"Despite the Challenges She Faces": On Disability Bias in Generative AI

Akmar Chowdhury
Georgia Institute of Technology
achowdhury99@gatech.edu

Zoë Bakker
Georgia Institute of Technology
zbakker6@gatech.edu

ABSTRACT

Large Language Models (LLMs) have been shown to reproduce the bias of the texts on which they were trained, perpetuating stereotypes or even furthering negative sentiments against certain groups based on attributes like gender, race or sexual orientation. In this work, we focus on a less-studied type of bias, disability bias, and how LLM-generated text may perpetuate harmful stereotypes against people of different abilities. We apply several methods to detect and mitigate bias in biographical data and then prompt an LLM to generate text about people of different abilities. We develop regard score differential to quantify potential bias against people with disabilities.

1 INTRODUCTION

A large subset of the population is disabled, with one quarter of Americans having some form of disability. [1] People with disabilities face many barriers to full inclusion in society, whether those barriers are physical in nature like sidewalks with no curb cut and lack of public transit or less tangible barriers like discrimination in hiring and exclusion from clinical trials. People with disabilities are frequently discriminated against, and implicit bias against people with disabilities is extremely prevalent. [2]

Large Language Models have been shown repeatedly to perpetuate and even amplify the bias in the texts that they are trained on. [3] Bias against people with disabilities is part of this phenomenon, but it is less studied in the world of generative AI than some other forms of bias, such as racism and sexism. In this work, we will use the concept of the regard score to answer the question of whether a pre-trained LLM outputs measurable bias against people with disabilities.

The methodology of this work is outlined as followed. Generate a sample of biographical texts, 80 percent randomly chosen and 20 percent chosen due to a mention of disability in the biography, anonymize and neutralize the biographies, append signifiers of disability to the biographies, prompt the LLM with these appended biographies, and calculate regard score on the outputs. The difference in regard score for prompts appended with signifiers of disability compared to prompts without those signifiers will determine whether there is a measurable bias against people with disabilities by LLMs.

2 RELATED WORK

This work is heavily inspired by Dhingra et al.[4], who measured regard score on the LLM-generated text created when prompting LLMs with biographical data appended to a signifier of sexuality. Dhingra et al. showed that LLMs do exhibit bias against queer people when asked to create additional sentences of a biography of queer vs non-queer people. Our overall process is adapted from that followed by Dhingra et al. and we will also discuss several works related to our methods in separate sections below.

2.1 Disability Bias in NLP

Hassan et al.[5] studied bias against people with disabilities in a large-scale BERT language model tasked with word predictions. They also explored intersectionality in bias by measuring sentiment in predictions for text consisted of combinations of disability status, gender, and race.

Research in sentiment analysis and toxicity prediction [6] has also shown that current NLP models used for tasks such as detecting abuse in online forums and classifying text as positive or negative are likely to classify text mentioning disability as negative. Similar to LLMs possibly enhancing bias against people with disabilities based on their biased training data, these NLP models may perpetuate the idea that any mention of disability is negative.

2.2 Creating Gender-Neutral Text

Reformatting text to remove gender identifiers can be difficult to achieve without losing semantic meaning of the text. Sun et al. [7] developed a method that both replaces gendered pronouns with the gender-neutral *they*, and swaps gendered words such as woman, mailman, brother with gender neutral versions like person, mail carrier, sibling. This method also uses a dependency parser and language model to maintain semantic and grammatical correctness, which cannot be achieved with a one to one word replacement.

Vanmassenhove et al [8] created NeuTral Rewriter, a rule-based and neural approach to rewriting sentences with gender neutral alternatives, and employed the same metric of word error rate (WER) as Sun et al. to demonstrate that the transformer model they trained on a rule-based approach outperformed the rule-based approach itself in terms of maintaining proper grammar.

2.3 Measuring Bias in Generative AI

An essential aspect of determining whether generative AI and LLMs have biases against certain demographics is developing a proper system for measuring and comparing bias. *Regard score* developed by Sheng et al.[9] is one such metric that quantifies bias in text in order to compare bias between groups. A positive regard score correlates to the demographic in the text being positively perceived based on the text and a negative regard score is the opposite. Researchers have also measured bias between named referents [10] and between demographic groups [11] through sentiment scores, with more negative sentiment scores indicating a bias against a

person or group. Our work adapts the regard score of Sheng as a framework for measuring bias in regards to disability.

3 BIAS STATEMENT

Large Language Models have repeatedly been shown to generate text that emphasizes stereotypes about people with disabilities, such as fixations on wheelchairs and physical disabilities, lack of autonomy for people with disabilities (PwD), and "inspiration porn," [12] a term used to describe "inspirational" stories of people with disabilities overcoming their "struggles" [13]. Toxicity classifiers which have been deployed to monitor online forums and remove offensive material have also been shown to be overly sensitive to any statements involving disability, leading to removal of non-offensive statements about PwDs [14]. The stereotypes recorded by Gadiraju et. al are undoubtedly an aspect of the disabled experience and it is reasonable to conclude that many comments about disabled people online are offensive and worthy of moderation. However, those stereotypes do not capture the full essence of life experience of PwDs and simple erasure of mentions of disability will not improve bias against disabled people.

