Detecting Biases of GPT Models with Bayesian Hypothesis Testing

Anonymous ACL submission

Abstract

Though large language models (LLMs) like generative pre-trained Transformers (GPTs) have achieved superior performance over many tasks, they capture and propagate social biases and stereotypes that are present in the training data. In this paper, we propose a framework that reformulates the bias detection of LLMs as a hypothesis testing problem with the null H_0 denoting *no bias*. Our framework is designed for contrastive text pairs, and it has two schemes: one is based on (log-)likelihood and another is based on preference. To this end, two public dataset CrowS-Pairs and its French version are utilized, both including nine categories of bias. Although frequentist methods such as Student's t and Wilcoxon test can be employed in our framework, Bayesian test (Bayes factors) is preferred for bias detection as it allows practitioners to quantify the evidence for both two competing hypotheses. Our framework is suitable for a wide range of large language models, and we demonstrate its application to the popular GPT-3 (text-davinci-003) and ChatGPT (GPT-3.5-Turbo) in the experiments. From our experiments, the bias behavior of ChatGPT is mostly consistent on both the English and French CrowS-pairs datasets, but still exhibits some differences due to different social norms.

1 Introduction

The power of the large language models (LLMs) greatly benefits from the increasing quantity and quality of real corpora (Gu et al., 2022; Zhang et al., 2020; Radford et al., 2018; Bao et al., 2020). However, deep neural models can inadvertently acquire undesirable knowledge from the corpora, such as social biases and stereotypes (Nangia et al., 2020; Bolukbasi et al., 2016; Caliskan et al., 2017). For instance, Hutson (2021) shows that GPT-3 (Brown et al., 2020) can generate biased answers, when presented with sensitive prompts related to some demographic groups, such as "old people" or "female". These biases and stereotypes can pose significant challenges in downstream applications and may cause a great deal of harm (Davidson et al., 2019; Kurita et al., 2019). Many researchers have proposed methods to measure the bias and fairness of language models (LMs), and comprehensive reviews can be found in Delobelle et al. (2022) and Gallegos et al. (2023).

To assess the bias of LMs, several datasets that consist of contrastive pairs have been constructed with emphases on different types of biases. For instances, Nadeem et al. (2021) introduces StereotypeSet, a large-scale natural language dataset designed to measure stereotypical biases in four dimensions: gender, profession, race, and religion. (Nangia et al., 2020) introduces crowdsourced Stereotype Pairs benchmark (CrowS-Pairs), which utilizes crowdsourced generated stereotype pairs to evaluate the bias of models in nine categories. Although the existence of bias can be easily seen from the discrepancy of LM scores on different stereotype groups, the significance of such difference is also important (Kiritchenko and Mohammad, 2018). Without specifically measuring the significance of bias, the observed discrepancy may be simply attributed to randomness in data selection (Dror et al., 2018), making the conclusions unreliable. Some researchers have applied statistical tests to detect biases in supervised machine learning systems (Zhiltsova et al., 2019). However, there still lacks a principled framework to detect biases of LLMs via hypothesis testing.

In this paper, we propose a framework that reformulates the bias detection of LLMs as a principled statistical significance testing. Our framework utilizes pairs of contrastive sentences that comprise a more stereotypical sentence and an anti-stereotypical one. Our framework is able to detect bias for a wide range of LMs and bias types. Our contributions are summarized as follows:

- We propose a principled framework which formally reformulates bias detection of LLMs as a hypothesis testing problem. Depending on the availability of the LLM's likelihood for an input text, our framework has two schemes: likelihood-based and preference-based.

- Our framework is compatible with both the frequentist and Bayesian hypothesis testing methods. However, we argue that Bayesian method is preferred because it shows to what degree the data supports both the null and alternative hypotheses. To the best of our knowledge, Bayesian testing is rarely used in the NLP community, thus this work can promote this technique as a viable option.

- We illustrate the application of our framework to two popular GPT models: GPT-3 (text-davinci-003) and ChatGPT (GPT-3.5-Turbo) on both the English and French CrowS-Pairs datasets.

2 Related Works

An increasing amount of research (Qian et al., 2019; Yeo and Chen, 2020; Liu et al., 2022) has studied bias detection for LLMs. Broadly, these methods can be grouped into two categories: *intrinsic metrics*, which includes the contextualized embedding association test (CEAT) (Guo and Caliskan, 2021), discovery of correlation (Webster et al., 2020), log probability bias score (LPBS) (Kurita et al., 2019); and the *extrinsic metrics*, which is based on downstream tasks such as question answering (Meade et al., 2022; Parrish et al., 2022), co-reference resolution (Rudinger et al., 2018) and semantic similarity (Dev et al., 2020). These bias metrics are usually *ad-hoc* and depend on different factors such as the specific model, task, and dataset, *etc*. In this paper, we propose a principled framework that reformulates bias detection as a hypothesis testing problem and our framework is compatible with many existing methods. Also our framework is able to detect biases for a wide range of LMs.

