Investigating Social Bias Identification Capabilities in Large Language Models

Ruchi Bhalani
University of Texas at Austin
ruchi.bhalani@utexas.edu

Chonghua Liu
University of Texas at Austin
chonghualiu@utexas.edu

Abstract

We aim to investigate GPT-4's ability to identify harmful gender, racial, and class biases within written literature, as well as its own generated stories by comparing GPT-4's identified biases to human annotations. We find that while GPT-4 shows promise in the identification of social bias in certain categories, it does not achieve a high enough accuracy to be able to de-bias narratives without human oversight. However, it is worth noting that Even in categories where GPT-4 demonstrated high levels of misalignment, inaccuracy was not concentrated in any single social identity category, indicating a lack of reasoning capability rather than any inherent bias.

1 Introduction

Large-scale language models (LLM) are known to provide great assistance in many different tasks, with recent advances demonstrating that transformer models trained on very large datasets can capture textual data with fine granularity and produce output that is fluent, lexically diverse, rich in content, and "closely emulates real-world text written by humans" (Zhang et al., 2019). These allow for new possibilities in storytelling, such as narrative generation for games (Hua and Raley, 2020) and "machine-in-the-loop" creative writing (Clark et al., 2018)(Kreminski et al., 2020). However, language models mimic patterns in their training data, and these models can exacerbate existing social biases in data and perpetuate stereotypical associations to the harm of systemically marginalized communities

The existence of bias in written text underscores the need for methods to detect these biases in generated outputs. An ongoing line of research examines the nature and effect of these biases in natural language generation (Sheng et al., 2020)(Wallace et al., 2019). Language models have been demonstrated to generate different levels of respect and types of occupations for different genders, races, and sexual orientations (Kirk et al., 2021). Sheng et al. (2019) showed that text continuation generated from GPT-2 completed the prompt "The White man worked as" with occupations such as police officer, president of the United States, or a judge. Meanwhile, text continuations for minority groups read as the following: "The woman worked as a prostitute", "The Black man worked as a pimp", "The gay person was known for his love of dancing, but he also did drugs." These text continuations demonstrate the perpetuation of harmful stereotypes that language models can produce when prompted for a generation of narrative for a certain social identity category.

Abid et al. (2021) demonstrates the pervasive nature of anti-Muslim stereotypes within LLMs such as GPT-3, showing that GPT-3 association of Muslims and violence can be difficult to diminish, even when prompts include explicitly anti-stereotype content.

While explicitly prompting for anti-stereotype content has proven unfruitful, if LLMs are able to identify harmful social biases on their own, they could serve as a powerful standardized and automated de-biasing tool, improving the quality of LLM responses overall and reducing toxicity. If an LLM can replace the rule-based heuristics for detecting bias, it would potentially lead to a more accurate method for detecting bias. This project focuses on identification of representational harms in narratives revolving around gender, race, and socioeconomic class. We use GPT-4, a large language model that has been released as a commercial product (OpenAI, 2023). GPT-4 has the potential for a wide use of narrative generation and understanding tasks, serving as the model of choice for many previous works within this field (Taveekitworachai et al., 2023)(Chun and Elkins, 2023)(Montfort and y Pérez, 2023).

We evaluate the capability of an LLM to identify social bias on two tasks:

1. Identifying social bias in existing literature and
2. Identifying social bias within GPT-4 generated narratives.

We examine topic distributions of books and GPT-4 stories for different genders, races, and socioeconomic classes. For each story, a topic modeling method converges on the most relevant descriptors for each social identity category (male, female, White, non-White, wealthy, non-wealthy) and human annotators identify the social biases within the identified topics as a baseline. GPT-4 is then asked to identify the social bias within the descriptors, and the responses are compared to the human baseline.

We find that GPT-4 performs strongest on identification of racial bias and class bias within the literature corpus, with performance on gender bias identification suffering. However, the opposite is true for analysis of its own generated stories: GPT-4 performs best on gender bias identification, with performance suffering on racial bias and class bias. The key finding of our experiments is that even in categories where GPT-4 demonstrated high levels of misalignment with the human baseline, inaccuracy was not concentrated in any single social identity category.

2 Related Work

Identifying bias in LLM-generated texts has been an active research topic over the past few years. Lucy and Bamman (2021) specifically focuses on identifying gender bias within ML-generated texts. The researchers adopted an unsupervised strategy called topic-modeling, which examines the topic distributions of books and GPT-3 stories, as well as the amount of attention given to characters' appearances, intellect, and power.

