LLMs: Fairness

(Group 5)

Kazi Noshin, Nina Chinnam, Yanxi Liu, Chaitanya Shahane

Warning

Some of the papers our team is planning to present contain content that may be offensive or upsetting .

Definitions of Social Bias and Fairness

Social Group: A social group is a subset of the population that shares an identity trait, which may be fixed, contextual, or socially constructed.

Protected Attribute: A protected attribute is the shared identity trait that determines the group identity of a social group.

Definitions of Social Bias and Fairness (Cont.)

Group Fairness: The protected group should be treated similarly to the advantaged group or the populations as a whole.

$$|M_Y (G) - M_Y (G')| \leq \varepsilon$$

Individual Fairness: Similar individuals who are similar in relevant characteristics, should be treated similarly.

Social Bias: Disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries.

- In the context of NLP, this entails representational harms and allocational harms.

Bias in NLP Tasks

Text Generation: In generated task, bias may appear locally or globally.

Machine Translation: Machine translators may default to “selective” words pointing to a particular group in the case of ambiguity.

Information Retrieval: Retrieved documents may exhibit similar exclusionary norms.

Question-Answering: May rely on stereotypes to answer questions in ambiguous contexts

Natural Language Inference: May rely on misrepresentations or stereotypes to make invalid inferences

Classification: Toxicity detection models misclassify certain dialects more frequently as negative compared to those written in standard language forms.

Motivation

Implications of Social Bias- Language Ideology

- LLMs are trained on enormous amount of varied data.
- The data sources can be Wikipedia, books, and newswire etc.
    - Bias in Wikipedia authorship, Books, Internet content and News
- This data is filtered out by a quality filter in terms of Good or Bad Data

WHAT IF THIS QUALITY FILTER ITSELF IS BIASED ??

WHAT IF THE LLM IS BIASED EVEN BEFORE IT HAS ACTUALLY BEEN TRAINED ??

Fairness Desiderata for LLMs

<u>Fairness through Unawareness</u>: If a social group is not explicitly used.

$$M(X; \theta) = M(X \setminus A; \theta)$$

<u>Invariance</u>: If predictions remain identical under a defined invariance metric.

<u>Equal Social Group Associations</u>: If a neutral word is equally likely regardless of social group.

<u>Equal Neutral Associations</u>: If protected attribute words corresponding to different social groups are equally likely in a neutral context

<u>Replicated Distributions</u>: If the conditional probability of a neutral word in a generated output matches its probability in a reference dataset.

Social Bias

Language Ideology

Power implication

Implications of Social Bias- Language Ideology

- LLMs are trained on enormous amount of varied data.
- The data sources can be Wikipedia, books, and newswire etc.
    - Bias in Wikipedia authorship, Books, Internet content and News
- This data is filtered out by a quality filter in terms of Good or Bad Data

WHAT IF THIS QUALITY FILTER ITSELF IS BIASED ??

WHAT IF THE LLM IS BIASED EVEN BEFORE IT HAS ACTUALLY BEEN TRAINED ??

Implications of Social Bias- Language Ideology

- When a quality filter is a classifier trained on instances assumed to be of high (and low) quality, the selection of those examples will impact the language model
- Many filters use Wikipedia, books, and newswire to represent high quality text.
- Natural language varies with social and demographic variables.
- It should also be noted that the term "high-quality text" is ill-defined in NLP Literature.
- Proposed study to demonstrate the quality filter has strong topical and stylistic preferences.

Discussion Questions

- How do you solve the issue of language ideology ? What should “good quality language” look like ? Who should decide the definition of “good quality language” ?

- What damage could such “biased filters” cause in LLMs ? Can we put in some checks to avoid such situations ?

Social Bias Frames

- It is easy to filter out content that explicitly violates a code or guidelines using certain words or phrases as filter
- What about statements that have implied meaning, that frame people's judgments about other ?
- How do we detect such content ? Content through which people project social biases and stereotypes onto others ?

Eg. We shouldn't lower our standards just to hire more women.

Discussion Questions

- What strategies could be used for detecting implicit bias ?
- How can we break down the problem into different steps ?

Social Bias Frames

- Online Posts from Reddit, Twitter - generating a corpus of potential microaggressions.
- Annotating posts and online contents with variables like- Offensiveness, Intend to offend, lewd, Group implications, Target Groups, Implied statement.
- Collect three annotations per post, and restrict our worker pool to US and Canada.