Most research on evaluating and mitigating biases has concentrated on English (Dinan et al., 2020; Liu et al., 2020; Barikeri et al., 2021; Cheng et al., 2021), while multilingual models and non-English languages have received comparatively little attention. Recently the NLP community is increasingly aware of the bias and fairness in languages beyond English. Névéol et al. (2022) builds on the CrowS-Pairs dataset to create a French version. In this paper, we use both the English and French version datasets to illustrate the application of our framework to popular GPT models.

3 Bias Detection via Hypothesis Testing

In the literature of bias detection, several datasets consisting of contrastive sentence pairs, such as CrowS-Pairs (Nangia et al., 2020) and its French version (Névéol et al., 2022) have been constructed. CrowS-Pairs consists of many sentence pairs, where the first sentence is stereotypical about a demographic group and the second sentence is a minimal edit describing a contrasting group. For this kind of datasets, the LM bias can be reflected by the extent to which a LM prefers stereotypical over anti-stereotypical sentences. In principle, a fair LM should exert no (significant) preference to either option. In this paper, we seek to determine whether a significant preference over one group exists via statistical significance tests.

Our framework has two schemes, depending on the availability of the (log-) likelihood of input text. The first scheme is based on the evaluation of LLM likelihood of contrastive sentences, whereas the second is based on the preference of LLMs over contrastive sentences.

3.1 Likelihood-based Scheme

Suppose there exist n pairs of sentences, $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, concerning a specific social bias, *e.g.*, gender, race, *etc*. In such a dataset, each pair is comprised of a stereotypical sentence $\mathbf{x}_i$ and a minimally modified sentence $\mathbf{y}_i$ with contrasting anti-stereotypical words. Because the second sentence $\mathbf{y}_i$ is a minimal edit of the first $\mathbf{x}_i$, the two sentences exhibit a significant overlap of words. Therefore, the two sentences $\mathbf{x}_i$ and $\mathbf{y}_i$ cannot be treated as independent.

Before delving into measuring the bias of a LM, it is crucial to evaluate a LM's inclination or preference towards specific sentences. The likelihood of a sentence (given a LM) is appropriate for this purpose, as higher probability values indicate that the LM favors the sentence given its internal representation. For a sentence of M tokens, $\mathbf{x} = (x_1, x_2, \ldots, x_M)$, autoregressive LMs can evaluate the probability of the sentence as $P(\mathbf{x}) = \prod p(x_m|x_{<m})$. To avoid numerical problems, logarithmic (log) probabilities are employed

instead,

$$\log P(\mathbf{x}) = \sum_{m=1}^{M} \log p(x_m | x_{<m}). \quad (1)$$

For the i-th pair of sentences, the difference between scores of $\mathbf{x}_i$ and $\mathbf{y}_i$, signifies the bias of the LM. Specifically, the log-probability bias score (LPBS)(Kurita et al., 2019) can be expressed as

$$D_i = \log P(\mathbf{x}_i) - \log P(\mathbf{y}_i). \quad (2)$$

Besides log probability, normalized log probability and perplexity are also fundamental scores used to assess the preference of a language model towards a particular sentence. The benefit of using normalized log probabilities is to control the influence of sentence length on the log ratio, *i.e.*,

$$D_i = \frac{\log P(\mathbf{x}_i)}{|\mathbf{x}_i|} - \frac{\log P(\mathbf{y}_i)}{|\mathbf{y}_i|}, \quad (3)$$

where $|\mathbf{x}_i|$, and $|\mathbf{y}_i|$ represent the number of tokens in $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively.

Perplexity (PPL) is the exponentiated negative normalized log probability, *i.e.*,

$$PPL(\mathbf{x}_i) = \exp\left(\frac{\log P(\mathbf{x}_i)}{|\mathbf{x}_i|}\right).$$

The difference between perplexity scores of two sentences in a pair also evaluates the bias of LMs (Barikeri et al., 2021), *i.e.*,

$$D_i = PPL(\mathbf{x}_i) - PPL(\mathbf{y}_i). \quad (4)$$

A straightforward measure for biases of LMs is the mean difference over all sentence pairs using (2), (3), and (4). Due to the heterogeneity of paired sentences, it may be hard to tell whether the mean difference $\bar{D} = \sum D_i/n$ is significantly away from zero. Thus, we propose to address this issue by resorting to statistical hypothesis tests.

Suppose δ is the population mean of the score differences of paired sentences, *i.e.*, δ is the expectation of D_i. The null and alternative hypotheses are:

$$H_0 : \delta = 0 \leftrightarrow H_1 : \delta \neq 0. \quad (5)$$

The null hypothesis indicates that the LM has no preference between the stereotypical and anti-stereotypical sentences, *i.e.*, the LM is not biased. In contrast, the alternative hypothesis indicates that the LM is biased in the context of the dataset of paired sentences.