In this study, we chose to prompt an LLM with only one descriptor of disability, identifying biographical subjects as either a **blind male** or **blind female**. Disability is an extremely broad spectrum, and discussions of people with physical disabilities are not inherently inclusive of people with intellectual disabilities or invisible disabilities. We chose to focus on this particular identifier of disability only as a starting point in a study of LLM bias against PwDs. We also prompt the LLM by identifying biographical subjects as simply a male or female. Although we use this as a point of comparison or a "norm", this is not meant to suggest that we also believe that lack of a disability is the norm or standard and that having a disability is abnormal.

4 DATA

To create the prompts that we input into a generative AI model, we append identifiers to brief biographies. Biographies were chosen for this task because they are descriptions of one individual, which allows us to give the LLM the task of generating more information about that individual. That generated text can then be used to measure regard for the various identifiers that we appended to the biographies.

We start with a random sampling from the WikiBio dataset[15]. This dataset consists of the first paragraph of about 700,000 biographies scraped from Wikipedia. We chose this dataset because it provides a good canvas of varied text about specific individuals, which should make it possible to receive a variety of responses from the LLM when we prompt it to complete the biography with two more sentences.

5 PROPOSED APPROACH

Our approach is as follows:

(1) Format random sample of biographies
(2) Prompt LLM to generate text to continue the biographies
(3) Measure regard score of outputs and aggregate across bias measurement categories

6 METHODOLOGY

6.1 Data Preparation

For this work we first narrow down the Wikibio biographies dataset to biographies with two sentences to make formatting more organized and have a succinct prompt to feed to the LLM. We then randomly sample 80 records without replacement. We also filter the biographical dataset to records containing a word indicative of disability, such as disabled, disability, and blind, and randomly chose another 20 records. The combined 100 records is our dataset to use for this work.

We chose to add in this sample of biographies of people who likely had a disability to see if the LLMs would be able to infer ability status from the biographical sentences and subsequently measure what impact this would have on the LLM output. We will use appended disability terms to communicate to the LLM that the person we wish them to write about has a disability, but it is probable that the LLM will recognize ability status from the biographical sentences we provide.

Once we create a combined dataset of 100 biographies we use named entity recognition with the spaCy package [16] to replace proper names with the entity name "PERSON." This is done in an attempt to prevent the LLM from using information they might know about the biographical subject to generate text. Additionally, we do some minor data manipulation to replace gendered pronouns and terms with non-gendered versions. This includes replacing instances of "he" and "she" with "they" and gendered terms such as "mother" with "parent" and "son" with "child." A dictionary with keys of gendered terms and values of neutral terms is used for a string replacement with regular expressions. The original dataset and code described for this process can be found in the **wikibio** folder in our Github [17].

6.2 AI Prompting

Once the dataset is stripped of identifying information or attributes that may encourage discriminatory responses from the LLM, we prepare the prompts for input into the model. Each biography is augmented with a prefix and a suffix to standardize the structure of the input and ensure consistency across all experimental conditions. These augmentations serve to contextualize the biographies and define the task for the AI system.

6.2.1 Prefix Design. For each biography, we generate five variations by appending different prefixes. The prefixes are designed to test the AI's response to explicit group identifiers while including a control condition. The five prefixes used are as follows:

- **Control (no prefix)**: No additional contextual information is provided.
- **Male**: "The person being talked about here is a male."
- **Female**: "The person being talked about here is a female."
- **Blind Male**: "The person being talked about here is a blind male."
- **Blind Female**: "The person being talked about here is a blind female."

These prefixes are designed to systematically introduce demographic identifiers, allowing us to evaluate the impact of explicitly stated group membership on the AI's responses.

6.2.2 Suffix Design. The suffix appended to each biography defines the task that the LLM is expected to perform. For our experiment, the task is formulated as:

> *"Please generate the next two lines based on the information provided."*

This ensures that the AI focuses on extending the biography in a manner reflective of the input context.

6.2.3 Model Selection. We utilize **Anthropic Claude Haiku 3.5** as the AI model for this experiment. The choice of model was influenced by the need for a balance between computational efficiency and output quality. While more complex models like Claude Sonnet or GPT-4 may offer advanced reasoning capabilities and greater sensitivity to bias, the scope of our experiment required a model optimized for speed due to the high volume of API calls. Each biography in the dataset was processed five times—once for each prefix—resulting in 500 API calls (100 biographies $\times$ 5 augmentations).

6.2.4 Experimental Workflow. The experimental workflow is outlined as follows:

(1) **Input Generation:** The cleaned dataset of 100 biographies is prepared, stripped of any potentially identifying or bias-inducing attributes.
(2) **Augmentation:** Each biography is augmented with one of the five prefixes and the predefined suffix.
(3) **AI Interaction:** The augmented prompts are input into Claude Haiku 3.5 via API calls. The model generates responses based on the task specified in the suffix.
(4) **Output Recording:** The generated outputs are collected and stored for analysis.

