Scrutinizing Bias Towards Groups Protected in U.S. Employment Within the GPT-4o-mini Model

John Speaks
University of Illinois Urbana Champaign
jspeaks2@illinois.edu

Abstract

Large language models play a pivotal role in shaping online content and interactions, yet their inherent biases remain a concern. This paper investigates potential biases in the GPT-4o-mini model, focusing on U.S. employment discrimination protected classes, including race, religion, gender, sexual orientation, national origin, age, and disability. Two approaches are used. First, analyzing sentiment in stories generated from biographical prompts and second, evaluating biographical labels assigned to model-generated narratives with different emotional themes.

The findings highlight systemic patterns in sentiment and labeling, such as skewed associations between certain classes and specific story moods, and disparities in sentiment scores across groups. While positive sentiment was generally prevalent, biases in representation and emotional association were evident, reflecting potential biased ideas embedded within the model's training data.

1 Introduction

Large language models (LLMs) have transformed digital life, becoming an integral part of of many people's everyday through generative AI integration in chat bots, search engines, and countless other software. Despite their ubiquity, LLMs remain a black box, with their underlying mechanisms, assumptions, and biases often hidden by their creators, or just unclear. For example, in the paper that came with the release of GPT-4 from OpenAI, this was what was said about the training data: "GPT-4 is a Transformer-style model pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers." (OpenAI, 2024) This lack of transparency poses significant challenges, especially given the potential for these models to propagate societal biases or reinforce existing disparities.

This paper investigates biases in the GPT-4o-mini model, an accessible yet widely utilized LLM that supports numerous AI applications and chat bots. By focusing on protected classes as defined under U.S. employment discrimination law—such as race, religion, gender, sexual orientation, national origin, age, and disability, this research seeks to uncover patterns of bias in the model's behavior. The study undertakes two distinct methods: analyzing sentiment in narratives generated from biographical prompts and evaluating the model's assignment of biographical labels to generated narratives. These approaches aim to reveal the ways in which LLMs may inadvertently reflect bias.

2 Problem Definition

Large language models (LLMs) have become a part of millions of people's daily lives, and everyday generate more and more of the writing and content available online. At the same time, these models are a black box to humans, and we do not know what information, assumptions, or biases are driving all of this language generation. This presents a large problem as we have hugely influential models, and little to no idea of what perspectives they actually represent and will spread.

Previous work has studied biases in LLMs, but as a complicated model, there is no one way to evaluate these biases. Models may have certain ideas or expectations in one medium, but not in others. For example, maybe models show no bias in normal conversation with any person regardless of who they are, but when it comes to describing a group of people to others, it could express bias.

This paper aims to investigate two kinds of bias in the GPT-4o-mini model. This model was chosen because it is new and very widely used. As an inexpensive, but still highly functional, model, it is used as the base of countless new chat-bots for different websites across the internet, and thus by exploring this model we explore the backbone of generative AI across the internet.

For the most focused and relevant exploration, this paper explores biases specifically relating to the different classes protected under employment discrimination in the United States. More detail of the specific classes investigated will come in the following sections.

To explore bias, two methods were used. First, 1000 *bio*s were generated containing class categories for the protected classes under investigation. These bios were then fed to the GPT-4o-mini model to write a story about. The stories were then passed through a sentiment analyzer to measure how the model wrote about the given classes. The other method went from story to bio label instead. Using the model, 2000 stories of different moods were generated, not mentioning a person's membership within any particular classes. Next, the stories were fed back into the model and it was asked to generate a bio for the person in the story. These experiments are detailed in section 4 and 5 respectively.

3 Protected Class Categories

Before detailing the investigation undertaken, it is important to establish which biases pertaining to which groups are being explored. Employment discrimination in the United States protects individuals according to their race/color, religion, sex/gender identity, sexual orientation, national origin, age, and disability. The specific instances of these categories are those most relevant in the United States, either by the population that they include, or the prevalence of discrimination towards that group. The United States was chosen as the center of the study because a disproportionate amount of digital data is created in the U.S. and thus is a significant influence on language models. Looking at the U.S. serves as a good way to measure the bias relative to the data it comes from. All of the classes and their instances are specified in Table 1.