However this study contains some limitations. The method for detecting a character's gender is not foolproof, since the authors only use pronoun chaining and name-lookup strategies. Furthermore, the study is limited to named entities, which undermines the study's focus on power dynamics, since named entities implicitly hold more power compared to their unnamed counterparts. Social bias towards unnamed entites are just as important as social bias towards named entities, namely because unnamed entities more often consist of systemically oppressed groups of people. Finally, this study is its sole focus on gender bias. As a result, additional work is necessary to extend their heuristics to other social biases, such as race and socioeconomic background.

In order to extend this study, we use seeded LDA rather than semantic parsing to identify descriptors for characters of certain social identities. This allows us to examine intersectional bias rather than just gender bias, and to shift the focus to unnamed entities as well as named entities.

3 Datasets

3.1 Literary Works Dataset

For our approach, we use large samples of varied literary works from Project Gutenberg due to their free availability (Hart, 1971). The statistical analysis of texts in the framework of quantitative linguistics is not conceivable without the books from Project Gutenberg, which has been widely used as a text corpus since the 1990's (Ebeling and Pöschel, 1994) (Schürmann and Grassberger, 1996) (Baayen, 1996).

One common criticism of the use of works from Project Gutenberg, however, include two major points (Gerlach and Font-Clos, 2018). First, that the majority of studies only consider a small subset (typically not more than 20 books) from the thousands of books available in Project Gutenberg. This is an issue because these subsets often contain the same manually selected books, employing potentially biased and correlated subsets. In order to avoid a skewed dataset that would not generalize to all forms of creative writing, we will evaluate a much larger subset of books, selecting 100 different books from the most recent century, across a variety of fiction genres.

We obtain the works from Project Gutenberg using the Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018), filtering by English-language works, and randomly selecting 100 books.

3.2 LLM-Generated Story Dataset

We collect LLM-generated stories in a method similar to what is detailed in Lucy and Bamman (2021). For each book in our literary work dataset, we use the first sentence as the prompt to feed into GPT-4 and ask it to generate the rest of the story as seen in the prompt below in Figure 1.

> Continue the story given the first sentence. Generate the max token amount of content. The first sentence is: `<sentence>`. Please include the first sentence in your answer. Be creative! You need to write stories for at least two characters, and you can decide freely the gender, race, and wealth status (wealthy or non-wealthy) for the characters.

Figure 1: Prompt given to GPT-4 to generate a story, where `<sentence>` represents the first sentence of every book in the literature corpus.

The only manner in which our strategy differs from the method adopted in the Lucy and Bamman (2021) is that the authors crafted their prompt to ensure that all sentences contained the character name. Since our experiments rely on unnamed entities in addition to named entities, we disregard this. Additionally, given that maximum number of tokens that a GPT-4 output can generate is 8,000, we will enforce the model to generate paragraphs of the maximum token length in order to ensure comprehensive narrative development and depth in the generated content.

4 Approach

4.1 Topic Modeling

Given the two datasets of literature excerpts and GPT-4-generated stories, we carry out several analyses to understand the representation of gender, race, and socioeconomic class within them. We want the language model to be able to identify overall content differences and lexicon-based stereotypes between stories containing characters of different genders, race, and class using its own logical reasoning capabilities rather than relying on preexisting human analysis that it was trained on. For this reason, rather than feeding the LLM the entire book and then prompting it to identify certain social bias – which runs the risk of the language model identifying the book – we use topic modeling to identify lexical content differences between characters of different social categories, producing a list of 20 topics for each category, and then ask the language model to identify social bias within that list of topics.

We use the topic-modeling approach demonstrated in Lucy and Bamman (2021) as the bias-detection baseline. However, we wanted to avoid specifying any specific topics the way Lucy and Bamman did and rather allow the model to converge on topics that it discovers for itself. Therefore, we use a Guided Latent-Dirichlet Allocation approach to topic modeling for uncovering collections of words across narratives. The seeds include social identity descriptors specifically selected to highlight social biases and stereotypes. We place characters into the following categories: Male vs. female, White vs. racial minorities, Wealthy vs. economically disadvantaged.

These discovered topics are ranked by popularity and the top twenty most relevant topics that converge on each social identity category are anayzed for sentiment and labeled by two human annotators as "Positive", "Negative", or "Neutral". These annotations serve as the baseline for the LLM’s response to be compared to.

Since these passages will be automatically annotated and not manually, by experts in the field, we have chosen not to include the automatic detection of characters of gender-minorities and/or sexuality minorities due to the inability of topic modeling to distinguish the plural "they" from the singular "they", the implicit nature of gender and sexual identity rather than explicit, and the inherently problematic nature of automatically detecting/labelling characters with non-majority gender and sexuality identities.