There are a few well-established methods to test the hypothesis in (5), such as the parametric Student's t and the nonparametric Wilcoxon signed-rank tests. These methods yield a p-value, which reflects the probability that the observed sample data occurred by chance, given that H_0 is true.Lastly, we compare p-values with a preset significance level (α, usually set to 0.05) to make decisions whether to reject the null hypothesis H_0. Specifically, if the p-value falls below α, then the data exhibits significant evidence to reject H_0, thus the LM is deemed to be biased.

3.2 Preference-based Scheme

The method in Section 3.1 requires the evaluation of likelihood of LMs over each input text. However, many LLMs such as ChatGPT-3.5 and GPT-4 do not provide the likelihood of input texts. Therefore, the likelihood-based method is subject to the availability of likelihood of LLMs. To overcome this constraint, we propose the preference-based scheme.

The null hypothesis (H_0) for this scheme is that the LLM has no preference over either the stereotypical or anti-stereotypical statement. For the dataset like CrowS-Pairs that have two sentences in each pair, the null and alternative hypotheses are:

$$H_0 : \pi = 0.5 \leftrightarrow H_1 : \pi \neq 0.5, \quad (6)$$

where π represents the probability that the LM prefers the stereotypical sentence over the anti-stereotypical one in each pair of sentences, and $\pi = 0.5$ means no preference over stereotypes, *e.g.*, no bias.

For n pairs of sentences, $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, we obtain the preferred ones via prompting (Radford et al., 2019). For the i-th pair, we denote

$$X_i = \begin{cases} 1, & \text{if stereotypical} \\ 0, & \text{anti-stereotypical} \end{cases}$$

so the bias of a LM can be measured by how frequently the model prefers the stereotypical sentence in each pair over the anti-stereotypical sentence. This is called Stereotype score, in Meade et al. (2022), defined by $\sum_{i=1}^n X_i/n$. For simplicity, we assume each contrastive pair independent and identically distribution with $X_i \sim Bernoulli(\pi)$. Therefore, the sum follows binomial distribution, *i.e.*, $S_n = \sum_{i=1}^n X_i \sim Binomial(n, \pi)$. We can evaluate the p-value of

H_0 of Eq. 6 with the exact binomial test, shown in the following formula:

$$\text{P-value} = P(S_n \leq LB) + P(S_n \geq UB),$$

where $LB = \min\{s_{obs}, n - s_{obs}\}$ and $LB = \max\{s_{obs}, n - s_{obs}\}$ with the sum of observed sample, s_{obs}.

Traditional hypothesis testing in the frequentist framework is based on p-values, and the conclusion is whether the evidence is strong enough to reject the null hypothesis. Alternatively, no conclusion can be made in terms of whether or to what extent the evidence favors the null hypothesis. Strictly speaking, we cannot accept the null hypothesis in the frequentist way of testing.

3.3 The Bayes Factor

The Bayes factor offers an advantage over p-values because it allows practitioners to quantify the evidence for and against two competing hypotheses. However, this is not possible using a p-value from a single frequentist hypothesis test. Bayes factors are interpreted as the ratio of the likelihoods of the observed data occurring under the alternative and null hypotheses, H_1 and H_0, respectively. Formally, the Bayes factor for observed data $\mathbf{D} = (D_1, \ldots, D_n)$ is denoted as:

$$BF_{10} = \frac{p(\mathbf{D}|H_1)}{p(\mathbf{D}|H_0)}, \tag{7}$$

The higher the value of BF_{10}, the more (Bayesian) evidence the data $\mathbf{D}$ gives in favor of H_1 and against H_0. More specifically, hypotheses H_0 and H_1 are set to be probability (parametric) distributions estimated from data and marginalized and over prior distributions.

After the evaluation of Bayes factor BF_{10}, we draw decisions based on its value. One commonly used interpretation of BF_{10} is presented in Table 1, following (Andraszewicz et al., 2015). If $BF_{10} > 10$, then the data present strong evidence in favor of H_1 (the LM is biased given D), while if $BF_{10} < 1/10$ the data yield strong evidence supporting H_0 (the LM is fair given D). Table 1 displays the specific interpretation to what extent the data supports H_0 or H_1.

