REQUAL-LM: Reliability and Equity through Aggregation in Large Language Models*

Sana Ebrahimi
University of Illinois Chicago
sebrah7@uic.edu

Nima Shahbazi
University of Illinois Chicago
nshahb3@uic.edu

Abolfazl Asudeh
University of Illinois Chicago
asudeh@uic.edu

Abstract

The extensive scope of large language models (LLMs) across various domains underscores the critical importance of responsibility in their application, beyond natural language processing. In particular, the randomized nature of LLMs, coupled with inherent biases and historical stereotypes in data, raises critical concerns regarding reliability and equity. Addressing these challenges are necessary before using LLMs for applications with societal impact.

Towards addressing this gap, we introduce REQUAL-LM, a novel method for finding reliable and equitable LLM outputs through aggregation. Specifically, we develop a Montecarlo method based on repeated sampling to find a reliable output close to the mean of the underlying distribution of possible outputs. We formally define the terms such as reliability and bias, and design an equity-aware aggregation to minimize harmful bias while finding a highly reliable output. REQUAL-LM does not require specialized hardware, does not impose a significant computing load, and uses LLMs as a black-box. This design choice enables seamless scalability alongside the rapid advancement of LLM technologies. Our system does not require retraining the LLMs, which makes it deployment-ready and easy to adapt.

Our comprehensive experiments using various tasks and datasets demonstrate that REQUAL-LM effectively mitigates bias and selects a more equitable response, specifically the outputs that properly represents minority groups.

1 Introduction

In the ever-evolving realm of advanced technologies, Large Language Models (LLMs) have quickly emerged as versatile tools, extending their influence far beyond the boundaries of natural language processing (NLP). Many of the traditionally challenging tasks with decades of research in various fields of computer science are finding more effective resolutions with the help of LLMs. Let us consider Example 1 as a motivating example for subset selection using LLM.

Example 1: (Part 1) *Selecting a subset of candidates from a pool, based on a set of criteria is common across multiple applications ranging from journalism, to college admissions and job hiring. For example, consider the HR department of a sales company who wants to select a set of candidates for the performance award based on multiple criteria such as* SALES *and* CUSTOMER-SATISFACTION. *Passing the performance information of the employees, they can ask the LLM to select the candidate set.*

LLMs are sequential randomized approaches based on estimations learned from large textual datasets. In particular, based on the prompt and the sequence of tokens generated so far, each word (token) in the dictionary is assigned a probability. Then, the next token is generated probabilistically (proportional to the probabilities of the top-k or top-p%) using the parameter temperature. Consequently, the output may vary when the LLM is queried again. As a result, a valid concern, particularly for a decision maker, is *whether they should rely on the LLM's output* for taking action. In settings similar to Example 1, the reliability question is further significant, since a method to combine the performance criteria has not been specified, while small changes in the combination details may significantly change the output (Guan et al., 2019).

Another challenge that makes a single query to the LLMs unreliable arises for the *symmetric* settings, where the ordering between the input does not matter, i.e., shuffling the input should not impact the output. For instance, in Example 1 the ordering based on which the employees are passed to the LLM should not impact the output. Conversely, LLMs receive an input as a (ordered) sequence. As

*This work was supported in part by NSF 2107290.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 549–560

a result, as it was observed in (Gao et al., 2023), the output of the LLMs for symmetric problems vary when the input is shuffled. We also observed the same behavior in our experiments on a subset selection task, where the entities that are placed at the beginning of the list had a higher chance of being returned as the output.

To resolve these issues we introduce REQUAL-LM that, instead of relying on a single query to an LLM, follows a Monte-carlo method (Hammersley, 2013) based on repeated sampling. Particularly, viewing each LLM output as a sample from the underlying distribution of possible outputs, it identifies the centroid of a collection of samples as its estimation of the mean of the distribution, and returns the closest output to the centroid as the most reliable one. To further clarify this, let us consider Example 1 once again.

Example 1: (Part 2) *Observing the dependency of the LLM output with the input ordering, and to possibly consider various combinations of performance criteria, the HR department does not rely on a single output of the LLM. Instead* REQUAL-LM *enables issuing multiple queries to the LLM, each time shuffling the list of the employees. It then returns the "closest-to-centroid" of the obtained samples as the most reliable output.*

While being effective in practice, data-driven technologies have been heavily criticized for machine bias (Angwin et al., 2022), and LLMs are not an exception when it comes to bias. As a result, another valid concern when using LLMs for decision making is neutrality: to ensure that impact of historical biases and stereotypes are minimized and that values such as diversity are promoted.

Example 1: (Part 3) *The HR department would likes to maximize diversity in the selected awardees. In particular, they would like to prevent selecting a male-only list of employees.* REQUAL-LM *allows specifying two or more demographic groups and it minimizes the output bias (measured as the cosine-similarity difference of its output's embedding with different groups' representations).*

LLMs are among the fast-growing technologies, with new and advanced versions regularly emerging, while many of these systems are "black-box". Our system design is not dependent on a specific LLM, which makes it *a ready-to-apply wrapper* that works on top of *any* of the current and future closed-source and open-source LLMs. REQUAL-LM does not require pre-training or fine-tuning, is task-agnostic, and can handle non-binary demographic groups.

In the following, first in § 2 we carefully discuss the problem setting, introduce notations, and formally define terms such as reliability and bias. Next, in § 3 we review the architecture of REQUAL-LM, and develop our methodology for finding an equitable centroid and return the closest output to it, the one that is both equitable and reliable. The experimental evaluations, related work, and the discussions of the benefits and limitations of REQUAL-LM are provided in § 4, § 5, § 6, and § 8, respectively.