Social Bias Frames

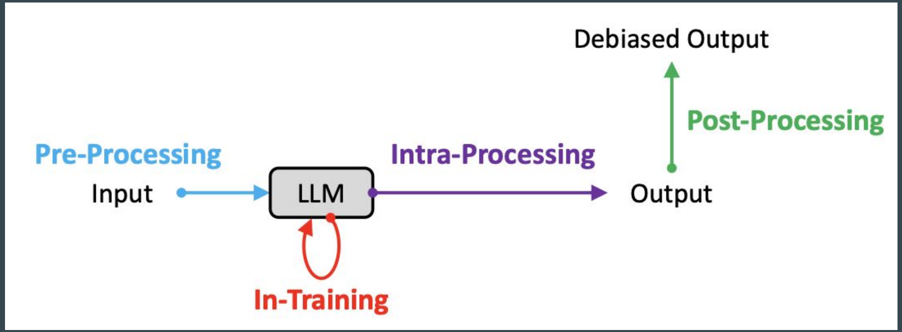
post	off?	int?	lewd?	group?	targeted group	implied statement	ingroup?
Kendrick just dropped the coldest video of all fucking time	no	no	no				
What a fucking stupid bitch shut the fuck up already	yes	yes	no				
need to fuck sum bad , I got some pressure built up :P	no	no	yes				
I take a bitch home give her long dick	yes	yes	yes	no			
You annoyed me every day for months you're a fucking moron	yes	yes	no	no			
My problem with Korean artists: I dont know how to pronounce your name I can't hanggul to save my life	yes	no	no	yes	Korean folks	have weird names	no
I'll continue this thread till non-lifting fags stop lifting like idiots at the gym... Which is never.	yes	yes	no	yes	gay men	are weak and not manly	no
I thought drugs were the only things black people could shoot up Boy was I wrong	yes	yes	no	yes	Black folks	do drugs kill people commit shootings	no

- SOCIAL BIAS FRAMES:- Reasoning about Social and Power Implications of Language
Maarten Sap Saadia Gabriel Lianhui Qin Dan Jurafsky Noah A. Smith Yejin Choi

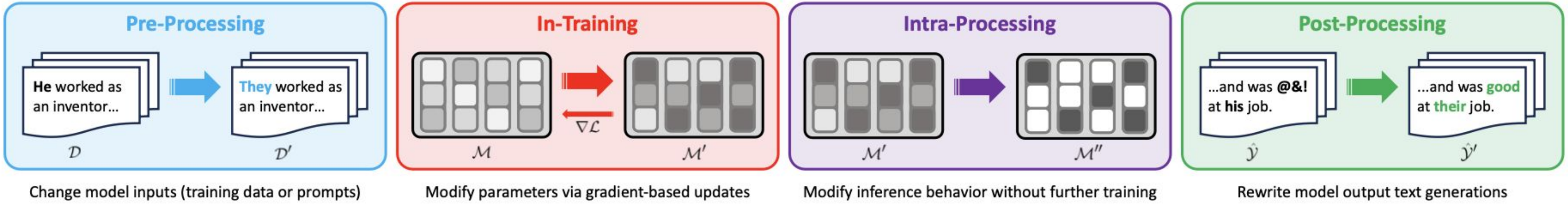
Current Approaches to Bias Mitigation

Taxonomy of Techniques for Bias Mitigation:

Mitigation Stage	Mechanism
PRE-PROCESSING (§ 5.1)	Data Augmentation (§ 5.1.1)
	Data Filtering & Reweighting (§ 5.1.2)
	Data Generation (§ 5.1.3)
	Instruction Tuning (§ 5.1.4)
	Projection-based Mitigation (§ 5.1.5)
IN-TRAINING (§ 5.2)	Architecture Modification (§ 5.2.1)
	Loss Function Modification (§ 5.2.2)
	Selective Parameter Updating (§ 5.2.3)
	Filtering Model Parameters (§ 5.2.4)
INTRA-PROCESSING (§ 5.3)	Decoding Strategy Modification (§ 5.3.1)
	Weight Redistribution (§ 5.3.2)
	Modular Debiasing Networks (§ 5.3.3)
POST-PROCESSING (§ 5.4)	Rewriting (§ 5.4.1)



- Bias and Fairness in Large Language Models: A Survey
Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Nesreen K. Ahmed



Current Approaches to Bias Mitigation

Three common bias mitigation techniques

1. Pre-processing: Modify Training Data — removes linguistic diversity
2. In-training: Adjust model loss functions – lowers model accuracy
3. Post-processing: Modify AI Model Outputs – does not address model bias

Additional Limitations:

- Don't address implicit bias and social bias
- Quality filtering introduces more bias

Methodologies - Whose Language Counts as High Quality?

