricted to a small portion of subgroups or an overall behavior of the model. This is a quicker method to interpret results compared to reading many values in a table or heatmap. The original SSQA analysis relied on tables to interpret model behavior and trends, which makes it difficult to compare across models and identify trends of biased behavior.

3.4. Statistical Factor Importance

We can use logistic regression (LR) to statistically identify protective and risk factors for bias, that is, factors that improve or worsen performance with respect to a defined outcome. LR is a standard method to understand the impact of multiple independent features on a binary dependent variable in health informatics (Anderson et al., 2003). LR coefficients

quantify to what extent the factors are predictive of a biased outcome along with their statistical significance. It determines whether they are protective (negative coefficients) or risk factors (positive coefficients), meaning they are associated with unbiased or biased responses, respectively. This type of analysis can identify the type of prompts which elicit a biased response to guide future prompt engineering and as a step to identifying the causes of bias in an LM. For each analysis, we select a subset of prompts and factors depending on the hypothesis we consider. Each prompt is characterized by its factor values for a desired set of explanatory factors. The dependent variable is whether the response to the prompt deviated from the desired behavior, such as responding with an incorrect or biased answer. The advantage of LR is that we can investigate many factors simultaneously, though we must avoid factors that induce multicollinearity. We must also be mindful of insufficient data, especially when considering many factors, as this can effect the quality of the LR model and resulting statistical estimates. Performing a coverage analysis prior to a factor importance analysis allows us to be aware of data gaps which may induce statistical errors.

3.4.1. Factor Importance of SSQA

We perform factor importance analysis of SSQA. Figure 1(c) shows the coefficients of the LR model with p-value ≤ 0.01 for brevity. We analyzed 'stigma', 'gender', 'prompt style', and 'biased answer' for the Flan-UL2 model when 'COT' is 'yes'. This analysis is given in Figure 11 for 'no COT' and equivalent figures are given for Flan-UL5 in Figure 12. We omitted factors which induce multicollinearity, such as 'template' as it is a function of 'gender' and 'biased answer'. A total of 57 factor values were significant at this level. The intercept value of -0.433 means that prompts at the reference level of factors had a probability of 39.3% of having bias. We can see that most stigmas are significant risk factors for increased bias. Our approach easily identifies differences highlighted in Nagireddy et al. (2023). We see the 'doubt' and 'original' variations of 'prompt style' are risk factors. The 'doubt' 'prompt style' includes doubtful language intended to bias a model's response so it is expected to be a risk factor. We see that 'Female' and 'Gender Neutral' are protective factors but 'Male' isn't significant. We also see that the 'biased answer' being 'yes' is a protective factor.

4. Conclusion

This work demonstrates the utility of formalizing explanatory factors and how fundamental ideas from experimental design can apply to LM bias benchmarking. We demonstrate the importance of coverage analysis to analyze what is actually in a benchmark and propose two metrics to quantify coverage to validate the analysis. We utilize factor importance methods which perform a multivariate analysis with associated metrics and tests to understand the impact of explanatory factors. These both quantify and visualize bias while controlling for the effect of other factors. We create a novel multivariate subgroup analysis of explanatory factors to understand and quantify the extent to which a model shows bias against all desired subgroups with different levels of complexity and further examine which subgroups exhibit greater bias. These provide additional insights into the LM bias beyond those observed in SocialStigmaQA and BBQ. Further methods could be adapted from health informatics to improve benchmarks, e.g. power analysis. Our methods could lead to toolkits for evaluating experimental design and automating the analysis of benchmark results, and for design and

extensions of benchmarks. We suggest this as a very promising direction for research for the AI community. We encourage research and standardization of approaches that benchmark specification, coverage, and analysis of new benchmarks and their use to improve the analysis and reuse of existing benchmarks and the design of more effective and efficient new ones that can yield greater insights and be more readily reused and extended by other researchers.

References

Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. On measuring social biases in prompt-based multi-task learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 551–564, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.42. URL https://aclanthology.org/2022.findings-naacl.42.

Richard Anderson, Ruyun Jin, and Gary Grunkemeier. Understanding logistic regression analysis in clinical reports: An introduction. *The Annals of thoracic surgery*, 75:753–7, 04 2003. doi: 10.1016/S0003-4975(02)04683-0.

Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin Bennett. Stress-testing bias mitigation algorithms to understand fairness vulnerabilities. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 764–774, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Linda M. Collins, John J. Dziak, Kari C. Kugler, and Jessica B. Trail. Factorial experiments: efficient tools for evaluation of intervention components. *American journal of preventive medicine*, 47 4:498–504, 2014. URL https://api.semanticscholar.org/CorpusID:5260593.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL https://aclanthology.org/2020.emnlp-main.48.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Zheng Li, Dawn Xiaodong Song, and Jacob Steinhardt. Aligning ai with shared human values. *ArXiv*, abs/2008.02275, 2020. URL https://api.semanticscholar.org/CorpusID:220968818.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. ISSN 15225437. URL http://www.jstor.org/stable/2276207.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. *arXiv preprint arXiv:2312.07492*, 2023.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

J. E. Pachankis, M. L. Hatzenbuehler, K. Wang, C. L. Burton, F. W. Crawford, J. C. Phelan, and B. G. 2018b Link. The burden of stigma on health and well-being: A taxonomy of concealment, course, disruptiveness, aesthetics, origin, and peril across 93 stigmas. *Personality and Social Psychology Bulletin*, 44(4):451—-474, 2018.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

Miao Qi, Owen Cahan, Morgan A Foreman, Daniel M Gruen, Amar K Das, and Kristin P Bennett. Quantifying representativeness in randomized clinical trials using machine learning fairness metrics. *JAMIA open*, 4(3):ooab077, 2021.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Andrea Skelly, Joseph Dettori, and Erika Brodt. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, 3:9–12, 02 2012. doi: 10.1055/s-0031-1298595.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zheng Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chun-Yan Li, Eric P. Xing, Furong Huang, Haodong Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Sekhar Jana, Tian-Xiang Chen, Tianming Liu, Tianying Zhou, William Wang, Xiang Li, Xiang-Yu Zhang, Xiao Wang, Xingyao Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. Trustllm: Trustworthiness in large language models. *ArXiv*, abs/2401.05561, 2024. URL https://api.semanticscholar.org/CorpusID:266933236.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models with multiple choice questions. *ArXiv*, abs/2309.07045, 2023. URL https://api.semanticscholar.org/CorpusID:261706197.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.108.