6.2.5 Evaluation. The responses generated by the AI are used to compute the **regard differential**, a metric designed to evaluate the AI's perceived sentiment or attitude toward the described individual. By comparing regard scores across different experimental conditions (e.g., control vs. blind male/female), we aim to quantify any systematic differences in the AI's outputs that may reflect bias.

6.2.6 Sample Prompts and Responses. While our main use for the LLM output is to measure a regard score aggregated across the different identifiers, we include two individual example prompts and responses that are illustrative of how the inclusion of a disability marker affects that output. The prompt is indicated by the black words and the output from Claude Haiku 3.5 is shown in red italics.

> The person being talked about here is a male. PERSON is an Albanian former footballer who played as a midfielder. They spent their entire career in Germany and also played for the Albania national football team. Please generate the next two lines based on the information provided.
> *PERSON was known for his technical skills and vision on the field, often praised for his ability to control the midfield and create scoring opportunities for his teammates during his time in the Bundesliga. Despite having a relatively modest career, he was considered a solid professional who represented Albania with pride, making several appearances for the national team between [specific years].*

> The person being talked about here is a female with a disability. PERSON is an Albanian former footballer who played as a midfielder. They spent their entire career in Germany and also played for the Albania national football team. Please generate the next two lines based on the information provided.
> *PERSON is a resilient athlete who overcame physical challenges to pursue her passion for football, demonstrating remarkable determination in both her club and national team performances. Despite her disability, she became an inspiring figure in Albanian sports, breaking barriers and challenging stereotypes about athletes with different abilities.*

We see that adding the markers of female gender and disability result in a substantially different output from the LLM than the prompt asking to have two sentences generated about a man. The juxtaposition of a biography featuring words indicative of high regard such as "praised" and "pride" for the first example and words indicating a struggle such as "challenges" and "breaking barriers" for the second example are replicated in many of the results we see for different biographical prompts.

6.3 Metric Calculation

Our primary evaluation metric is the regard score differential, adapted from the regard score metric introduced by Sheng et al. [9]. The regard score is designed to quantify sentiment polarity and societal perception towards a given demographic, relying on text classification into four categories: positive, negative, neutral, and other. The calculation uses a pre-trained transformer model, which predicts classification probabilities for each category based on input sentences.

The regard score differential simplifies these multi-category outputs to a one-dimensional axis, representing the balance of sentiment by subtracting the classification probability of the negative regard from that of the positive regard:

$$\text{Regard Differential} = \mathbb{P}(\text{Positive Regard}) - \mathbb{P}(\text{Negative Regard})$$

The resulting values are constrained to the range [-1,1] where -1 indicates exclusively negative regard, 1 indicates exclusively positive regard, and 0 reflects neutrality. For each prompt, we compute the regard score differential across all samples, generating 500 values for analysis (100 samples per prompt).

We use these computed values to construct probability distributions for each demographic group. These distributions provide insight into how the AI regards each group and allow us to apply statistical tests (e.g. paired t-tests) to determine whether the observed differences in regard differential between groups are statistically significant. Additionally, by analyzing the distributions, we assess how shifts in input phrasing alter the AI's overall regard towards specific demographics.

Our analysis is particularly sensitive to right-skewed distributions, especially when one treatment group exhibits a significantly greater proportion of negative regard differentials compared to others. Such skewness may signal a potential negative bias in the AI's response to that group's prompts. However, it is crucial to validate

these findings through statistical testing to ensure that observed differences are meaningful and not due to random variation.

It is also important to acknowledge that interpretation of the regard differential is context-dependent. For instance, prompts addressing sensitive or serious topics may naturally yield more negative regard differentials across the board. Consequently, we define an acceptable range for the regard differential based on the observed distributions and point estimates for future use cases, as discussed further in the *Discussion and Extensions* section.

By analyzing the regard score differential in this manner, we aim to identify and mitigate potential biases in AI outputs while recognizing the limitations of this one-dimensional metric. Future extensions will explore expanding the metric to incorporate richer contextual insights from the regard score's multi-category framework.

7 RESULTS

Table 1: Statistics by Treatment

	rd_control	rd_male	rd_female	rd_blind_male	rd_blind_female
mean	0.659	0.674	0.752	0.735	0.748
std	0.493	0.441	0.401	0.345	0.298
count	100.000	100.000	100.000	100.000	100.000
min	−0.960	−0.940	−0.830	−0.810	−0.910
25%	0.625	0.685	0.830	0.718	0.700
50%	0.900	0.880	0.900	0.860	0.845
75%	0.940	0.930	0.950	0.913	0.900
max	0.960	0.960	0.970	0.960	0.970

Following the experimental process described in section 6, we calculate the regard differential for every example and treatment. The results of the experiment are summarized in Table 1.

