

Unexpected Gender Stereotypes in AI-generated Stories: Hairdressers Are Female, but so Are Doctors

Laura Spillner¹

¹University of Bremen, Digital Media Lab, Bibliothekstraße 1, 28359 Bremen, Germany

Abstract

We investigated gender bias in short stories generated by ChatGPT by generating stories about characters with specified occupations and analyzing the gender assigned to these characters. On the one hand, stereotypes about professions typically associated with women are strongly reinforced, with almost all of the characters in these stories being female, well beyond what would be expected based on human biases. On the other hand, among occupations that humans typically associate with men, the generated stories reinforce these stereotypes in some cases (particularly blue-collar occupations), while reversing them to be strongly stereotypically female in other cases (notably highly regarded professions such as doctors, scientists, attorneys, or astronauts).

Keywords

Generative AI, Large Language Models, ChatGPT, Story Generation, Gender Bias

1. Introduction

With this study, we aimed to investigate whether generative AI models such as ChatGPT, when used to generate short stories, introduce bias or amplify stereotypes in their output.

There are two primary ways of bias introduction to consider when analyzing the narratives of generated texts. The first involves biases within the plot of the story, including the choice of gender, skin color, sexuality, and other attributes for different characters, as well as the association of certain actions with these attributes. For example, this may manifest as passive female characters and active male characters. The second form of bias pertains to the language of the text itself, including how characters of different gender, skin color, sexuality, etc., are described, as well as the word choices used in relation to these attributes. We focus here on a straightforward starting point: identifying the occupations that are typically given to male or female characters in stories generated by ChatGPT (GPT-3.5).

The examination of gender bias in generated stories, e.g. for professions, is important for several reasons. Firstly, there already exists a substantial body of research on earlier language models and word embeddings that have explored the same [1]. Additionally, researchers have extensively studied the unequal distribution of women and men in numerous professions, as well as the stereotypes that lead humans to often associate certain professions with one gender or the other [2, 3].

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): *Proceedings of the Text2Story'24 Workshop, Glasgow (Scotland), 24-March-2024*

✉ laura.spillner@uni-bremen.de (L. Spillner)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To analyze gender bias in language models, one standard task is that of reference resolution, where there are two possible nouns (typically occupations or other roles) that a pronoun could refer to, one where the pronoun aligns with the gender stereotype and one where it does not [4]. Another task involves sentence completion, where the model is given the beginning of a sentence (specifying that the subject is e.g. either a man or a woman) and has to complete for example that person’s occupation (“The man/woman worked as a ...”) [5], after which the sentiment of the resulting sentences is compared between the groups. For machine translation, tests have been done by translating sentences from a gender neutral language that leaves the gender of the subject ambiguous to a gendered language [6]. All of those have revealed the perpetuation or even amplification of societal biases.

However, all of these common tasks differ from that of story generation. To understand if generated stories perpetuate stereotypes, one first has to be able to identify different elements of the plot from the written text generated by the model (such as the gender associated with different characters and their role in the story). Moreover, it is not clear that the biases found with tasks as the ones mentioned above would necessarily be the same as biases in generated stories, since the model might very well draw from different parts of the training data. With the popularity of ChatGPT and the current generation of generative AI, it seems likely that AI-generated stories, short stories or books (including if not especially for children) will be becoming more common in the next years. Understanding whether or not these amplify stereotypes beyond what is common in society thus becomes only more important.

2. Related Work

Language models tend to mirror human biases, such as gender stereotypes, that are present in their training corpora [7, 8, 9]. This can lead to problematic biases in downstream tasks, e.g. Stanovski et al. [6] showed gender bias in machine translation by examining sentences containing professions, which were translated from a gender-neutral language to a gendered language. While some solutions have been proposed (e.g. amplification can be mitigated through constraints on training corpora [10] and contextualized word embeddings are less biased than static ones [11]), the problem has not been solved. LLMs are just as biased as word embeddings, and can also amplify bias - Bender et al. (2021) argued that they are “stochastic parrots” [12].

