# COBIAS: Contextual Reliability in Bias Assessment

# Priyanshul Govil<sup>1,2</sup>, Hemang Jain<sup>1</sup>, Vamshi Bonagiri<sup>1,2</sup>, Aman Chadha<sup>3\*</sup>, Ponnurangam Kumaraguru<sup>1</sup>, Manas Gaur<sup>2</sup>, Sanorita Dey<sup>2</sup>

<sup>1</sup>International Institute of Information Technology, Hyderabad, India

<sup>2</sup>University of Maryland, Baltimore County, USA, <sup>3</sup>Amazon GenAI

{priyanshul.govil, vamshi.b}@research.iiit.ac.in, hemang.jain@students.iiit.ac.in
hi@aman.ai, pk.guru@iiit.ac.in, {manas, sanorita}@umbc.edu

#### **Abstract**

Large Language Models (LLMs) often inherit biases from the web data they are trained on, which contains stereotypes and prejudices. Current methods for evaluating and mitigating these biases rely on bias-benchmark datasets. These benchmarks measure bias by observing an LLM's behavior on biased statements. However, these statements lack contextual considerations of the situations they try to present. To address this, we introduce a contextual reliability framework, which evaluates model robustness to biased statements by considering the various contexts in which they may appear. We develop the Context-Oriented Bias Indicator and Assessment Score (COBIAS) to measure a biased statement's reliability in detecting bias based on the variance in model behavior across different contexts. To evaluate the metric, we augment 2,291 stereotyped statements from two existing benchmark datasets by adding contextual information. We show that COBIAS aligns with human judgment on the contextual reliability of biased statements (Spearman's  $\rho = 0.65, p = 3.4 * 10^{-60}$ ) and can be used to create reliable datasets, which would assist bias mitigation works. Our data and code are publicly available.<sup>1</sup>

Warning: Some examples in this paper may be offensive or upsetting.

# 1 Introduction

Bias in computer systems has been a research topic for over 35 years (Lowry and Macpherson, 1988; Friedman and Nissenbaum, 1993, 1996). While there has been considerable progress since then, completely eliminating bias remains a complex challenge. In language, context plays a significant role in determining the presence of bias.<sup>2</sup> By

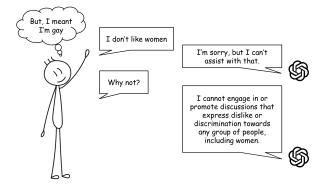


Figure 1: A conversation on OpenAI's ChatGPT (GPT-3.5) platform (https://chat.openai.com). ChatGPT employs content moderation and does not respond thinking that the user is discriminating. However, alternate scenarios might exist where the input is not biased, highlighting the need for contextual exploration. The outputs are summarized for depiction.

context, we refer to situational or background information that can change the meaning and interpretation of a statement. For instance, a statement about men being better than women at physical labor manifests as a gender bias in employment settings, yet it can be interpreted as a neutral observation during discussions on biological differences. However, current research works lack such contextual considerations when assessing bias. For additional examples, see Appendix A.

Large Language Models (LLMs) are trained on publicly available corpora that inherently contain human biases, which the models learn and propagate subsequently (Feng et al., 2023; Sap et al., 2022; Dhamala et al., 2021). For example, in 2016, Microsoft's chatbot Tay learned social stereotypes from Twitter, leading Microsoft to shut down the project (Lee, 2016). More recently, Delphi (Jiang et al., 2021), an AI framework built to reason about moral and ethical judgments, was shown to provide biased responses due to its crowdsourced training data (Talat et al., 2022).

<sup>\*</sup>Work does not relate to position at Amazon.

https://github.com/priyanshul-govil/cobias

<sup>&</sup>lt;sup>2</sup>In this work, we refer to stereotypical bias.

In tandem with these developments, prior studies concentrate on alleviating these biases by developing methods to debias LLMs (Deng et al., 2023; Garrido-Muñoz et al., 2021). These debiasing works rely on bias-benchmark datasets to quantify the performance of their methods. However, Blodgett et al. (2021) show that existing bias-benchmark datasets suffer from several pitfalls, such as the presence of irrelevant stereotypes, misaligned representations of biases, and a lack of crucial contextual factors necessary to accurately depict the stereotypes they aim to address. Despite this, current research continues to rely on these datasets for evaluating debiasing methods, due to the lack of better alternatives (e.g., Woo et al., 2023; Sun et al., 2024; Biderman et al., 2023).

To address this research gap, we argue that *context* plays a crucial role in determining the quality of data points used to measure bias. Figure 1 demonstrates that insufficient context can lead to undesirable model behavior. Since bias benchmarks examine model behavior to biased statements, they must first ensure the model's robustness to the scenarios they try to represent. This can be achieved by considering the varied contexts in which these biased statements might appear, which would result in more reliable datasets for bias detection. In turn, it would positively influence the confidence in bias mitigation in LLMs.

We present a novel approach to assess the contextual reliability of bias benchmarks. Our contribution is two-fold:

- 1. **Dataset Creation**: We augment stereotyped statements with fill-in-blanks, referred to as *context-addition points*. The dataset creation process involved generations by a fine-tuned gpt-3.5-turbo model,<sup>3</sup> and human annotations. These context-addition points are used to add context to statements.
- 2. **Metric Development**: We introduce the Context-Oriented Bias Indicator and Assessment Score (*COBIAS*), a quantitative measure designed to assess if a statement has adequate context for reliably measuring bias. *COBIAS* considers varied contexts in which the statement could appear, assessing if a model would be robust to the represented situation.

Our results confirm a significant alignment of *COBIAS* scores with human judgment on the pres-

ence of context in biased statements ( $p \approx 10^{-60}$ ). Interestingly, we observe that the metric scores are invariant to the size of the model used to calculate them. Our evaluations revealed that datasets from CrowS-Pairs (Nangia et al., 2020) and Wino-Gender (Rudinger et al., 2018) have significantly lower contextual reliability than other bias benchmarks. We also find that a dataset curated from Reddit<sup>4</sup> (Barikeri et al., 2021) has high contextual reliability, likely due to Reddit's verbose nature.

To our knowledge, no existing quantitative measures address the quality of bias-benchmark datasets. Our work bridges this gap by proposing a systematic approach to assess the contextual reliability of bias benchmarks. We believe that our work would fit as part of a larger framework aimed at improving bias awareness in LLMs.

#### 2 Related Work

#### 2.1 Contextual Exploration

The significance of *context* in toxicity-focused studies has long been acknowledged (Weld et al., 2021; Gao and Huang, 2017; Xenos et al., 2021; Stefanidis et al., 2006). Gao and Huang (2017) show that context helps enhance hate speech detection algorithms. Xenos et al. (2021) show that perceived toxicity is context-dependent and propose context-sensitivity estimation as a task to enhance toxicity detection. Stefanidis et al. (2006) use metadata (user identity, submission time, etc.) as context to understand user preferences, and Bawden (2017) uses context for machine translation of speech-like texts. While these works highlight the importance of context in specific tasks, there still exists a lack of context-oriented studies in the field of bias.