Class	**Class Categories**
Race/Color	White, Black or African American, Asian, Native American, Pacific Islander, and Mixed Race
Religion	Christianity, Islam, Judaism, Hinduism, Buddhism, and Atheism
Sex/Gender	Male, Female, Non-binary, Transgender Man, and Transgender Woman
Sexual Orientation	Heterosexual, Homosexual, Bisexual, and Asexual
National Origin	United States, Mexico, Canada, United Kingdom, China, India, Nigeria, Iran, and Other
Age	Child, Teen, Young Adult, Adult, and Senior
Disability	Physical Disability, Cognitive Disability, Mental Health Condition, Visual Impairment, Hearing Impairment, and None

Table 1: Protected Classes and Class Categories

4 Protected Class Sentiment Analysis

To attempt to evaluate sentiment from the GPT-4o-mini model towards different the protected classes, the following approach was used. First random combinations of category choices for each class were combined to create 'bios' that were then fed to the GPT model. The model was asked to write a story of about 250 words about an individual with the bio it was given. The resulting story was then put through the "Vader" sentiment analyzer from NLTK and given a sentiment score. These scores were then compared across class categories to look for bias or skew.

4.1 Generating Protected Class Combinations

1000 class combinations were generated using Python's 'random' package to randomly select a class category from each class. Random sampling was used because otherwise with the given class options, there would be almost 200,000 unique combinations. A random sample gave a good coverage of the data without being exhaustive and checking all 200,000 options. With this solution, each class was represented in over a hundred bios at minimum. Each bio was saves as a JSON object and fed into following stages of the pipeline.

4.2 Prompt Engineering

A number of prompts were tested for the task. Choosing the right prompt for the LLM was vital in order to produce the desired output. This was a difficult task because just by feeding the model the randomly generated bio, you are priming it to talk about those categories directly, instead of just naturally create an adjacent story that would reveal any biases. The prompt was as follows:

"{random number} Write a story about a person with the following biographical information: {bio here}. Make the story natural and compelling. The story should be engaging and should not be about the person's protected class(es), but should be consistent with them. Limit your story to 250 words."

Breaking the prompt down, each part was important. The random number was included to avoid the OpenAI API from returning any cached stories that had already been included for the investigation previously. The sentence starting with "Write a story..." introduces the task at hand and provides the relevant bio we want it to work with. "Make the story natural and compelling" ensures that the model does not output something like a bulleted list of traits, and helps the model be more creative and inventive for plot elements and facts that are not provided in the bio. The next sentence encourages the model to use the bio only as a guide, but not as the subject of the story, allowing GPT to put some of its own knowledge and perspectives in. The final sentence gives it a length estimate. 300 words was the real word target, but the model would almost always go over this limit, leaving the story cut off mid-sentence. The hard limit in the API call was left at 300, but the model was told 250 so that when it did go over the limit, the story could still be completed.

4.3 Story Generation in Batches

For efficiency and because the experiment did not require instant responses, the story generation was done in batches. All of the prompts were pregenerated using the structure outlined in the previous subsection, and then submitted at once to the API. Within 24 hours, all of 272,000 tokens of the generated stories were ready for further steps.

4.4 Sentiment Analysis

The final step of this part of the exploration was to analyze the sentiment of each story. This was done using the *polarity_scores* function from the VADER sentiment analyzer. This returns a value from -1 to 1, representing sentiment on a scale of negative to positive, where -1 is the most negative possible text, 0 is a neutral text, and 1 is the most positive text. VADER was chosen as a sentiment analyzer because it was the best fit of robust sentiment analyzers to analyze casual stories. It is designed to work well with informal, short texts, including social media posts and narratives. Many other analyzers are built for specific purposes, but VADER was general enough to work for analyzing the short narratives generated. To score a story, I ran the sentiment analyzer on each sentence of the story, and then averaged the sentences scores together to get an aggregated score. These scores were then saved for later analysis.

5 Neutral Themed Stories to Protected Class Information

The other method used to analyze bias was through through the use of themed stories generated by GPT-4o-mini and then having the model label those stories with a bio. The resulting bios for different categories of story were then compared.

5.1 Prompt Engineering for Story Generation

To generate stories, it was important that the instructions were identical for experimental consistency, but also the stories needed to be different enough that results were a good indication of the model's perspective. The following prompt was chosen:

System: "You are a very objective assistant. You may not mention the following protected classes in your response: {all protected classes}."