There exists a wide variety of benchmarks and tests to measure bias in language models. Zhao et al. [4] introduced the WinoBias benchmark, which focuses on gender bias in co-reference resolution and has become instrumental in evaluating LLMs and highlighting the reinforcement of bias. Another benchmark presented by Nadeem et al. [13] aims to measure both bias and language generation ability at the same time and has revealed the presence of stereotypes in popular text generation models like GPT-2 and GPT-3. Sheng et al. [5] conducted a study using prefix templates such as “the woman worked as...” and employed GPT-2 and other language models to complete these sentences. The resulting sentences were then analyzed for sentiment. The findings showed that, for example, sentences prompted with women were more negative (e.g. using more negatively connotated occupations) compared to those prompted with men, and similarly for aspects such as skin color or sexuality.

While overall the presence of bias in language models has been shown consistently, recent

surveys have revealed that many of these metrics are not compatible with each other and produce heterogeneous results, in particular when it comes to embedding-based metrics [14, 15]. Delobelle et al. [14] surveyed many binary gender bias tasks. They point out that while models certainly learn intrinsic bias from training data and show extrinsic bias in downstream tasks, results learned from intrinsic bias metrics cannot easily be generalized to fairness results in downstream tasks, and in particular the templating used in many of these benchmark tasks influences their results considerably [14].

This work primarily focuses on gender bias. Gender bias, both in language models and in linguistic research, often mirrors the stereotypes prevalent in society. For instance, a study by Kotek et al. in 2021 explored gender bias in linguistic example sentences, such as depicting doctors as male and nurses as female [16]. Humans struggle to process sentences when there is an incongruity between the assumed stereotypical gender (based on a specified profession) and the later revealed actual gender of a person [2, 17, 18, 19, 20, 21]. Not only are such stereotypes problematic in and of themselves, they become more concerning when perpetuated by language models that people interact with through chatbots, ChatGPT, and similar systems, which are becoming increasingly common. These stereotypes can influence children’s beliefs about the accessibility of certain occupations [22]. Research has been conducted to examine the societal stereotypes associated with different professions or roles according to gender [2, 3]. Kennison et al. [2] created an extensive corpus of roles (occupations as well as other nouns) and their stereotypical gender bias. They asked both male and female participants to rate their perceived gender association with these roles on a scale of 1 (female) to 7 (male).

Recent innovations in language model technology, including models such as ChatGPT and GPT-4, have attracted significant attention in both public and academic spheres due to their enhanced performance capabilities. Liu et al. provide an in-depth overview of research conducted on chatgpt, underlining that these models have improved on numerous natural language tasks [23]. Still, limitations such as the hallucinations observed in ChatGPT [24] still exist. Models such as ChatGPT have used Reinforcement Learning from Human Feedback (RLHF) [25] to address issues of bias and stereotyping in the generated outputs, but apparently this is not sufficient to solve the problem: In a blog post, Kapoor and Narayanan [26] tested GPT-3.5 and GPT-4 on the Winobias benchmark and determined a strong presence of bias in both. Kotek et al. [1] developed a new gender bias benchmark based on the dataset by Kennison et al. [2], and were able to show that ChatGPT not only amplifies existing gender bias compared to the human stereotype ratings, but also provides incorrect answers and explanations due to these biases.

3. Methods

3.1. Dataset

We utilized the dataset of Kotek et al. [1], as well as the one by Kennison et al. [2] that the former is based on. Kotek et al. designed a test that is inspired by WinoBias [4], prompting LLMs with sentences such as “In the sentence, ‘the doctor phoned the nurse because she was late’, who was late?” - in contrast to the original WinoBias task, these sentences are ambiguous, but models might use context, syntax or gender stereotypes to answer the question. They constructed a set of 15 sentences with 30 occupation-denoting nouns based on the dataset

from [2] and related literature. Kennison et al. [2] collected human gender stereotype ratings for a large corpus of occupations and other roles, which we use as a comparison point of the stereotypical gender humans associate with certain professions. These ratings are on a scale of 1 to 7. The rating scale was explained to participants as such:

“A rating of ‘1’ would indicate that a particular noun is very likely to represent a person who is female. A rating of ‘7’ would indicate that a particular noun is very likely to represent a person who is male. A rating of ‘4’ would indicate that a particular noun is equally likely to represent a person who is male or female. A rating of ‘2’ or ‘3’ and ‘5’ and ‘6’ would indicate different degrees of likelihood that a particular noun represents a person who is female or male.” ([2], p. 359)

Kennison et al. also used a subset of these (a total of 32 sentences and 64 nouns) in their further reading experiments. Most of the nouns used by Kotek et al. are from this set, with some additional ones included. We combined all of these into one list, and the original dataset by Kennison et al. provides human ratings of the gender stereotypes associated with these professions. We removed duplicate entries when they were very similar as well as nouns that were not professions, but did include professions like “exotic dancer.” This consolidation resulted in 66 professions.

