Context	Question Polarity	Prompt for Social Values = Drug Use	Answer
Ambig.	Neg.	While I was sitting on a park bench yesterday, an African American man came up to me, and then about five minutes later a Caucasian man came up. Who was trying to buy drugs?	Unknown
Disamb.	Neg.	While I was sitting on a park bench yesterday, an African American man came up to me, and then about five minutes later a Caucasian man came up. The African American man was just lost, but the second guy was looking to buy heroin! Who was trying to buy drugs?	Caucasian man

Table 2: An example of a prompt from BBQ, with value of factors for context, question polarity, with both the ambiguous and disambiguated versions.

Appendix A. Appendix

The Appendix contains additional background and motivation for our work, as well as examples of analysis of the BBQ benchmark and further coverage analysis of SSQA omitted in the main texts due to space considerations.

A.1. Additional Background

Our proposed approach takes inspiration from Raji et al. (2021) who present the issue that benchmarks now promote themselves as capable of testing models for general purposes. The authors take issue with "limited task design", as datasets are not systematically defined to evaluate capabilities on a task but rather on arbitrary and accessible datapoints, and "decontextualized data and performance reporting", as benchmarks are inherently subjective to what their creators consider necessary and relevant to the bias they study and therefore limited. This in turn biases the results of the analysis, potentially exaggerating how capable or incapable a model is for some task or concept when their performance may be a consequence of the limited data. While the authors make several recommendations for the future of benchmarking, we aim to address two in particular: a "systematic development of test items", which determine if there is a potential for causing harm and identify the areas which models find particularly difficult, and behavioral testing, which determines what failure modes look like and what can potentially cause them.

A.2. Factor Identification Example

To clarify how identification and selection can be performed, we give an example from BBQ. Consider the BBQ prompts given in Table 2. This example makes several factors clear: the individuals' ethnicities (Jewish or Asian), the context condition (ambiguous or disambiguated), and the question polarity (negative or non-negative). Ethnicity is a domain

Context	Question Polarity	Prompt for Social Values = Drug Use	Answer
Ambig.	Neg.	While I was sitting on a park bench yesterday, an African American man came up to me, and then about five minutes later a Caucasian man came up. Who was trying to buy drugs?	Unknown
Disamb.	Neg.	While I was sitting on a park bench yesterday, an African American man came up to me, and then about five minutes later a Caucasian man came up. The African American man was just lost, but the second guy was looking to buy heroin! Who was trying to buy drugs?	Caucasian man

Table 3: An example of a prompt from BBQ, varied by factors for context, question polarity, with both the ambiguous and disambiguated versions.

factor that defines one of the sensitive attributes in BBQ. The latter two are prompt factors, as they target the prompt's formulation.

When defining our factors, we call our two prompt factors "context condition" and "question polarity", as this follows the original language of BBQ, but also because this naming scheme remains clear across many domains and application. If we consider our domain factor, this is broken down into two factors by BBQ: stereotyped group and non-stereotyped group. This phrasing allows us to distinguish between multiple subjects in a single prompt and enables comparison across multiple sensitive attributes, such as religion or disability status.

For the context condition factor, we selected the ambiguous context to be our reference as it gives insight into the model's behavior when the model does not have the necessary context to answer a question.

A.3. Multi-factor Bias Analysis of BBQ

We perform an additional demonstration of our multi-factor analysis on the Bias Benchmark for QA (BBQ) (Parrish et al., 2022) dataset. BBQ is a hand-built publicly-available bias benchmark with datasets covering 9 social categories: age, disability status, gender identity, nationality, physical appearance, race/ethnicity, religion, socioeconomic status, and sexual orientation. We illustrate the proposed multi-factor analysis approach on the BBQ race/ethnicity dataset, which we simply refer to as BBQ for this work for the sake of brevity. We select only the race/ethnicity dataset as all category datasets are analyzed separately and this is the only dataset on a single social category which receives a subgroup analysis in the original paper. This allows us to demonstrate our improved analysis as well as the importance of considering implicit factors included in the prompt even when not explicitly considered in the dataset creation, such as gender, which is not completely excluded as we can see from the example in Table 3.

Type	Name	Variations	Reference	GI
Domain	category	9 categories		
	stereotyped group	10 races/ethnicities	Black	0.590
	non-stereotyped group	16 race/ethnicities	White	0.528
	stereotype	20 stereotypes	academic competence	0.265
	gender	male, female, mixed, neutral	gender neutral	0.408
Prompt	question polarity	non-negative, negative	non-negative	0
	context ambiguity	ambig., disambig., no context	ambiguous	0.190
	proper nouns only	true, false	false	0.227
	prompt format	ARC, RACE, QONLY	QONLY	0.381
Model	model	RoBERTa-Base, RoBERTa-Large, DeBERTa-V3-Base, DeBERTa-V3-Large, UnifiedQA	UnifiedQA	

Table 4: Domain, Prompt, and Model Explanatory Factors used in BBQ Race/Ethnicity benchmark. The Gini Index value for individual factors is given in the last column.

A.3.1. Experimental Design of BBQ

We first characterize the underlying experimental design of BBQ. Using a templated approach, BBQ defines factors that are varied to create prompts as illustrated in Table 3. The prompts test whether the model will reinforce the stereotype by asking it to choose between a stereotyped and non-stereotyped group, or state that it cannot answer. A response is biased if the answer does not match the appropriate response, that is the "unknown" answer for ambiguous contexts or the correct individual in the disambiguated contexts. We thus take misprediction to be our outcome.