2 Preliminaries

– (Input) *Task:* We consider a task, such as subset selection, sentence completion, assembling a team of experts, etc., described as a prompt: LLM.
– (Input) *Demographic Groups:* We assume the existence of at least one sensitive attribute (e.g., `sex`) that specify the demographic groups $\mathcal{G} = \{\mathbf{g}_1, \cdots, \mathbf{g}_\ell\}$ (e.g., `{male, female}`). The demographic groups are used to specify the output bias.
– *LLM:* We assume access to (at least) one LLM, which is used for task answering. The LLM is randomized, i.e., the tokens are sequentially drawn based on the underlying distribution of the (top-k or top-p%) token-probabilities. We treat the LLM as a black-box oracle that upon querying generates an *output* based on the input prompt. Treating the LLM as black-box allows the adaptation of REQUAL-LM both for closed-source and open-source LLMs.
– *Text Embedding:* We rely on an external text embedding model that transforms a text into an embedding vector. Specifically, given a text O_i, it generates the vector representation $\vec{v}(O_i) = \vec{v}_i = \langle v_1, v_2, \cdots, v_d \rangle$. Our system, REQUAL-LM, is agnostic to the choice (but limited to the performance) of the embedding model, and can adapt any state-of-the-art text embedding technique. Without loss of generality, we use INSTRUCTOR – a method for generating task-specific embeddings in accordance with provided instructions (Su et al., 2023).

Given two text phrases O_i and O_j and their corresponding embeddings $\vec{v}_i$ and $\vec{v}_j$, the similarity between O_i and O_j is measured as the cosine similarity between their embeddings, i.e., $\mathcal{S}_{im}(O_i, O_j) =$

$\cos\angle(\vec{v}_i, \vec{v}_j)$. Similarly, the distance between O_i and O_j is defined as $\Delta(O_i, O_j) = 1 - \mathcal{S}_{im}(\vec{v}_i, \vec{v}_j)$.

Definition 1 (Reliability)**.** *Given a prompt I, let $\mathcal{O}_I$ be the universe of possible-to-generate outputs for I. Furthermore, let ξ be the probability distribution of outputs for I. That is, $\forall O \in \mathcal{O}_I$, $Pr_\xi(O)$ is the probability that O is generated for I. Let $\vec{\mu}_\xi$ be the mean of ξ in the embedding space. Then the reliability of an output $O \in \mathcal{O}_I$ is defined as its similarity to $\vec{\mu}_\xi$. That is,*

$$\rho(O) = \mathcal{S}_{im}(\vec{v}(O), \vec{\mu}_\xi)$$

Let $O \in \mathcal{O}_I$ be an output generated for the prompt I comprising a sequence of $|O|$ tokens $\langle t_1^O, t_2^O, \cdots t_{|O|}^O\rangle$ sequentially generated by the LLM. At each iteration i, let $Pr(t_i^O)$ be the probability of generating the token t_i^O. Then $Pr_\xi(O)$ can be computed as the product of its token probabilities. That is, $Pr_\xi(O) = \prod_{i=1}^{|O|} Pr(t_i^O)$.

Definition 2 (Bias)**.** *Consider a set of demographic groups $\mathcal{G} = \{\mathbf{g}_1, \cdots, \mathbf{g}_\ell\}$ and their corresponding vector representation*[1] *$\{\vec{\mathbf{g}_1}, \cdots, \vec{\mathbf{g}_\ell}\}$. The bias of an output O for a prompt I is computed as the maximum similarity disparity of the demographic groups with O. Formally,*

$$\beta(O) = \max_{\mathbf{g}_i, \mathbf{g}_j \in \mathcal{G}} \left| \mathcal{S}_{im}(\vec{v}(O), \vec{\mathbf{g}_i}) - \mathcal{S}_{im}(\vec{v}(O), \vec{\mathbf{g}_j}) \right|$$

Bias is sometimes inherent to the task at hand and is not harmful. For example, when the task involves summarizing or rephrasing a paragraph that is particularly written about a specific gender, the resulting output tends to be naturally biased towards that gender. We call this type of output bias as the *inevitable bias*. Formally, we say a bias level ε is inevitable if there is no valid output $O \in \mathcal{O}_I$ with a bias less than ε. In other words, for any output O' where $\beta(O) < \varepsilon$, we can say $O' \notin \mathcal{O}_I$. Therefore, we define the inevitable bias as $\beta_n(I) = \min_{O \in \mathcal{O}} \beta(O)$. We consider any bias that is not inevitable, discriminatory. Harmful stereotypes are in this category. We call this type of output bias as the *harmful bias*. Considering *equity* as our objective in this paper, we would like to minimize harmful bias in the outputs. The harmful bias of an output can be computed by subtracting its bias from the inevitable bias, i.e., $\beta_h(O) = \beta(O) - \beta_n(I)$.

After defining the terms and notations, we are able to formulate our problem: *given a task presented in the form of a prompt I, and including the demographic groups $\mathcal{G}$, the objective is to identify an output $O \in \mathcal{O}_I$, such that it maximizes $\rho(O)$ and minimizes $\beta_h(O)$.*

[1]Please refer to Appendix A for the details of obtaining the vector representations for the demographic groups.

3 Technical Details

3.1 Architecture Overview

Figure 1 shows the architecture of REQUAL-LM. Following the Monte-carlo method described in § 3.2, the first step is to obtain a set of iid output samples by issuing m independent queries to the LLM. The results are subsequently fed into the text embedding model, INSTRUCTOR, to obtain the vector representations $\{\vec{v}(O_1), \cdots \vec{v}(O_m)\}$. Next, the vector representations, as well as the vector representations of the demographic groups, pass on to the *aggregation function* (referred to as AVG in the figure). The aggregation function generates the vector representation that corresponds to the average of $\vec{v}(O_1)$ to $\vec{v}(O_m)$. Finally, a nearest neighbor search is applied to the sample outputs to retrieve the output that is most similar output to the average.