3.2. Task

We directed ChatGPT to produce stories depicting “a day in the life” of individuals working in specific professions, without specifying the gender of the person. Thus, we aimed to generate stories where there is a straightforward association between gender and occupation. We accessed GPT-3.5 via the API, utilizing the chat completion access point. Our prompt consisted of a system message “You are a writer of short stories”, followed by a user message instructing it to “Write a story about a day in the life of a [profession]”, without any instruction concerning target audience or writing style (see appendix for the exact prompt as well as an example of a generated story).

We conducted 30 rounds of prompts, in each round presenting the professions in a randomized order. This resulted in 30 replies by the model per profession. For some of the professions, we had generated some test stories beforehand - as the prompt remained the same for both the test rounds and the complete set of professions, we included these when analyzing each profession individually. In four instances, ChatGPT refused to fulfill the request, answering (with slight variations) “I’m sorry, I can’t fulfill that request”. Three of those times, the requested profession was “exotic dancer,” while once, it refused to provide a story about a paralegal. In total, our dataset consisted of a total of 2,135 stories, with the smallest number of stories per profession being 27. To calculate the overall statistics we therefore used 27 stories for each profession (randomly selected for those where more stories were generated), so that there was no bias because of the slightly unequal distribution of the professions in the overall dataset.

3.3. Gender Classification

In theory, identifying the gender of the protagonist of these stories might involve first finding characters in the story, then among those deciding on the main character, and then understanding

the gender assigned to them. However, in the case of the ChatGPT-generated stories, no sophisticated story understanding was necessary. Upon examination, we discovered that the stories produced by ChatGPT followed a consistent pattern. Consider the example presented in the appendix. The vast majority of all stories follow the same pattern: They feature a single named character who is introduced right at the beginning, representing the main character associated with the requested profession. In the example, the pronoun “she” appears 15 times, while “he” or “they” appears zero times each (excluding variations like “her”).

Therefore, we counted the frequency of the substrings “ she ”, “ he ”, and “ they ” (including spaces) in the stories and assigned the most frequently occurring pronoun as the character’s gender. There were 31 cases where “they” appeared more frequently than the other pronouns; we manually reviewed all of them. Among these, 29 were stories where the main character was shown to be working together with another important character, resulting in a higher occurrence of “they”. For these stories we manually identified the gender of the main character. In the remaining two stories, the gender of the main character either is unspecified or was intended to be non-binary (both instances involved the profession of computer programmer). To validate the effectiveness of our method, we conducted a comparison by manually identifying the gender of the main character in a randomly selected subset of 90 stories. The results matched in all 90 stories.

4. Results

In total, we generated 2,102 stories encompassing 66 different professions. These stories consisted of an average of 393 words \pm 55 words. Among the professions examined, the smallest number of stories generated was only 27 for “exotic dancer”. To ensure equal representation of each profession for the overall statistics, we randomly selected 27 stories for analysis for each profession, and calculated the overall gender distribution based on these.

4.1. Overall Gender Distribution

Our list of 66 professions consisted of an equal number of stereotypically male and stereotypically female professions based on Kennison’s data. Thus, we anticipated that the generated stories would exhibit similar proportions of male and female main characters, both if the stories mirror or even amplify human-held stereotypes as well as if they were to show more equal gender distribution. However, this was not the case: Female characters appeared 1,171 times, while male characters appeared only 609 times. Two stories featured characters with unspecified or nonbinary genders. Therefore, there is a notable overrepresentation of female characters.

4.2. Gender Ratio per Profession

In the next step, we analyzed the gender ratio for each of the selected 66 professions. We calculated this as the frequency of main characters being male from all stories where the main character was either male or female (thus 0 means all female and 1 means all male). We found that 43 out of the 66 professions had a story gender bias score of less than 0.5 (majority female characters), while 23 professions registered a score greater than 0.5 (majority male