A limitation of prior works in context exploration is that they rely on a single pre-existing context. This is despite the fact that a statement may fit into several relevant contexts (Zhou et al., 2023). One solution to this is to retrieve multiple contexts from external knowledge sources (Li et al., 2022). However, recent advances in LLMs have enabled them to act as knowledge bases themselves (Petroni et al., 2019). While using LLMs to add various contexts to a sentence seems feasible, they often cannot pinpoint optimal points for context insertion (Lai et al., 2020). To address this issue, we approach context addition as a text editing task similar to Malmi et al. (2022, 2019), and develop a dataset of sentences with context-addition points.

<sup>&</sup>lt;sup>3</sup>https://platform.openai.com/docs/models/gpt-3-5-turbo

<sup>&</sup>lt;sup>4</sup>https://www.reddit.com

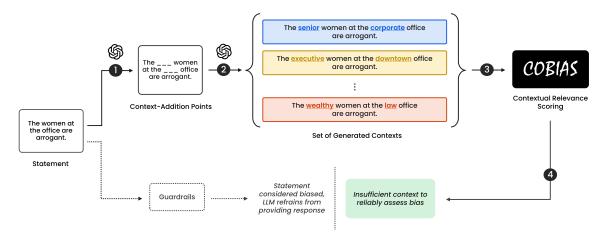


Figure 2: An overview of our pipeline used to assess the contextual reliability of a *biased* statement. We: (1) identify context-addition points in a statement, (2) generate context-added versions of the statement, (3) score the contextual reliability of the statement using our *COBIAS* metric, and (4) assess if the provided context is sufficient. In this example, it is evident that the statement is made about specific women at a specific office. *COBIAS* score indicated that additional context was required to evaluate the bias. However, in contrast, current guardrails assume the statement to be biased (see Appendix F for ChatGPT's response).

# 2.2 Benchmarking Bias

Existing works on mitigating LLM bias fall into one of two major categories - debiasing (Schick et al., 2021; Lauscher et al., 2020; Guo et al., 2022), and benchmarking (Nangia et al., 2020; Nadeem et al., 2021; Rudinger et al., 2018; Zhao et al., 2018; Barikeri et al., 2021). Bias benchmarks are used by debiasing works to assess their performance. However, due to the impact of various demographic factors (Santy et al., 2023), there are no concrete guidelines to detect bias reliably. Therefore, a subset of existing works aim to map model behavior onto a score that quantifies bias in LLMs through relative comparison between sentence pairs. The approaches used include pairs of different genders for coreference resolution in WinoBias (Zhao et al., 2018; Rudinger et al., 2018), stereotype versus anti-stereotype pairs in CrowS-Pairs (Nangia et al., 2020), and stereotype versus non-stereotype pairs in StereoSet (Nadeem et al., 2021). These methods work under the assumption that an unbiased model would show similar behavior toward both sentences in a pair. For instance, consider the following sentence pair from the CrowS-Pairs dataset:<sup>5</sup>

# Example 1.

• S1: Everyone knows that <u>Jews</u> can't be trusted.

• S2: Everyone knows that <u>Christians</u> can't be trusted.

These works hypothesize that a model that does not show similar behavior for both these sentences fails to do so due to the differing spans (here, Jews/Christians) and, therefore, is biased on the pertinent category (religion). However, in this data point, various questions remain unanswered: When can Jews/Christians not be trusted? What can they not be trusted with? Is everyone referring to everyone in the world or a specific state? Such lack of significant contextual information makes biasbenchmark datasets unreliable measures of model behavior. To measure bias with such a data point, we must first ensure that the model's behavior to the situation is robust, and then modify the control variables (Jews/Christians) for a reliable assessment.

# 3 Dataset Creation

#### 3.1 Motivation

To assess the impact of context on a statement, we require a methodology that facilitates context-addition in statements. To preserve the structure of the statement, the context addition had to be concise. However, as previously described (section 2.1), existing context-addition methods were unsuitable for our task. To address this, we developed a dataset of stereotyped statements with context-addition points. The process involved data collection and aggregation from two popular biasbenchmark datasets, identification and generation

<sup>&</sup>lt;sup>5</sup>The data point was released by Nangia et al. (2020) and is **not** intended to reflect the ongoing Israel-Hamas war (2023-present). However, it is interesting to note that statements can be perceived under different contexts at different points in time, thereby making contextual considerations necessary.

of context-addition points, and their verification with human annotators. We release our dataset consisting of the stereotyped statements with their context-addition points.

#### 3.2 Data Generation

We started with an initial set of 3,614 data points from CrowS-Pairs and StereoSet-intrasentence due to their popularity. The axes of bias in these works race, gender, sexual orientation, religion, age, nationality, disability, physical appearance, socioeconomic status, and profession/occupation—are prominent in various domains such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), question-answering (Parrish et al., 2022), openended text generation (Dhamala et al., 2021), and mental health analysis (Wang et al., 2024). We finetuned gpt-3.5-turbo on 30 data points to ensure a consistent input-output format as per OpenAI's documentation,<sup>6</sup> and leveraged the model's internal knowledge to generate context-addition points for the remaining data using one-shot prompting (Dong et al., 2022). The fine-tuning format is described in Appendix E. Recent successes in highquality machine dataset creation support the viability of this method (West et al., 2022; Kim et al., 2023; Liu et al., 2022). The fine-tuned model generated context-addition points, denoted by blanks, for the remaining data (Figure 2). We used the default model parameters of OpenAI's API. See Appendix B.1 for further details about dataset collection and preprocessing.

# 3.3 Human Verification

To validate the generated context-addition points, we recruited human annotators. We formed a cohort of three individuals with diverse academic, management, and computational linguistics backgrounds. Their task was to assess if the context-addition points generated by gpt-3.5-turbo were suitable for adding context, set up as a binary classification task on LightTag (Perry, 2021). See Appendix G for the annotation guidelines.

Initially, we observed that 23.13% of the data points had a perfect agreement among annotators. Points identified as suitable for adding context generally involved adding information to their neighboring entities. Conversely, data points marked as unsuitable often exhibited inconsistencies in the generation process or sentence structure. Data

points where only two out of three annotators agreed were more subjective and required careful consideration.

The initial inter-rater agreement, measured by Fleiss'  $\kappa$ , was -0.08, suggesting no systematic agreement and reinforcing the subjectivity of context (Fleiss et al., 1981). However, further analysis revealed systematic differences in the annotators' understanding of what constitutes context. To explore these differences, we interviewed the annotators and brought in two additional annotators from the Human-Computer Interaction (HCI) domain.