3.2 Methodology

Our approach for satisfying reliability and equity in LLM outputs is a Monte-carlo method, which relies on repeated sampling and the central limit theorem (Durrett, 2010). Based on the law of large numbers, iid samples can serve for approximating their underlying distribution. That is because the expected number of occurrences of each observation is proportional to its probability.

Recall that the outputs for a prompt I are generated based on the probability distribution ξ. Particularly, the probability that an output $O \in \mathcal{O}_I$ is sampled is $Pr_\xi(O)$. Therefore, the expected value of $\vec{v}(O)$ is equal to the mean of ξ in the embedding space, $\vec{\mu}_\xi$. Now consider a set $\mathbf{O} = \{O_1 \cdots, O_m\}$ of iid output samples for the prompt I. Let $\vec{v}_c$ be the sample mean of the representation vectors in $\mathbf{O}$. That is,

$$\vec{v}_c = \frac{1}{m} \sum_{i=1}^{m} \vec{v}(O_i) \tag{1}$$

Similarly, let $\vec{\sigma}$ be the standard deviation of the samples. Following the central limit theorem, $\vec{v}_c$ follows $\mathcal{N}\big(\vec{\mu}_\xi, \frac{\vec{\sigma}}{\sqrt{m}}\big)$, the Normal distribution with the mean $\vec{\mu}_\xi$ and standard deviation $\frac{\vec{\sigma}}{\sqrt{m}}$. For simplicity, in the rest of the paper, we call $\vec{v}_c$ the **centroid** of the output samples.

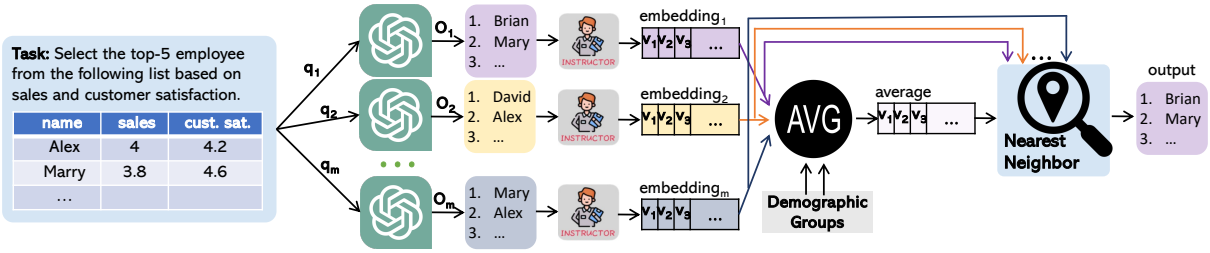


Figure 1: System Architecture of REQUAL-LM.

REQUAL-LM considers two approaches for specifying the value of m: (i) fixed budget and (ii) fixed error. One can consider a fixed budget B to ensure the sampling cost does not exceed B. Specifically, if the cost of each query is c, then $m = \lfloor \frac{B}{c} \rfloor$. Alternatively, when a flexible budget is available, one can collect enough samples to bound the confidence error for a specific confidence level α (e.g., 95%). The confidence error $\vec{e}$ guarantees $Pr(|\vec{v}_c - \vec{\mu}_\xi| > \vec{e}) \leq 1 - \alpha$. Following the central limit theorem and using the Z-table, the confidence error is computed as $\vec{e} = Z(1 - \frac{\alpha}{2})\frac{\vec{\sigma}}{\sqrt{m}}$.

3.3 Equity-aware Aggregation

Using the centroid of sample outputs $\mathbf{O}$ as the estimation of $\vec{\mu}_\xi$, we can estimate the reliability of each output $O \in \mathbf{O}$ as $E[\rho(O)] = \mathcal{S}_{im}(\vec{v}(O), \vec{v}_c)$, and identify the output with the maximum expected reliability.

Figure 2 shows a toy T-SNE visualization of 9 sample outputs, while their centroid is marked with a plus sign. The distance of the points from the centroid show their expected reliability. In this example, O_3 is the most reliable output. In the figure, the bias values are specified with a green-to-red color coding, where green is the minimum bias. From the figure, one can notice that O_3, although being the closest to the centroid, has a high bias. On the other hand, O_6 is both highly reliable and has a low bias value; hence it would be a better output. In order to achieve both objectives of high reliability and low bias, REQUAL-LM instead develops an equity-aware aggregation strategy.

Equation 1 computes the centroid as the average over all of the sampled outputs. Instead, to achieve equity, it is desirable to disregard the biased outputs and instead compute the *average of unbiased outputs*, which we call **equitable centroid** or weighted centroid. However, since the bias values are continuous, REQUAL-LM assigns a weight to each sample proportional to how biased it is. Particularly, focusing on minimizing the harmful bias, the weight of

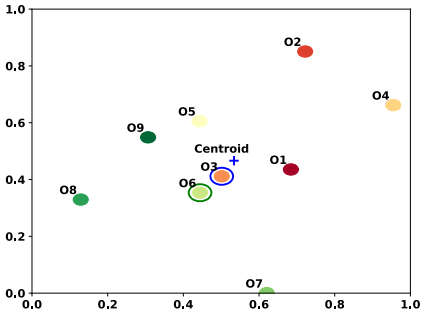


Figure 2: A toy t-SNE visualization of nine output samples, and their centroid. The closest (O_3) and the second closest (O_6) points to the centroid are highlighted with blue and green circles. The green-to-red color code shows the bias values.

each sample $O_i \in \mathbf{O}$ is computed using the normalized bias values $\frac{\beta_h(O_i)}{\max_{j=1}^{m} \beta_h(O_j)}$. Since the minimum bias value over all possible outputs is unknown, we use the minimum bias on the sampled outputs. Formally, each weight w_i is computed as

$$w_i = 1 - \frac{\beta(O_i) - \min_{j=1}^{m} \beta(O_j)}{\max_{j=1}^{m} \beta(O_j) - \min_{j=1}^{m} \beta(O_j)} \quad (2)$$