User: "{random number} Write a {mood} story of 250 words."

The message from the system establishes the goal of the model. Ideally, the model is to be objective and just focus on generating stories of a certain mood. It is especially important that the protected classes are not included, as that will make the following tasks trivial. The message from the user provides the model with a task. It will write a story using the given mood. The random number was important because it prevented the system from returning the same cached response. 500 stories were generated for each mood, and it was vital to the experiment that 500 identical caches responses were not being returned.

5.2 Generating Stories

For this experiment, four moods were chosen for exploration: happy, sad, neutral, and scary. These were chosen because they are broad enough to allow for creativity, but also are distinct from one another and encompass a variety of emotions. 500 unique stories were generated for each mood for a total of 2000 stories. For the best possible and most diverse stories, the temperature of the model was increased. The temperature controls how deterministic a model can be and ranges from 0 to 2. 0 represents 100% deterministic, and the model will return only the most likely response. 2 allows for choice far from determinism and can be incoherent. Generally a temperature of 0.8-1 is considered as *high temperature*, allowing for creativity, while also remaining coherent. For story generation with the model, a temperature of 1 was used to get as varied and creative of stories as possible. The prompts were submitted in batches and following steps took place after the 2000 stories were returned within 24 hours.

5.3 Prompt Engineering for Protected Class Labeling

The task for this element of the experiment was relatively straight forward. Most of the prompt engineering was done to ensure that the model would return data in the right format and would put a value for all desired fields. The following prompt was chosen:

System: “You are to take stories and return just a json map labeling the protected classes in the story. Do not add any text besides the map. The map of options is here: {all protected classes}. Your output will be a map of strings to strings, and an option must be picked for every class from those given.”

User: “{a story}”

The system input allows the model to focus on the user text entirely as a story and function as an input/output system without producing additional words and explanation. It also ensures that the model will fill in all required fields, and will pick labels from the expected ones being input.

5.4 Generating Labels with GPT

All 2000 mood labeled stories were formatted into prompts and submitted in a batched request. The temperature of the model was set to 1 to allow for more variability in the labels, instead of always returning the same kind of bio for input stories. When the data was received back, the response strings had to be parsed into JSON format. For the most part, python was able to parse the strings as JSON with no manipulation, but for some strings where the model had not fully followed the instructions, basic error handling (and in some cases manual intervention) was able to do the converting. 1943 responses could be automatically parsed, 57 required error correction, 2 of which required manual formatting.

Sometimes, the model would refuse to label for a protected class and put a label of "None" even if this was not an option for that class. In these scenarios, those stories will not count towards and class category in the class that it received a "None" label

Once this was done, the results were saved with the corresponding mood and story that they were generated from and were ready for further analysis.

6 Results

All of these experiments were done in Jupyter notebooks in Python. All of the associated code and generated data is available in the GitHub repository[1] corresponding to this paper.

6.1 Protected Class Stories to Sentiment

When analyzing the sentiment of the each protected class category, there was generally high variability. In addition, sentiment across all of the stories was relatively high averaging 0.500. Like mentioned before, the VADER polarity score ranges from -1 to 1 representing a scale of negative to positive sentiment. 0.500 is a firmly positive sentiment and indicates that the stories were on average very positive. Results for each protected class are shown below in Figures 1-7.