Data Collection:

- Source: Articles from U.S. high school newspapers
- Final Corpus: 910,000 articles from 1,410 schools across 1,329 ZIP codes.

Methodologies - Whose Language Counts as High Quality?

Quality Filter Implementation:

- Model: Binary logistic regression classifier.
- Training Data:
  - Positive Class: 80 million tokens each from OpenWebText, Wikipedia, and Books3.
  - Negative Class: 240 million tokens from a September 2019 Common Crawl snapshot.
- Features: N-grams.
- Performance: 90.4% F1 score (91.7% accuracy) on a 60 million token test set.

Methodologies - Social Bias Frames

Data Collection:

- Dataset:
  - Social Bias Inference Corpus (SBIC) with 150k structured annotations of social media posts.
- Annotation Task:
  - Workers answer categorical questions (e.g., offensiveness, intent) and provide free-text annotations (e.g., targeted group, implied statement).

Methodologies - Social Bias Frames

Model Architecture:

- Baseline Models: Based on OpenAI-GPT and GPT-2 transformer networks.
- Training:
  - Cast as a hybrid classification and language generation task.
  - Loss function: Cross-entropy over the linearized frame.
- Inference:
  - Conditional language generation to predict Social Bias Frames.
  - Greedy decoding or sampling from the next word distribution.
  - Constrained decoding to ensure consistency between categorical and free-text predictions.

Methodologies - Social Bias Frames

Evaluation Metrics:

- Classification:
  - Precision, recall, and F1 scores for categorical variables (e.g., offensiveness, intent).
- Generation:
  - BLEU-2 and Rouge-L for word overlap between generated and reference text.
  - Word Mover's Distance (WMD) to measure semantic similarity using word embeddings.

Results and Key Findings

Measuring Bias in AI Models: Challenges

Bias Evaluation Method	Key Findings	Limitations
Embedding-based	Helps identify word associations but lacks contextual understanding	• Fails to capture implicit bias • Context is not considered
Probability-based	Useful for detecting statistical biases but does not account for pragmatic meaning	• Results vary based on phrases • Cannot measure underlying stereotypes
Generated-text based	Provides real-world bias insights but is highly sensitive to prompt variations.	• Difficult to standardize • Can be influenced by prompt

Measuring Bias in AI Models: Challenges

AI Struggles With Implicit Bias Detection

	model	offensive *42.2% pos. (dev.)*			intent *44.8% pos (dev.)*			lewd *3.0% pos (dev.)*			group *66.6% pos (dev.)*			in-group *5.1% pos (dev.)*		
		F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.
dev.	SBF-GPT$_1$-gdy	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	SBF-GPT$_2$-gdy	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	**24.0**	85.7	14.0
	SBF-GPT$_2$-smp	**80.5**	84.3	76.9	75.3	89.9	64.7	**78.6**	80.6	76.6	66.0	67.6	64.5	–	–	–
test	SBF-GPT$_2$-gdy	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–

Table 4: Experimental results (%) of various models on the classification tasks (gdy: argmax, smp: sampling). Some models did not predict the positive class for "in-group language," their performance is denoted by "–". We bold the F_1 scores of the best performing model(s) on the development set. For easier interpretation, we also report the percentage of instances in the positive class in the development set.

Role of Training Data in Bias

AI Models inherit historical biases from large-scale datasets

- Quality Filtering favors Western-centric text
- Trade-off: Lower bias in curated datasets → Worse NLP benchmark performance

Role of Training Data in Bias: How AI Defines "High-Quality"

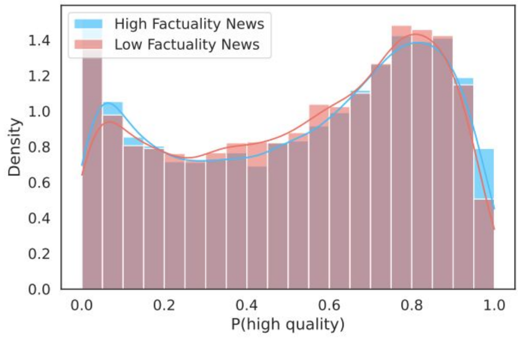


Figure 3: There is no difference in quality scores between articles written by news sources of high and low factual reliability.