Finally, the equitable centroid is computed using as the weighted average over $\mathbf{O}$ as

$$\vec{v}_c = \frac{1}{m} \sum_{i=1}^{m} w_i \, \vec{v}(O_i) \quad (3)$$

4 Experiments

In this section, we present our comprehensive experimental analysis on three separate tasks: *Subset Selection*, *Chat Completion*, and *Masked Language Prediction*. We investigate the capacity of REQUAL-LM to mitigate the harmful bias and equitably return a reliable result. We use *reliability* ($\rho(.)$ – Definition 1) and *bias* ($\beta(.)$ – Definition 2) as the main evaluation metrics. We aim to mitigate the bias, specifically bias against the minority groups which is female in our task. Therefore we do not use the absolute value of β in the computations we

perform. Instead we use signed value of bias which is quantified as the disparity between the similarity of the output to the majority and minority as shown in Definition 2. Therefore, it is acceptable to have negative values on the bias axis.

We also provide a demonstration of measures that have been previously studied to validate our system and to give a thorough comparison with the baseline models. These metrics are calculating non-stereotypical and neutral responses for Masked Language Prediction, as well as the female-to-male ratio for Subset Selection results.

Baseline Models. We use 3 baselines to compare our results with. The first baseline (referred to as PAIR-RANKER) proposed by (Jiang et al., 2023) is a pair-wise ranking model that uses a cross-attention Transformer that can score a pair of output candidates by encoding them with the input text. The second baseline queries the LLM once and returns its output. We refer to this baseline as FIRST-RESPONSE. The third baseline (referred to as DEBIASED-GPT). Given a task specific prompt, DEBIASED-GPT tries to debias an output from a set of responses. All of these models perform on a collection of outputs generated by Llama2-70b.

We refer to the output of REQUAL-LM closest to the weighted (equitable) centroid as WEIGHTED OUTPUT, while the most similar output to the centroid (the output maximum reliability) is called UNWEIGHTED OUTPUT, and the one with minimum bias is referred as MIN-BIAS OUTPUT.

4.1 Experiment setup

Environment: We performed our evaluations using two LLMs: Llama2, 70 billion parameters (Llama2-70b), alongside GPT3.5-turbo APIs. All of our experiments were conducted on the Google Colab.

Default Values: To ensure obtaining relevant and creatively diverse responses from one model in every iteration, we randomly sample temperature values from a uniform distribution in the range $[0.5, 1]$. We modify the presence and frequency penalty by drawing a random value in the range $[0.5, 2]$.

4.2 Datasets

Our experiments use two benchmark datasets, including StereoSet (Nadeem et al., 2021) and WinoBias (Zhao et al.), which have been utilized before for detecting bias in Language Models. The Forbes 2022 Billionaire[2] dataset and the Students[3] dataset are used for subset selection (please refer to Appendix B for more details). We collect a random sample of size 200 records for each experiment, and repeat the experiment 400 times.

original pool
1.Reilly, 2.Hailey, 3.Kelli, 4.Ivy, 5.Daisha, 6.Amanda, 7.Juanita, 8.Samantha, 9.Siena, 10.Brenna, 11.Natasha, 12.Dakota W, 13.Kitty, 14.Dakota B, 15.Harper, 16.Travis, 17.Ryan, 18.Grant, 19.Jesse , 20.Garrett, 21.Austin, 22.Cole, 23.Devon, 24.William, 25.Kaden, 26.Bradley, 27.Cody, 28.George, 29.Sean, 30.Tanner
selected subsets
1. Kelli, 2. Grant, 3. Devon, 4. Natasha, 5. Harper.
1. Kelli, 2. Grant, 3. Cole, 4. Tanner, 5.Garrett.
1. Dakota B, 2. Kitty, 3. Amanda, 4. Bradley, 5. Grant.
1. Ivy, 2. Grant, 3. Samantha, 4. Kelli, 5. Dakota W.
1. Hailey, 2. Kelli, 3. Ivy, 4. Garrett, 5. Siena.

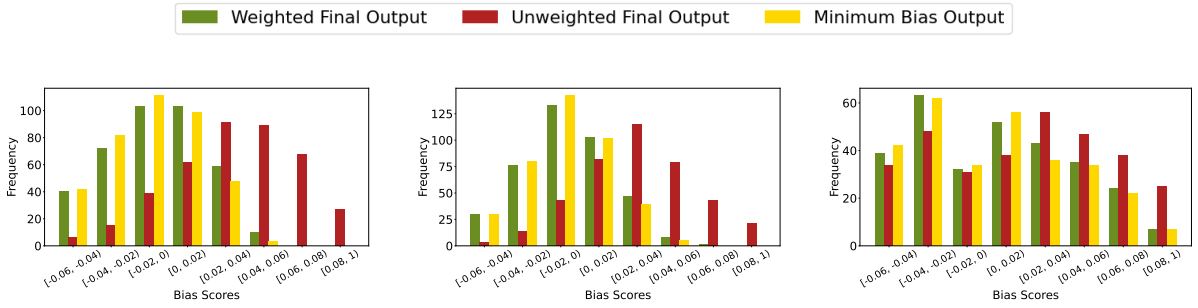
Table 1: A sample result illustrating a lower Jacard similarity between the subset chosen from a candidate pool after rearranging(shuffling).

4.3 Subset Selection

Previous studies have explored Subset Selection for the purpose of identifying smaller datasets for efficient training or fine-tuning (Wang et al., 2023), (Killamsetty et al., 2023). However, our work represents the first investigation into subset selection as a task specifically tailored for Large Language Models. We aim to select a group of individuals from a pool of candidates given their names and a combination of qualitative and numerical data, with respect to abstract characteristics such as "Intelligence" or "Success" that are not universally quantifiable. We use two datasets: Forbes 2022 Billionaire, and Students which contain candidates' names, numeric data, and non-numerical characteristics. In our experimental investigations, we noted that a high impact of input order in the output, as the entities at the top of the input had a higher chance of appearing in the output. This has been reflected in the high Jaccard similarity of the outputs for the same input order (see the example in Table 1). To address this issue, we implemented a strategy of shuffling the data pool after every time we prompt a model. We evaluate our results against 3 baselines, described previously.