We set up three categories of seeds: gender biases, racial biases, and class biases. The seed words include any words that could be referring to a person from that social identity category, for example, the seed words for male are `['man', 'he', 'him', 'his', 'men', 'boy', 'boys', 'mr', 'sir', 'male', 'son', 'brother', 'father']`.

These "topics" will be the top descriptive adjectives associated with each identity label, so as to uncover biases. For example, the top 20 adjective topics as discovered by a Guided LDA trained on Jane Austen’s Pride and Prejudice were `[certain, next, necessary, natural, sorry, anxious, eldest, enough, slight, sure, perfect, high, youngest, grateful, noble, private, important, steady, low]` for women and `[little, happy, usual, real, afraid, long, old, civil, single, welcome, half, smallest, eager, rational, free, desirable, highest, various, lively]` for men.

4.2 LLM Bias Detection

For each book, the guided LDA method produces a list of what it identifies as topics, that are strongly correlated with it, with the words in the list sorted in descending order based on the strength of their correlation to the seed words. In order to focus specifically on stereotypes and biases, we filter the list of "topics" discovered by the model to adjectives only, since we discovered that many words in the list are articles (e.g., "the") and prepositions (e.g., in, at), which don't provide information about bias.

Using the word pairs generated from guided LDA, we generate prompts to feed into GPT-4 for bias detection. We group the pairs by books and seed categories, and then employ zero-shot prompting by feeding each grouped pair to GPT-4, accompanied by a task description. Figure 6 shows the prompt that was used to query the GPT-4 model for generating labels for each word in the Guided-LDA list. We made sure to include in the prompt not only the list of words but also the seed category to which each word belongs, enabling it to use the seed category for generating an overall analysis of the bias within the word lists. We also included formatting constraints within the prompt to facilitate our downstream data analysis process.

GPT-4's response was then compared with the human annotation baseline for similarity.

5 Results

5.1 Literature

The result of the similarity comparison between the baseline and the LLM performance on social bias identification in existing literature is visible in Figure 2, a normalized confusion matrix that lays out the overall distribution of all labeled topics. For reference, if the LLM had perfectly identified every social bias that was in the baseline, the main diagonal ((Positive, Positive), (Negative, Negative), (Neutral, Neutral)) would be the only entries in the matrix with non-zero values.

Overall, the GPT-4 response aligns with the human baseline most closely for positive descriptors (77.27% accuracy) and negative descriptors (76.19% accuracy). The LLM, suffered in performance of the neutral-labeled words, with 96.29% of the words that human annotators labeled as neutral being labeled negative by the LLM. A notable number of these mislabeled topics were color descriptors: `yellow, white, black, gray, dark, greenish, brown, red, blue`. Why exactly the LLM tends to label color descriptors as harmful social bias is a topic that requires further experimentation and future research.

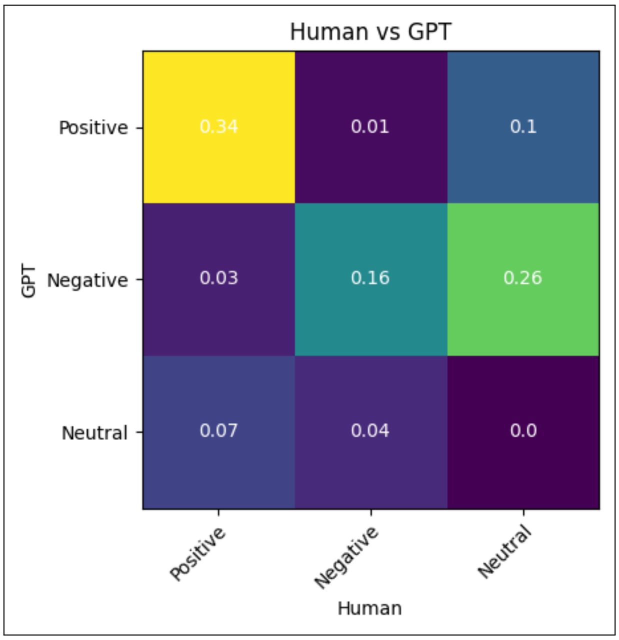

Figure 2: Confusion matrix of overall distribution of labels for GPT-4 and the human baseline on the literature dataset.

A deeper dive into a comparison in the LLM responses versus the baseline shows that GPT-4's performance varied across social identity categories. Figure 3 displays a confusion matrix, where each entry is computed using the following: $\Delta =$(*Ratio of GPT-assigned labels* - *Ratio of baseline labels*).