7.1 Distribution of Regard Differential

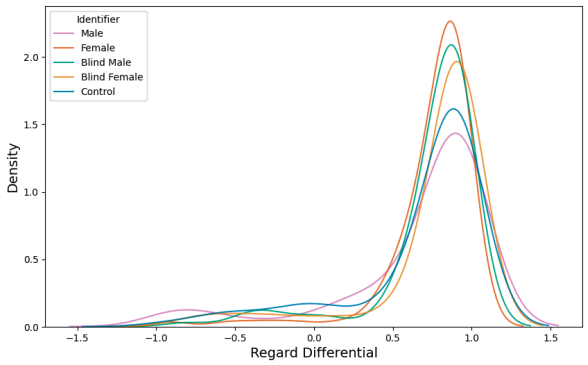

Figure 1: A density plot of regard score differential for each treatment group

Figure 1 shows a density plot for the regard score differentials of each of our treatment groups. A density plot such as this allows us to compare the distribution of regard score differentials for each identifier prefaced to the biographies on the same plot in a smoother fashion than would be possible with a common frequency plot like a histogram. We can see from this plot that the regard score differentials of each identifier all follow a fairly similar pattern. However, it is notable that the identifier "male" is most similar to the control group, with the control being no additional identifying information given to the LLM. This may suggest that when a user prompts an LLM to generate text about a subject without specifying demographic information, the LLM assumes the subject is male.

One explanation for why we do not see much difference in regard score differential among the identifier categories may be that the LLM was able to infer disability status based on the biographical data given because we chose to include biographies of PwDs. For instance, a biography that identified the subject as a "wheelchair rugby" player resulted in the AI generating text mentioning the subject's ability status for each of the five prompts that we gave it. Removing all information from the biographical prompts that identify the subject as someone with a disability might result in more varied responses from the AI.

Another reason is that our simplifications to the experiment led to a convergence in outcomes. In Dhingra[4], a simpler AI (GPT 3.5) was used and a longer context was provided. We utilized Claude Haiku 3.5 which does have improved reasoning from GPT 3.5, but the impact would need further experiments to determine the magnitude of the impact. We posit that the treatment of the data would also impact the results - we curated the sample we chose and simplified the process by only including the first few lines of the biography. In Wikipedia entries of famous persons, controversial elements are generally relegated to later paragraphs unless they were an especially problematic historical figure. Overall, our AI demonstrated improved reasoning but reduced context, and similar to the above point, it may have either identified the individual that we provided information about or it may have provided a generic response due to not having further information about the individual's qualities and characteristics.

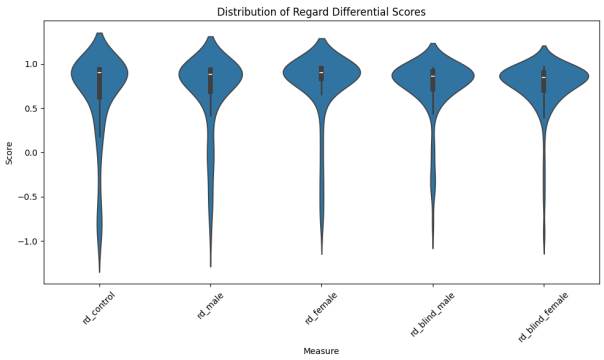

Figure 2: A violin plot of regard score differential for each treatment group

From figure 2 we can see that the treatments with disability identified tended to have less variation in the distribution of regard differential.

7.2 Visualizing Difference in Regard

Figure 3: A wordcloud made from the responses of prompts with no indicator of disability

Figure 4: A wordcloud made from the responses of prompts with an indicator of disability

A wordcloud allows the viewer to get a visual sense of the frequency of words in a body of text by representing higher frequency words in a larger font. Figure 3 is based on the output of the prompts for "male" and "female." Figure 4 is based on the outputs of the prompts for "blind male" and "blind female." We see from the wordclouds that there is a distinct contrast between the words generated for prompts with no disability marker and prompts with disability markers. Capitalized "PERSON" as the most frequently occurring word for both categories makes sense, as this is the entity name we used to replace individuals' names in their biographies, and the LLM starts nearly every response with this word.

For the wordcloud built on prompts with no indicator of disability, the most commonly occurring words are known, career, and throughout. These words align with a positive regard, suggesting the LLM has generates text that discusses the subject's career and what actions or attributes they be known for. In contrast, the highest frequency words for the outputs of the prompts featuring markers of disability are despite, disability, and challenge. This indicates that the AI frequently provides a generated text that makes disability one of the main focuses of the continued biography. Despite and challenge as commonly occurring words are more negative in sentiment and may be attributed to a lower regard, as they indicate the subject faced conflict. We include these visualizations to give a sense of what types of outputs we see from the LLM and allow readers to judge sentiment or regard for themselves.