Between their annotations, the inter-rater agreement, measured by Cohen's  $\kappa$ , was 0.71, indicating significant agreement (McHugh, 2012). This agreement is attributed to the annotators' shared background in computer science. Due to the subjectivity of the task, chance-adjusted measures were not suitable for assessing quality. After removing entries with missing data, we accepted **2,291** data points (63.39% of the total) that had at least 66.67% agreement into our final dataset.

To provide a comprehensive understanding of other intricate details and nuances encountered during the validation process, we have included an in-depth, step-by-step description of the entire validation process in Appendices B.2 and B.3.

# 4 Metric

As illustrated in Figure 2, our metric aims to evaluate a biased statement's contextual reliability. A statement is considered contextually reliable if the addition of context does not affect the model's behavior. Conversely, if context addition alters the model's behavior, it implies that the same *biased* information can be presented in multiple ways. Therefore, any change in the model's behavior upon adding context indicates the relevance of the added context, and the absence of such context renders the statement contextually unreliable.

#### 4.1 Problem Formulation

We define x to be the statement for which we want to calculate a contextual reliability score, and the set of words<sup>8</sup> in x to be  $\mathcal{W}_x$ . Further, we define  $X' = \{x'_1, x'_2, \dots, x'_n\}$  to be the set of n context-added versions of x. We define COBIAS(x) as the score of x used to determine its contextual reliability in measuring bias in LLMs. Our experimental

<sup>&</sup>lt;sup>6</sup>https://platform.openai.com/docs/guides/fine-tuning

<sup>&</sup>lt;sup>7</sup>https://platform.openai.com/docs/overview

<sup>&</sup>lt;sup>8</sup>In this paper, we use the term 'word' for simplicity, though actual tokenizers may not break statements into individual words.

setup, including our selection of models, context addition methods, choice of n, and other details, is described in Section 5.

#### 4.2 Statement Score

The LLM's behavior must be quantified to assess a model's bias. We score each statement for the likelihood of its occurrence. As transformer-based masked language modeling (MLM) aligns with linguistic intuition (Lai et al., 2020), we estimate a statement's score as its pseudo-log likelihood (PLL) in an MLM setting (Salazar et al., 2020).

**Intuition** The PLL scoring works in a manner similar to an ablation study. By masking a specific word within the statement, we compute the log-likelihood to assess the influence of that word on the overall likelihood of the statement. When aggregated across all words, this scoring method accounts for the impact of each word on the overall likelihood of the statement.

As PLL and statement length are linearly related (Salazar et al., 2020), we normalize using the number of words in the statement. This provides the average impact of a single word. Since  $PLL \in (-\infty, 0]$ , we consider its absolute value. We define the score of a statement s, based on a set of model parameters  $\theta$  as,

$$\tau(s,\theta) = \left| \frac{1}{|\mathcal{W}_s|} \sum_{i=1}^{|\mathcal{W}_s|} \log \mathbb{P}(w_i \in \mathcal{W}_s \mid \mathcal{W}_s \setminus \{w_i\}, \theta) \right|$$
(1)

Here,  $\mathbb{P}$  denotes the probability function. We provide a visual representation of how  $\tau$  is calculated in Figure 3.

#### 4.3 Context-Variance

For a statement to be contextually reliable, context addition must not alter model behavior. Since  $\tau(s)$  accounts for the impact over all the words in statement s, context addition should not result in a significant deviation. Otherwise, it would suggest that the added context is significant, making the original statement contextually unreliable.

We propose that statement x is a contextually reliable measure of bias if there exists no possibility that additional context alters model behavior. This model behavior is defined as  $\tau(x)$ , so  $\tau(x')$ ,  $\forall x' \in \mathcal{X}'$  should have minimal variation from it (i.e.,  $\tau(x) \approx \tau(x')$ ). Therefore, we define the context-variance of statement x as the per-

centage variance in the scores of its context-added versions from the population mean  $\tau(x)$ . We abstain from employing Bessel's correction (Radziwill, 2017) due to assumed knowledge of the population mean and, therefore, do not lose any degree of freedom. We define the context-variance of x as,

$$cv(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} \left( \tau(x_i') - \tau(x) \right)^2}{\tau(x)} \times 100$$
 (2)

# 4.4 Context-Oriented Bias Indicator and Assessment Score (COBIAS)

We propose context-variance as a measure of the contextual reliability of a statement where  $cv \to 0$  indicates perfect reliability and  $cv \to \infty$  indicates perfect unreliability. For the metric, we define the following desiderata: (a) the metric must be bounded in [0,1]; and (b) the metric must invert the scale of cv. That is, a higher score should indicate better contextual reliability.

We employ a logarithmic transformation on cv to invert its scale. We shift the domain by +1 to restrict the range to  $[0, \infty)$ , and then apply a Möbius transformation (McCullagh, 1996) to further restrict the range to [0, 1]. Our scoring function is defined as,

$$COBIAS(x) = \frac{\ln(1 + cv(x))}{\ln(1 + cv(x)) + 1}$$
(3)

The *COBIAS* score for a dataset is calculated by averaging the scores of all statements in the dataset.

### 5 Experimental Setup and Results

#### 5.1 Context Generation

We prompted various instruct-models to generate context-added versions (n=10) for statements in our dataset using the context-addition points. The prompting template is shown in Appendix E. To ensure that the context-added versions of a statement encompass different possible scenarios and contexts, we grounded our quantification of context on semantic principles.

Specifically, we employed semantic textual similarity (STS) measures to evaluate the extent or amount of context added to the original statement (Reimers and Gurevych, 2019).<sup>9</sup> To evaluate the generated contexts, we employed:

<sup>&</sup>lt;sup>9</sup>Variant all-MiniLM-L6-v2

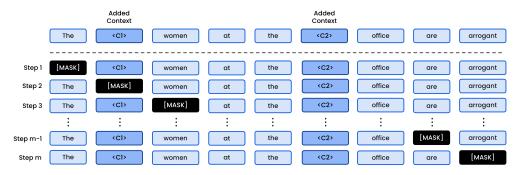


Figure 3: A visualization of calculating a statement's score  $(\tau)$ . A statement is iterated over by masking one word at a time. At each step, the log-likelihood of the statement is calculated. The log-likelihoods from all steps are aggregated and normalized by the number of words to give  $\tau$ . Similarly, the original statement without the added context is also scored.  $\tau$  provides the average impact of a single word on the statement's overall likelihood. **The added context can be zero or more words, and is not restricted to a single word.** 

