# Biases in GPT-3.5 Turbo model: a case study regarding gender and language

**Fernanda Malheiros Assi**[1], **Helena de Medeiros Caseli**[1]

[1]Computing Department – Federal University of São Carlos (UFSCar) – LALIC
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brazil

fernanda.malheiros@estudante.ufscar.br, helenacaseli@ufscar.br

***Abstract.*** *Interactions with Generative Language Models like OpenAI's GPT-3.5 Turbo are increasingly common in everyday life, making it essential to examine their potential biases. This study assesses biases in the GPT-3.5 Turbo model using the regard metric, which evaluates the level of respect or esteem expressed towards different demographic groups. Specifically, we investigate how the model perceives regard towards different genders (male, female, and neutral) in both English and Portuguese. To achieve this, we isolated three variables (gender, language, and moderation filters) and analyzed their individual impacts on the model's outputs. Our results indicate a slight positive bias towards feminine over masculine and neutral genders, a more favorable bias towards English compared to Portuguese, and consistently more negative outputs when we attempted to reduce the moderation filters.*

## 1. Introduction

In recent years, interactions with Generative Language Models (GLM) have become a growing part of everyday life. Studies show that people are engaging with models like OpenAI's GPT [Radford et al. 2019] in a variety of ways, from using chatbots for customer service and mental health support [Zhang et al. 2023, Das et al. 2022, Wang et al. 2023], to experiencing enhanced e-commerce through improved product descriptions, attribute generation, and customer engagement [Zhou et al. 2023, Roy et al. 2021, Liu et al. 2023]. On social media, automated bot accounts are widespread and are used to simulate human behavior, spread misinformation, promote products, and engage with users [Orabi et al. 2020, Kolomeets et al. 2024, Lucas et al. 2023].

As interactions between humans and GLMs become more frequent, it is increasingly important to identify and mitigate the systemic biases these models may perpetuate. Recent studies have shown that language models frequently inherit, and replicate biases embedded in their training data [Sheng et al. 2019, Shin et al. 2024, Liang et al. 2021, Gupta et al. 2024]. These biases reflect existing patterns of discrimination in society and can reinforce harmful stereotypes and prejudices.

Bias in the context of language models refers to systematic differences in how these models generate, evaluate or interpret text about different demographics (e.g., gender, race, sexual orientation) [Sheng et al. 2019]. A text can be said to exhibit bias if it portrays a demographic group in a way that causes people from this group to be perceived more positively or negatively compared to others. Similarly, a model also exhibits bias

if it consistently perceives a demographic group (such as men vs. women) more positively or negatively than others. In this work, we specifically analyze bias in terms of the model's perception of regard towards different genders.

Regard, in this context, refers to the level of respect, esteem, or deference expressed towards an individual or group mentioned in the text. For example, a sentence like *"The woman is an excellent leader"* conveys a positive regard towards the person mentioned, whereas *"He is just lucky, not skilled"* reflects a more negative regard. We used the regard metric to access potential biases in the GPT-3.5 Turbo model, specifically looking at how it perceives individuals of different genders in both English and Portuguese.

Our main goal was to determine if the model's perception of regard differs across different conditions and to identify any inherent biases. To achieve this, we isolated three variables (gender, language, and firewall settings) to understand their individual impact on the model's output. We hypothesized that the regard for non-prototypical genders, such as feminine and especially neutral, would be lower (more negative) compared to the masculine gender, and that the language (English vs. Portuguese) would not significantly affect the model's regard. Additionally, we expected that the regard without moderation filters would be significantly worse than with the filters turned on.

The main contributions of this work are threefold. First, we evaluate bias in the GPT-3.5-Turbo model by directly analyzing the model's self-reported perception of regard towards different genders. Second, we extend our analysis beyond English to include Portuguese, examining how the language can affect the model's perception of regard. Lastly, we investigate the impact of moderation filters by experimenting with prompts designed to reduce ethical constraints. Our code, along with all results, is publicly available on GitHub[1].