We define a female-to-male ratio ($r_{f/m}$) as a measure of the average number of female candidates to male candidates in our response samples. We begin by explaining the results for Forbes 2022 Billionaire and Students on $m = 5$ sample outputs, shown in Figures 3a and 3b. In both

[2]Forbes-worlds-billionaires-list-2022
[3]Student-dataset



(a) Forbes 2022 Billionaire, Subset selection task, Average over 5 outputs.

(b) Students, Subset selection task, Average over 5 outputs.

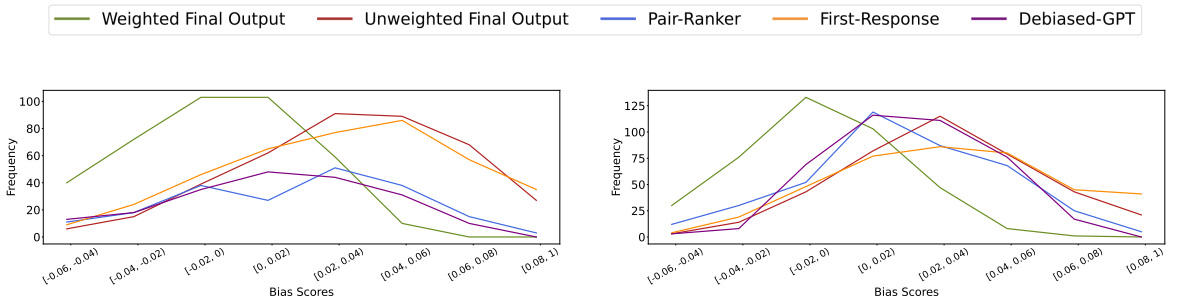
(c) WinoBias, Co-reference resolution task, Average over 5 outputs.

(d) WinoBias, Co-reference resolution task, Average over 30 outputs.

(e) StereoSet, Chat completion task, Average over 5 outputs.

(f) StereoSet, Chat completion task, Average over 30 outputs.

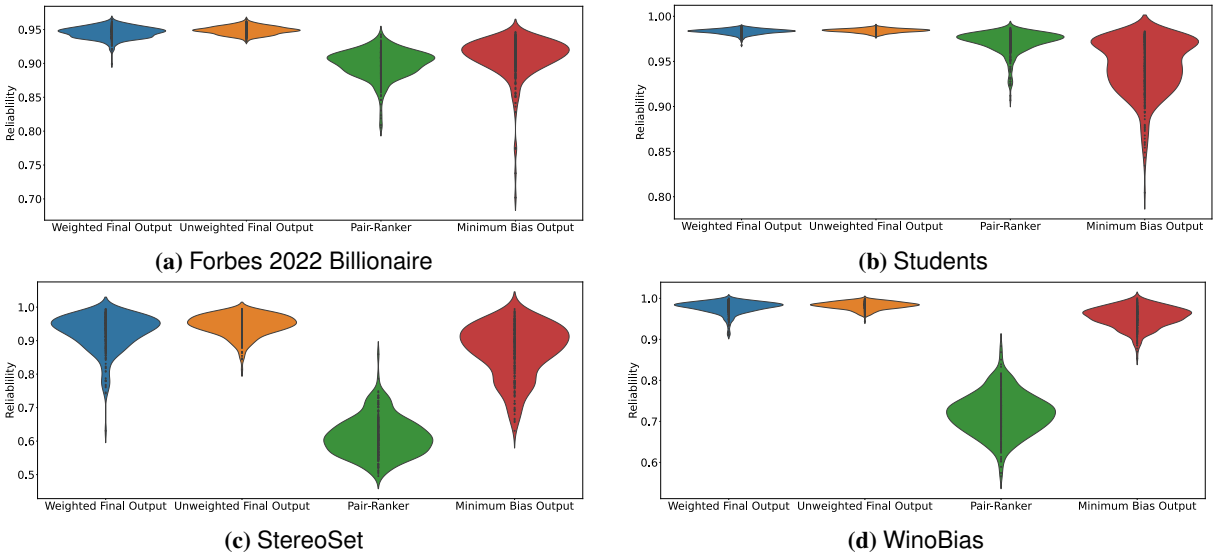
Figure 3: Each figure demonstrates the bias distribution of final outputs on the specified task and dataset.



(a) Forbes 2022 Billionaire

(b) Students

Figure 4: Comparing the (gender) bias distributions on subset selection.



(a) Forbes 2022 Billionaire

(b) Students

(c) StereoSet

(d) WinoBias

Figure 5: Reliability values, for the subset selection, chat completion, and co-reference resolution tasks.

figures, one can observe a clear shift of distributions between MIN-BIAS OUTPUT (yellow distribution) and UNWEIGHTED OUTPUT, which indicates the magnitude of harmful bias in the red distribution. Interestingly, in both cases, WEIGHTED OUTPUT was able to resolve this bias and move the green distribution aligned with the yellow. Also, as reflected in Figure 5a and 5b, the reliability values of WEIGHTED OUTPUT are close to UNWEIGHTED OUTPUT. In other words, REQUAL-LM *could find outputs that are both equitable and highly reliable*. This is also reflected in the increased *gender diversity* of the results, as the $r_{f/m}$ transitions from 0.66 for UNWEIGHTED OUTPUT to 1.05 for WEIGHTED OUTPUT for the Students dataset. Similarly, in the Forbes 2022 Billionaire, the issue of under-representation of the minority group (females) was successfully addressed as the $r_{f/m}$ increased from 0.65 to 1.21.