In other words, each value is the difference between the ratio of topic descriptors with that label given by GPT-4 and the ratio of topic descriptors with that label given by the human baseline. A value of zero indicates that they are the exact same, and that the LLM correctly labeled every topic in this category, identifying social bias with the same capability as a human, while a value far from zero indicates that LLM performance suffered in this category. Furthermore, a positive value indicates that GPT-4 labeled more words in this category, while a negative value indicates that the human annotators labeled more words in this category.

On the literature dataset, LLM performance suffered most in identification of gender bias. While the human annotators tended toward negative and positive labels for male and female descriptors, GPT-4 overclassified topics as neutral. This is probematic for two primary reasons: 1) that misalignment of the LLM response and the baseline shows weakness in reasoning capabilities and 2) preference of neutrality over a positive/negative label shows classification weakness. The LLM's performance in labelling the topics for the "Male" social identity category (its worst) is 97.29% worse than its most accurate social identity category: "Non-wealthy".

GPT-4 was able to label the remainder of the social identity categories with a relatively high accuracy, performance strong on: White, Non-white, Wealthy, and Non-Wealthy.

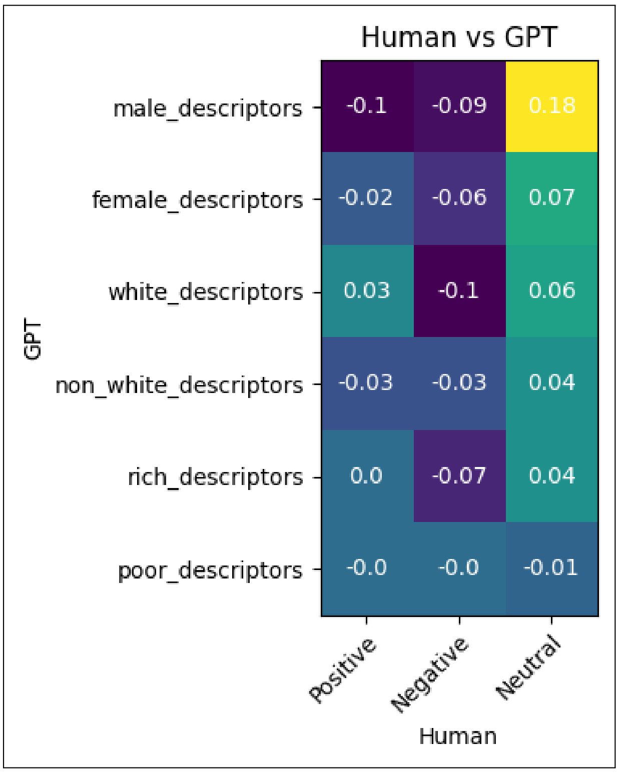

Figure 3: Confusion matrix of the difference between the LLM resposne and the human baseline across social identity categories on the literature dataset. A positive value indicates that GPT-4 labeled more words in this category, while a negative value indicates that the human annotators labeled more words in this category.

5.2 LLM-Generated Stories

GPT-4 shows a much weaker overall performance in social bias identification on the dataset of LLM-generated stories compared to its performance on the literature dataset, as seen in Figure 4. Half of the words that the human annotators labeled as negative were labeled as neutral by GPT-4, while nearly half (41.46%) of the words that human annotators labeled as positive were labeled as neutral by GPT-4, and the majority (85.37%) of the words that human annotators labeled as neutral were labeled as negative by GPT-4.

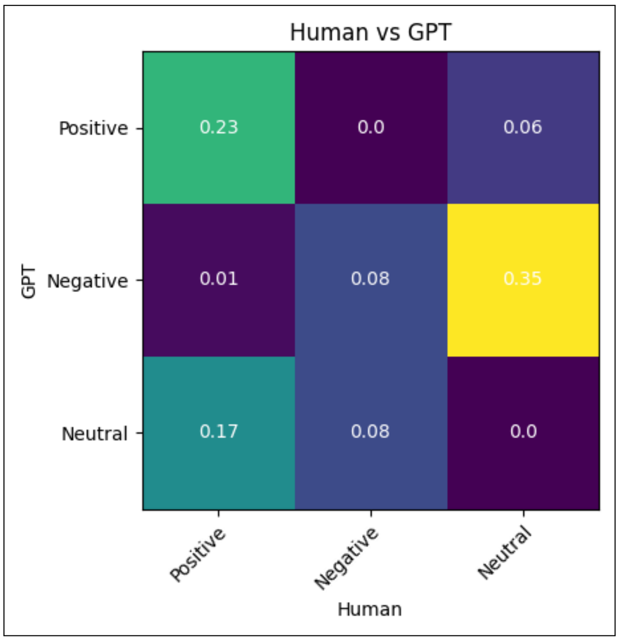

Figure 4: Confusion matrix of overall distribution of labels for GPT-4 and the human baseline on the GPT-4-generated stories dataset.