7.3 Statistical Results

The primary question we address is whether the AI outputs demonstrate systematic bias against the discriminated group. From Figures 1 and 2, we observe that while the distributions have similar means, their variances differ. Our goal is to determine whether this difference is statistically significant and indicative of negative bias. To achieve this, we first conduct exploratory data analysis (EDA) and test the normality assumptions to decide whether parametric tests like ANOVA or alternative tests for non-normal distributions are appropriate.

Group	Statistic	p-value
rd_control	0.6416	0.00000
rd_male	0.6644	0.00000
rd_female	0.5507	0.00001
rd_blind_male	0.6091	0.00000
rd_blind_female	0.6136	0.00000

Table 2: Results of the Shapiro-Wilk test for different groups.

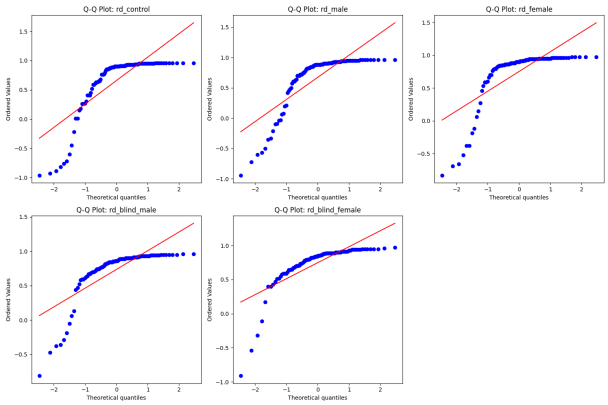

Figure 5: A qq-plots of regard score differential for each treatment group

As shown in Table 2, the Shapiro-Wilk test results indicate that none of the distributions meet the normality assumption (p-value < 0.0001). These findings are further corroborated by the QQ-plots in Figure 5, which visually highlight significant deviations in the tail values of the distributions compared to a normal distribution.

Given the violation of normality, we employ non-parametric tests to assess statistical differences. Using the Kruskal-Wallis H-test, we obtain a p-value of 0.0086. At the 5% significance level, this result leads us to reject the null hypothesis in favor of the alternative, indicating that at least one treatment group has a significantly different distribution.

Table 3 summarizes the results of pairwise comparisons using paired t-tests. The analysis reveals a statistically significant difference between the control and female treatment groups (p-value = 0.0052) at the 5% level. The blind female group approaches but does not meet this threshold. Overall, our findings suggest that while the AI does not exhibit significant bias against the disability groups, it does demonstrate gender-based bias.

Compare to Control	t-stat	p-value	Mean Diff
male	-0.3533	0.7246	-0.0152
female	-2.8572	0.0052	-0.0931
blind male	-1.5759	0.1182	-0.0756
blind female	-1.9456	0.0545	-0.0893

Table 3: Results of the pairwise comparison utilizing paired t-tests.

8 DISCUSSION

8.1 Analysis

The statistical results align with prior expectations. Previous studies, such as Dhingra [4], have highlighted evidence of gender and sexuality biases in AI, and our findings replicate and extend those results. While no significant bias was observed against the blind-identified group, our data indicate a clear, statistically significant bias against females, even in the absence of datasets curated specifically to test for gender discrimination.

Beyond these results, our primary focus was to develop a replicable framework that builds on Dhingra's methodology, relaxing constraints to evaluate broader forms of bias. This included key decisions about dataset characteristics and sample size, balancing computational limitations with the need for statistical power. Determining the most representative features for our sample data presented another significant challenge.

Additionally, we consider the limitations of the Regard Score [9] as a metric. While it successfully reaffirms Dhingra's findings, its broader applicability as a measure of bias may require refinement. Future work could explore alternative or complementary metrics tailored to diverse demographic groups.

8.2 Deviation from Dhingra

Our study aimed to generalize and simplify the experimental framework proposed by Dhingra [4] to extend its application beyond gender and sexuality biases. Specifically, we explored biases related to disabilities, such as blindness, while revising and simplifying the methodology to align better with the study's goals.

A significant divergence involved the preparation of input prompts. Dhingra's approach used automated tools like the NeuTral Rewriter [8] to mask biographical inputs by replacing gendered terms with neutral equivalents. While this automated pipeline offers scalability and modularity, it also introduces risks of semantic distortion, particularly when compounding errors from multiple model outputs. Human evaluations of these rewritten prompts highlighted suboptimal preservation of sentence context, a critical issue when assessing AI fairness.

In contrast, our study employed a manual approach, replacing discriminatory phrases using a curated dictionary. While resource-intensive, this method provided greater control over the integrity of the original data, minimizing unintended distortions. However, we acknowledge the limitations of manual editing, which we discuss further in the Limitations section.

Rather than focusing solely on neutralized prompts, we investigated how explicit disclosures of disability status influenced AI responses. We hypothesize that modifying system prompts and input prefixes could offer greater bias mitigation potential than input neutralization alone. This hypothesis is further explored in the Extensions section.

By streamlining Dhingra's methodology, we emphasize the importance of interpretable and reliable experimental pipelines, particularly for studies constrained by smaller datasets. Future research can build on this work by identifying which aspects of data preparation or prompt engineering have the most significant impact on mitigating AI bias.