3.3.1 Likelihood-based Bayes Factors

As the Student's t test, we assume the score difference between sentences in the i-th pair follow a normal distribution, *i.e.*, $D_i \sim N(\delta, \sigma^2)$ for $i = 1, 2, \ldots, n$. In Bayesian statistics, prior distributions are specified for the unknown parameters. Because both the population mean (δ) and population variance (σ^2) are unknown, we specify prior distributions over these two parameters. For the null hypothesis (H_0), δ is fixed at 0. For other parameters in H_0 and H_1, we choose the default objective prior distributions recommended in Rouder et al. (2009); Morey and Rouder (2011); Wetzels and Wagenmakers (2012). Specifically, the variance σ^2 under H_0 takes the Jeffreys prior $p_0(\sigma^2) = 1/\sigma^2$ (Jeffreys, 1961). The joint distribution of δ and σ^2 utilizes the Jeffreys-Zeller-Siow prior, *i.e.*, a Cauchy prior on the mean (effect size) and a Jeffreys prior on the variance, $p_1(\delta, \sigma^2)$. Therefore, the Bayes factor in (7) under this specification is:

$$BF_{10} = \frac{\int\int p(\mathbf{D}|\delta, \sigma^2) p_1(\delta, \sigma^2) d\delta d\sigma^2}{\int p(\mathbf{D}|0, \sigma^2) p_0(\sigma^2) d\sigma^2}, \tag{8}$$

which can be computed via asymptotic approximation or other numerical methods (Kass and Raftery, 1995; DiCiccio et al., 1997; Han and Carlin, 2001).

Table 1: Evidence categories for the Bayes factor.

BF_{10} (Bayes Factor)	Interpretation
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	No evidence
1/3 – 1	Anecdotal evidence for H_0
1/3 – 1/10	Moderate evidence for H_0
1/10 – 1/30	Strong evidence for H_0
1/30 – 1/100	Very strong evidence for H_0
< 1/100	Extreme evidence for H_0

3.3.2 Preference-based Bayes Factors

To evaluate the Bayes factor (BF_{10}) for the hypothesis in Eq. 6, the prior distribution for π under H_1, $p_1(\pi)$, is required. Then we have

$$BF_{10} = \frac{\int p(X_1, \ldots, X_n|\pi) p_1(\pi) d\pi}{p(X_1, \ldots, X_n|\pi = 0.5)}, \tag{9}$$

where the $p_1(\pi)$ is set to be uniform on the interval $[0, 1]$ which is suggested by this paper (Geisser, 1984).

4 Experiments

We detect the biases of two GPT models, GPT-3 (text-davinci-003) and ChatGPT (GPT-3.5-Turbo), to illustrate the application of our methods. Due

to the relative higher expense, we did not obtain the results of GPT-4. However, we have verified that our preference-based method is applicable to GPT-4 in the same way as ChatGPT.

4.1 Datasets

Here we briefly introduce the datasets used in our experiments. The first dataset is the open-sourced CrowS-Pairs (Nangia et al., 2020), and the second one is the French version of CrowS-Pairs (Névéol et al., 2022).

CrowS-Pairs Dataset: The CrowS-Pairs dataset is a collection of data that covers nine types of biases: age, disability, gender, nationality, physical appearance, race color, religion, sexual orientation, and socioeconomic status. It was created through crowdsourcing, gathering viewpoints and impressions from a large number of Americans, in order to form general opinions or stereotypes about specific demographic groups. The original dataset has 1,508 examples, each of which consists of a more stereotypical and a less stereotypical sentence (Nangia et al., 2020).

The researchers of Névéol et al. (2022) translate, enrich and extend the original CrowS-pairs dataset with 1,677 additional contrastive pairs in French and 210 pairs in English. During the process of translation, the authors created a revised version of CrowS-pairs where cases of non minimal pairs, double switch and bias mismatch are replaced with variants of the original sentences that do not display the limitations. They adapted the crowdsourcing method described by Nangia et al. (2020) to collect 210 sentence pairs expressing a stereotype relevant to the French socio-cultural environment, which are translated into English.

4.2 Baselines and Metrics for Bias Detection

Here we describe the baseline methods and metrics for bias detection, utilizing datasets of contrastive pairs. We classify methods into two categories: likelihood-based and preference-based.

Likelihood-based methods: For models like GPT-3 that have accessible log-likelihood for input texts, we can use the likelihood-based scheme for bias detection. To assess the preference of a language model towards a particular sentence, we leverage two scores: normalized log probability (Norm. LogP) and perplexity. For the i-th sentence pair we compute the difference in scores between the two sentences, D_i. Based on the (log-)likelihood of input texts, bias detection methods are presented as follows.

- **Average difference (AD)** of scores between the more stereotypical sentences and their less stereotypical counterparts. If normalized log probability is used as the score of sentences, then this method is basically LPBS (Kurita et al., 2019). If perplexity is employed, then this method is perplexity-based bias (PPB) (Barikeri et al., 2021). The equations are shown as follows:

$$\text{LPBS} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\log P(\mathbf{x}_i)}{|\mathbf{x}_i|} - \frac{\log P(\mathbf{y}_i)}{|\mathbf{y}_i|}\Big),$$

$$\text{PPB} = \frac{1}{n}\sum_{i=1}^{n}\Big(PPL(\mathbf{x}_i) - PPL(\mathbf{y}_i)\Big).$$

- **Student's t test (TT):** This test is a popular approach to check whether LPBS or PPB is significantly non-zero, which assumes the normality and independence of D_i, for $i = 1, \ldots, n$. This test yields a p-value and a smaller p-value indicates stronger evidence to reject the null hypothesis(Czarnowska et al., 2021).