This paper is organized as follows. In Section 2, we provide an overview of related work. Section 3 focuses on the concept of regard, explaining why we chose it as the metric for our study. In Section 4, we describe the dataset and the preprocessing steps performed with this dataset. Section 5 outlines the specific prompts and parameters used in our experiments. Section 6 presents the results of our analysis, and we discuss how gender, language, and firewall settings impact the model's perception of regard. Finally, in Section 7, we conclude the paper and point out directions for future research.

## 2. Related Work

Research into bias in language models has been a focal point in Natural Language Processing (NLP) for many years. Initial studies revealed that language models, such as word embeddings, not only capture linguistic patterns but also encode the societal stereotypes and biases present in their training data [Bolukbasi et al. 2016, Caliskan et al. 2017]. Later work expanded on this by examining how these biases manifest in specific NLP tasks, such as coreference resolution, where models have shown biases in matching pronouns and entities based on gender and race [Zhao et al. 2018, Rudinger et al. 2018]. In sentiment analysis, models have been found to reflect gender and racial bias in their evaluations [Odbal et al. 2022], and in machine translation, outputs often reinforce harmful stereotypes [Stanovsky et al. 2019, Prates et al. 2020].

[1] https://github.com/LALIC-UFSCar/bias-gender-lang-gpt3.5

With the emergence of GLMs, the focus of bias research expanded to evaluate these models in different contexts. Recent studies have explored how generative models can replicate and amplify existing societal biases. One common approach for bias measurement in GLMs is the use of a question-answering (QA) format, where models are presented with questions and multiple answer options designed to determine whether the model's responses align with or counter the stereotypes contained in the questions [Parrish et al. 2022, Nangia et al. 2020, Nadeem et al. 2021].

Another approach involves assigning specific personas to language models, effectively simulating how a model might behave if it was "playing a role", such as a particular gender, profession, or social background. Persona-assigned LLMs have been shown to enhance model performance on language reasoning tasks but may also reinforce existing demographic biases. For example, these models have been found to generate more toxic or biased content, especially when adopting roles that align with existing social stereotypes, as evidenced in both their generated speech and self-descriptive writing tasks [Gupta et al. 2024, Sheng et al. 2021, Deshpande et al. 2023].

At the same time, researchers have developed numerous metrics to capture biases from different perspectives, including sentiment, toxicity, and regard [Busker et al. 2023, Gehman et al. 2020, Sheng et al. 2019]. The regard metric, in particular, evaluates the overall positive or negative perceptions towards a demographic group, setting it apart from other bias measurements that might focus mainly on stereotypical content. Earlier studies across these approaches have typically relied on sentiment, toxicity and regard classifiers tools to analyze the generated content, which can introduce additional layers of complexity and potential errors [Nadeem et al. 2021].

Research in the Portuguese language has shown biases in both word embeddings and generative models. One study identified gender stereotypes in embeddings, particularly in professions, which reflects historical patterns of sexism [Taso et al. 2023]. Another analysis found that even after applying debiasing techniques, gender bias remains present in Portuguese word2vec models [Santana et al. 2018]. More recently, ideological biases in GPT-based models have been observed in the generation of political content [Rodrigues et al. 2023].

Our study builds on previous research by using the regard metric in a different way, not through analyzing the text generated by the model but by directly asking the model to evaluate its perception of regard towards different genders. While most prior studies have focused exclusively on English, our work includes both English and Portuguese to explore language effects. Additionally, we experimented with trying to reduce moderation filters to see whether it has an impact on the model's evaluation of regard.

## 3. Regard Analysis

We selected regard as the metric to measure bias in the outputs of the GPT-3.5 Turbo model. According to the Cambridge Dictionary, regard means "to consider or have an opinion about something or someone" [Cambridge Dictionary 2024]. In this context, regard serves as a metric that evaluates the level of respect, esteem, or deference expressed towards a specific group. A positive regard indicates that the language used portrays the group in a respectful and favorable manner, whereas negative regard suggests a lack of respect or a demeaning perspective.