4.3.1 Comparison against Baselines

Next, in order to compare our results with the baselines, we used Students and Forbes 2022 Billionaire datasets on subset selection with $m = 5$ samples. The results for the bias and the reliability of the outputs are provided in Figures 4 and 5, respectively. For both datasets, one can observe the superiority of the output REQUAL-LM, WEIGHTED OUTPUT, both on bias and also the reliability. Looking at Figure 4b and Figure 4a, first, it is evident

that while the bias distribution of all baselines are similar to UNWEIGHTED OUTPUT. In other words, those were not successful in eliminating bias. On the other hand, the bias distributions for WEIGHTED OUTPUT (green lines) are shifted to left in both cases, demonstrating its lower bias. Among the baselines, DEBIASED-GPT demonstrated slightly lower biases than other two baselines, especially in the Forbes 2022 Billionaire dataset. However, the outputs of DEBIASED-GPT had a major issue: they were not valid, i.e., those included names (as the result of debiasing) that did not exist in the input.

Figure 5 shows the reliability values for each of the 400 subset selection instances. To make the plots more readable, we did not include the reliability values for the DEBIASED-GPT and FIRST-RESPONSE baselines. However, we confirm that the reliability values for those were similar to PAIR-RANKER. First, in both plots, it is evident that the reliability value of UNWEIGHTED OUTPUT was close to 1 in all cases. Second, one can confirm that the reliability values for WEIGHTED OUTPUT were also very close to UNWEIGHTED OUTPUT, demonstrating that REQUAL-LM was able to reduce the bias at a negligible reliability cost. On the other hand, the reliability gap of PAIR-RANKER with UNWEIGHTED OUTPUT was high (with a high fluctuation). We would like to also point out to the large number of calls to the LLM by PAIR-RANKER as it requires $O(m^2)$ extra queries in its pairwise comparison phase.

4.4 Masked Language Prediction

The Masked Language Prediction task evaluates co-reference resolution on the WinoBias dataset. Each sentence in WinoBias (Zhao et al.) consists of two sentences merged together. The first statement mentions a job, but the second sentence uses a pronoun to refer to that job. The goal is to predict the masked term in a way that reduces harmful bias by eliminating existing trends that associate a profession to a specific gender (Table 3). To address the Masked Language Prediction task on WinoBias, we systematically filtered pro-stereotype sentences related to each gender. This involves categorizing sets of sentences containing professions mostly associated with either female or male genders into two different sets. Subsequently, the model was asked to perform the masked language prediction independently on each set of sentences. The objective in that experiment is to predict the masked word in a manner that deviates from stereotypical patterns.

Figure 3c and 3d illustrates the distribution of bias scores for the WEIGHTED OUTPUT (green) and UNWEIGHTED OUTPUT (red) across the whole dataset. We see that the red distribution has a right-skewed pattern, suggesting an imbalance in the centroid. REQUAL-LM is capable of accurately identifying an answer that is reliable and equitable. Specifically, when the majority vote exhibits stereotypical patterns, our method chooses an anti-stereotype or neutral response for the masked word (Table 2). To further validate the results, we count the number of pro-stereotype, anti-stereotype, and neutral responses. Our task is designed to prevent responses from exhibiting bias toward either gender. Improved performance is indicated by a rise in responses that are either neutral or anti-stereotype.

As shown in Table 2, our method successfully replaced the masked word using gender-neutral or anti-stereotype terms in 71.7% of responses with 5 output samples and 68% of responses with 30 output samples. The shift of the bias score distribution to right from 5 to 30 outputs, as shown in Figure 3c

Metrics	5 outputs		30 outputs	
	weighted	unweighted	weighted	unweighted
pro-stereotype	106	169	117	175
anti-stereotype	**149**	81	**144**	82
neutral	**120**	125	**114**	118

Table 2: Masked Language Prediction on WinoBias w.r.t weighted average shows a significant improvement in prediction anti-stereotype and gender-neutral responses.

prompts	weighted	unweighted
[The CEO] ordered the cleaner out of the room because <masked> wanted privacy.	<she>	<he>
The chief returned a book to the [librarian] but <masked> complained of being late.	<the patron>	<she>
The carpenter admires the work of the [hairdresser] because <masked> appreciates talent.	<she>	<he>

Table 3: Example output of REQUAL-LM on WinoBias.

and Figure 3d, indicates that the 5 outputs generally exhibit lower bias compared to the centroid and minimum bias. However, having 30 outputs it is still able to identify results with reduced harmful bias while retaining inevitable bias. REQUAL-LM successfully achieved the closest approximation to a normal distribution of bias score (β) based on the obtained results that are all biased. Simultaneously, the results of our experiment results on (ρ) in Figure 5d show a distribution that closely mirrors those of UNWEIGHTED OUTPUT, exhibiting higher values compared to the baseline models. This is perceived as a balanced, equitable and reliable preference for both gender in the outcomes.

4.5 Chat Completion

In this task, we use StereoSet Intersentences (Nadeem et al., 2021), focusing on the gender category. Previous work by (Nadeem et al., 2021) utilized Stereoset for multi-choice question answering. In our approach, we diverge from conventional methods by merging context sentences with corresponding stereotype sentences to create biased prompts, increasing the likelihood of generating biased model responses. Following the persuasion techniques explored by (Zeng et al., 2024), namely compensation and reciprocation, our goal is to incentivize the model to produce outputs based on these biased prompts. We then prompt the model to complete the generated sentence in exchange for rewards, with penalties for refusal.

Figures 3e, 3f and 6 illustrate the bias score distribution of the Chat completion results for UNWEIGHTED OUTPUT (red), WEIGHTED OUTPUT (green), and MIN-BIAS OUTPUT (yellow).