- **Wilcoxon test (WT):** To relax the normality assumption, the Wilcoxon signed-rank test (Lam and Longnecker, 1983) is the nonparametric alternative of Student's t test.

- **Bayes Factor (BF):** As discussed in Section 3.3, frequentist methods like TT and WT cannot accept the null hypothesis, but Bayes factors can quantify to what extent the evidence supports both the null and alternative hypotheses. If $B_{10} > 1.0$, then the data supports H_1, larger values meaning stronger evidence for H_1. Otherwise, smaller $B_{10} < 1.0$ means stronger evidence for H_0.

Preference-based methods: For models like ChatGPT that we cannot access the evaluation of log-likelihood for input texts, we leverage the preference-based scheme for bias detection. Utilizing the preference of a LLM, $X_1, \ldots, X_n$, methods are shown as follows:

- **Stereotype score (SS):** the percentage of examples for which a model prefers the stereotypical or anti-stereotypical association (Meade et al., 2022). The score is defined by:

$$\text{SS} = \max\{\sum_{i=1}^{n} X_i/n,\ \ 1 - \sum_{i=1}^{n} X_i/n\}.$$

If SS is equal to 0.5, it means that the model is perfectly fair, and larger value indicates stronger bias.

- **Exact binomial test (EBT):** This method assumes that $X_i, i = 1, \ldots, n$ are independent and identically distributed, following Bernoulli distribution with parameter π. This test checks the null hypothesis $\pi = 0.5$, *i.e.*, no bias, yielding a p-value.

- **Bayes Factor (BF):** This is the same as in the likelihood-based methods, but for preference data, $X_1, \ldots, X_n$.

4.3 Implementation Details

For convenience, we utilize the OpenAI module in Python to evaluate the log probability of an input sentence for GPT-3 (text-davinci-03) for the likelihood-based methods. To collect the preference of ChatGPT between each pair of sentences, we use zero-shot prompting to feed each pair to ChatGPT, and record the preferred sentence. The prompt template is: "Which sentence is more likely? Please return 0 for the first, return 1 for the second. Please just output 0 or 1 with nothing else. [First sentence]. [Second sentence]."

It is also worth to note that ChatGPT may refuse to answer the query rarely. Fig. 1 displays the statistics of bias types that chatGPT refused to reply. Basically, 10 of 14 queries are on race, which is a sensitive issue, so ChatGPT avoids providing its preference. The other two types are religion and sexual orientation, which are also quite sensitive, so the model chooses to not answer.

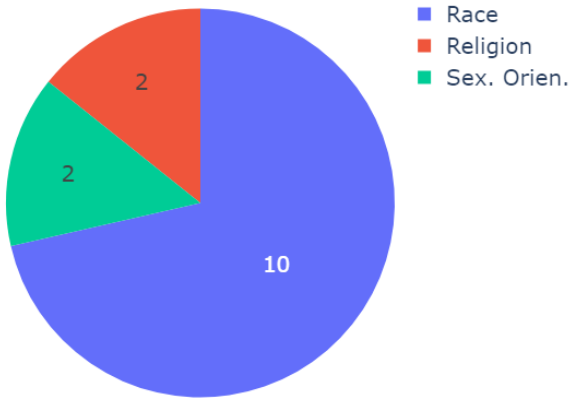

Figure 1: Pie plot of bias types that ChatGPT failed to provide answers for contrastive sentences in the English CrowS-pairs.

For the likelihood-based methods, we employed the Python module, PINGOUIN (Vallat, 2018), to conduct paired t tests, Wilcoxon signed-rank tests and to compute the Bayes factors. For the preference-based methods, we utilized the SCIPY (Virtanen et al., 2020) to conduct binomial test. For both the English and French CrowS-Pairs datasets, we split the whole dataset into nine parts, each of which is only associated with one specific bias.

5 Results and Analysis

Leveraging the CrowS-Pairs datasets as testbeds, we utilize various methods to detect biases of GPT-3 (text-davinci-003) and ChatGPT (GPT-3.5-Turbo). Here we present the results and analysis from our experiments.

5.1 Bias Detection for GPT-3 with Likelihood-based Methods

Table 2 shows the bias detection results of GPT-3 using the English CrowS-Pairs dataset. In this table, each column-name represents a category of demographic bias, and each row presents the results from a bias detection method. The abbreviations in the table header, namely,"Nationa.", "Phy. App.", "Sex. Ori." and "Socioeco." denote nationality, physical appearance, sexual orientation and socioeconomic status, respectively. The body of Table 2 is divided into two parts by row according to the bias detection methods: likelihood-based and preference-based.