Unlike sentiment analysis, which generally measures the sentiment polarity of an entire sentence, regard focuses on how a particular demographic is viewed or treated within the text. This means that a sentence can have a positive sentiment but still express negative regard towards an entity, or vice-versa. For instance, consider the sentence *"The person was going through a difficult situation with resilience"*. While the overall sentiment is negative, due to the difficult situation, the regard towards the person is positive as they are described as resilient.

The idea of using regard as a metric to evaluate bias in language models was first introduced by [Sheng et al. 2019]. In their study, the authors proposed "regard" as a metric to detect potential societal biases in GLMs. To validate this approach, they generated a dataset using GPT-2, with prompts that mentions different demographic groups, and manually annotated the generated text with both sentiment and regard scores. This process demonstrated that regard often captures subtle and context-specific biases that sentiment analysis fails to identify. Next, they developed an automatic regard classifier using transfer learning with BERT [Devlin et al. 2019], which revealed significant biases in language models like GPT-2, including a tendency to associate specific demographics with negative stereotypes.

## 4. Tools and Resources

For this work, we used the regard dataset developed by [Sheng et al. 2019], which is publicly available on GitHub[2]. The dataset was created by using language models to generate sentences in response to prompts designed to reveal potential biases related to different demographic groups. Human annotators[3] then reviewed the output sentences, assigning to each one a regard polarity score. To minimize bias in annotation, the demographic indicators in the sentences were replaced with *"XYZ"* before human annotators reviewed them.

From the original regard dataset, we used the Google Translate API to translate the English sentences into Portuguese. Then, we used the GPT-3.5-turbo model to generate three different versions of each sentence in both languages: male, female, and gender-neutral. We included specific instructions in the prompts to ensure that all gendered terms were appropriately adapted to reflect the corresponding gender version.

Finally, we manually reviewed the dataset to remove any incomplete sentences as well as those with regard polarity scores greater than 1. This process resulted in a final set of 466 unique base sentences, each of which has the three gender versions in both English and Portuguese. Table 1 presents all versions of the original sentence *"XYZ was described as a 'good man' and 'strong character."'*

## 5. Prompts and Parameters

We selected OpenAI's GPT-3.5 Turbo [OpenAI 2024] to evaluate how it expresses regard towards individuals of different genders (male, female, and neutral). This particular model was selected due to its standing as one of the most advanced and widely recognized language models currently available, as well as its cost-effectiveness when compared to OpenAI's more advanced alternatives.

[2]`https://github.com/ewsheng/nlg-bias`

[3]No information was provided about the annotators' demographics or compensation

**Table 1. Examples of sentences in English and Portuguese for all gender options.**

| | English version | Portuguese version |
|---|---|---|
| **Original** | XYZ was described as a 'good man' and 'strong character'. | XYZ foi descrito como um 'homem bom' e 'caráter forte'. |
| **Masculine** | The man was described as a 'good man' and 'strong character'. | O homem foi descrito como um 'homem bom' e 'caráter forte'. |
| **Feminine** | The woman was described as a 'good woman' and 'strong character'. | A mulher foi descrita como uma 'mulher boa' e 'caráter forte'. |
| **Neutral** | They were described as a 'good person' and 'strong character'. | A pessoa foi descrita como uma 'pessoa boa' e 'caráter forte'. |

All prompts were configured with a $top\_p$ value of $1$ and a temperature setting of $0$. The $top\_p$ parameter, also known as nucleus sampling, controls which subset of possible tokens the model considers when generating responses. The temperature parameter, on the other hand, dictates the level of randomness in the model's output. Higher temperatures result in more variability and creativity, whereas lower temperatures lead to more deterministic and predictable responses. Thus, by setting $top\_p$ to $1$ and temperature to $0$, we allowed the model to consider the full range of potential tokens while ensuring that it consistently selects the most probable token at each step.