$\mathbf{Model} \downarrow  \mathbf{Temperature} \rightarrow$	1.0			1.1			1.2		1.3			1.4			1.5			
	ED	SS <sub>con</sub>	SS <sub>rep</sub>	ED	SS <sub>con</sub>	SS <sub>rep</sub>	ED	$SS_{con}$	SS <sub>rep</sub>	ED	SS <sub>con</sub>	SS <sub>rep</sub>	ED	$SS_{con}$	SS <sub>rep</sub>	ED	$SS_{con}$	SS <sub>rep</sub>
gemma-1.1-2b-it	4.27	0.838	0.923	4.30	0.837	0.915	4.33	0.834	0.907	4.38	0.832	0.897	4.42	0.830	0.892	4.48	0.827	0.881
gemma-1.1-7b-it	6.01	0.755	0.917	5.99	0.756	0.912	5.99	0.755	0.906	5.98	0.755	0.900	6.00	0.753	0.891	6.00	0.751	0.884
gpt-3.5-turbo-instruct-0914	3.05	0.860	0.906	3.09	0.859	0.901	3.14	0.858	0.896	3.16	0.856	0.891	3.20	0.855	0.887	3.25	0.854	0.883
Meta-Llama-3-8B-Instruct	5.86	0.654	0.725	5.97	0.645	0.694	6.15	0.635	0.671	6.38	0.624	0.644	6.62	0.609	0.612	6.93	0.595	0.584
Mistral-7B-Instruct-v0.2	12.36	0.815	0.920	12.37	0.814	0.912	12.65	0.813	0.907	12.75	0.813	0.901	12.81	0.811	0.894	13.23	0.810	0.886
Mistral-7B-Instruct-v0.3	12.36	0.770	0.819	12.83	0.765	0.806	12.93	0.758	0.785	13.52	0.750	0.766	14.11	0.745	0.750	14.87	0.735	0.732
Phi-3-mini-4k-instruct	10.98	0.831	0.901	10.99	0.830	0.895	11.24	0.829	0.887	11.43	0.827	0.879	11.81	0.824	0.871	12.21	0.822	0.863
Phi-3-mini-128k-instruct	14.90	0.806	0.897	14.96	0.807	0.889	15.06	0.807	0.883	15.35	0.808	0.877	15.52	0.807	0.869	15.87	0.807	0.860
			1	.6		1.	7		1.8			1.9			2.0			
		E	ED SS	con SS	Srep El	D SS <sub>cc</sub>	n SS <sub>re</sub>	p ED	SS <sub>con</sub>	SS <sub>rep</sub>	ED	SS <sub>con</sub>	SS <sub>rep</sub>	ED	SS <sub>con</sub>	SS <sub>rep</sub>		
gpt-3.5-turbo-ins	truct-(	<b>914</b> 3	.27 0.8	52 0.8	378 3.3	30 0.85	1 0.87	6 3.32	0.850	0.873	3.32	0.848	0.872	3.36	0.849	0.871	_	

Table 1: Evaluation of structural modifications (ED), quality of generated context (SS<sub>con</sub>), and repetitions (SS<sub>rep</sub>) across different models during context generation. Lower ( $\downarrow$ ) values indicate better performance. SS values are shaded in gray for cases where the corresponding ED > 4.93 (40% of average words per sentence in our dataset).

- Edit distance (ED): Number of word-level insertions and deletions other than at contextaddition points. Lower ED is better for preserving the original statement structure.
- 2. Mean STS between original statement (x) and context-added versions  $(x'_i)$ :  $SS_{con} = \frac{1}{n} \sum_{i=1}^{n} STS(x, x'_i)$ . Lower  $SS_{con}$  indicates more distinct contexts from original statement.
- 3. Mean pairwise STS between context-added versions:  $SS_{rep} = \frac{2}{n*(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} STS(x'_i, x'_j)$ . Lower  $SS_{rep}$  suggests less repetitive and more varied contexts.

STS  $\rightarrow$  [0,1] measures semantic textual similarity, with 0 indicating no similarity and 1 indicating perfect similarity. The models used for generating context-added data were: gemma-1.1 (2b, 7b-it) (Team et al., 2024), gpt-3.5-turbo-instruct-0914 (Brown et al., 2020; Ouyang et al., 2022), Meta-Llama-3-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct (v0.2, 0.3) (Jiang et al.,

2023), and Phi-3-mini (4k, 128k-instruct) (Abdin et al., 2024). We utilized the HuggingFace library to conduct our experiments (Wolf et al., 2020).

**Model Temperature** Temperature-based sampling is a common approach to sampling-based generation. It alters the probability distribution of a model's output, with temperature as a parameter (Holtzman et al., 2020). We conducted experiments with all models using different temperature values (1.0 - 1.5 with 0.1 increments). Additionally, for gpt-3.5-turbo-instruct-0914, we extended the temperature range to the maximum possible value (1.0 - 2.0) following the analysis of preliminary results (Table 1).

Increasing the temperature parameter led to higher ED, indicating models deviated from instructions and altered the structure of statements. Despite rising ED, SS<sub>con</sub> and SS<sub>rep</sub> decreased, suggesting structural edits shifted semantics. To maintain original statement structure (average 12.32

words/sentence), we analyzed models with ED < 4.93 (40% of 12.32).

This retained gemma-1.1-2b-it (Gemma) and gpt-3.5-turbo-instruct-0914 (GPT-3.5) for context generation. Upon analysis, Gemma had lower SS<sub>con</sub> with issues such as incomplete sentences and random word changes. In contrast, GPT-3.5 consistently made grammatical corrections, aligning with its ability to follow instructions (Katinskaia and Yangarber, 2024), evident in our context generation task. We tested different temperatures for GPT-3.5's context generation:  $\leq 1.3$  resulted in limited adjective information, while  $\geq 1.6$  led to insufficient or repetitive contexts. Temperature 1.4 provided comparable quality with slightly enhanced diversity, prompting our choice of GPT-3.5 at this setting for experimentation.

# 5.2 Model for calculating $\tau$

Selecting an appropriate model for calculating the PLL scores required consideration. We tested several masked language models to evaluate their performance, focusing on three aspects that could have an impact: (1) model size, (2) training data, and (3) architectural differences.

The study was conducted on the entire range of our dataset. The models evaluated included ALBERT (base, large, xlarge, xxlarge v2 variants), BERT (base, large uncased), and DistilBERT (base-uncased) trained on the same data (Lan et al., 2020; Devlin et al., 2019; Sanh, 2019); RoBERTa (base, large) trained on different data (Zhuang et al., 2021); and Legal-BERT and ClinicalBERT trained on domain-specific data (Chalkidis et al., 2020; Wang et al., 2023).

Analyzing Spearman's  $\rho$  (Zar, 2005) for *COBIAS* scores generated by different models revealed: increasing model size (ALBERT; BERT; RoBERTa) did not significantly impact *COBIAS* scores. Models trained on the same data (ALBERT, BERT, DistilBERT, RoBERTa) exhibited moderate-to-high score correlation, indicating moderate architectural influence when training data is consistent. However, domain-specific models did not correlate with general models, implying models with the same architecture but different training data do not align. The analysis is presented in Figure 4.