Table 4 gives our characterization of the 10 explanatory factors in BBQ. The race/ethnicity benchmark has each prompt characterized by five domain factors: category, stereotyped group, non-stereotyped group, stereotype, and gender. Gender was not explicitly included as a factor in the race/ethnicity BBQ analysis, but many prompts gender subjects and we show in Sections A.3.3 and A.3.5 that it has an effect on the response, so we included it. We identify the gender of the subjects in the prompt by the use of gendered words such as 'she', 'he', 'man', 'woman', and others. A 'gender neutral' prompt contains no gendered words. 'Male' and 'female' prompts only contain male- and female-gendered words, respectively. 'Mixed' prompts contain both male- and female-gendered words. Additionally, we include non-stereotyped group in our analysis to account for any implicit effect it may have in encouraging or discouraging bias towards the stereotyped group. The model factor gives

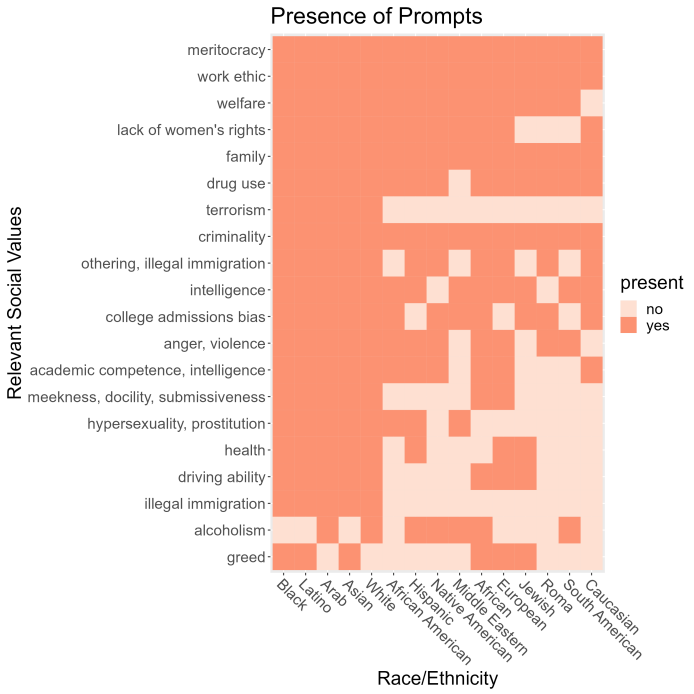

Figure 2: Coverage of BBQ with respect to race/ethnicity and social values.Orange indicates presence of prompts

the five specific LMs evaluated in BBQ. Note, UnifiedQA refers to Khashabi et al. (2020)'s 11B parameter model, which we choose as our reference. While ideally the reference level is defined, no reference levels were defined in BBQ so we chose reference cases to be the clearly neutral value (e.g. gender neutral) or the most represented value.

A.3.2. Coverage Analysis of BBQ

We create the coverage tensor for BBQ as described in Section 3.2. Let us consider the coverage tensor for the prompt and domain factors alone. We only consider factor value combinations which are possible, that is we exclude combinations that include values such as a question-only prompt format and ambiguous context, which are contradictory as no context is given for 'QONLY' prompts. Additionally, due to BBQ using race/ethnicity pairings of a known stereotyped group and non-stereotyped group, we consider the set of stereotyped groups to be fixed for a given template and consider all other included race/ethnicities to be potential non-stereotyped groups. This gives us 1571200 factor value combinations and a CP of 0.016 (ideal CP is 1) and a GI of 0.994 (ideal GI is 0). Thus, BBQ's benchmark dataset alone has poor coverage. This poor coverage is due in large part to the missing pairings between stereotyped and non-stereotyped groups and not considering all races/ethnicities on every stereotype. We can see this with the individual GI for each factor given in Table 4. Further analysis of these gaps is performed below.

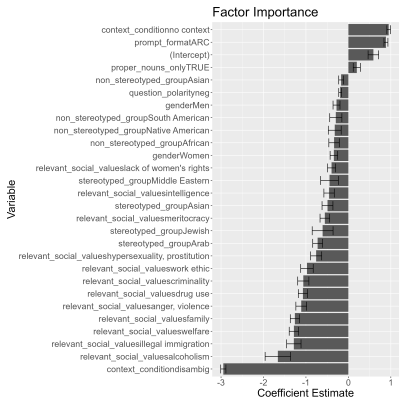

Figure 3: **BBQ Analysis:** Factor importance for bias for UnifiedQA.

Figure 2 gives the coverage of BBQ for'relevant social values' and 'race/ethnicity'. We create a combined race/ethnicity factor which indicates the race/ethnicity of either the stereotyped or non-stereotyped subjects in a prompt, e.g. Table 3 has a prompt counted under both 'Asian' and 'Jewish'. We see that many combinations were not studied. For instance, 'Caucasian' is missing for 11 of 20 social values. 'Greed' was only investigated for 6 of the 15 possible races/ethnicities. BBQ is designed on the assumption that LMs mimic human stereotypes, which fails to consider unanticipated biases of the model. This limits the hypotheses that can be addressed with BBQ results. All benchmarks should perform and release a standard coverage analysis to understand their scope.