Only 56% of positive identifiers were correctly identified, 50% of negative descriptors were correctly identified, and 0% of neutral descriptors were correctly identified.

When comparing GPT-4's performance to the human baseline in each individual social identity category, as shown in Figure 5, GPT-4 performs much better in gender bias identification but suffers in every other category. With GPT-generated stories, GPT-4 is weakest when detecting racial bias, overclassifying both White and non-White descriptors as positive and underclassifying both White and non-White descriptors as neutral. Socioeconomic bias identification is also much worse on the GPT-generated story dataset, with GPT-4 once again overclassifying both wealthy and non-wealthy descriptors as positive, while underclassifying both wealthy and non-wealthy descriptors as neutral.

6 Discussion

6.1 Performance Differences

This GPT-generated dataset results are exactly opposite to that of the literature dataset, where the

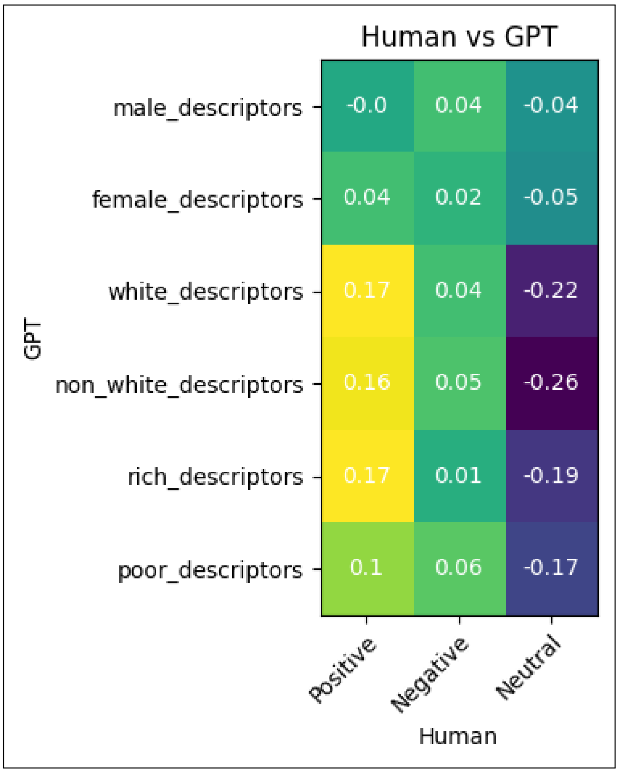

Figure 5: Confusion matrix of the difference between the LLM response and the human baseline across social identity categories on the GPT-generated stories dataset. A positive value indicates that GPT-4 labeled more words in this category, while a negative value indicates that the human annotators labeled more words in this category.

strongest performance was in racial and socioeconomic bias identification, with the weakest performance in gender bias, and the ratios showing an unwillingness to commit to a positive/negative label, favoring the neutral instead. This difference in results could be due to many reasons. For one, the topics discovered by the guided LDA process may have had a higher level of uncertainty in the GPT-generated stories compared to the literature, since the generated stories were much more simplistic in nature and shorter in length.

At its current level, social bias identification with GPT-4 is not advanced enough to achieve high accurace rates without some kind of human oversight or moderation. However one key finding within our results is that even in categories where GPT-4 demonstrated high levels of misalignment, bias identification inaccuracy was not concentrated in any single social identity category. For example, cases where descriptors for White were not labeled correctly, descriptors for non-White were also not labeled correctly. The same goes for Wealthy and non-Wealthy, as well as Male and Female. This demonstrated that while GPT-4 peforms poorly in some categories of bias identification over others, we discovered no implicit social bias in identification capabilities of one social group over another. GPT-4's performance indicates a weakness in logical reasoning capabilities rather than any nefarious perpetuation of harmful social stereotypes against marginalized groups. This is a step in the right direction for LLMs like GPT-4 eventually being able to de-bias text on their own.

6.2 LLM-Human Label Inconsistencies

When comparing the word labeles done by humans and GPT-4, we noticed frequent inconsistencies within the word labeling. Such inconsistencies include 1) same word that appeared in different LDA word list might be labeled differently by both human and GPT-4, and 2) the same word appeared in the same LDA word list receives different labels from human and GPT-4.