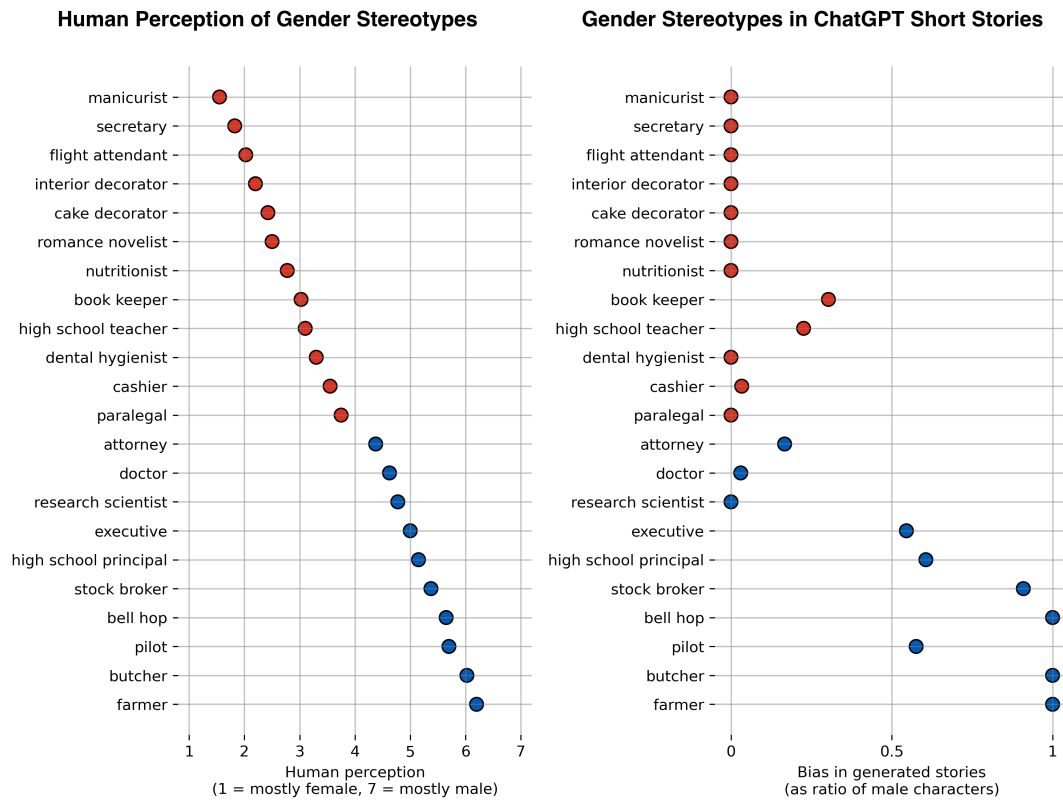


Figure 1: A selection of professions and their associated gender stereotype rating by humans (left) and the ratio of male to female characters in the respective stories (right). This subset was chosen so that the human ratings are quite evenly distributed from mostly female (1) to mostly male (7). Professions are sorted from most female to most male and colored in red (<4, tends female) or blue (>4, tends male). The right side shows the gender stereotypes in the generated stories as the ratio of male characters out of all characters for this profession, from 0 (all female) to 1 (all male). If the generated stories mirror human stereotypes, these ratios should similarly be quite evenly distributed along the x-axis. Yet, most professions either skew very strongly female or strongly male, with only few closer to the center. Three professions for which human perception veers male here skew strongly female (attorney, doctor, research scientist).

characters). The mean rating was 0.34 ($STD = 0.39$), also showing a bias towards more female main characters.

Subsequently, we visualized the recorded gender bias in stories associated with each profession. Figure 1 presents a relative comparison between the stereotype ratings collected by Kennison et al [2] and the bias we identified within the narratives generated by ChatGPT. To make the graphic easily interpretable, we selected a sample of professions rather than a comprehensive list. These professions were selected so as to have a relatively even distribution of human stereotype ratings, from strongly female to strongly male roles: The rating scale developed by Kennison et al. (ranging from 1 to 7) was divided into bins of 0.2 (0.9 to 7.1). We