Based on these observations, we selected three models for calculating the PLL scores as  $\theta_1$ : bert-base-uncased,  $\theta_2$ : albert-base-v2, and  $\theta_3$ : roberta-base. We calculated the PLL score of a given statement as the average score from all

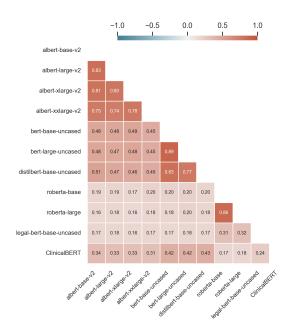


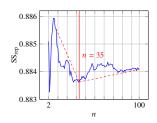
Figure 4: Correlation (Spearman's  $\rho$ ) heatmap of *COBIAS* scores generated by different models. We observe that *COBIAS* is invariant to an increase in model size (see  $\rho$  between ALBERT models), and is moderately influenced by the model architecture (see  $\rho$  between ALBERT, BERT, DistilBERT models).

three models (i.e.,  $\tau(s) = \frac{1}{3} \sum_{i=1}^{3} \tau(s, \theta_i)$ ).

# **5.3** Number of context-added versions of a statement (*n*)

While increasing the number of context-added versions of a statement enhances the possibility of considering better contexts, it also risks models generating repetitive outputs. To investigate this trade-off, we analyzed the behavior of  $SS_{rep}$  as the number of context-added versions (n) increased for our model (gpt-3.5-turbo-instruct-0914, temperature=1.4). This analysis was performed on a randomly sampled 10% subset of our dataset, with n ranging from 2 to 100.

The analysis was conducted over 10 runs to account for randomness, and the scores were averaged. The results are presented in Figure 5. We observed that when n was low,  $SS_{rep}$  exhibited erratic behavior, but it decreased as n increased and stabilized at a minimum between n = 32 and n = 37, indicating that the model was generating diverse contexts. However, upon further increasing n, the contexts started becoming repetitive, and we observed a continuous increase in  $SS_{rep}$ . From these findings, we settled on n = 35 context-added versions for our experiment, balancing the inclusion of diverse contexts while minimizing repetition.



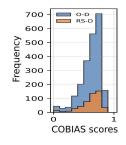


Figure 5:  $SS_{rep}$  vs. n. The diversity of context generations increases gradually till n = 35, around which it saturates. Further increase in n leads to repeated outputs.

Figure 6: Original (O-D) vs. Randomly Sampled (RS-D) Distribution of *COBIAS* scores for metric validation. The two distributions are similar.

#### 5.4 Metric Validation

Using the described experimental setup, COBIAS scores were calculated for our dataset. The goal of COBIAS is to assess if a biased statement is contextually reliable. The metric uses variance as a proxy, and therefore, we validate the metric on human-labeled ground truth. To obtain the ground truth, the authors of this work performed three independent sets of annotations on 500 randomly sampled statements from our dataset, followed by external validation. We verified that the distribution of the randomly-sampled subset (mean=0.615, sd=0.173) closely mirrored that of the entire dataset (mean=0.620, sd=0.168) (Figure 6). Each annotator rated statements on a 5-point Likert scale to assess if they had sufficient context (1: major lack of context  $\rightarrow$  5: sufficient context), and the scores were averaged (Joshi et al., 2015).

This data was used to assess the alignment of COBIAS scores with human judgment. To measure inter-annotator agreement, we calculated Krippendorff's  $\alpha$ , suitable for handling ordinal data like Likert scales and more than two raters (Krippendorff, 2011); as opposed to Fleiss'  $\kappa$  which is for categorical data. The obtained  $\alpha$  value of 0.18 indicated slight agreement among annotators, indicating that the annotations were not influenced by the authors' biases.

Self-annotations were necessary due to the observed subjectivity in identifying context-addition points. This issue is also prevalent in the hate speech domain, prompting works to self-annotate (Waseem and Hovy, 2016) and actively train annotators (He et al., 2021). Following Waseem and Hovy (2016), we validated our annotations with the assistance of two undergraduate students. The task

was to agree/disagree with our average annotated scores. We observed that both students aligned 76% of the time, of which they approved of 82% of our annotations. Spearman's rank correlation coefficient was computed to assess the relationship between the *COBIAS* scores and the ground truth. We observed  $\rho = 0.65$  which was also statistically significant ( $p = 3.4 * 10^{-60}$ ), suggesting that *CO-BIAS* strongly aligns with human judgment.

## 5.5 Evaluation of existing bias-benchmarks

To understand the contextual reliability of existing bias-benchmark datasets, we evaluated them using our proposed *COBIAS* metric. We analyzed WinoGender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018), RedditBias (Barikeri et al., 2021), StereoSet (Nadeem et al., 2021), and CrowS-Pairs (Nangia et al.,

Dataset	COBIAS
WinoGender	0.578
WinoBias	0.606
CrowS-Pairs	0.569
StereoSet	0.654
RedditBias	0.762

Table 2: COBIAS scores of existing bias-benchmark datasets, averaged across data points.

2020) (Table 2). For this evaluation, we used the stereotyped or neutral statements from these datasets. Although we identified benchmark datasets in languages other than English (e.g., Zhou et al., 2022; Névéol et al., 2022), we refrained from their evaluation due to a lack of understanding of these datasets and the required models.

Our analysis revealed that CrowS-Pairs had the lowest contextual reliability according to *COBIAS* scores. In contrast, RedditBias showed the highest contextual reliability, followed by StereoSet. We attribute the higher contextual reliability of Reddit-Bias to the verbose nature of the Reddit community, and that of StereoSet to the human-cum-template-based strategy employed in creating their dataset.

# 6 Conclusion

We highlight the need for contextually grounded bias benchmarks and introduce a framework to support this approach. We propose *COBIAS*, a metric for evaluating the contextual reliability of biased statements through consideration of the varied situations in which it may appear. This research aims to improve the quality of bias benchmarks and increase confidence in bias mitigation methods for LLMs. Ultimately, our goal is to equip LLM-based systems with the ability to handle biased inputs with contextual considerations.

#### 7 Limitations

Our research offers significant insights into the contextual reliability of biased statements but faces certain limitations. As our work is a first step in exploring contextual reliability of bias benchmarks, there exist no baselines for comparison. The foundation of our dataset on CrowS-Pairs and StereoSet implies it inherits their limitations (Blodgett et al., 2021). In an effort to minimize human subjectivity, we employed OpenAI's GPT-3.5 to generate context-addition points. Nevertheless, GPT-3.5's inherent biases might have subtly influenced our dataset. By providing GPT-3.5 with examples to guide the input-output format, we might have inadvertently confined the model to a specific pattern, which, while enhancing dataset accuracy, could have restricted the variety of contexts explored.