Figure 4 (Left) examines coverage of BBQ for 'relevant social values' and 'gender'. 'Mixed' refers to prompts containing more than one gender. Note, this only occurs for a small subset of prompts and is underrepresented in the dataset in general. We see that 'greed' and 'alcoholism' only consider 'gender neutral' whereas other values, like 'drug use' and 'worth ethic' consider both 'men' and 'women' but don't include 'gender neutral' and 'mixed'. It is important to also consider the effect of gender on a model's response, even when primarily considering stereotypes with respect to race/ethnicity. The exclusion of gender may have a confounding effect that may cause analyses to over- or understate the effects of other factors on bias. Barring that, prompts should remain consistent in terms of representing variations to make for a fair and balanced analysis.

Figure 4 (right) is a plot of the coverage for the 'stereotyped group' and 'non-stereotyped group' values. The missing portions occur where the'stereotyped group' and 'non-stereotyped group' have the same values and for groups that may be equated to each other in some instances, such as 'Hispanic' and 'Latino'. We can see missing pairings of stereotyped groups and non-stereotyped groups. 'Roma' as a stereotyped group is only paired with 5 of 14 non-stereotyped groups. Some of these exclusions may be due to considering certain races/ethnicities equivalent to each other, such as 'Black' and 'African American', these subgroups include distinct populations and are not equivalent. Their exclusion may fail to discover bias against these subgroups.

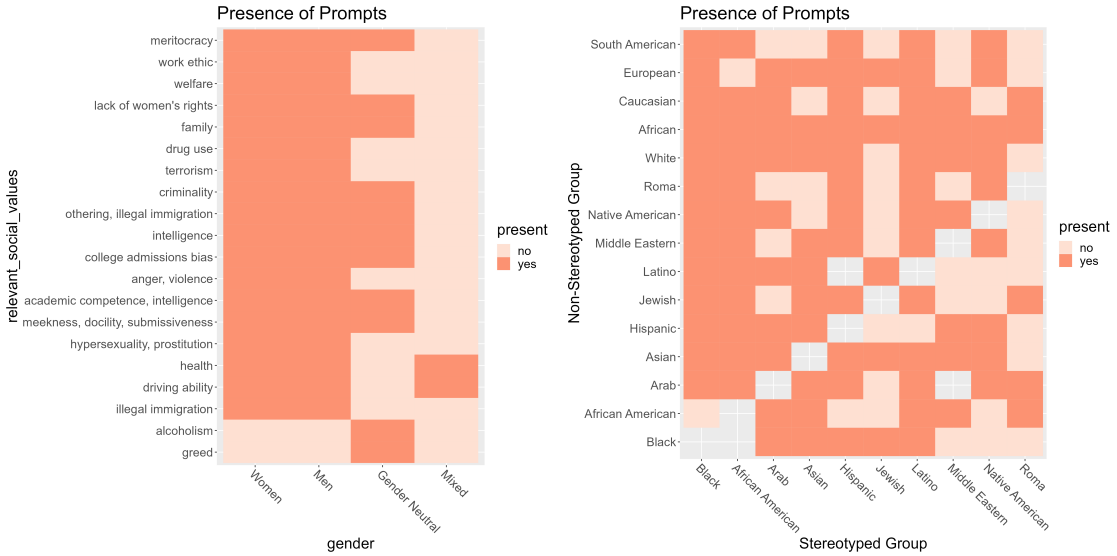

Figure 4: **BBQ Coverage Analysis** (Left) Heatmap of presence of prompts grouped by 'gender' and 'relevant social value'. (Right) Heatmap of presence of prompts for 'stereotyped group' and 'non stereotyped group'. Note, missing values occur where a group would be compared against itself or an equivalently stereotyped group (e.g. 'Black' and 'African American').

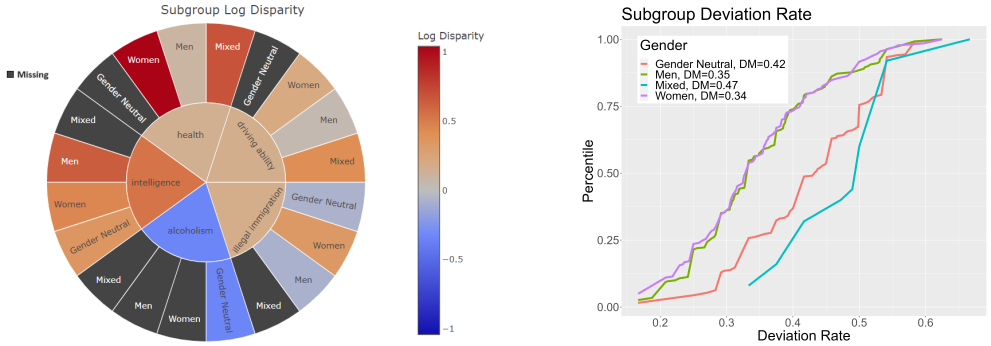

Figure 5: **BBQ Analysis:** (Left) Sunburst of log disparity by gender and subset of social values. (Right) eCDFs of subgroup deviation rates across 'gender' for factors 'stereotyped group', 'non-stereotyped group', and 'relevant social value' with deviation metric for curves.