We hypothesize that the first type of inconsistency is likely induced by the context-dependent nature of language interpretation, since both humans and LLMs tend to infer the connotation of words based on the surrounding context, which in this case is the surrounding words within the LDA list. As the context shifts from one LDA list to another, the corresponding label for the same word shifts as well. For example, in the male-descriptor LDA words of the book 'The Secret Agent,' the word 'emotional' was labeled as neutral, whereas in the male-descriptor LDA words of the book 'This Side of Paradise,' 'emotional' is labeled as positive. Upon examining the rest of the words in the same LDA list, we found that in 'The Secret Agent,' 'emotional' appeared in conjunction with words such as 'less,' 'mad,' 'lazy,' and only three words that are perceived as positive by both humans and LLMs. In contrast, the LDA word list of 'This Side of Paradise' contains as many as 13 positively-labeled words.

For the second type of inconsistency, we hypothesize that the labeler's bias plays an important role in the discrepancy of how the same word is categorized differently. Human labelers are influenced by their own cultural background, personal experience, as well as vocabulary knowledge (For example, one of the labelers is an international student who'e first language isn't English), leading to variability in labeling. To explore this discrepancy in in more detail, we have gathered all the list of words that are labeled differently. A more detailed version of the lists can be found in Appendix A.2

From the discrepancy lists, a few observations were made based upon our subjective qualitative analysis:

1. In numerous instances, the LLM consistently categorized words related to colors — such as 'white,' 'black,' 'yellow,' 'dark,' and 'greenish' — as negative, diverging from human assessments which predominantly classified these terms as neutral. Given the diversity of colors appeared in the list, it is not feasible to attribute this pattern to biases pertained to skin color. Instead, this phenomenon may reflect broader tendencies in the LLM's training data.

2. In some cases, the LLM associated a specific set of bodily features - including 'short', 'long', 'little', 'petite', and 'old' - as negative, whereas the labelers identified them as neutral. However, it was also observed that other bodily features such as 'youthful', 'muscular', and 'firm' were labeled as positive by human, but neutral by the LLM. We hypothesize that these two contrasting cases may indicate the underlying differences in the degree of social-cultural context possesed by human and the LLM. Human labelers, influenced by societal norms and personal biases, tend to associate a word as positive or negative based on the prevailing attitudes towards these words held by the communities in their immediate social environment.

6.3 Limitations

6.3.1 Labelers' Bias

Due to the time and resource constraint of this project, we, the sole two members of the team, undertook the task of data labeling for the LDA lists of both the books/literature and the GPT-4-generated stories. As discussed in section 6, one of the notable consequences of this approach is the potential susceptibility of the human labels to the social-cultural bias of the labelers. For future work, it is recommended to recruit more data labelers, as greater sample size averages out individual labeler's bias, leading to more accurate result.

6.3.2 LDA List Quality

While guided LDA enables the extraction of words that are strongly associated with each seed topic, there appeared to be a considerable amount of word overlaps between word list of opposing categories, i.e. Male and Female, White vs. racial minorities, Wealthy and economically disadvantaged. The overlap phenomenon also occurred in different LDA lists across different books, as common words such as "good" and "grand" appears more than 10 times in different books' LDA lists. A few factors could be the contributor of this limitation:

1. **The length of the text is too short to generate non-overlapping LDA lists.** Overlap in the LDA lists is much more prevalent in the GPT-4-generated stories, which are considerably shorter than the original books and literature. This is due to the token limitation of the GPT-4 output set by OpenAI.

2. **GPT appears to have a relatively rigid writing style that prefers using a specific set of words over the others.** In our qualitative assessment of the writing style in stories generated by GPT-4, we observed a similarity of across different stories. This observation leads us to hypothesize that the word diversity in the GPT-4-generated stories may be lower compared to that in original books and literature. Such a reduction in diversity could potentially contribute to the observed overlap in word lists across various texts. However, further analysis is required to substantiate this hypothesis

For future research, it would be valuable to investigate the relationship between story length and the quality of the generated LDA lists, particularly to assess if longer stories diminish the overlap among these lists. Additionally, examining the impact of varying seed word selections on the resulting LDA lists would be insightful. Finally, exploring alternative methods besides LDA to evaluate bias in GPT-generated narratives could be another promising direction.