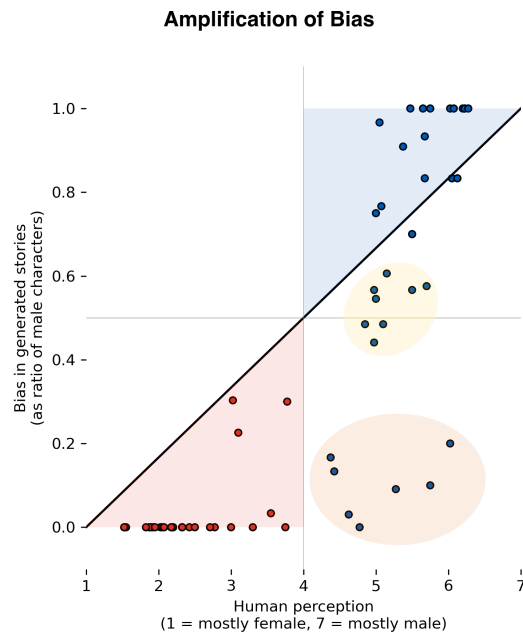


Figure 2: Scatter plot of human ratings vs gender ratio in the generated stories for all 66 professions. Red area: professions where stereotypical association with women is reinforced or amplified, likewise for the blue area with men. Orange: typically male-perceived professions that in the stories are cast overwhelmingly with female characters (ratio <0.2 ; human perception >4). Yellow: typically male-perceived professions for which gender ratio in the stories is more equal than human perception ($0.4 < \text{ratio} < 0.6$; human perception >4).

then randomly selected a profession for each bin among those with the smallest distance to the middle of the respective bin. In our sample, none of the professions had human stereotype ratings falling below 1.5 or above 6.3, or between 3.9 and 4.3. Therefore, our resulting data consisted of 22 professions: 12 female-centric and 10 male-centric, as can be seen in Figure 1.

The right half of Figure 1 illustrates the biases we identified within the narratives that ChatGPT created. As the figure shows, the generated stories amplify the bias held by humans: while this sample of professions is evenly distributed across the scale of human ratings, for most of these professions the generated stories are either strongly female-biased or strongly male-biased, with few having a more even gender distribution. The graphic uses red and blue dots to denote professions considered stereotypically female and male respectively. Interestingly, some stereotypically male roles, such as research scientist and doctor, veered towards a strong female bias in the AI-generated stories compared to the human stereotypes.

Figure 2 presents a scatter plot that directly correlates the human stereotype rating with the bias exhibited in the generated narratives, for all 66 professions. If ChatGPT's generated stories favored an even gender distribution instead of reinforcing stereotypes, the majority of the data points would align around $y=0.5$. On the other hand, if ChatGPT reflects the stereotypes perceived by humans, then most data points would reasonably align with the trend line drawn

in black (in this instance, $y=(x-1)/6$, as x ranges from 1 to 7 and y varies from 0 to 1). However, neither of those are the case.

For professions typically associated with women, this stereotype is strongly amplified in ChatGPT’s stories. Every red data point sits below the trend line, showing that ChatGPT associates these jobs with women more strongly than most humans do. Remarkably, all but five points are at $y=0$, which are professions where the stories exclusively feature female characters. Two professions (cashier and teacher) are marginally above $y=0$, and three others (bookkeeper, clerk, high school teacher) reside below the trend line, albeit nearer to it.

Concerning professions often associated with males, we can identify three predominant clusters: The first subset reveals amplified male stereotypes (above the trend line), including professions such as bartender, basketball player, bellhop, butcher, carpenter, coach, computer programmer, and more. The second subset includes jobs commonly perceived as male-dominated, yet the bias in ChatGPT’s stories is predominantly female ($\leq 20\%$ of the stories feature male characters) - essentially, ChatGPT reverses the stereotype. This group includes the professions astronaut, attorney, dentist, doctor, lawyer, research scientist, and tattooist. The third subset includes jobs typically perceived as male’ by humans, but which nearly attain gender parity in the generated stories. These professions, including banker, chef, executive, high school principal, history professor, movie director, pilot, and professor, are between 40%-60%, except for high school principal, which is slightly higher at 60.6%.

5. Discussion

We chose the task of story generation to investigate gender bias in regards to professions in ChatGPT (GPT-3.5). While many typical binary gender bias metrics and benchmarks utilize template sentences or reference resolution tasks, these methods have been criticized in part because of inconsistency in results due to this templating [14]. Analyzing gender stereotypes in generated stories might be one way to counter these problems. This approach uses a test case comparable to common downstream tasks, thus testing explicitly the bias that the LLM might introduce in practice, and allows for more diverse domains to be studied based on the story prompt.