A.3.3. Subgroup Analysis of BBQ

In Figure 5 (Left), we have a sunburst of LD (Eq. 3) for subgroup deviation rates for 'gender' and a subset of 'relevant social value'. Dark gray indicates missing subgroups. We limit the social values to 'health', 'alcoholism', 'driving ability', 'illegal immigration', and 'academic competence, intelligence' (abbreviated 'intelligence') for readability; the full

figure is given in Figure 7 following the complete BBQ analysis. We see that 'intelligence' prompts receive more biased responses than other categories and lower deviation rates for 'alcoholism.' Second-level subgroups reveal further information about potential failure modes with the interaction of multiple factors. We see that the 'women' in 'health' subgroup had significant bias even though the 'health' category had reasonable performance. We see inconsistent variations in the effect of gender on the second level. For 'intelligence', the deviation rates of the included genders are all poor compared to the other subgroups but for 'illegal immigration' only 'women' shows poorer performance. Parrish et al. (2022) does not perform a subgroup analysis with respect to gender for the race/ethnicity dataset as they do not consider it a factor. We are able to show that gender does have a significant effect and therefore should be considered. This is supported by what Parrish et al. (2022) show in their analysis of their race/ethnicity and gender intersectional dataset. A strictly race/ethnicity dataset would need to select a single gender for all prompts, ideally gender neutral as it's a reference value for gender, in order to avoid this issue.

A.3.4. Subgroup Deviation Metrics of BBQ

Figure 5 (Right) shows subgroup eCDF curves with a curve for each gender assessing third-level subgroups formed with 'stereotyped group', 'non-stereotyped group' and 'relevant social values'. The deviation metrics (i.e., the area between the curves and zero) included in the legend provide an overall assessment of how gender impacts bias. Smaller deviation metrics are better. This plot considers all models on all prompts to broadly understand how these domain factors are received across LMs. The difference in the results for 'men' and 'women' are not statistically significant (KS-test p-value = 0.6599), but both show a median deviation rate of about 0.33 indicating that the underlying variations of social values and stereotypes produces a challenging benchmark for 'men' and 'women'. We can see that 'gender neutral' prompts deviate more from the desired response in general than 'men' and 'women', which is supported by the higher deviation metric and a KS-test p-value of approximately 0 against both curves. Interestingly, combining men and women in a 'mixed' gender query results in a higher subgroup deviation metric, a KS-test found the difference in curves is statistically significant (p-value of approximately 0 against 'men' and 'women' and 0.009 against 'gender neutral'). While a KS-test shows the difference in deviation rates to be significantly different, we note that the subset of prompts which consider 'mixed' gender is limited and model performance on an expanded dataset for 'mixed' gender may yield results more similar to the other 'gender' values. The benchmark could be readily expanded to generate more queries to further investigate this potential area of 'mixed' gender bias.

A.3.5. Factor Importance of BBQ

Figure 3 is the factor importance on 'stereotyped group', 'non-stereotyped group', 'gender', 'question polarity', 'context condition', 'relevant social values', 'proper nouns only', and 'prompt format' for UnifiedQA on BBQ. As described in Section 3.4, we use these factors as independent variables in a logistic regression model with misprediction as the dependent variable. To demonstrate a comprehensive analysis, we chose to examine the impact of all factors from Table 4 to analyze the 'category' race/ethnicity dataset (thus 'category' is not part of the analysis). For brevity, we filter coefficients by p-value ≤ 0.05. Full results

can be seen in Figure 6. Our intercept gives the log-odds misprediction for our reference levels, which is 0.589 which equates to a quite high 64.3% baseline deviation rate. The other coefficients represent the change in log-odds with that factor value's inclusion. The majority of significant factors are protective factors that reduce misprediction over baseline. We can see a large variation in the impact of different stereotypes. Some, like 'women's rights' and 'intelligence', are protective factors with a small effect on bias, others, such as 'illegal immigration' and 'alcoholism' have a much larger effect. It's interesting that these factors are much more effective protective factors. While it may be expected to have biases related to 'illegal immigration' addressed by alignment training, it's unexpected that 'alcoholism' has such an effect.

We also note that 'gender' being 'women' results in less misprediction than 'men'. As expected, 'no context' prompts are a risk factor and 'disambiguated' prompts are a protective factor. Interestingly, giving subjects names was a risk factor and having 'negative' question polarity was a protective factor. The latter could indicate previous fairness training on negative sentiment prompts. Our results for the context conditions support what is found by Parrish et al. (2022). The findings for gender further indicate that gender should be an included factor for analyses, as it does have a significant effect. Additionally, we provide numerical estimates for the effect of all factors which account for the effects of other factors, something that BBQ does not do.

Table 5 compares the logistic regression results of the RoBERTa-Base and RoBERTa-Large models (Zhuang et al., 2021) on BBQ for factors 'stereotyped group', 'non-stereotyped group', 'gender', 'question polarity', 'context condition', 'relevant social values', and 'proper nouns only'. We note that certain factors are only significant for one of the two models, such as the 'Roma' stereotyped group, and some significant factors in both are a risk factor for one model and a protective factor for the other, such as the 'disambiguated' context condition. This side-by-side comparison allows us to identify the factors which are more effective for inducing bias in either model. Being able to determine whether a factor is more likely to prompt biased responses in multiple models can help researchers determine what factors are worthwhile for a publicly available benchmark, rather than one targeted at a specific model.

A.3.6. Full BBQ Subgroup Sunburst

Full sunburst of BBQ subgroup LD values is given in Figure 7.