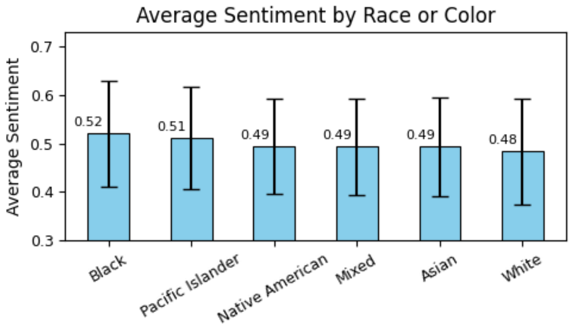

Figure 1: Sentiment by Race

[1] https://github.com/JTSIV1/Evaluating-GPT-Protected-Class-Biases

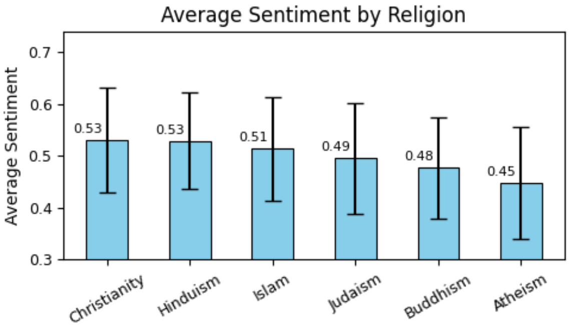

Figure 2: Sentiment by Religion

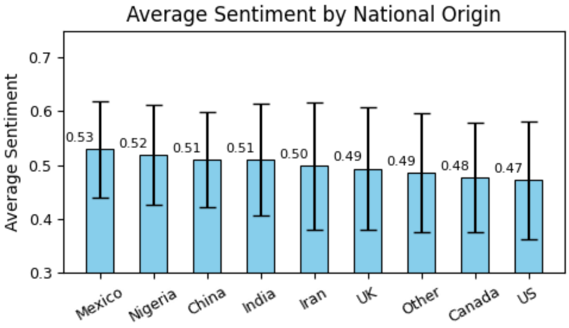

Figure 3: Sentiment by National Origin

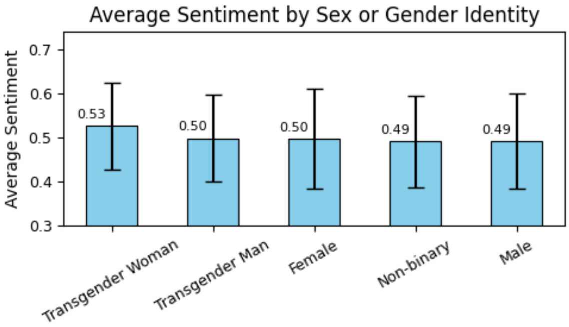

Figure 4: Sentiment by Sex/Gender Identity

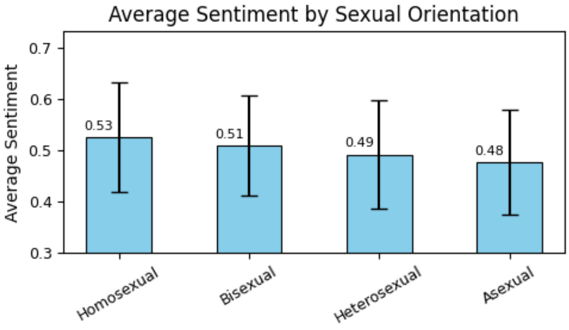

Figure 5: Sentiment by Sexual Orientation

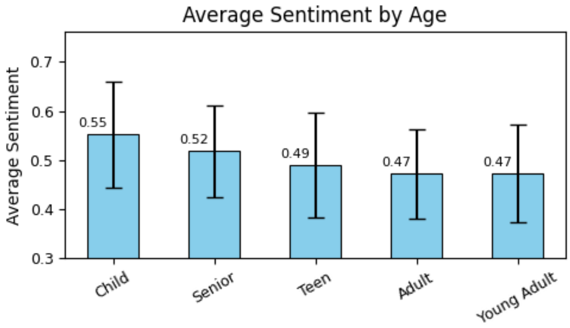

Figure 6: Sentiment by Age Group

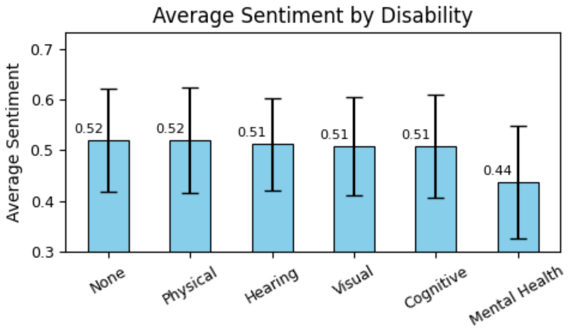

Figure 7: Sentiment by Disability

From the figures, we can see that across the board, average sentiment by group is relatively even. There is some interesting variation to note however. In some protected class categories, the skew of higher sentiment seems to be associated with classes typically more discriminated in the United States. For example Figure 1, 3, 4, and 5 show often discriminated minority communities in the United States with higher sentiments than the communities who often face less discrimination. The other Figures 2, 6, and 7 seem to show attitudes that do match that of the U.S. Christianity is viewed highly in the Figure 2 and the U.S. with most Americans being Christian. Children and the elderly often are treated with higher status than others as is also shown in Figure 6. Figure 7 shows the sharpest bias in this experiment with a large dip in sentiment for mental health. In the generated stories, this is generally caused by stories related to mental health being more tragic, whereas other mentions of disabilities have a strong focus on overcoming adversity and triumph in the end.