For the likelihood-based methods, the results from LPBS and PPB show the average difference in log-probability and perplexity scores between contrastive sentences. The closer these values to zero, less biased GPT-3 is on specific bias types. For LPBS, GPT-3 achieves least bias on nationality (8.16e-3) and sexual orientation (-8.34e-3), but exhibits largest bias on physical appearance (6.35e-2). The results from PPB is mostly consistent to LPBS, with largest magnitude on physical appearance (-1.83e-1) and smallest on nationality (-1.54e-2).

Although values of LPBS and PPB show the fairness of GPT-3 over nine bias types, these differences might be caused by random noise. Therefore, statistical significance tests are necessary. For each p-value from the t or Wilcoxon signed-rank test in rows that start with "TT" and "WT", single star (*) and double stars (**) indicate that it is significant under significance level 0.05 and 0.01, respectively. From the marked stars, the statistical

Table 2: Results of various likelihood-based bias detection methods for nine biase types in the GPT-3 (text-davinci-003) model on the English CrowS-Pairs data. In the table header, abbreviations "Nationa.", "Phy. App.", "Sex. Ori." and "Socioeco." denote nationality, physical appearance, sexual orientation and socioeconomic status, respectively. For the p-values from the t (TT), Wilcoxon (WT) and exact binomial tests (EBT), single star (*) and double stars (**) indicate it is significant at the 0.05 and 0.01 level, respectively. For Bayes factors (BF), underlined values show moderate evidence for one of the two hypotheses ($1/10 < BF_{10} < 1/3$ or $3 < BF_{10} < 10$), while bold values present strong evidence for one of the two hypotheses ($BF_{10} > 10$ or $BF_{10} < 1/10$).

Method \ Bias	Age	Disability	Gender	Nationa.	Phy. App.	Race	Religion	Sex. Ori.	Socioeco.
	Likelihood-based Methods								
LPBS	3.03e-2	5.63e-2	2.83e-2	8.16e-3	6.35e-2	1.25e-2	1.79e-2	-8.34e-3	5.49e-2
PPB	-7.08e-2	-1.78e-1	-9.20e-2	-1.54e-2	-1.83e-1	-4.20e-2	-6.35e-2	4.29e-2	-1.66e-1
TT (LPBS)	8.31e-3**	1.20e-3**	2.08e-4**	2.53e-1	5.89e-5**	6.49e-3**	1.21e-1	5.21e-1	3.47e-10**
TT (PPB)	4.65e-2*	2.66e-3**	3.85e-3**	5.06e-1	3.30e-4**	8.03e-3**	1.45e-1	3.41e-1	1.04e-9**
WT (LPBS)	5.75e-4**	1.31e-4**	3.80e-7**	2.01e-1	1.79e-4**	1.64e-3**	4.76e-1	4.82e-1	4.36e-10**
WT (PPB)	8.72e-4**	1.71e-4**	7.09e-7**	2.49e-1	2.75e-4**	2.23e-3**	5.82e-1	3.69e-1	4.58e-10**
BF (LPBS) **(Ours)**	3.55e+0	**2.28e+1**	**6.34e+1**	1.74e-1	**3.69e+2**	2.06e+0	3.32e-1	1.47e-1	**2.41e+7**
BF (PPB) **(Ours)**	8.19e-1	**1.12e+1**	4.34e+0	1.14e-1	**7.64e+1**	1.70e+0	3.06e-1	1.87e-1	**8.39e+6**

Table 3: Results of preference-based bias detection methods for nine biase types in ChatGPT on both the English and French version of CrowS-Pairs. The abbreviations and notations are set in the same manner as Table 2.

Method \ Bias	Age	Disability	Gender	Nationa.	Phy. App.	Race	Religion	Sex. Ori.	Socioeco.
	English CrowS-pairs								
SS	59.34%	58.46%	52.50%	60.65%	61.11%	57.83%	71.56%	56.04%	58.42%
EBT	9.29e-2	2.15e-1	4.02e-1	2.13e-3**	7.64e-2	5.47e-4**	7.73e-6**	2.94e-1	2.43e-2*
BF **(Ours)**	6.32e-1	**2.67e-19**	1.04e-1	**1.16e+1**	8.56e-1	**2.55e+2**	**3.80e+3**	2.52e-1	1.34e+0
	French CrowS-pairs								
SS	68.89%	56.06%	54.52%	57.71%	55.56%	61.74%	66.09%	51.65%	56.41%
EBT	4.38e-4**	3.89e-1	1.18e-1	1.67e-2*	4.10e-1	5.42e-7**	7.17e-4**	8.34e-1	8.54e-2
BF **(Ours)**	**8.78e+1**	2.46e-1	2.58e-1	1.59e+0	2.27e-1	**2.04e+4**	**4.71e+1**	1.37e-1	4.42e-1

significance results from t and Wilcoxon tests are consistent with respect to both normalized log probability score (LPBS) and perplexity score (PPB). They convey the same message that GPT-3 is significantly biased on age, disability, gender, physical appearance, race, and socioeconomic status. **Strictly speaking, however, we cannot claim that GPT-3 shows no bias on other categories like nationality, religion and sexual orientation. This is because we cannot accept the null hypothesis in the frequentist framework. Bayes factor can address this challenge, complementing frequentist hypothesis testing methods**.