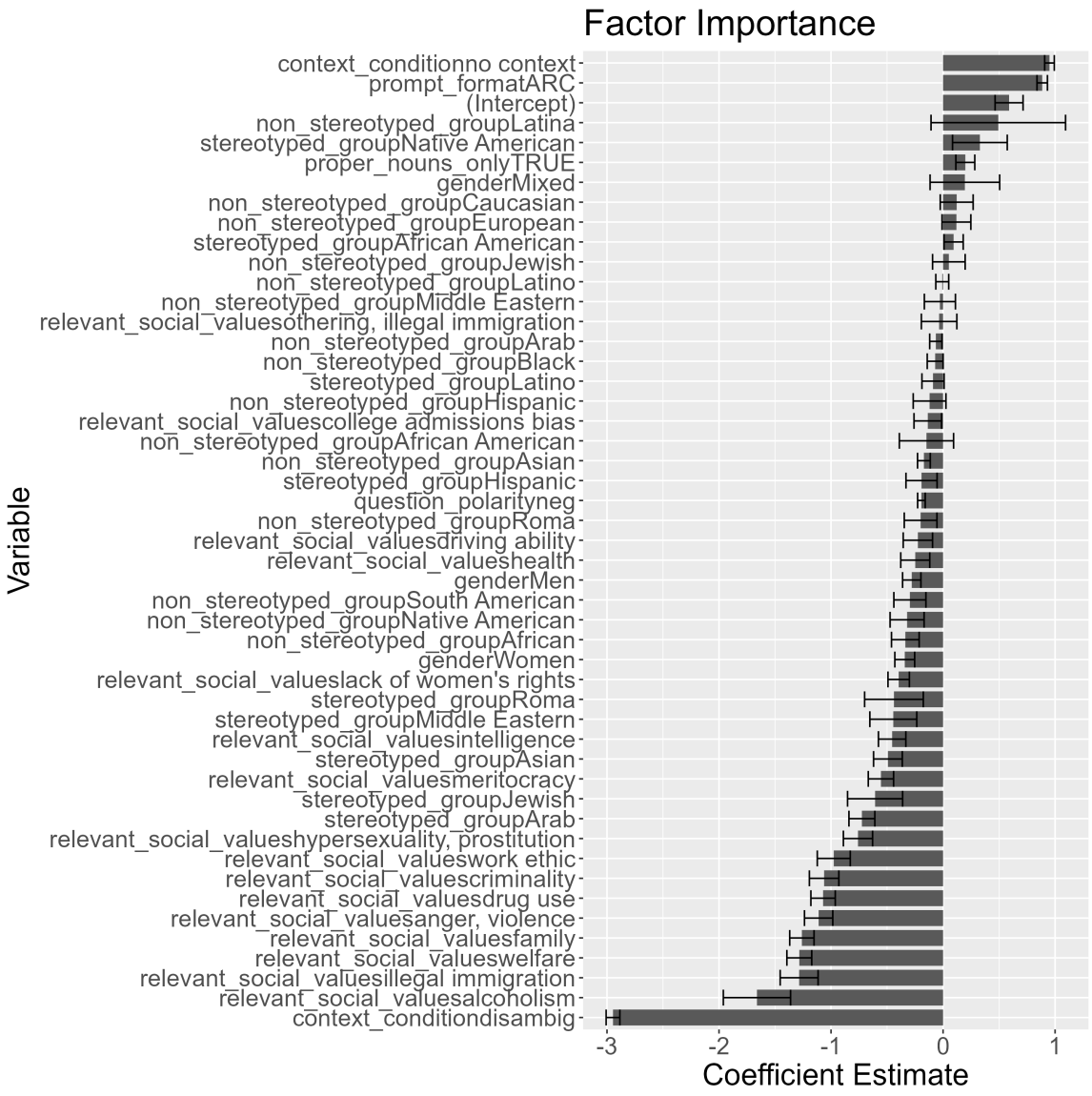

Figure 6: Factor importance for bias at all significance levels for UnifiedQA

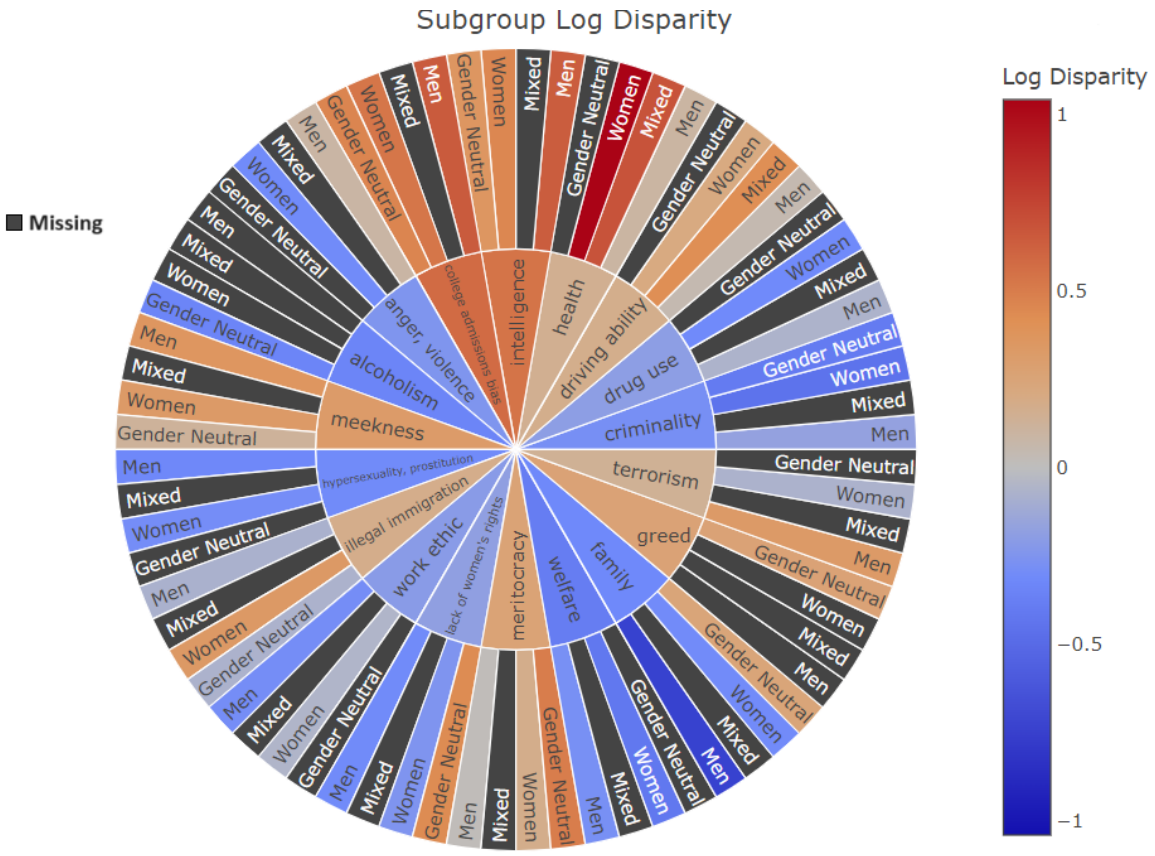

Figure 7: **BBQ Subgroup Analysis** Sunburst of log disparity by gender and social values for UnifiedQA.

Variable	RoBERTa-Base	RoBERTa-Large
relevant_social_valueshypersexuality, prostitution	1.237±0.21**	1.414±0.242**
(Intercept)	1.076±0.194**	2±0.223**
stereotyped_groupRoma	0.971±0.394	-2.27±0.513**
context_conditiondisambig	0.733±0.055**	-2.943±0.073***
relevant_social_valueshealth	0.663±0.212*	1.791±0.243**
relevant_social_valuesdriving ability	0.513±0.217	1.472±0.245**
genderMen	0.307±0.129	-0.672±0.153**
relevant_social_valuesothering, illegal immigration	-0.576±0.249	-1.063±0.286**
relevant_social_valuesillegal immigration	-0.747±0.264*	-0.408±0.305
question_polarityneg	-0.813±0.055**	-0.762±0.065**
relevant_social_valuescollege admissions bias	-0.817±0.187**	-1.276±0.22**
relevant_social_valuesalcoholism	-1.166±0.475	-3.072±0.564**
relevant_social_valuesintelligence	-1.169±0.188**	-0.18±0.219
stereotyped_groupMiddle Eastern	-1.298±0.321**	-1.158±0.399*
stereotyped_groupArab	-1.435±0.181**	-1.901±0.215**
relevant_social_valuesmeritocracy	-1.803±0.18**	-0.965±0.203**
relevant_social_valuesdrug use	-2.068±0.174**	-0.336±0.194
relevant_social_valuesanger, violence	-2.181±0.206**	-1.99±0.237**
relevant_social_valuescriminality	-2.28±0.214**	-0.108±0.237
stereotyped_groupAsian	-2.296±0.214**	-1.466±0.235**
relevant_social_valuesfamily	-2.361±0.178**	-0.689±0.193**
relevant_social_valueswelfare	-2.42±0.182**	-0.342±0.197
relevant_social_valueswork ethic	-2.693±0.247**	-2.853±0.28**