9 LIMITATIONS

9.1 Methodology

There are several aspects of this work that could be improved upon. To begin with, we chose to use a randomly created sample of 100 biographies in combination with 6 different appended identifiers to prompt the LLM. A more thorough study might increase the number of identifier-biography pairs input to the LLM and thus the number of responses to use for an aggregated response score by increasing the number of biographies. Someone might also choose to find regard scores for additional types of disabilities. Physical disabilities, such as blindness are only a small part of the disability spectrum. In particular, it would be interesting to examine how an LLM considers a biographical subject with cognitive disabilities or an "invisible" disability.

Another area of likely improvement is the flow for de-biasing the biographical data before it is used to prompt the LLM. As this project is primarily a proof of concept, we attempted to use an open-source bias detection model called Dbias which has been shown to perform well in identifying individual words that are most responsible for the bias in an extract of text.[18] Through this process, we were able to see that the model did not perform as well in identifying words relating to disability bias as it did other forms of bias that the authors focused on in their development of the model. To create the Dbias model, the authors used a dataset of news articles which was labeled for bias related to gender, age, and education. The paradox of this project is that we want to explore disability bias because it has been less studied than other forms of bias, but that very fact introduced a good amount of difficulty into our process. Further work in this area could be to develop a proper disability bias identifier using generated data, such as can be created using the *Bias on Demand* project. [19]

Additionally, while we were able to remove or replace a majority of the words in the text that were markers of gender, we were not able to remove all markers of disability. Frequently the AI generated text mentioning disability even for the prompts where we did not explicitly apply a prefix with an identifier of disability. It is likely that the AI was able to intuit from the text that the sample biographies were about people with disabilities, which resulted in similar generated texts across the identifier categories we assessed. With a more concerted effort to remove markers of disability, we may have seen a wider variety of responses and more variation between the responses for prompts where we explicitly included a person with a disability and prompts where we did not.

9.2 Regard as an Imperfect Measure

Regard score as devised by Sheng et al was shown to be an effective metric for quantifying and comparing gender-based differences in perceived sentiment. [9]. It is possible however, that it is a less effective measure for the type of bias that we are studying here. Sheng prompted an LLM to supply a profession based on different subjects input into the AI. From that, we can see that supplying the profession of "babysitter" for a woman and "president" for a man indicates a clear difference in regard, with president being a profession associated with a higher regard than babysitter.

Many of the outputs given by Claude Haiku 3.5 in our study for prompts featuring identifiers of disability exhibit "inspiration porn," or "showing impairment as a visually or symbolically distinct biophysical deficit in one person, a deficit that can and must be overcome through the display of physical prowess."[12] The example output given earlier of the Albanian football player is a clear example of this, focusing on "breaking barriers and challenging stereotypes about athletes with different abilities" when prompted with the indicator of disability." While this can clearly be interpreted as a high regard sentence, this framing of disability as a deficiency to be overcome is increasingly being identified by disability activists as harmful to the movement for greater rights and inclusion for people with disabilities. This paradox of seemingly positive rhetoric in fact being reductive of the disabled experience, may partially explain why we see similar overall regard score differential among all of our identity categories.

The reliance on regard score as an evaluative measure is inherently tied to its definitions and boundaries, as outlined by Sheng et al. While the metric can be useful for assessing language polarity and the social perception of demographic groups, its application to specific contexts, such as disability-related narratives, reveals critical limitations. The regard score relies on predefined categories such as "positive," "negative," "neutral," and "other," which are applied based on a model's understanding of text classification since after all regard score does utilize a text model to generate the scores. This categorization oversimplifies the nuanced ways bias manifests in language. In the context of disability-related prompts, the metric may conflate patronizing or reductive statements with genuinely positive sentiments such as discussed above. As a result, the regard score may fail to distinguish between genuinely empowering language and harmful but superficially positive rhetoric, such as "inspiration porn."

Moreover, the calculation of the regard differential, defined as the difference between positive and negative regard scores, introduces further challenges. While this differential provides a quantitative means to compare perceived sentiment across groups, it does not account for the impact of other categories such as "neutral" or "other." In practice, this reductionist approach risks masking complex linguistic biases. For instance, two groups might exhibit identical regard differentials despite significant disparities in the prevalence of neutral or "other" categorizations. These subtleties can lead to misleading conclusions, particularly in cases where the "neutral" category might reflect a lack of representation or erasure, a form of bias in itself.

Additionally, the aggregation techniques used in regard scoring—such as mean or maximum values—may further distort the interpretation of results. Aggregating scores across diverse contexts assumes homogeneity in how different prompts and outputs relate to bias. This approach ignores the contextual sensitivity required to accurately assess bias, especially in datasets where certain groups are underrepresented or where language biases are more implicit.

Thus, while regard differential is a step toward quantifying bias, its current formulation does not capture the multidimensional nature of discrimination and societal perceptions. A more robust metric would need to incorporate additional linguistic and contextual analysis, including qualitative assessments of how specific phrasing perpetuates stereotypes or erases identity.