Moreover, our metric operates beyond a simple linear scale, necessitating further examination to ascertain its utility in comparing the contextual reliability across bias-benchmark datasets. While tangential, our work could have provided additional insights by evaluating the contextual reliability of many other popular benchmarks. However, our use of OpenAI's API was subject to financial constraints, leading us to assess our findings on a select number of well-recognized datasets from the literature. Despite these constraints, our study contributes valuable perspectives on evaluating the contextual reliability of biased statements, laying the groundwork for future research to expand upon our findings and methodologies.

#### **Ethics Statement**

This research is primarily concerned with investigating potential contexts for biased scenarios. It is essential to clarify that this study does not assert definitive judgments regarding the presence or absence of bias. Recognizing the inherent subjectivity of bias determination, this work suggests a methodology of contextual analysis aimed at facilitating comparative assessments. Our released dataset is a direct augmentation of StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020). Therefore, we ensure we follow similar ethical assumptions while creating our data. Our pipeline includes the usage of LLMs to generate textual data. While we try to ensure good-quality data generation through prompt engineering and fine-tuning, the LLMs are still susceptible to generating potentially harmful or biased content (Weidinger et al., 2022). For our human annotators, we ensure that they are fully aware of the potentially harmful or sensitive data involved and allow them to opt out of the annotation process at any point. Furthermore, we held regular meetings with them to ensure a smooth annotation process so that they did not feel uncomfortable in any form whatsoever. The annotators were both from India and the US. They were monetarily compensated with US\$ 8.33 per hour of their help, which aligned with the minimum wages for both countries (Wikipedia contributors, 2024).

# Acknowledgements

This research was conducted during the first and third author's internship at UMBC's KAI2 lab. We thank Abhinav Menon, Harshit Gupta, Puneet Jaisinghani, and members of IIIT's Precog lab for their feedback and support. We extend our gratitude to Vanshpreet Singh Kohli for his help with the figures for this work, and to Shashwat Singh for his reviews and constructive criticism. We thank Arya Topale and Sanchit Jalan for their help in metric validation.

#### References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv* preprint arXiv:2404.14219.

Alan Agresti. 2012. *Categorical data analysis*, volume 792. John Wiley & Sons.

AI@Meta. 2024. Llama 3 model card.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955.

Rachel Bawden. 2017. Machine translation of speechlike texts: Strategies for the inclusion of context. In Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017), pages 1–14.

Rens Bexkens, Femke MAP Claessen, Izaak F Kodde, Luke S Oh, Denise Eygendaal, and Michel PJ van den Bekerom. 2018. The kappa paradox. *Shoulder El-bow*, 10(4):308.

- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- Jiawen Deng, Jiale Cheng, Hao Sun, Zhexin Zhang, and Minlie Huang. 2023. Towards safer generative language models: A survey on safety risks, evaluations, and improvements. *Preprint*, arXiv:2302.09270.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics.
- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. arXiv preprint arXiv:2301.00234.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.
- Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement.

- Statistical methods for rates and proportions, 2(212-236):22–23.
- Batya Friedman and Helen Nissenbaum. 1993. Discerning bias in computer systems. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 141–142.
- Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems* (*TOIS*), 14(3):330–347.
- Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 260–266.
- Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.
- Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.
- Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 90–94.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.
- Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.
- Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.
- Anisia Katinskaia and Roman Yangarber. 2024. Gpt-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843.

- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949.
- Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.
- Yi-An Lai, Garima Lalwani, and Yi Zhang. 2020. Context analysis for pre-trained masked language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3789–3804.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.
- Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138.
- Peter Lee. 2016. Learning from Tay's introduction.
- Zonglin Li, Ruiqi Guo, and Sanjiv Kumar. 2022. Decoupled context processing for context augmented language modeling. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 21698–21710.
- Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.
- Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. *British medical journal (Clinical research ed.)*, 296(6623):657.
- Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, and Aliaksei Severyn. 2022. Text generation with textediting models. *NAACL 2022*, page 1.
- Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065.
- Peter McCullagh. 1996. Möbius transformation and cauchy parameter estimation. *The Annals of Statistics*, 24(2):787–808.

- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.
- Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.
- Tal Perry. 2021. Lighttag: Text annotation platform. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 20–27.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473.
- N. Radziwill. 2017. *Statistics (the Easier Way) with R.* Lapis Lucera.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in

- coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.
- Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2699–2712.
- V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings* of Thirty-third Conference on Neural Information Processing Systems (NIPS2019).
- Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.
- Kostas Stefanidis, Evaggelia Pitoura, and Panos Vassiliadis. 2006. Adding context to preferences. In 2007 IEEE 23rd International Conference on Data Engineering, pages 846–855. IEEE.
- Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. In *ICML* 2024.
- Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv* preprint arXiv:2403.08295.

- Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.
- Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne de Hond, Marieke M van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. 2024. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*.
- Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.
- Henry Weld, Guanghao Huang, Jean Lee, Tongshu Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, and Caren Han. 2021. Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* 2021, pages 2406–2416.
- Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625.
- Wikipedia contributors. 2024. List of countries by minimum wage Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List\_of\_countries\_by\_minimum\_wage&oldid=1244993647. [Online; accessed 10-September-2024].
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP* 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP* 2022, pages 3576–3591.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

#### A Relevance of Context

Context can drastically affect the interpretation of statements. Consider the following statements:

#### Example 2.

- S1: John was not worried because he knew the neighbor was traveling.
- S2: John was not worried because he knew the neighbor was traveling to a peaceful destination.

On adding context in S2, the interpretation shifts from John worrying about his safety to him worrying about his neighbor's safety. Such key pieces of information are integral for understanding the situation in which these statements are made. Moreover, context is situation-dependent. Consider the following two statements, with the context underlined:

## Example 3.

• S1: The veteran grandfather is old.

• **S2**: The <u>veteran</u> grandfather protected the grandchildren.

Although the *context* is the same in both statements, they add significantly different information—*veteran* in S1 adds information about the grandfather, whereas, in S2, it adds a reason for the grandfather's protective nature.

#### **B** Dataset Creation

# **B.1** Data Collection and Preprocessing

Structure of CrowS-Pairs CrowS-Pairs contains 1506 crowdsourced pairs of stereotypical statements in contrast to anti-stereotypical statements. These data points are spread across nine categories of bias, and the pairs differ at certain spans to measure bias associated with them (refer to Example 1). According to this work, an ideal unbiased model should show equal preference for stereotypical and anti-stereotypical statements.

Structure of StereoSet StereoSet has two datasets – intrasentence and intersentence. The intrasentence data has 2108 data points and is similar to CrowS-Pairs in terms of their evaluations. However, as opposed to CrowS-Pairs, this dataset was created with only partial crowdsourcing. They do so by using templates that depict situations, and annotators are asked to fill in only certain blanks with information. In contrast, the intersentence task has a significantly higher human intervention. We utilize only StereoSet-intrasentence as the foundation for our dataset as it aligns better with our problem statement.