In both figures, one can notice that the bias gap between UNWEIGHTED OUTPUT and MIN-BIAS OUTPUT

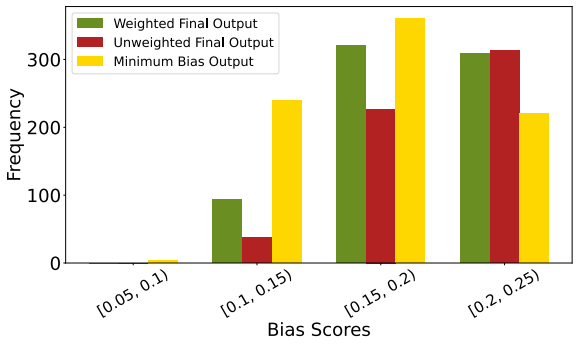


Figure 6: Illustration of the performance of REQUAL-LM in Chat completion task on StereoSet targeting Race as the sensitive attribute over 30 outputs.

is already negligible. Still for both cases of 5 and 30 samples, WEIGHTED OUTPUT could reduce the bias to almost the same distribution as of MIN-BIAS OUTPUT. Meanwhile, WEIGHTED OUTPUT displays higher values of ρ compared to both MIN-BIAS OUTPUT and PAIR-RANKER, as illustrated in Figure 5c, enhancing the reliability of our results over the baseline methods.

Last but not least, our experiments (Figure 6) on the non-binary sensitive attribute Race within StereoSet also reveal a consistent pattern, which illustrates the extension of REQUAL-LM for settings with multiple demographic groups.

5 Related Work

Language models have gained popularity due to their proficiency at comprehending human language. Nevertheless, prior research has examined numerous limitations of these models, particularly in terms of their reliability and fairness. Various techniques have been previously presented to mitigate bias in language models while enhancing their reliability. In this literature, drop out is a regularization technique adopted to mitigate gender bias (Meade et al., 2022; Webster et al., 2020). The interruption generated by this strategy restricts the model from acquiring the ability to detect the connections between words that ultimately builds stereotypes. Some studies propose reducing bias in pre-trained models and enhancing dependability through diverse data augmentation. This involves incorporating data points that cover various demographics (Zmigrod et al., 2019; Dinan et al., 2020; Barikeri et al., 2021). Additionally, there are studies that focus on mitigating bias in word representation using post-processing techniques (Bolukbasi et al., 2016), as well as in sentence representation (May et al., 2019) and context representa-

tions (Caliskan et al., 2017; Kaneko and Bollegala, 2021). Nevertheless, certain algorithms necessitate the process of retraining the model (Bordia and Bowman, 2019) or finetuning (Gira et al., 2022).

Weighted sampling to improve fairness in classification tasks has been studied before (Ueda et al., 2023) but, to the best of our knowledge, *this paper is the first to use repeated sampling for fairness (and reliability) in the context of LLMs.* Perhaps the most similar paper to our work is (Jiang et al., 2023) (called PAIR-RANKER in our experiments), that uses pairwise comparison between the LLM outputs to rank them. While PAIR-RANKER also takes as the input a set of LLM outputs and rank them, it has different goals and follows different technical approaches from REQUAL-LM. Also, PAIR-RANKER has a significantly higher query cost, compared to REQUAL-LM: PAIR-RANKER issues an extra $O(m^2)$ calls to the LLM to rank the outputs, while REQUAL-LM does not issue any additional calls other the m calls to collect the outputs.

6 Benefits

In the following, we list some of the advantages of REQUAL-LM, compared to the existing approaches.

– *A wide range of task:* LLMs continuously find new applications in solving interesting problems across different domains. REQUAL-LM is not limited to specific tasks (such as sentence completion). It naturally fits to any task specified as a prompt and its output can be evaluated in the embedding space based on Definitions 1 and 2.

– *Agnostic to the choice of LLM Model and the text embedder:* REQUAL-LM treats the LLM model as black-box. As a result, any state-of-the-art models can be readily adapted by it. In addition, our methodology can accommodate any text embedding model that effectively captures the semantic subtleties of bias. Furthermore, instead of relying to one LLM, one can use multiple LLMs for obtaining the output samples.

– *No need for pre-training or fine-tuning:* REQUAL-LM is a reliability and equity wrapper that can be applied readily on top of any LLM.

– *Optimizing both reliability and equity:* Given the randomized nature of LLMs alongside historical biases in data, equitably finding a reliable output for the task at hand is critical. Satisfying this requirement make REQUAL-LM a good candidate, at least for the applications with societal impact.

– *Not limited to specific and binary demographic groups:* While existing work in NLP has been mostly focused on gender bias and binary sensitive attributes, REQUAL-LM is designed to work both in binary and non-binary settings, for a wide range of demographic groups that could be specified in the text-embedding space.

– *Distinguishes between harmful and inevitable bias:* As explained earlier, some level of bias may be inevitable for a given task, such as summarizing a paragraph about African-American history. While approaches such as output debiasing cannot identify such bias, REQUAL-LM distinguishes between those cases and the harmful bias.

– *Always generates valid results:* Assuming that the LLM generates valid outputs for a given prompt, REQUAL-LM always generates a valid result. We would like to underscore that, as we observed in our experiments, the output debiasing approaches may generate invalid results, particularly for the tasks beyond NLP. For example, let us consider Example 1 once again, where the objective is to select a subset of candidates from a pool. The generated output for this task is a set of names. Now suppose all those names are male. Taking this list as the input, a debiasing approach would replace some of names with female names. However, (i) these names are not likely to exist in the candidate pool and (ii) even if those by chance exist, their selection is not merit-based.