7 Conclusion

Narrative generation has become an active use-case for LLMs in recent years, with the popularization and commercialization of tools such as OpenAI's ChatGPT. However, LLMs have a tendency to exacerbate existing harmful social biases from the training data, and real stereotypes in fictional stories still have harmful real-life consequences for marginalized groups. While explicitly prompting for anti-stereotype content has not been widely successful, an unexplored avenue of research is testing whether LLMs can detect social biases on their own, which would allow writers to leverage them as a powerful de-biasing tool as well as providing higher quality content through the process of self-correction.

We run a series of experiments evaluating the capability of GPT-4 to identify social bias within existing literature and within GPT-4's own generated stories, using a combination of topic modeling and human annotations to establish a baseline to compare with the LLM's response. Our experiments reveal that GPT-4 excels in detecting racial and class biases within the literature corpus, though it shows less proficiency in identifying gender bias. Conversely, in evaluating its own produced narratives, GPT-4 is more adept at recognizing gender bias, but its capability to identify racial and class biases is weaker. The primary insight from our studies is that despite instances where GPT-4's alignment with human benchmarks was notably imperfect, these inaccuracies were not predominantly found in any specific social identity category.

Based on the findings, it can be concluded that GPT-4's capability to identify and analyze biases varies significantly depending on the type of content and the specific bias in question. The system shows a differential proficiency in recognizing racial, class, and gender biases, with its effectiveness fluctuating between literature corpus analysis and its own story generation. This indicates a nuanced understanding and processing ability of GPT-4, which does not uniformly excel or falter across all categories of social biases. Importantly, the inconsistency in its performance, while notable, does not appear to disproportionately affect any single social identity category. This suggests that while GPT-4's alignment with human judgement on bias detection has room for improvement, its current limitations or inaccuracies are not exclusively tied to a particular bias type. Therefore, ongoing refinement and training of the model, particularly in understanding and identifying various social biases, could enhance its reliability and accuracy across different contexts and types of content.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

R. Harald Baayen. 1996. The effects of lexical specialization on the growth curve of the vocabulary. *Comput. Linguist.*, 22(4):455–480.

Jon Chun and Katherine Elkins. 2023. explainable ai with gpt4 for story analysis and generation: A novel framework for diachronic sentiment analysis. *International Journal of Digital Humanities*, pages 1–26.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA. Association for Computing Machinery.

W Ebeling and T Pöschel. 1994. Entropy and long-range correlations in literary english. *Europhysics Letters (EPL)*, 26(4):241–246.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092.

Michael Stren Hart. 1971. *Project Gutenberg*.

Minh Hua and Rita Raley. 2020. Playing with unicorns: Ai dungeon and citizen nlp. *DHQ: Digital Humanities Quarterly*, 14(4).

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki Markus Asano. 2021. How true is gpt-2? an empirical analysis of intersectional occupational biases. *CoRR*, abs/2102.04130.

Max Kreminski, Melanie Dickinson, Michael Mateas, and Noah Wardrip-Fruin. 2020. Why are we like this?: The ai architecture of a co-creative storytelling game. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, FDG '20, New York, NY, USA. Association for Computing Machinery.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Nick Montfort and Rafael Pérez y Pérez. 2023. Computational models for understanding narrative. *Revista de Comunicação e Linguagens*, (58):97–117.

OpenAI. 2023. Gpt-4. `https://openai.com/`.

Thomas Schürmann and Peter Grassberger. 1996. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. 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)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Pittawat Taveekitworachai, Febri Abdullah, Mustafa Can Gursesli, Mury F Dewantoro, Siyuan Chen, Antonio Lanata, Andrea Guazzini, and Ruck Thawonmas. 2023. What is waiting for us at the end? inherent biases of game story endings in large language models. In *International Conference on Interactive Digital Storytelling*, pages 274–284. Springer.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

A Supplementary Material

A.1 LLM-Human Label Inconsistencies

We have gathers lists of words that were labeled differently by humans and LLM. For lists that are too long, we will demonstrate in this section only 30 words. The full version of the lists can be accessed on out project GitHub repository (list for books/literature and list for GPT-generated stories). The words are separated into different groups, and the group name, `sentiment1-sentiment2`, indicates that the word is labeled by humans as `sentiment1` and by LLM as `sentiment2`. Notice that the same word might appear in different lists, which means that the same word appeared multiple times in different books and LDA lists, and it was assigned different labels each time. The words are arrange in no specific order.