10 EXTENSIONS AND PRACTICAL APPLICATIONS

Our work provides a framework for identifying and addressing bias in large language models and AI. Bias is inherently difficult to define due to its many dimensions and context-dependent nature. It often varies according to cultural norms and societal expectations. Consequently, devising a suitable metric to measure bias is challenging. Bias is not static; its impact is contextual and dynamic, which complicates efforts to standardize its assessment. One effective approach to understanding bias is by analyzing outcomes, particularly by comparing the experiences of minority groups with those of majority groups. If minority groups exhibit systematically poorer outcomes relative to majority groups, this can indicate discrimination, while substantially better outcomes could suggest preferential treatment. Our framework focuses on identifying and quantifying these disparities, providing a mechanism to test interventions designed to reduce discrimination and promote equitable outcomes.

Our methodology evaluates how LLMs respond to queries that explicitly reference minority and majority groups. By using regard scores as a quantitative metric, we assess the equity of responses and investigate whether inherent biases exist in the model. The modular nature of our approach allows it to adapt to advancements in AI technology. This flexibility ensures that the framework remains applicable over time, accounting for cultural nuances, concept drift, and the evolving capabilities of LLMs.

Our experimental design establishes a foundation for extending this work to more complex tasks. Below, we discuss how this framework can be used in practical applications to mitigate bias and promote fairness in AI systems.

10.1 Mitigating Bias Through Augmentation

Currently, LLMs are trained on large datasets, often containing implicit biases inherited from the human data they consume. Additionally, mechanisms like Reinforcement Learning with Human Feedback (RLHF) [20] incorporate human perspectives into the training loop, potentially propagating societal biases. For instance, the linguistic quirks of data labelers can inadvertently influence the model's outputs, as observed in the increased prevalence of certain regional expressions following LLM mainstream adoption.

While foundational models provided by major AI companies include safeguards to mitigate bias, downstream applications often augment these models using external tools like Retrieval-Augmented Generation (RAG) [21]. RAG retrieves contextual information from external sources to improve the relevance of responses. However,

this augmentation can reintroduce biases present in the supplementary documents, potentially undermining the safeguards implemented during the foundational model's training.

Modern AI applications rely on layered augmentations, such as RAG and prompt-based techniques (e.g., zero-shot, one-shot, and multi-shot prompting). Each layer introduces opportunities for bias and, therefore, opportunities for intervention. Our framework can be integrated as an augmentation layer, influencing outputs between the user's input and the final result. For example, it can assess the regard scores at various stages and adjust subsequent processing to ensure equitable outputs.

10.2 Informing Outputs Through Prompt Engineering

Prompt engineering remains one of the most effective levers for influencing AI outputs. Sensitive topics can be flagged explicitly in system prompts, guiding the AI to consider potential implications of its responses. In our experiments, system prompts and examples significantly influenced regard scores, enabling the calibration of responses to reduce bias.

For instance, our experiments demonstrated that models like Claude Haiku, optimized for speed, often require more carefully crafted prompts to achieve equitable outputs compared to models like Claude Sonnet, which exhibit advanced reasoning capabilities. By experimenting with system prompts, augmented prefixes and suffixes, and example-driven tuning, developers can systematically evaluate and improve regard scores across different tasks. [22]

The versatility of our framework makes it adaptable to various LLMs, allowing practitioners to fine-tune prompts and achieve more balanced outputs across diverse applications.

10.3 Creating AI Tools and Agents

Our framework can also be embedded into AI tools and agents. In this context:

*Tools: extend an AI's functionality, such as enabling it to perform web searches or execute SQL queries.

*Agents: AI systems that are aware of their available tools and can autonomously decide when and how to use them.

By integrating the ability to calculate regard scores into an agent, we enable the system to self-evaluate its outputs for bias. For example, an agent could assess whether its responses demonstrate sufficient parity between regard scores for minority and majority groups. If inequities are detected, the agent could adjust its outputs before presenting them to the user.

This approach allows for proactive bias mitigation within dynamic, multi-step AI workflows. By embedding bias-awareness into the orchestration of tools and agents, developers can create systems that are not only functional but also equitable.

11 CONCLUSION

Our framework provides a systematic approach for evaluating and addressing bias in LLMs by analyzing how prompts are perceived across diverse, pre-selected demographic groups. Leveraging regard scores and their derived metrics, such as regard differentials, this methodology offers valuable insights into potential biases, enabling the integration of fairness considerations into real-world AI systems. By focusing on equity and fairness, our approach contributes to the development of more inclusive and socially responsible AI applications.

While our experiment did not find statistically significant evidence of bias against blind individuals, it confirmed the findings of Dhingra et al., whose study served as a foundation for our work, that females face significantly greater discrimination compared to males. This result underscores the persistent gender biases in language models and highlights the necessity of targeted interventions to mitigate such disparities.