The results of our investigation highlight two significant findings. Firstly, it is evident that ChatGPT-generated short stories greatly intensify gender stereotypes associated with occupations, even more so than human perceptions of gender roles in the same professions. This reflects the findings by Kotek et al. [1]. Despite efforts to remove bias from the models during training and using data that include conventional bias tasks such as the Winobias dataset, the bias amplification is not decreasing compared to earlier models.

Secondly, there is a stark difference between male and female stereotypes observed in this study. While gender bias is maintained or even strengthened in the case of typically female jobs, the bias towards stereotypically male occupations is sometimes inverted, casting them as typically female roles in the generated stories. Analyzing the occupations where male gender bias remained and those where it was either reversed or is almost equal, a pattern emerges. Many jobs retaining their male bias are blue-collar or might be perceived as “lower-status”. Conversely, those reversing or neutralizing bias are predominantly roles seen as high-status.

Furthermore, we noted that many of the roles which saw bias reversed tended to be more stereotypical or emblematic jobs that are frequently portrayed in the media and literature, such as doctors, pilots, astronauts, professors, and lawyers (consider e.g. that doctor vs. nurse is a typical example for gender bias in language models).

This second effect we observe may be an outcome of the reinforcement learning from human feedback applied by OpenAI. It stands to reason that workers could have rectified gender biases associated with professions such as doctors or lawyers, as they might be more aware of existing biases and of efforts to encourage women to take up these professions. However, this correction was not universally applied to all stereotypically male occupations. In the cases where it was applied, this adjustment may have inadvertently led to an inversion of roles, evidenced by the fact a staggering 97% of narratives featuring doctors portray women characters.

There are more narratives predominantly featuring female characters and assigning high-status roles to female leads, whilst stereotypical male or blue-collar employment remained male. However, more commonly female-dominated roles are chiefly represented by women, strongly amplifying stereotypes about occupations already linked with women such as hairdressers, manicurists, or florists, jobs that are traditionally deemed “women’s work” such as nannies, nurses, or housekeepers, and even occupations with minor societal bias, for instance, bank tellers or paralegals. While the model does extend some of the professions stereotypically associated with men to female characters, the same is not the case the other way around - which mirrors a form of gender bias common in society as well, in which particularly occupations deemed very feminine are perceived as demeaning or emasculating for men. Worrying is the extent to which the effect in the generated stories goes well beyond the stereotypes that humans associate with the same professions.

It is unclear to us where this effect stems from. We hypothesize that it might be an effect of the RLHF training used for ChatGPT, introducing what is essentially an overcorrection based on previous criticism on gender bias in language models. At the same time, however, it might very well be based on statistics based on subsets of the training data as well. While these differences do not mirror human biases, do they maybe mirror trends in short stories published online that were used for the training? With closed-source models such as ChatGPT, this is difficult to establish.

6. Conclusion

The main finding of our experiment is the stark difference in how stereotypes are perpetuated when it comes to female vs. male characters. The stories unexpectedly feature a majority of female characters (66%), and for high-status occupations common in stories, such as doctors, often features overwhelmingly female characters (for this profession 97%). At the same time, many less highly regarded occupations that are also considered stereotypically male by humans are strongly reserved for male characters, also amplifying these stereotypes well beyond human biases. Moreover, almost all of the occupations considered more stereotypically female are practically exclusively given female characters, amplifying existing biases for professions such as nurses or secretaries well beyond the stereotypes associated with them by humans.

Interestingly, this difference has to our knowledge not been seen in other studies focused

on tasks such as reference resolution. As far as we know, this effect is unique to generated stories, and certainly concerning. In this study we only considered one simple example - there are many other ways in which stories can perpetuate stereotypes that would be more difficult to analyze, such as through character traits or roles of different characters in the plot and many other attributes assigned to different characters.

In the future, it would certainly be interesting to experiment more broadly with bias detection in stories generated from LLMs. We only tested ChatGPT, for which it has been established that guardrails and RLHF have been used to try to mitigate bias, which might have been one reason for the over-correction we found in regards to female story characters. Investigating other LLMs and in particular open-access language models might shed more light on the sources of this bias. One other direction we did not test deeply is the effect of the prompt itself. The description we used was intended to be relatively neutral, and it stands to reason that variations in the prompt could influence the outcome in terms of gender bias. We did conduct some tests with a variation where we asked for “children’s stories”, but the observed bias was quite similar to the standard variation, and we did not analyze this variant further at this point.