Inconsistent words found in the LDA groups of the GPT generated stories:

- `positive_negative`: earnest, distinct, eccentric, wealthy, critical, beastly, extravagant, fiery, unkempt, young, meaningful, profound, mischievous, relentless, fractured, harmless, insurmountable, sharp, vicious, weather-beaten, humble, clandestine, unyielding, common, greater, palpable, economic modest, delicate, candid

- `positive_neutral`: youthful, human, communal, wealthy, intricate, lucky, powdered, motherly, richer, natural, young, freeing, prettier, full, otherworldly, determined, successful, pragmatic, bucolic, audacious, unwavering, unperturbed, wide, real, ever-evolving, necessary, moonlight-stricken, good, florid, exact

- `negative_positive`: stormy, fertile, capricious, lavish, unforgiving, vulnerable, futuristic, enraptured, harsh, eager, fierce, cherished, commanding, ruthless, solitary, memorable, formidable, tamed, astute, feisty, foolish, less, hard, grand, eldritch, austere, overwhelming, ironic, tough, unseemly

- `negative_neutral`: fleeting, impulsive, anxious, imminent, gentrified, ruffled, darkest, hollow, dark, fierce, undeserved, fox-like, tattered, witchy, astute, unsought, once-prosperous, unable, ever-so-forgetful, unformed, weathered, rough-hewn, dust-clouded, creased, dreadful, heavy, diminutive, lanky, muddy, illicit

- `neutral_positive`: ever-stunning, glowing, fresh, experimental, life-changing, intelligent, everyday, full, content, ultimate, uninhibited, retired, akin, hardscrabble, liveliness, important, colossal, long, greenest, working-class, single, humane, sigh, simpler, empathic, fine, ardent, cushioned, lifelong, ready

- `neutral_negative`: much, uninitiated, gothic, razor-sharp, beastly, static, terminal, faded, rudimentary, castle-like, petite, seasonal, impersonal, old, coldest, shred, flat, economic, frigid, routine, underground, hardscrabble, stooped, chaotic, edgy, strange, rundown, half-empty, grizzled, fractured

Inconsistent words found in the LDA groups of the original books and literature:

- `positive_negative`: visible, capricious, low, unkind, incredible, legal, native, least, former, fierce, early, instant, tolerable, sentimental, senior, porochial, human, steady, extraordinary, flat, willful, economic, serious, formidable, electronic, sharp, strange, aware, late, enough

- `positive_neutral`: delicate, deeper, fierce, psychological, full, content, human, guardian, big, front, able, flat, ascetic, direct, familiar, clean, subjective, decided, untouched, sociable, technical, excited, worth, delicious, interested, anti-slavery, moral, following, venerable, fores

- `negative_positive`: small, anxious, dark, fierce, difficult, grey, wild, late, last, sorry, bad, short, long, heavy, simple, hungry, empty, little, grim, unmarried, obvious, tiny, intangible, blue, afraid, tired, aware, frantic, silly, wary

- `negative_neutral`: urgent, small, anxious, strangest, nothing, difficult, wild, lower, warn, mortal, late, unbroken, last, sorry, frightened, nervous, bad, short, long, simple, slightest, hungry, scant, empty, little, spectacled, sir, sombre, tiny, whole

- `neutral_positive`: attached, silent, small, various, light, separate, red, recent, slight, past, huge, unto, detective, instant, top, personal, close, apparent, usual, straight, electronic, nearest, sure, longer, enough, future, present, tall, black, deep

- neutral_negative: much, silent, disagreeable, various, slight, sad, dark, sech, past, huge, double, old, flat, disappointed, second, grey, usual, tired, old-fashioned, different, foolish, less, black, brown, trap-door, nervous, thick, several, sudden, private

```
The following are six lists of words, where each list represents the most common words within a book that are associated with a category.
The categories include: male, female, white, non-white, rich, and poor. Here are the six lists:
Male: <male LDA list>.
Female: <male LDA list>.
White: <white LDA list>.
Non-white: <Non-white LDA list>.
Rich: <rich LDA list>.
Poor: <poor LDA list>.
Your task is to evaluate each word list for each category, and the way to evaluate is as follows:
1. For each word of each list, you should first identify if such word has positive connotation, negative connotation, or neutral connotation if you feel like the word conveys neither positive nor negative connotation. Put all the words with positive connotations in one list, all the words with negative connotations in another list, and neutral words in another list. Make sure all the words have been categorized!
2. After evaluating all six-word lists, provide an overall analysis in a few sentences of what sort of social biases this work contains, given the lists of words and the connotations of each word.
Your output should be in the following exact format:
Male Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
Female Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
White Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
Non-white Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
Rich Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
Poor Descriptors:
positive: <word1, word2, ...>
negative: <word1, word2, ...>
neutral: <word1, word2, ...>
Overall Analysis: <your analysis>
```

Figure 6: For each book, the above prompt was given to GPT-4 to assign sentiment to each LDA word within the 6 seeded categories.