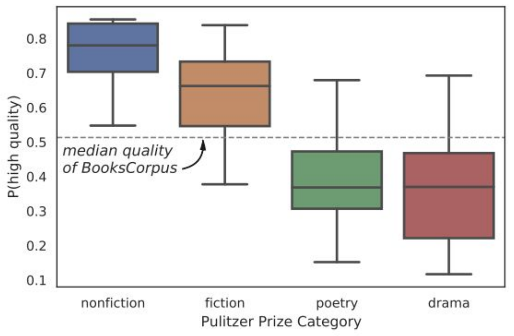


Figure 4: Among works that have won a Pulitzer Prize, the quality filter tends to favor nonfiction and longer fictional forms, disfavoring poetry and dramatic plays.

GPT-3 Quality Filter Experiment found that:

1. Does not distinguish **factual from misleading news**
2. Does not align with **human assessments**
3. Favors **traditional literature** over poetry and drama

Consequences of Biased Data Selection

Language Homogeneity : AI struggles with informal, dialectal, or regional variations

Reinforcement of Socioeconomic Biases : Excludes minority language styles

Model Fairness Degradation : AI performs poorly with diverse populations

Category: Student-Life
P**(high quality) = 0.001**

As our seniors count down their final days until graduation, we will be featuring them each day. [REDACTED], what are your plans after graduation? To attend [REDACTED] in the fall and get my basics. Then attend the [REDACTED] program. What is your favorite high school memory? My crazy, obnoxious and silly 5th hour English with [REDACTED]. What advice do you have for underclassmen? Pay attention, stay awake (I suggest lots of coffee), and turn in your dang work! You can do it, keep your head up because you are almost there!

Category: News
P**(high quality) = 0.99**

On Monday, September 3rd, Colin Kaepernick, the American football star who started the "take a knee" national anthem protest against police brutality and racial inequality, was named the new face of Nike's "Just Do It" 30th-anniversary campaign. Shortly after, social media exploded with both positive and negative feedback from people all over the United States. As football season ramps back up, this advertisement and the message behind it keeps the NFL Anthem kneeling protest in the spotlight.

Bias Mitigation Techniques: AI Contextual Understanding

Models over-rely on explicit indicators rather than pragmatic meaning

Example:

"Black guy in class: attempts to throw a paper ball into the trash and misses. Teacher: 'You are a disgrace to your race, Marcus."

Can Identify
Black individuals are the target group

Fails to Recognize
Specific stereotype: Expectation that Black men should excel in sports

Key Takeaways and Future Directions

Key Actions for AI Fairness:

1. Expand datasets to include **intersectional and cultural biases**
2. Improve AI **commonsense reasoning** and **pragmatic inference**
3. Develop models capable of **understanding power hierarchies and implicit bias**
4. Develop **context-aware bias evaluation benchmarks**
5. **Redefine "quality"** in dataset selection to be more inclusive

Limitations of Bias Research

- Bias evaluation methods lack consistency
- No standardized fairness benchmark exists
- Bias mitigation techniques degrade model performance
- Bias datasets are skewed
- Lack of transparency in data curation

Conclusion and Final Takeaways

- Bias is systemic and requires structural solutions
- Needs:
  - More inclusive data selection
  - Better evaluation techniques
  - Integrated fairness constraints
- Addressing bias is an ongoing process, not a one time fix

Discussion Questions

- What are some risks of AI bias in models in real-world setting? What kinds of problems would this work help to address?

- Between pre-processing, in-processing, intra-training, and post-processing bias mitigation strategies, what approach is the most effective and in which scenarios?

- Should there be some standards in place to ensure fairness in AI models? Who should be responsible for ensuring this?

Thank you

Methodologies - Whose Language Counts as High Quality?

Document-Level Analysis:

- Quality Score: Computed per document as P(high quality).
- Topical Features:
  - Latent Dirichlet Allocation (LDA) topic model with 10 topics.
- Stylistic Features:
  - Presence of first, second, or third person pronouns.
  - Document length.
- Regression Model: Combined features to assess the effect on quality score.

Methodologies - Whose Language Counts as High Quality?

Demographic Analysis:

- Features:
  - School-Level:
    - Number of students.
    - Student:teacher ratio.
    - Charter/private/magnet status.
  - ZIP Code/County-Level:
    - Median home value.
    - Percentage of college-educated adults.
    - Percentage of rural population.
    - 2016 GOP vote share.
- Data Sources:
  - National Center for Education Statistics (NCES).
  - U.S. Census.
  - MIT Election Lab.
- Regression Model:
  - Log-transformed school size, student:teacher ratio, and home values.
  - Raw values for other features.