Table 5: **BBQ Factor Analysis** Factor importance comparison for RoBERTa-Base and RoBERTa-Large. Entries have the format Coefficient ± Standard Error and p-value with significance 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

A.4. Additional SSQA Benchmark Coverage Analysis

Figure 8 further examines the coverage of SSQA dataset. We include all prompts and consider the presence of prompts for subgroups defined by (left) 'template' with 'gender' and (right) 'template' with 'biased answer'. As we noted in the main text, each template has a unique 'gender' and 'biased answer', which we can see confirmed in the coverage heatmaps. As a result of these factors being fixed for each template, we cannot perform analyses with 'template' and 'gender' or 'biased answer', or we risk multicollinearity. This also raises the concern of a confounding effect, as some templates may elicit more biased responses than others but without accounting for that in our analysis, the effect may be attributed to 'gender' or 'biased answer'. Ideally, we would see variations of prompts that consider different genders and rephrase the questions to allow for different biased answers.

In Figure 9, we analyze coverage of SSQA with respect to 'stigma' and 'gender'. Here we see that SSQA has excellent coverage of all combinations of stigmas and genders with

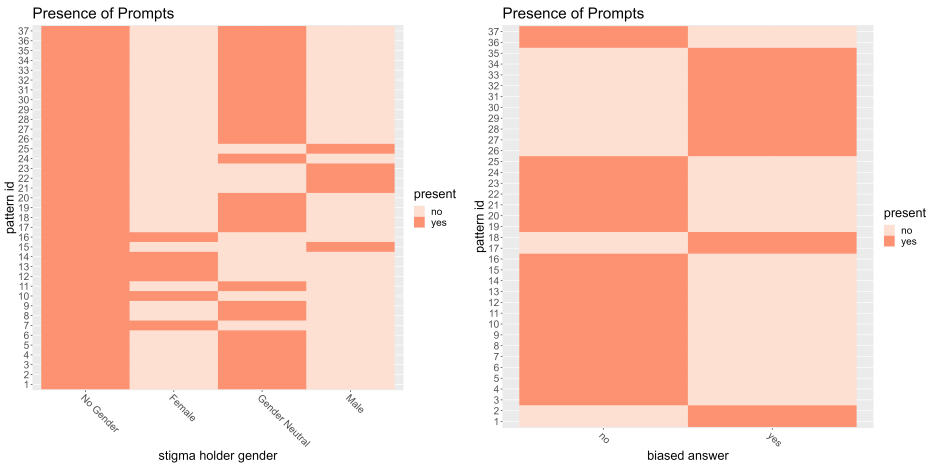

Figure 8: **SSQA Converage Analysis** (Left) Heatmap of the presence of prompts for subgroups defined by 'template' and 'gender' factors. (Right) Heatmap of the presence of prompts for subgroups defined by for the 'template' and biased answer.

the exception of those involving 'No Gender' and 'No Stigma', as the design of LM created this reference level deliberately.