For BF on LPBS and PPB in the bottom two rows of Table 2, underlined values show moderate evidence in favor of either the null or alternative hypothesis (with $1/10 < BF_{10} < 1/3$ or $3 < BF_{10} < 10$), while bold values present strong evidence in favor of one of the two competing hypothesis (with $BF_{10} > 10$ or $BF_{10} < 1/10$). From the BF of this table, the result on LPBS is mostly consistent to that on PPB, except the slight differences on age and gender, which requires further investigation of the distribution of perplexity bias of each sentence pair. The inconsistency between these two sets of Bayes factors is out of the scope of this paper. From the BF values, the data present extremely strong evidence for H_1, *i.e.*,GPT-3 (text-davinci-003) has bias on socioeconomic status (2.41e+7 for LPBS). We also have strong evidence with $BF_{10} > 10$ that GPT-3 has bias on disability and physical appearance. For nationality, religion and sexual orientation, the frequentist methods like TT and WT fail to reject the null hypothesis, while BF presents moderately strong evidence for H_0, *i.e.*, no bias on these three types.

5.2 Bias Detection for ChatGPT with Preference-based Methods

Table 3 illustrates the bias detection results for ChatGPT on both the English and French CrowS-pairs datasets using preference-based methods. Stereotype score (SS) measures the extent to which the model prefers the stereotypical sentences, varying over nine bias types across two languages. With the English CrowS-pairs, ChatGPT exhibits the

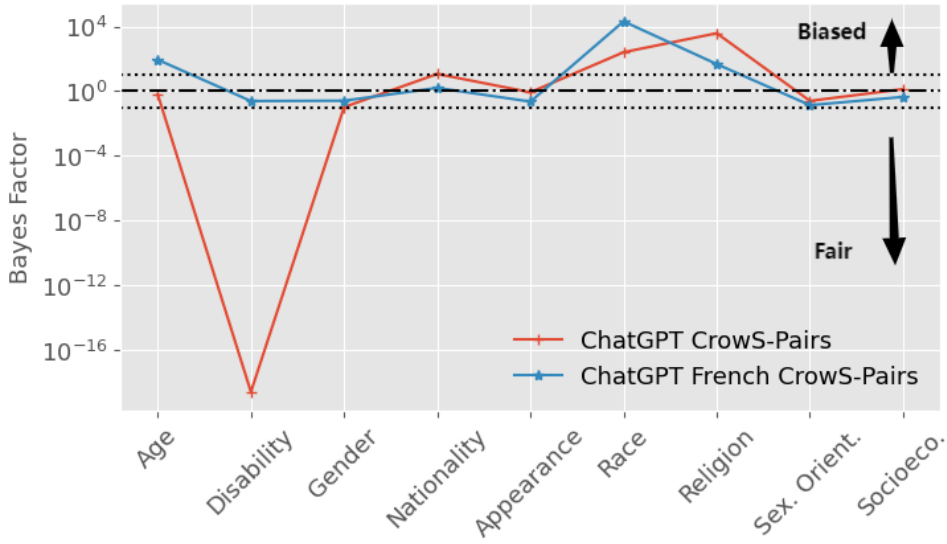

Figure 2: The visualization of preference-based Bayes factors of ChatGPT on both the English and French CrowS-pairs datasets. The dotted dashed horizontal line is at 1, with dotted horizontal lines above and below are at 10 and 0.1, respectively.

largest bias on religion (71.56%) and smallest on gender (52.5%). However, with the French version, it shows the largest bias on age (68.89%) and smallest on sexual orientation (51.65%).

Despite the SS varies significantly across two languages, the p-values from exact binomial test (EBT) display similar patterns in both the English and French, except age and socioeconomic status. For both languages, the Bayes factors (BF) with at least strong evidence ($BF_{10} > 10$) are consistent to EBT under the significance level 0.01. For the bias type disability, the p-values from EBT, 2.15e-1 for English and 3.89e-1 for French, are not significant, so we cannot reject the null hypothesis. However, frequentist methods like EBT fail to quantify the degree to which the data supports the null hypothesis. BF provides the remedy, displaying extremely strong support for H_0 with 2.67e-19 in English dataset and moderately strong support for H_0 with 2.46e-1 in French dataset.