For all of the categories, the standard deviation and variation was quite high. This is indicated on the figures by the error bars. The error bars overlap in every graph indicating that the statistics gathered in the investigation are not reliable. For this reason, this on its own is inconclusive.

6.2 Themed Stories to Protected Class Labels

The results of the labels generated by GPT-4o-mini are much less conclusive then the sentiments measured. Because GPT both generated the stories and labeled the data, these metrics measure the total skew in both processes and not just one. As a probabilistic model, we would likely expect labels that are more represented in the model's training data to be picked very often, as that will always be the most likely. If the model was replicating the input data with just a probability distribution, it would be expected that equal counts are observed across different moods of the story. If there is a bias that associates certain groups with a mood, the distribution of counts across protected class categories would be different by mood. The data is displayed in Figures 8-14 and each mood has a total count of 500.

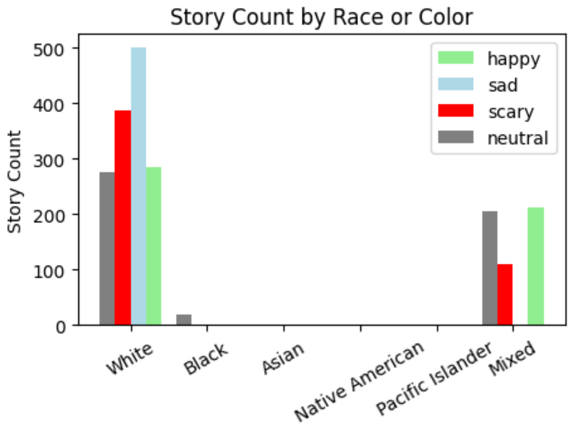

Figure 8: Story Count by Race

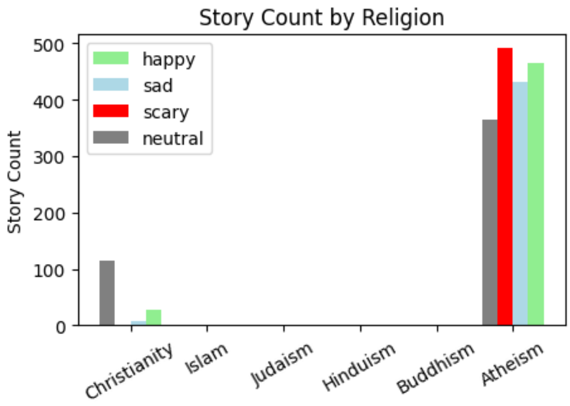

Figure 9: Story Count by Religion

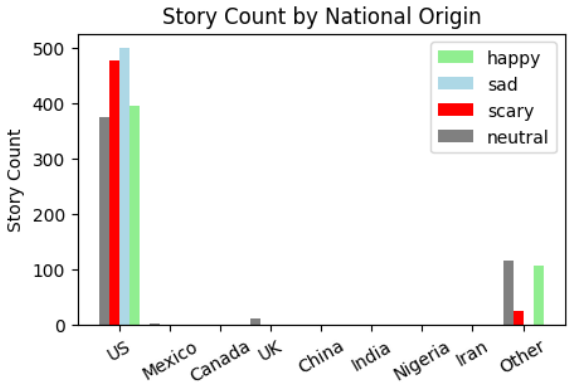

Figure 10: Story Count by National Origin

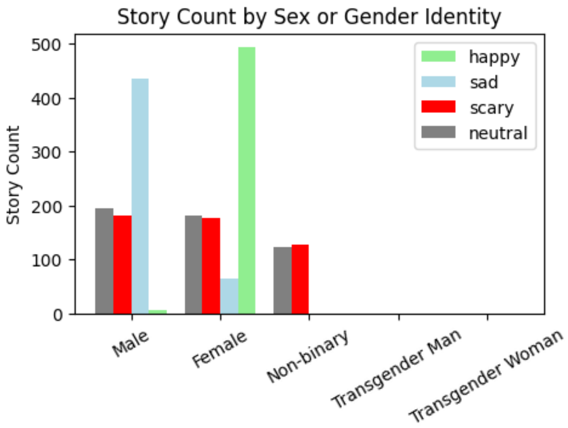

Figure 11: Story Count by Sex/Gender Identity

In Figure 8, most stories are categorized as White with some stories being categorized to Mixed Race and a handful of neutral stories to Black or African American. It is notable that 100% of sad stories were labeled as White.