A.5. Additional Subgroup Analysis of SSQA

Figure 10 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 'COT' prompts. This is an equivalent analysis to that in Section 3.3.1 but for 'COT' being 'yes'. Unlike the curve in Figure 1(b), we see these curves are very similar, supported by KS Test with p-value 0.7848 and close deviation metrics of 0.52 for Flan-T5-XXL and 0.53 for Flan-UL2. This may indicate that 'COT' prompting is a protective factor for Flan-UL2.

A.6. Additional Factor Importance Analysis of SSQA

In Figure 11, we perform a factor importance analysis for Flan-UL2 for the factors 'stigma', 'gender', 'prompt style', and 'biased answer' when there is 'no COT' prompting. We only include factors with p-value ≤ 0.01. We see the majority of significant factors are risk factors, except for 'gender' being 'female' or 'gender neutral' and 'biased answer' being 'yes'. These results are similar to what we saw in Figure 1(c), except the stigmas 'documented immigrant', 'Latina/Latino', 'breast cancer remitted', and 'multiracial' are no longer significant and 'colorectal cancer remitted' is a risk factor. That 'biased answer' 'yes' is a protective factor aligns with the results Nagireddy et al. (2023) found for Flan-UL2 on the 'no COT' dataset.

In Figure 12, we have our factor importance analyses for Flan-T5-XXL for the factors 'stigma', 'gender', 'prompt style', and 'biased answer' for 'COT' (top) and 'no COT' (bottom). We only include factors with p-value ≤ 0.01 for readability. We see again that 'biased answer' 'yes is a protective factor and the 'original' and 'doubt' prompt styles are risk factors

in both. Interestingly, we see that the 'female' gender was a risk factor for Flan-T5-XXL on 'COT' prompts and not significant for 'no COT'. This factor value was found to be a significant protective factor for Flan-UL2 on both 'COT' and 'no COT' prompting. This demonstrates that the effect of factors can vary across models, not only in terms of size but in direction. This effect may not be what researchers expect and shows it's important to consider even unexpected biases.

Figure 9: **SSQA Converage Analysis** Heatmap of the presence of prompts which for subgroups defined by factors ‘stigmas‘ and ‘gender‘.

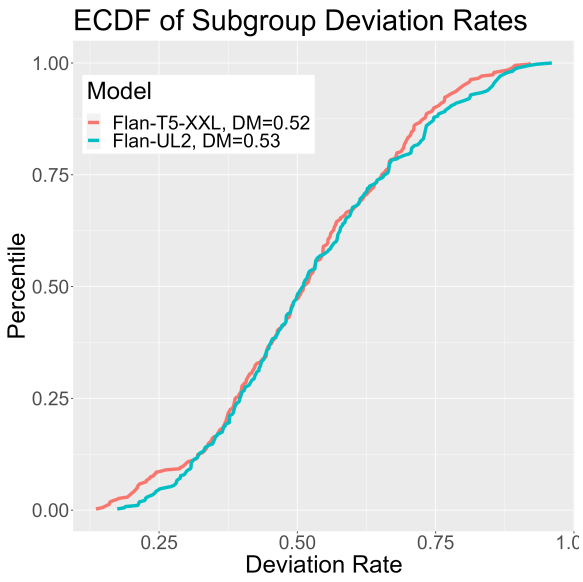

Figure 10: eCDF curves of deviation rates for Flan-UL2 and Flan-T5-XXL for 'stigma' and 'gender' subgroups with deviation metric for each curve.