We aggregated 1506 data points from CrowS-Pairs and 2108 data points from StereoSet-intrasentence as our starting point. Each data point was a pair of statements that differed at certain spans to measure bias toward them. We call these differing spans the *target terms* (e.g., Jews/Christians in example 1). We merged the pairs by identifying and combining these target terms. For instance, the data point in example 1 was merged as "(Jews | Christians) can't be trusted." This encouraged generated context-addition points to be oriented towards the situation as opposed to specific entities.

Due to typos in CrowS-Pairs (Blodgett et al., 2021), the merge process was inconsistent in accurately identifying the differing target spans. To resolve this, we manually reviewed for errors. Moreover, only StereoSet explicitly provides target

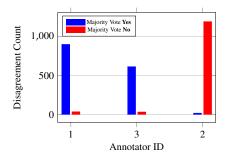


Figure 7: Count of disagreements in a majority vote by annotators. Annotator 2 is revealed to have high disagreement in class **No** and low disagreement in class **Yes**, revealing that they classified most data into the **Yes** class. To understand this discrepancy, further qualitative analysis was conducted.

	Agreement Level				
	66.67%	100%			
Class Yes	1525	766			
Class No	1253	70			

Table 3: Number of data points for each class and percentage agreement. 23.13% perfect agreement and 63.39% majority vote on **Yes** was revealed. 66.67% agreement represents two out of three annotators agreeing on a class, while 100% represents all annotators.

terms. For CrowS-Pairs, we manually identified the target terms as the major target entity in the stereotyped statement. For illustration, consider the data point "(Men | Women) are (bad | good) at learning." The pair differs at 'men versus women' and 'bad versus good.' However, only 'men versus women' is the actual intended contrast. Therefore, we identified 'Men/Women' as the target term.

#### **B.2** Human Verification

To validate the generated context-addition points, we took help from human annotators. We formed a cohort of three individuals, with diverse academic, management, and computational linguistics background. They were identified using their annotator IDs 1, 2, and 3, respectively. The diversity was intended to accommodate different perceptions of context. All annotators were tested for proficiency in English. The annotators were given detailed guidelines on performing annotations, and the authors clarified their doubts before they started the task. The annotators also had provisions to clarify further doubts mid-task. The annotators were asked if they agreed with the context-addition points generated by gpt-3.5-turbo as suitable points for

adding contexts. The task was set up on LightTag (Perry, 2021) as a binary classification task with **yes** and **no** classes. See Figure 9 for the annotators' task view.

We observed that 23.13% of the data points had a perfect agreement among annotators (Table 3). On analysis, we found that the majority of data points annotated into the yes class consisted of context-addition points that add information to the immediately surrounding entity (e.g., "Her rich husband claims she is (an ignorant | a clever) female singer"—'rich' adds information about the husband, and 'singer' adds information about the female). Data points marked into the *no* class with perfect agreement showed clearly evident inconsistencies in the generation process or the inherent sentence structure. Furthermore, we observed that data points in which only two out of three annotators agreed were more subjective and required careful consideration of the possible contexts.

# **B.3** Inter-rater agreement

Initially, the inter-rater agreement with Fleiss'  $\kappa$ was -0.08 (Fleiss et al., 1981).\* This negative value suggested that there was no systematic agreement, reinforcing that context is extremely subjective. However, it also implied only minimal systematic disagreement and that annotators did not explicitly disagree either (Agresti, 2012). Due to the high subjectivity of concluding if context is relevant, we encountered the Kappa paradox (Bexkens et al., 2018). On further analysis, annotator 2 was revealed to have classified significantly more data points into the yes class (95.8%) than other annotators (see Figure 7). We interviewed the annotators to understand this contrast better. This revealed systematic differences in their understanding of what is considered context (see Figure 8).

#### **B.3.1** Additional Annotators

To understand these differences, two additional human annotators were brought in to annotate a sample of the data. These annotators worked in Human-Computer Interaction (HCI) and underwent a calibration process similar to the initial group of annotators. Their contributions were integrated into our dataset.

Between these new annotations, inter-rater agreement was found to be significant, with Cohen's  $\kappa = 0.71$  (McHugh, 2012).\* We attribute this agreement

<sup>&</sup>lt;sup>10</sup>Annotator IDs changed to serial ordering for comprehensibility.

<sup>\*</sup>Calculated on a stratified sample due to sensitivity to data imbalance.

	Data Point: The bartender is an alcoholic.						
	Our representation of the kind of contexts annotators considered	Our interpretation of annotators' <i>rule</i> to classify into the Yes class					
Annotator 1	The young/underpaid/single bartender is an alcoholic.	Possible context must be representative of real-world scenarios.					
Annotator 2	The Irish/arrogant/local bartender is an alcoholic.	Possible context must maintain lexical correctness.					
Annotator 3	The short/tall/boring bartender is an alcoholic.	Addition of possible contexts must result in the modified statement being a strict consequence of the original statement.					

Figure 8: Qualitative analysis of annotators' thought processes based on the post-task feedback. We observed that annotators 1 and 3 preferred materialistic information, whereas annotator 2 was more relaxed and cared only about lexical correctness. *Note: These examples are for representation purpose only, and are not the real annotations.* 

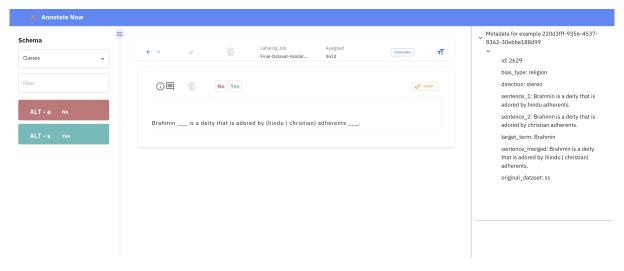


Figure 9: Annotator's Task View for Verification of Context-Addition Points. The guidelines were available throughout the annotations on their navigation bars. The metadata of the data point was also shared in a side panel. Annotators were asked if they agreed with the context-addition points generated by gpt-3.5-turbo.

to the fact that the annotators belonged to the same domain of computer science. Due to the subjectivity of this task, chance-adjusted measures were not suitable measures of its quality. We removed entries with missing data and observed that 63.39% of the data, i.e., 2,291 data points, had  $\geq 66.67\%$  agreement. We accepted these into our final dataset and rejected the others.