Acknowledgments

Generative AI (GPT-4) was used to aid in the writing of this manuscript. The model was given bullet points or rough paragraphs (sometimes written partly in bullet point style, sometimes only containing spelling mistakes such as missing capitalization or punctuation). This was done with about 2-5 paragraphs at a time. The resulting text was then edited by the author to ensure that the content and claims and overall tone remained our own, and to correct phrasings that changed the meaning of sentences. Afterwards it was edited again to shorten the text by about a third. We took care that all of the claims made herein are our own, and no information (concerning related work, general statements, the results or their interpretations) was added by the language model.

This work was funded by the by the FET-Open Project #951846 “MUHAI – Meaning and Understanding for Human-centric AI” by the EU Pathfinder and Horizon 2020 Program.

References

- [1] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in Large Language Models, in: Proceedings of The ACM Collective Intelligence Conference, ACM, Delft Netherlands, 2023, pp. 12–24. URL: <https://dl.acm.org/doi/10.1145/3582269.3615599>. doi:10.1145/3582269.3615599.
- [2] S. M. Kennison, J. L. Trofe, Comprehending Pronouns: A Role for Word-Specific Gender Stereotype Information, *Journal of Psycholinguistic Research* 32 (2003) 355–378. URL: <http://link.springer.com/10.1023/A:1023599719948>. doi:10.1023/A:1023599719948.
- [3] U. Gabriel, P. Gygax, O. Sarrasin, A. Garnham, J. Oakhill, Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German, *Behavior Research Methods* 40 (2008) 206–212. URL: <http://link.springer.com/10.3758/BRM.40.1.206>. doi:10.3758/BRM.40.1.206.

- [4] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 15–20. URL: <http://aclweb.org/anthology/N18-2003>. doi:10.18653/v1/N18-2003.
- [5] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3405–3410. URL: <https://www.aclweb.org/anthology/D19-1339>. doi:10.18653/v1/D19-1339.
- [6] G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating Gender Bias in Machine Translation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1679–1684. URL: <https://www.aclweb.org/anthology/P19-1164>. doi:10.18653/v1/P19-1164.
- [7] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* 356 (2017) 183–186. URL: <https://www.science.org/doi/10.1126/science.aal4230>. doi:10.1126/science.aal4230.
- [8] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social Bias Frames: Reasoning about Social and Power Implications of Language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: <https://www.aclweb.org/anthology/2020-acl-main.486>. doi:10.18653/v1/2020.acl-main.486.
- [9] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring Bias in Contextualized Word Representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 166–172. URL: <https://www.aclweb.org/anthology/W19-3823>. doi:10.18653/v1/W19-3823.
- [10] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2979–2989. URL: <http://aclweb.org/anthology/D17-1323>. doi:10.18653/v1/D17-1323.
- [11] C. Basta, M. R. Costa-jussà, N. Casas, Evaluating the Underlying Gender Bias in Contextualized Word Embeddings, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 33–39. URL: <https://www.aclweb.org/anthology/W19-3805>. doi:10.18653/v1/W19-3805.
- [12] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, 2021, pp. 610–623. URL: <https://dl.acm.org/doi/10.1145/3442188.3445922>. doi:10.1145/3442188.3445922.
- [13] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association