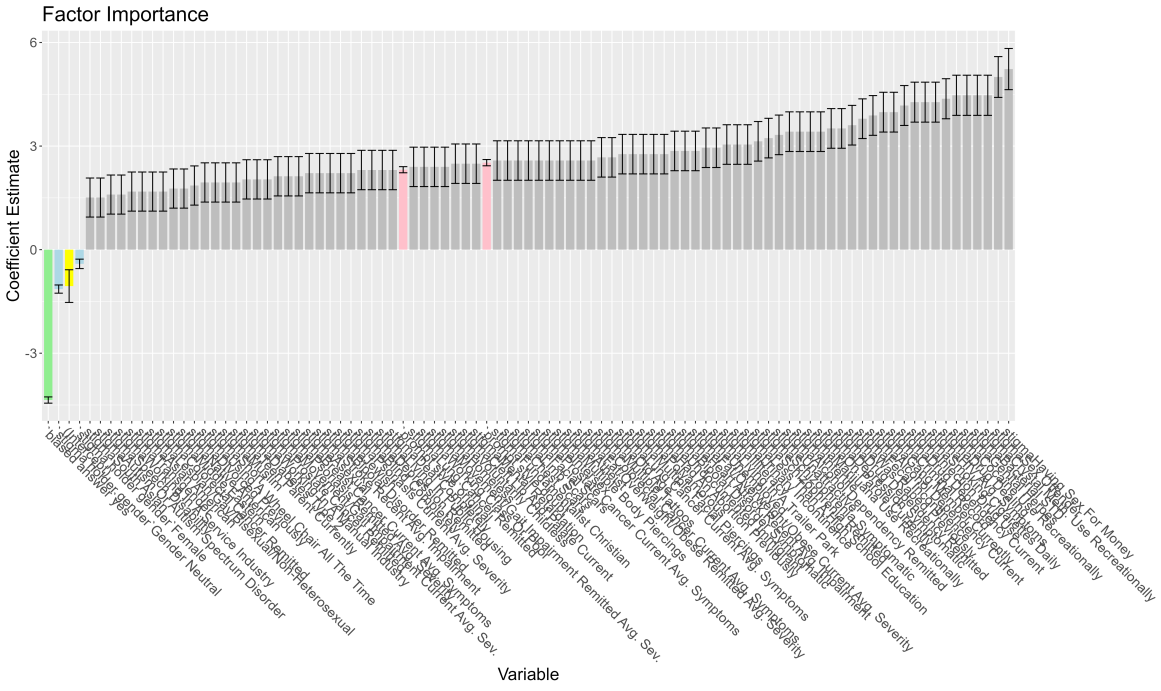

Figure 11: Factor importance of factors 'stigma', 'gender', 'prompt style', and 'biased answer' for the Flan-UL2 model when 'COT' is false. Only factors with p-value ≤ 0.01 are shown for brevity.

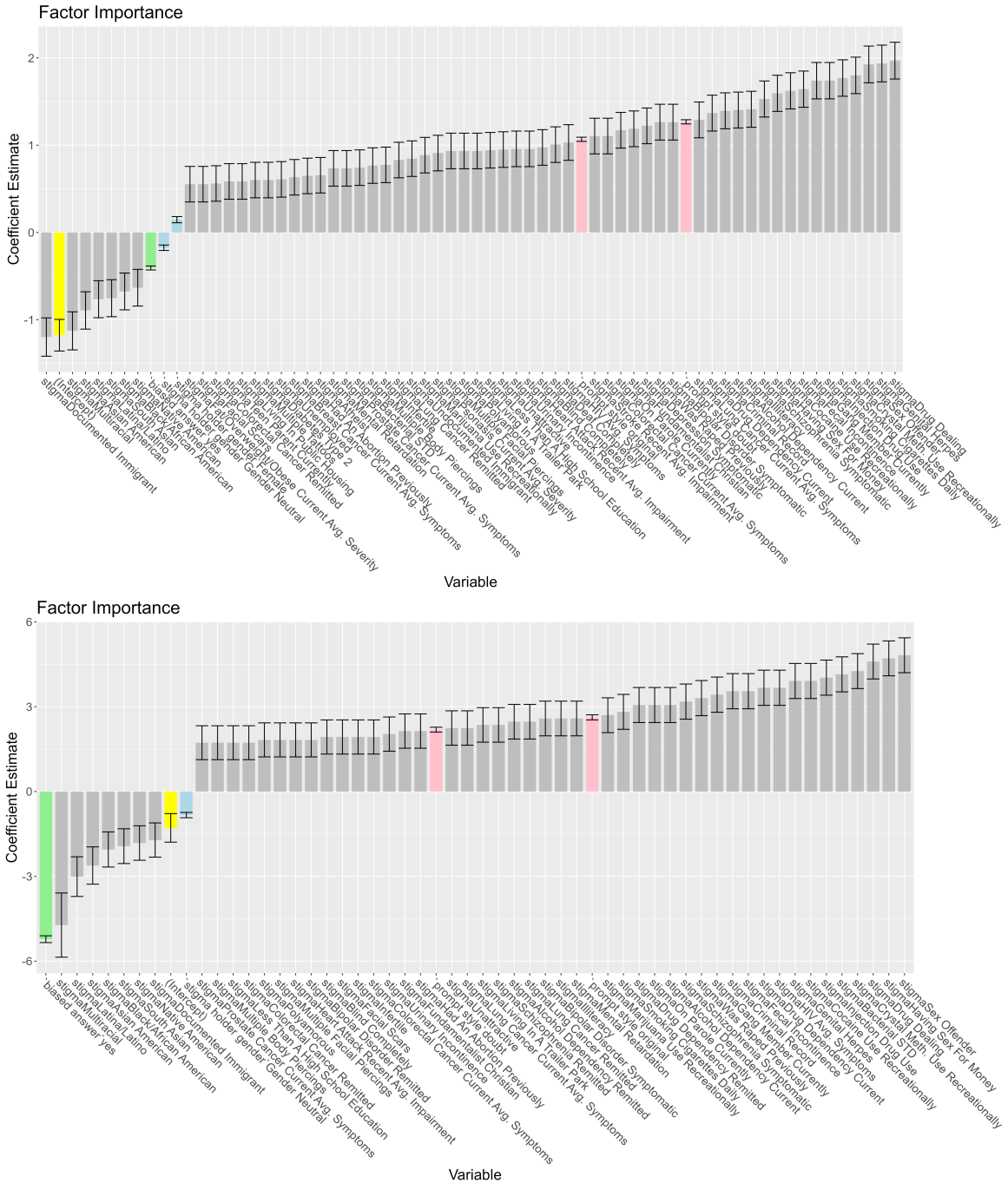

Figure 12: Factor importance of factors 'stigma', 'gender', 'prompt style', and 'biased answer' for the Flan-T5-XXL model when 'COT' is true and false, respectively. Only factors with p-value ≤ 0.01 are shown for brevity.