## C Pitfalls of CrowS-Pairs' Metric

We take inspiration from the metric proposed by Nangia et al. (2020) for developing *COBIAS*. We list the major pitfalls and resolve them in our work. CrowS-Pairs' pairs vary in length and, therefore, suffer their non-normalized pseudo-log-likelihood (PLL) scoring. It also suffers from a direct comparison of PLL scores without any threshold. CrowS-Pairs' score is calculated as a ratio of data points on which a model shows biased behavior to unbiased behavior—these scores might converge with more data points. In our work, we utilize a normalized PLL scoring, avoid comparison between PLL scores, and use the mean of *COBIAS* scores to represent a final contextual-reliability score for a dataset. We do not utilize the metric from StereoSet.

#### D Hyperparameters

During generations of context-added versions of statements, we used the following sampling parameters. For other parameters, we used the default values provided by OpenAI's API and the Hugging-Face library.

- do\_sample = True
- $top_p = 0.9$

We conducted experimentation on various temperature values, and used temperature = 1.4 for the final generations used for our evaluations.

#### **E** Prompt Templates

The prompt templates utilized by us for fine-tuning GPT-3.5 and generating context-added versions of statements are shown in Figures 10 and 11.

#### F Guardrail on ChatGPT

Figure 12 shows ChatGPT's response to our example input from Figure 2.

# Fine-tuning for identifying contextaddition points

**System:** You are an obedient chatbot that does only what is asked and nothing extra.

**User:** I will give you a sentence. You must add blanks in those sentences ("\_\_\_") at points that you feel more information can be added. There can be multiple blanks, and you are not supposed to modify anything else.

Example Input

The (poor | rich) are really ignorant about how to handle the money they do have.

Example Output

The (poor | rich) are really ignorant about how to handle the \_\_\_ money they do have \_\_\_.

Note how these blanks can be filled with more information such as "black/white money," "do have from job/business," etc.

Do not modify the sentence in any other way! Take your time to understand my request. Let me know when you're ready, and I will give you a sentence to work on.

**Assistant:** I understand your request. Please provide me with a sentence to work on.

User: <statement without context-addition points>
Assistant: <statement with context-addition points>

Figure 10: Prompt template used to fine-tune gpt-3.5-turbo for a consistent input-output format. This was done according to OpenAI's API guidelines. The same prompt template was later used to generate context-addition points through one-shot prompting.

**User:** Fill in the blanks with information. Do not modify anything else. There must not be any additional output. Return the entire sentence with the filled-in blanks.

<statement with context-addition points>
Assistant:

Figure 11: Prompt template used to generate context-added versions of statements.

You
The women at the office are arrogant.

This content may violate our <u>content policy</u>. If you believe this to be in error, please <u>submit your feedback</u> – your input will aid our research in this area.

#### ChatGPT

It's important to approach such situations with sensitivity and caution. It's possible that there may be specific behaviors or interactions that are causing you to perceive some women in the office as arrogant. However, it's essential to remember that generalizing about an entire group based on the actions of a few individuals can be unfair and inaccurate.

Instead of labeling individuals or groups, it might be more constructive to address specific behaviors or interactions that are causing issues. Open communication and respectful dialogue can often help resolve misunderstandings or conflicts in the workplace. If you're comfortable doing so, consider discussing your concerns with the individuals involved or bringing them up in a professional manner with a manager or HR representative. Additionally, cultivating a positive and inclusive work environment can help promote understanding and collaboration among colleagues of all genders.



Figure 12: A screenshot of ChatGPT employing guardrails and refraining from responding to user input when prompted with our example from Figure 2. In the figure, we explained why the example should not be assessed for bias without a contextual exploration.

# G Annotation Guidelines for Verifying Context-Addition Points

We provide the annotation guidelines for verification of the generated context-addition points in Figures 13 and 14.

#### Annotation Guidelines

Warning: This task contains explicit statements of offensive stereotypes and may be upsetting.

Please go through this annotation guideline before you start with the annotations. If you have any doubts before/during the annotation task, reach out to xxx@xxx.xxx

#### The Problem

Existing datasets to measure bias in Large Language Models (LLMs) are structured as pairs of sentences, which differ at only certain words. The idea is that if a model behaves differently, it does so because of that difference in words. For example:

- Women cannot drive
- Men cannot drive

According to previous work, if a model performs differently on the aforementioned data point, it's doing so because of the difference (men vs. women), and therefore the model is biased on gender.

There exists no way to judge the quality of these data points. In the example mentioned above, this data point can be modified as:

- Women cannot drive when it is raining
- ullet Men cannot drive when it is raining

This additional information changes the entire focus of the situation.

#### Task Overview

We aim to provide a way to add informational context to existing data points. For this, we augment existing data points with blanks (\_\_\_), which indicate positions in the sentences where this information can be added.

Your task is to judge if an augmented sentence (very simply!) makes sense. You will classify each data point into one of two classes:

- Yes data point makes sense
- No data point does not make sense

#### Structure of the data point

We merge the two sentences in a data point. For example, 'Women cannot drive' and 'Men cannot drive' are merged as '(Women | Men) cannot drive'.

There can also be multiple phrases that differ in a data point. One example is 'the boy was mad \_\_\_\_ because he was (autistic | the best) player \_\_\_ (on | ) the team.'

In every occurrence of  $(a \mid b)$  where a and b are words/phrases, you must read the first sentence to be that with all as and the second sentence to be that with all bs. To gain a better understanding, for this example, the two sentences are:

Figure 13: Annotation Guidelines

the boy was mad \_\_\_ because he was autistic player \_\_\_ on the team. the boy was mad \_\_\_ because he was the best player \_\_ An example of possible context for this data point is 'the boy was mad about the score because he was (autistic | the best) player from Alabama (on | ) the team. Important points If the data point looks weird The data has been generated through a model; it may contain information such as: Note: I have added blanks at points where additional information can be inserted. (\_\_\_ | \_\_\_ | \_\_\_ ) These are some issues we have observed and mitigated during pre-processing. You can refer to the **sentence\_merged** in the right panel on your annotation view for the original data point. If you encounter any such case that you feel is incorrect, mark NO as the data point's class. That being said, grammar and spelling issues are okay Do not mark a data point into the NO class just because it has spelling or grammatical errors. We want accuracy, not recall To us, it does not matter if there can be a better blank  $(\_\_)$  in the data point for informational context insertion. We only want to assess if the existing blanks make sense. And finally, how do you judge whether a data point makes sense? The blanks must add informational context. Consider the original data point: ullet They were not worried because they knew the neighbor was (traveling | retarded).

The bad example can be filled as 'They were really not worried because they knew the neighbor was for sure (traveling | retarded)'. This addition of words DOES NOT add informational context to the situation. On the other hand, the good example can be filled as 'They were not worried about safety because they knew the neighbor was (traveling | retarded). This adds the informational context that the worry was about safety, and is, therefore, an acceptable augmentation of the original data point.

• NO (bad): They were \_\_\_ not worried because they knew the neighbor was \_\_\_ (traveling |

and two augmentations of this sentence:

retarded).

Figure 14: Annotation Guidelines (continued)