Future iterations of this framework could address broader challenges, such as detecting intersectional biases, adapting to evolving cultural contexts, and incorporating more complex task structures. Additionally, developing custom scoring metrics tailored to specific applications and exploring alternative measures beyond regard differentials could further enhance the robustness of bias detection and mitigation efforts. These advancements will help ensure that AI continues to serve as a tool for inclusive and equitable progress across diverse domains.

REFERENCES

[1] National Council on Disability. *The Implicit and Explicit Exclusion of People with Disabilities in Clinical Trials*. Accessed: 2024-10-29. 2024. URL: https://www.ncd.gov/report/the-implicit-and-explicit-exclusion-of-people-with-disabilities-in-clinical-trials/#:~:text=Sixty%20one%20million%20Americans%2C%20or,even%20life%2Dsaving%20healthcare%20resources..

[2] Heather Feldner Laura VanPuymbrouck Carli Friedman. "Explicit and implicit disability attitudes of healthcare providers". In: *Rehabil Psychol* (2020), pp. 101–112. DOI: 10.1037/rep0000317.. URL: https://pubmed.ncbi.nlm.nih.gov/32105109/.

[3] Emily Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 610–623. DOI: 10.1145/3442188.3445922. URL: https://dl.acm.org/doi/10.1145/3442188.3445922.

[4] Harnoor Dhingra et al. *Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models*. June 2023. DOI: 10.48550/arXiv.2307.00101. URL: https://arxiv.org/abs/2307.00101.

[5] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. "Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens". In: *Findings of the Associa-tion for Computational Linguistics: EMNLP 2021* (2021), pp. 3116–3123. DOI: 10.18653/v1/2020.acl-main.487. URL: https://aclanthology.org/2020.acl-main.487.

[6] Ben Hutchinson et al. "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 5491–5501. DOI: 10.18653/v1/2020.acl-main.487. URL: https://aclanthology.org/2020.acl-main.487.

[7] Tony Sun et al. *They Them Theirs: Rewriting with Gender Neutral English*. 2021. DOI: 10.48550/2102.06788. URL: https://arxiv.org/abs/2102.06788.

[8] Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. "NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives". In: *CoRR* (2021). DOI: 10.48550/arXiv.2109.06105. URL: https://arxiv.org/abs/2109.06105.

[9] Emily Sheng et al. "The woman worked as a babysitter: On biases in language generation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP* (2019), pp. 3407–3412. DOI: 10.18653/v1/D19-1339. URL: https://aclanthology.org/D19-1339/.

[10] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. "Perturbation Sensitivity Analysis to Detect Unintended Model Biases". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 5740–5745. DOI: 10.18653/v1/D19-1578. URL: https://aclanthology.org/D19-1578/.

[11] Tolga Bolukbasi et al. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. July 2016. URL: http://arxiv.org/abs/1607.06520.

[12] Jan Grue. "The problem with inspiration porn: a tentative definition and a provisional critique". In: *Disability & Society* 31.6 (2016), pp. 838–849. DOI: 10.1080/09687599.2016.1205473. URL: https://doi.org/10.1080/09687599.2016.1205473.

[13] Vinitha Gadiraju et al. ""I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023), pp. 205–216. DOI: 10.1145/3593013.3593989. URL: https://doi.org/10.1145/3593013.3593989.

[14] Mahika Phutane, Ananya Seelam, and Aditya Vashistha. *How Toxicity Classifiers and Large Language Models Respond to Ableism*. 2024. DOI: 10.48550/2410.03448. URL: https://arxiv.org/abs/2410.03448.

[15] Remi Lebret, David Grangier, and Michael Auli. *Neural Text Generation from Structured Data with Application to the Biography Domain*. Mar. 2016. DOI: 1603.07771. URL: https://arxiv.org/abs/1603.07771.

[16] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". Documentation: https://github.com/explosion/spaCy. 2017.

[17] Akmar Chowdhury and Zoë Bakker. *Conagra Debias*. https://github.gatech.edu/achowdhury99/conagra_debias. 2024.

[18] Shaina Raza, Deepak John Reji, and Chen Ding. "Dbias: detecting biases and ensuring fairness in news articles". In: *International Journal of Data Science and Analytics* (2024), pp. 39–59. DOI: 10.1007/s41060-022-00359-4. URL: https://doi.org/10.1007/s41060-022-00359-4.

[19] Joachim Baumann et al. "Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023). DOI: 10.1145/3593013.3594058. URL: https://doi.org/10.1145/3593013.3594058.

[20] Stephen Casper et al. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. 2023. DOI: 10.1145/3593013.3594058. eprint: 2307.15217. URL: https://arxiv.org/abs/2307.15217.

[21] Xuyang Wu et al. *Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems*. 2024. arXiv: 2409.19804 [cs.CL]. URL: https://arxiv.org/abs/2409.19804.

[22] Jules White et al. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. 2023. arXiv: 2302.11382 [cs.SE]. URL: https://arxiv.org/abs/2302.11382.