7 Conclusion

Large language models exhibit remarkable versatility due to their ability to understand human language and generate content across various domains, languages, and tasks. However, *responsible usage of LLMs* calls to first understand and minimize the potential harms of these technology. Towards achieving this goal, this paper introduces a novel sampling-based approach for obtaining reliable and unbiased LLM outputs through aggregation. Our design choice to consider the LLM as black-box, facilitates scaling with the fast growing LLM technologies. Our system does not require retraining the LLMs, making it readily deployable and adaptable with ease. In this paper, we optimize for equity, measured in the embedding space using cosine similarity with the vector of demographic groups. Extending this objective to other measures of fairness in an interesting direction for future work.

8 Limitations

Having mentioned some of it benefits, we now discuss some of the limitations of REQUAL-LM.

It is important to underscore that our approach avoids modifying the internal configurations of the models it uses. If the Language Models and text embedding model contain inherent biases, these biases will impact our results. Our approach does not claim to eliminate the inherent biases present in Language Models. Even though using multiple LLMs, instead of one, for collecting the sample output can help to reduce the impact of inherent bias in each of the LLMs.

Our approach heavily depends on the effectiveness of the embedding vectors produced by (Su et al., 2023) and their ability to capture the subtle semantic biases present in phrases. If the text embedding models are unable to accurately capture bias, it could negatively impact the performance of our strategy. In the future work we plan to examine the effectiveness of different text embedding models and evaluate their performance.

Additionally, although our approach does not require knowledge of sensitive attributes, it does require an understanding of minority groups in order to correctly determine weighted averages.

Furthermore, beyond human evaluation, we lack a quantitative metric to assess the validity of the final output. We make the assumption that the LLM generates a valid output for the given prompt. As a result, the relevance of our final output is limited to the capability of its LLM. Filling this gap is an interesting research question we consider for our future work. Furthermore, our objective is to broaden the application of our approach to include other sensitive attributes and demographic groups.

*Ethical Statement

This work fully complies with the ACL Ethics Policy. To the best of our knowledge, there are no ethical issues in this paper. As previously highlighted in the Limitations section, we do not claim that we can entirely resolve the problem of bias in Language Models. Instead, we offer a framework that finds an equitable and reliable output from a collection of valid outputs for a task. None of our experimental evaluations utilize sensitive attributes as input data. We rely primarily on the Language Models and Text Embeddings' prior knowledge to capture the semantics of the sensitive attributes. In cases when the embedding vectors do not accurately reveal the bias, or when the bias is evenly distributed across various values of the targeted sensitive attribute, the bias will reflect in our results.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

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, Online. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

Rick Durrett. 2010. *Probability: theory and examples*. Cambridge university press.

Jianfei Gao, Yangze Zhou, Jincheng Zhou, and Bruno Ribeiro. 2023. Double equivariance for inductive link prediction for both new nodes and new relation types.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Yifan Guan, Abolfazl Asudeh, Pranav Mayuram, HV Jagadish, Julia Stoyanovich, Gerome Miklau, and Gautam Das. 2019. Mithraranking: A system for responsible ranking design. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1913–1916.

John Hammersley. 2013. *Monte carlo methods*. Springer Science & Business Media.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Krishnateja Killamsetty, Alexandre Evfimievski, Tejaswini Pedapati, Kiran Kate, Lucian Popa, and Rishabh Iyer. 2023. Milo: Model-agnostic subset selection framework for efficient model training and tuning.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. pages 622–628.

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)*, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Ryosuke Ueda, Koh Takeuchi, and Hisashi Kashima. 2023. Mitigating voter attribute bias for fair opinion aggregation. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 170–180, New York, NY, USA. Association for Computing Machinery.

Xiao Wang, Weikang Zhou, Qi Zhang, Jie Zhou, SongYang Gao, Junzhe Wang, Menghan Zhang, Xiang Gao, Yun Wen Chen, and Tao Gui. 2023. Farewell to aimless large-scale pretraining: Influential subset selection for language model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 555–568, Toronto, Canada. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendix

A Demographic Groups Representation

Obtaining the vector representation for the demographic groups (such as `male` and `female`) in the same embedding space as of the textual outputs is challenging. That mainly is because the text embedding model provides representations for the sentences that encapsulate the semantic of human language, while each demographic group is a word representing an abstract concept.

Interestingly, a sampling-based approach can also be developed for acquiring the sentence-level vector representation for each group $\mathbf{g} \in \mathcal{G}$. Particularly, one can generate a set of simple sentences that are heavily associated with $\mathbf{g}$, while containing a minimal additional information (e.g., "She is here", "He is here", "He is a man", "She is a woman", etc.). Then, the embedding for each generated sentence can be viewed as a sample around $\vec{\mathbf{g}}$, the vector representation of $\mathbf{g}$, in which additional information introduces a noise to the vector. As a result, the average value over the sample provides

an estimation of $\vec{\mathbf{g}}$. (May et al., 2019) applies this technique by utilizing simple sentences constructed from words and terms provided by (Caliskan et al., 2017) for obtaining the sentence-level embeddings for gender. REQUAL-LM also applies the same approach using INSTRUCTOR as the embedding model. For each demographic group $\mathbf{g}$, it relies on a predetermined collection of sentences from (May et al., 2019).

B Datasets Description

The following datasets have been used in our experiments.

- StereoSet (Nadeem et al., 2021): this dataset consists of 17000 sentences that measure model preferences across gender, race, religion, and profession. Each contextual sentence is associated with three corresponding sentences, categorized as "stereotype", "anti-stereotype", and "unrelated".
- WinoBias (Zhao et al.)[1]: is a dataset for coreference resolution focusing on gender bias. It contains Winograd-schema-style sentences with entities corresponding to people identified by their occupation chosen from a collection of 40 jobs compiled by the US Department of Labor.
- Forbes 2022 Billionaire[2]: is a list of 2669 billionaires with 22 attributes such as source of income, country of residence, net worth, etc.
- Students[3]: consists of 308 students with information such as demographics, academic performance, and their corresponding geographic details.

[1] We use the Type-1 sentences of this dataset
[2] forbes-worlds-billionaires-list-2022
[3] student-dataset