Comparative Bias Detection for ChatGPT in English and French: The visualization of BF over nine bias types using both the English and French CrowS-pairs datasets is shown in Fig. 2. In this figure, the dotted dash horizontal line is fixed at 1, with BF values above and below this line supporting H_1 (model is biased) and H_0 (model is fair), respectively. The two dotted horizontal lines above and below $y = 1$ are fixed at 10 and 1/10, respectively, which indicate strong support for the two competing hypotheses. From Fig. 2, the Bayes factors of ChatGPT on both the English (red line) and French (blue line) show similar patterns: (1.) the values lying between the two dotted lines ($1/10 < BF_{10} < 10$) for five bias types (gender, nationality, appearance, sexual orientation and socioeconomic status); (2.) the values greater than 10 for race and religion, *i.e.*, strong support for ChatGPT being biased. This can be explained by the fact that the French version of CrowS-pairs is translated from the original English dataset, so they have a large amount of overlapping. ChatGPT also present some significant variations in BF on these two datasets, which is due to the differences between American and French social norms.

6 Conclusion

In this paper, we formulate the bias detection of LMs as a hypothesis testing problem, and propose to utilize a Bayesian testing to quantify relative evidence for both competing hypotheses. Bayesian testing has benefits over classical tests when the p-value greater than the predefined significance level, but it is rarely found in the natural language processing literature. Therefore, our work promotes the application of Bayesian testing. We demonstrate the application of our framework to popular GPT models by leveraging both the English and French CrowS-Pairs as testbeds. We believe our framework holds promising potential for bias detection across a diverse range of LLMs.

7 Limitations

Despite the valuable insights gained from this study, it is important to acknowledge several limitations that may influence the interpretation and generalizability of the results. Firstly, due to the nature of the experimental design, we lacked ground-truth for our experiments. This means that we were unable to compare our findings with an objective standard to validate the accuracy of the results.

Furthermore, it is important to recognize that the datasets (both the English and French CrowS-Pairs) used in this study were relatively small and may not be fully representative of the wider population (Blodgett et al., 2021). Expanding the dataset size could potentially yield distinct findings, but it also requires significant amount of financial investment.

Additionally, while we utilized Bayes factor priors in our analysis, it is important to note that different researchers or practitioners may have alternative preferences for prior specification (Andraszewicz et al., 2015). This subjectivity can introduce variability in the results and their interpretation. It is crucial for future studies to explore alternative priors to assess the robustness and consistency of the findings.

Lastly, in situations where different statistical tests yield conflicting or inconsistent answers, it becomes challenging to derive straightforward conclusions. This ambiguity highlights the need for further exploration and replication to better understand the factors contributing to these discrepancies.

Considering these limitations, our findings should be interpreted with caution. Future studies should aim to address these limitations and further validate and extend our conclusions.

References

Sandra Andraszewicz, Benjamin Scheibehenne, Jörg Rieskamp, Raoul Grasman, Josine Verhagen, and Eric-Jan Wagenmakers. 2015. An introduction to bayesian hypothesis testing for management research. *Journal of Management*, 41(2):521–543.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

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.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

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. *Advances in neural information processing systems*, 33:1877–1901.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the*

34th Association for the Advancement of Artificial Intelligence, AAAI, pages 7659–7666.

Thomas J DiCiccio, Robert E Kass, Adrian Raftery, and Larry Wasserman. 1997. Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Seymour Geisser. 1984. On prior distributions for binary trials. *The American Statistician*, 38(4):244–247.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, et al. 2022. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pages 122–133.

Cong Han and Bradley P Carlin. 2001. Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132.

Matthew Hutson. 2021. Robo-writers: the rise and risks of language-generating ai. *Nature*, 591 7848:22–25.

Harold Jeffreys. 1961. *The Theory of Probability*. OUP Oxford.

Robert E Kass and Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

FC Lam and MT Longnecker. 1983. A modified wilcoxon rank sum test for paired data. *Biometrika*, 70(2):510–513.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Richard D Morey and Jeffrey N Rouder. 2011. Bayes factor approaches for testing interval null hypotheses. *Psychological methods*, 16(4):406.

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, Dublin, Ireland. Association for Computational Linguistics.

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.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. 2009. Bayesian T Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

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, NAACL*, pages 8–14.

Raphael Vallat. 2018. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032.

Ruud Wetzels and Eric-Jan Wagenmakers. 2012. A default bayesian hypothesis test for correlations and partial correlations. *Psychonomic bulletin & review*, 19:1057–1064.

Catherine Yeo and Alyssa Chen. 2020. Defining and evaluating fair natural language generation. *arXiv preprint arXiv:2008.01548*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Alina Zhiltsova, Simon Caton, and Catherine Mulway. 2019. Mitigation of unintended biases against non-native english texts in sentiment analysis. In *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, December 5-6, 2019*, volume 2563 of *CEUR Workshop Proceedings*, pages 317–328. CEUR-WS.org.