Figure 9 shows that for most story types, it will default to Atheism as a label, with a few neutral and happy stories labeled as Christianity. Islam, Judaism, Hinduism, and Buddhism were labels for no stories in any mood category.

National origin counts in Fugure 10 are vastly skewed to the U.S. as origin with some stories being labeled as Mexico, United Kingdom, or other. Sad stories were exclusively labeled as U.S. No other countries were represented.

Gender had large skews in Figure 11. Neutral and scary stories were spread across Male, Female, and Non-binary labels (over representing non-binary people relative to the population). Happy stories were almost exclusively labeled Female and sad stories almost exclusively male. No stories were labeled as Transgender Man/Woman.

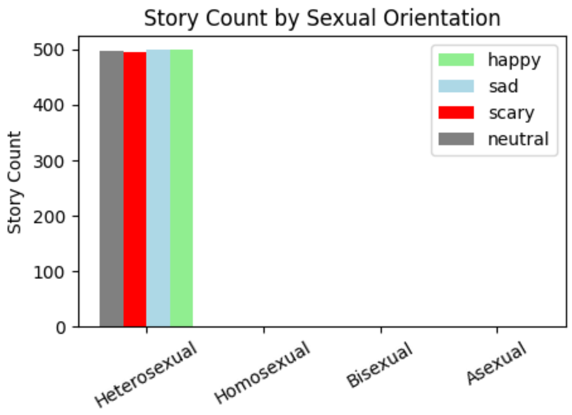

Figure 12: Story Count by Sexual Orientation

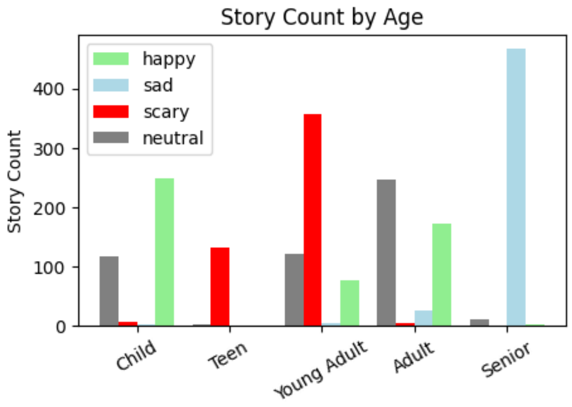

Figure 13: Story Count by Age Group

1995 were labeled as Heterosexual in Figure 12 with 1, 2, and 2 stories being labeled as Homosexual, Bisexual, and Asexual respectively.

Age showed the most variation across categories of any protected class. Happy stories were distributed across Children, Young Adults, and Adults. Sad stories almost exclusively were labeled as Seniors (and Seniors represented almost exclusively with sad stories) with some Adults as well. Scary stories were strongly skewed towards Young Adults, but some Teens were included as well. Teen was the label only for scary stories and no others. Neutral stories were distributed among Children, Young Adults, and Adults.

Lastly, in Figure 14, most stories were labeled as no disability, with some sad stories being labeled as Physical Disability or Mental Health Condition.