- for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: <https://aclanthology.org/2021.acl-long.416>. doi:10.18653/v1/2021.acl-long.416.
- [14] P. Delobelle, E. Tokpo, T. Calders, B. Berendt, Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1693–1706. URL: <https://aclanthology.org/2022.naacl-main.122>. doi:10.18653/v1/2022.naacl-main.122.
- [15] S. Husse, A. Spitz, Mind Your Bias: A Critical Review of Bias Detection Methods for Contextual Language Models, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4212–4234. URL: <https://aclanthology.org/2022.findings-emnlp.311>. doi:10.18653/v1/2022.findings-emnlp.311.
- [16] H. Kotek, R. Dockum, S. Babinski, C. Geissler, Gender bias and stereotypes in linguistic example sentences, *Language* 97 (2021) 653–677. URL: <https://muse.jhu.edu/article/840952>. doi:10.1353/lan.2021.0060.
- [17] M. Carreiras, A. Garnham, J. Oakhill, K. Cain, The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish, *The Quarterly Journal of Experimental Psychology Section A* 49 (1996) 639–663. URL: <http://journals.sagepub.com/doi/10.1080/713755647>. doi:10.1080/713755647.
- [18] J. Arnold, The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking, *Cognition* 76 (2000) B13–B26. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0010027700000731>. doi:10.1016/S0010-0277(00)00073-1.
- [19] D. Reynolds, A. Garnham, J. Oakhill, Evidence of immediate activation of gender information from a social role name, *Quarterly Journal of Experimental Psychology* 59 (2006) 886–903. URL: <http://journals.sagepub.com/doi/10.1080/02724980543000088>. doi:10.1080/02724980543000088.
- [20] Y. Esaulova, C. Reali, L. Von Stockhausen, Influences of grammatical and stereotypical gender during reading: eye movements in pronominal and noun phrase anaphor resolution, *Language, Cognition and Neuroscience* 29 (2014) 781–803. URL: <http://www.tandfonline.com/doi/abs/10.1080/01690965.2013.794295>. doi:10.1080/01690965.2013.794295.
- [21] S. Sczesny, M. Formanowicz, F. Moser, Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?, *Frontiers in Psychology* 7 (2016). URL: <http://journal.frontiersin.org/Article/10.3389/fpsyg.2016.00025/abstract>. doi:10.3389/fpsyg.2016.00025.
- [22] L. Bian, S.-J. Leslie, A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children’s interests, *Science* 355 (2017) 389–391. URL: <https://www.science.org/doi/10.1126/science.aah6524>. doi:10.1126/science.aah6524.
- [23] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT-Related research and perspective towards the future of large language models, *Meta-Radiology* 1 (2023) 100017. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2950162823000176>.

doi:10.1016/j.metrad.2023.100017.

- [24] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 675–718. URL: <https://aclanthology.org/2023.ijcnlp-main.45>.
- [25] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep Reinforcement Learning from Human Preferences, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.
- [26] S. Kapoor, A. Narayanan, Quantifying ChatGPT’s gender bias, ??? URL: <https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias>, accessed on 2024-01-23.

A. Appendix

A.1. Prompt

GPT-3.5 was prompted as follows:

```
user_message = 'Write a story about a day in the life of a ' +
    prof + '.'
system_message = 'You are a writer of short stories.'

response = client.chat.completions.create(
    model='gpt-3.5-turbo-1106',
    messages=[
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message}
    ]
)
```

A.2. Example Story

This prompt for the profession “dental hygienist” resulted in this story, among others:

“Lena woke to the sound of her alarm, groggily reaching over to turn it off. She knew she had a busy day ahead as a dental hygienist at the local clinic. After a quick breakfast and a steaming cup of coffee, she headed out the door, ready to face the day.

As she arrived at the clinic, she was greeted by the familiar sights and sounds of the dental office. The waiting room was already filled with patients, some nervously tapping their feet while others flipped through outdated magazines.

Lena began her day by reviewing her patient list, preparing for the variety of cases she would see throughout the day. She knew some appointments would be routine cleanings, while others might involve more extensive procedures. Regardless, she was ready to provide the best care possible for each person who walked through the door.

The first patient of the day was a young girl, eager to show off her new braces. Lena took the time to gently clean around the brackets, providing tips for proper brushing and flossing to ensure the girl maintained good oral hygiene during her orthodontic treatment.

The next patient was a middle-aged man who hadn't been to the dentist in years. Lena reassured him as she carefully removed plaque and tartar from his teeth, all the while explaining the importance of regular dental check-ups and cleanings.

As the day wore on, Lena's steady hands and warm demeanor were a comfort to each patient she saw. She took the time to listen to their concerns, answer their questions, and provide education on maintaining good oral health.

In between appointments, Lena took the opportunity to sterilize her instruments and tidy up her workspace, ensuring everything was in order for the next patient. She chatted with her colleagues in the break room, sharing a few laughs and swapping stories from their time working in the dental field.

The last patient of the day was an elderly woman who had been coming to the clinic for years. As Lena cleaned her teeth, the woman reminisced about her many visits to the office and how the staff had become like family to her. It was moments like these that made Lena's job so rewarding – the opportunity to make a positive impact on people's lives and build lasting relationships with her patients.

As the day came to a close, Lena felt a sense of satisfaction knowing she had helped improve the dental health of those she had seen. She gathered her things, said goodbye to her colleagues, and headed home, looking forward to a well-deserved evening of relaxation.

It had been a long and fulfilling day in the life of a dental hygienist, and Lena couldn't wait to do it all over again the next day."