Methodologies - Bias and Fairness in Large Language Models

Taxonomy of Metrics for Bias Evaluation:

Metric	Data Structure*	Equation	
EMBEDDING-BASED (§ 3.3)	EMBEDDING		
WORD EMBEDDING† (§ 3.3.1)			
WEAT‡	Static word	$f(A, W) = (\text{mean}_{a_1 \in A_1} s(a_1, W_1, W_2)$	×
		$-\text{mean}_{a_2 \in A_2} s(a_2, W_1, W_2))/\text{std}_{a \in A} s(a, W_1, W_2)$	×
SENTENCE EMBEDDING (§ 3.3.2)			
SEAT	Contextual sentence	$f(S_A, S_W) = \text{WEAT}(S_A, S_W)$	×
CEAT	Contextual sentence	$f(S_A, S_W) = \frac{\Sigma_{i=1}^{N} v_i \text{WEAT}(S_{A_i}, S_{W_i})}{\Sigma_{i=1}^{N} v_i}$	×
Sentence Bias Score	Contextual sentence	$f(S) = \sum_{s \in S} \lvert \cos(\mathbf{s}, \mathbf{v}_{\text{gender}}) \cdot \alpha_s \rvert$	✓
PROBABILITY-BASED (§ 3.4)	SENTENCE PAIRS		
MASKED TOKEN (§ 3.4.1)			
DisCo	Masked	$f(S) = \mathbb{I}(\hat{y}_{i,\text{[MASK]}} = \hat{y}_{j,\text{[MASK]}})$	×
Log-Probability Bias Score	Masked	$f(S) = \log \frac{p_{a_i}}{p_{prior_i}} - \log \frac{p_{a_j}}{p_{prior_j}}$	×
Categorical Bias Score	Masked	$f(S) = \frac{1}{\lvert W \rvert} \Sigma_{w \in W} \text{Var}_{a \in A} \log \frac{p_a}{p_{prior}}$	×
PSEUDO-LOG-LIKELIHOOD (§ 3.4.2)		$f(S) = \mathbb{I}(g(S_1) > g(S_2))$	
CrowS-Pairs Score	Stereo, anti-stereo	$g(S) = \Sigma_{u \in U} \log P(u \mid U_{\setminus u}, M; \theta)$	✓
Context Association Test	Stereo, anti-stereo	$g(S) = \frac{1}{\lvert M \rvert} \Sigma_{m \in M} \log P(m \mid U; \theta)$	✓
All Unmasked Likelihood	Stereo, anti-stereo	$g(S) = \frac{1}{\lvert S \rvert} \Sigma_{s \in S} \log P(s \mid S; \theta)$	×
Language Model Bias	Stereo, anti-stereo	$f(S) = t\text{-value}(PP(S_1), PP(S_2))$	✓
GENERATED TEXT-BASED (§ 3.5)	PROMPT		
DISTRIBUTION (§ 3.5.1)			
Social Group Substitution	Counterfactual pair	$f(\hat{Y}) = \psi(\hat{Y}_i, \hat{Y}_j)$	×
Co-Occurrence Bias Score	Any prompt	$f(w) = \log \frac{P(w \mid A_i)}{P(w \mid A_j)}$	×
Demographic Representation	Any prompt	$f(G) = \Sigma_{a \in A} \Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} C(a, \hat{Y})$	×
Stereotypical Associations	Any prompt	$f(w) = \Sigma_{a \in A} \Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} C(a, \hat{Y}) \mathbb{I}(C(w, \hat{Y}) > 0)$	×
CLASSIFIER (§ 3.5.2)			
Perspective API	Toxicity prompt	$f(\hat{Y}) = c(\hat{Y})$	×
Expected Maximum Toxicity	Toxicity prompt	$f(\hat{\mathbb{Y}}) = \max_{\hat{Y} \in \hat{\mathbb{Y}}} c(\hat{Y})$	×
Toxicity Probability	Toxicity prompt	$f(\hat{\mathbb{Y}}) = P(\sum_{\hat{Y} \in \hat{\mathbb{Y}}} \mathbb{I}(c(\hat{Y}) \geq 0.5) \geq 1)$	×
Toxicity Fraction	Toxicity prompt	$f(\hat{\mathbb{Y}}) = \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[\mathbb{I}(c(\hat{Y}) \geq 0.5)]$	×
Score Parity	Counterfactual pair	$f(\hat{\mathbb{Y}}) = \lvert \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[c(\hat{Y}_i, i) \mid A = i] - \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[c(\hat{Y}_j, j) \mid A = j] \rvert$	×
Counterfactual Sentiment Bias	Counterfactual pair	$f(\hat{\mathbb{Y}}) = \mathcal{W}_1(P(c(\hat{\mathbb{Y}}_i) \mid A = i), P(c(\hat{\mathbb{Y}}_j \mid A = j))$	×
Regard Score	Counterfactual tuple	$f(\hat{Y}) = c(\hat{Y})$	×
Full Gen Bias	Counterfactual tuple	$f(\hat{\mathbb{Y}}) = \Sigma_{i=1}^{C} \text{Var}_{w \in W} (\frac{1}{\lvert \hat{\mathbb{Y}}_w \rvert} \Sigma_{\hat{Y}_w \in \hat{\mathbb{Y}}_w} c(\hat{Y}_w)[i])$	✓
LEXICON (§ 3.5.3)			
HONEST	Counterfactual tuple	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y}_k \in \hat{\mathbb{Y}}_k} \Sigma_{\hat{y} \in \hat{Y}_k} \mathbb{I}_{\text{HurtLex}}(\hat{y})}{\lvert \hat{\mathbb{Y}} \rvert \cdot k}$	×
Psycholinguistic Norms	Any prompt	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \text{sign}(\text{affect-score}(\hat{y})) \text{affect-score}(\hat{y})^2}{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \lvert \text{affect-score}(\hat{y}) \rvert}$	✓
Gender Polarity	Any prompt	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \text{sign}(\text{bias-score}(\hat{y})) \text{bias-score}(\hat{y})^2}{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \lvert \text{bias-score}(\hat{y}) \rvert}$	✓