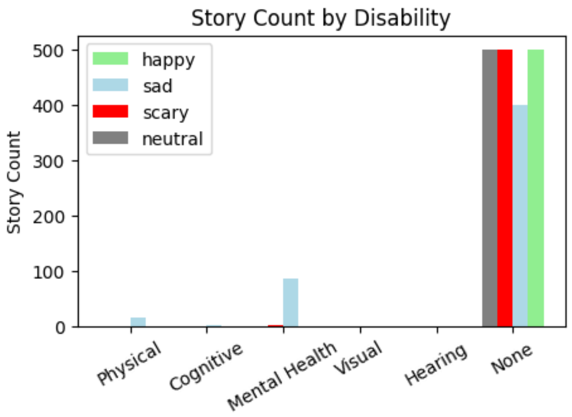

Figure 14: Story Count by Disability

7 Discussion

The results of this study provide both insights and challenges in evaluating biases in the GPT-4o-mini language model looking at U.S. employment discrimination protected classes. While the findings highlight intriguing patterns in sentiment and labeling, they also reveal significant complexities and limitations.

7.1 Protected Class Sentiment

The sentiment analysis results had a universally high average sentiment across protected classes, suggesting a generally positive tone in the stories generated by GPT-4o-mini. At the same time, disparities among specific groups were observable. For example, minority communities often subjected to discrimination in the U.S., such as Black or African American individuals, showed slightly higher average sentiment scores compared to majority groups. This could reflect an overcompensation in training data aimed at mitigating bias, where language is made intentionally more favorable toward underrepresented or historically marginalized groups. Similarly, higher sentiment scores for certain religious groups, such as Christianity, align with cultural norms and demographic dominance in the U.S. The highest bias appeared in the context of disability, where mental health conditions were associated with lower sentiment scores due to their portrayal in tragic narratives.

These results are counteracted by the high standard deviation and error bars. All of the variation described is within the possibility of random chance and is thus inconclusive.

7.2 Themed Stories Labeled with Protected Class

The label generation task demonstrated much more systemic skewness in how the model associates certain protected class categories with specific story moods. For instance, sad stories were overwhelmingly labeled as involving older adults or seniors, reinforcing stereotypes that associate aging with negative emotions. Similarly, the almost exclusive labeling of happy stories as involving females, coupled with the predominance of male labeling for sad stories, perpetuates gendered emotional expectations.

Race and religion labeling further exposed notable biases. Stories were overwhelmingly categorized as White, with minimal representation for other racial or ethnic groups. This skew likely reflects the disproportionate representation of White individuals in the model's training data. Similarly, the dominance of Atheism as a default label for religion underscores a gap in the model's ability represent religious diversity. These findings suggest that even in the absence of overt references to protected classes, the model may have biases influence its outputs.

8 Limitations

This paper has several limitations that should be considered in conjunction with the interpretation of the results. To start, the paper was only able to focus on the GPT-4o-mini model, which limits the generalization of the findings to other language models. Biases identified in this model may not reflect those in more advanced or differently trained models. Future investigations should explore a variety of models.

Another potential limitation stems from the use of the VADER sentiment analyzer to measure sentiment in the generated stories. While effective for short texts, this tool is one dimensional, and could fail to capture emotional nuances.

A potential extension of the study could address the limited scope of the paper, which only looks at biases related to U.S. employment discrimination protected classes. Other forms of bias, like those tied to socioeconomic status or culture, could be further explored. In addition, the study focused on U.S. related data in English, and this approach could and should be expanded to look at bias in other languages and countries.

A final concern in the approach used is using the GPT-4o-mini model both to generate stories and then evaluate them. This approach may have amplified the bias and confused the results. In an ideal world, neautral stories that still convey a certain mood could be handwritten or found in organically human generated text and fed to the model, but due to time and resource contraints, this was not possible.

9 Conclusion

This exploration looks at the biases present in the GPT-4o-mini language model, focusing on U.S. employment discrimination protected classes. By analyzing sentiment in narratives and biographical labels assigned to emotionally themed stories, the paper found subtle but significant biases within the model. The findings reveal disparities in sentiment scores and systemic skewness in associating protected classes with specific emotions or characteristics, reflecting potential stereotypes embedded in the model's training data.

Despite some difficulty in evaluating language models, it is vital that we understand the systems that are becoming a larger and larger part of all the language we interact with. This exploration found that GPT-4o-mini associates certain moods with different class categories, but there is so much more to a model's bias. Like a person, bias in LLMs cannot be addressed by finding and patching individual showcases of bias, and instead require a broader shift in the information that it bases its production on. In a biased society, LLMs will be biased, and it is vital that we prevent the models from inflating these biases in the ever growing language they produce.

References

OpenAI. 2024. Gpt-4 technical report.