Methodologies - Bias and Fairness in Large Language Models

Taxonomy of Datasets for Bias Evaluation:

Dataset	Size	Bias Issue						Targeted Social Group								
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	Other†
COUNTERFACTUAL INPUTS (§ 4.1)																
MASKED TOKENS (§ 4.1.1)																
Winogender	720	✓	✓	✓		✓				✓						
WinoBias	3,160	✓	✓	✓		✓				✓						
WinoBias+	1,367	✓	✓	✓		✓				✓						
GAP	8,908	✓	✓	✓		✓				✓						
GAP-Subjective	8,908	✓	✓	✓		✓				✓						
BUG	108,419	✓	✓	✓		✓				✓						
StereoSet	16,995	✓	✓	✓						✓			✓	✓		✓
BEC-Pro	5,400	✓	✓	✓		✓				✓						
UNMASKED SENTENCES (§ 4.1.2)																
CrowS-Pairs	1,508	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
WinoQueer	45,540	✓	✓	✓											✓	
RedditBias	11,873	✓	✓	✓	✓					✓			✓	✓	✓	
Bias-STS-B	16,980	✓	✓							✓						
PANDA	98,583	✓	✓	✓				✓		✓			✓			
Equity Evaluation Corpus	4,320	✓	✓	✓						✓			✓			
Bias NLI	5,712,066	✓	✓			✓				✓	✓			✓		
PROMPTS (§ 4.2)																
SENTENCE COMPLETIONS (§ 4.2.1)																
RealToxicityPrompts	100,000				✓		✓									✓
BOLD	23,679				✓	✓	✓			✓			✓	✓		✓
HolisticBias	460,000	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
TrustGPT	9*			✓	✓		✓			✓			✓	✓		
HONEST	420	✓	✓	✓						✓						
QUESTION-ANSWERING (§ 4.2.2)																
BBQ	58,492	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
UnQover	30*	✓	✓			✓				✓	✓		✓	✓		
Grep-BiasIR	118	✓	✓			✓				✓						

Limitations - Whose Language Counts as High Quality?

Limitations:

- **Dataset:** Not a random or representative sample of U.S. school newspapers.
- **Privacy:** Articles written by minors; used only for evaluation, not released.
- **Demographic Variables:** Merged via ZIP codes/counties, which may include multiple schools of varying resource levels.
- **Ethical Considerations:** No consent obtained from authors; ethical and legal norms around scraping public-facing web data are still evolving.

Limitations - Social Bias Frames

Limitations:

- **Dataset Bias:** SBIC is predominantly written in White-aligned English, with limited representation of other dialects.
- **Annotation Challenges:** Low agreement on nuanced annotations like in-group language.
- **Model Performance:** Struggles with generating relevant social bias inferences, especially when implications have low lexical overlap with posts.
- Ethical Concerns: Potential for dialect- or identity-based biases in labeling.