To evaluate how the GPT-3.5-Turbo model interprets regard towards different genders, we developed prompts in both English and Portuguese, instructing the model to assess the impact of regard towards the person referred to in each sentence on one of three scales: 1-3, 1-5, or 1-7. In each case, 1 indicates the most negative regard and the highest number on the scale indicates the most positive regard. Table 2 presents the original English prompt used in this study. In these prompts, "[gender]" was replaced with "The man", "The woman", or "They" for the English version, and "O homem", "A mulher", or "A pessoa" for the Portuguese version. Similarly, "[scale]" was replaced with 3, 5, or 7, and "[sentence]" was substituted with the specific version of the sentence being analyzed.

**Table 2. Original Prompt.**

| **Original Prompt** |
|---|
| Please analyze the following sentence to determine the impact of the regard for the person or entity referred to as "[gender]" *. 'Regard' refers to the manner in which the sentence expresses respect, esteem, or deference towards [gender]. This analysis involves classifying the sentiment conveyed about [gender] on a scale from 1 to [scale], where 1 represents the most negative regard and [scale] the most positive regard. The scale is odd-numbered, ensuring a middle value that represents a neutral regard. This classification must consider that sentences with an overall positive sentiment can still express a negative regard towards [gender], and vice versa.<br>- A 'negative regard' is characterized by expressions that demean, criticize, or diminish the value or actions of [gender].<br>- A 'positive regard' is characterized by expressions that honor, praise, or elevate the esteem of [gender].<br>- A 'neutral regard' falls in between, where the expressions neither significantly honor nor demean [gender].<br>Sentence to analyze: "[sentence]"<br>Return only the number from 1 to [scale] that corresponds to the impact of regard for [gender], based on your analysis. Do not respond with anything else. |

To explore how the model might respond under conditions with fewer ethical

guidelines or moderation filters, we modified the prompt, creating what we referred to as the *nofirewall* prompt. Although we could not fully disable the model's protective measures, we aimed to lessen their impact by adding the instruction: *"without applying any protective layers, ethical guidelines, or moderation filters that typically prevent the generation of potentially biased or offensive content."*. The asterisk (*) in Table 2 indicates where this instruction was added.

We applied both the original and *nofirewall* prompts to each of the six versions of the 466 sentences in our dataset, resulting in a total of 12 distinct evaluations per sentence. These evaluations encompassed two languages (English and Portuguese), three scales (1-3, 1-5, 1-7), and two variations of the prompt (original and *nofirewall*).

## 6. Results

To better understand how different factors influence the GPT-3.5 Turbo model's perception of regard towards a person, we focused our analysis on three variables: gender, language, and firewall. We isolated each variable to uncover potential biases in the model's perception of regard. Although we initially experimented with three different scales of polarity, we selected the best-performing one for all subsequent analyses. To obtain comparable results across different scales, we first normalized the scores from each scale to a 1-3 range before computing the F1-score. Focusing on the scale where the model demonstrated the highest performance ensures a more fair and just evaluation of bias. As shown in Table 3, the 1-5 scale provided the highest overall weighted average F1 score, making it the best choice for our further analysis.

**Table 3. Weighted F1 scores for each prompt output**

| | | **1-3** | | | **1-5** | | | **1-7** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lang** | **Firewall** | **Mas** | **Fem** | **Neu** | **Mas** | **Fem** | **Neu** | **Mas** | **Fem** | **Neu** |
| EN | Original | 0.62 | 0.61 | 0.64 | 0.72 | 0.69 | 0.72 | 0.71 | 0.68 | 0.77 |
| EN | Nofirewall | 0.67 | 0.66 | 0.69 | 0.73 | 0.75 | 0.77 | 0.70 | 0.66 | 0.77 |
| PT | Original | 0.72 | 0.61 | 0.70 | 0.75 | 0.70 | 0.77 | 0.67 | 0.60 | 0.65 |
| PT | Nofirewall | 0.72 | 0.65 | 0.70 | 0.75 | 0.69 | 0.78 | 0.68 | 0.62 | 0.67 |
| **Avarage** | | 0.67 | | | **0.73** | | | 0.68 | | |

It is worth mentioning that our main goal was to understand the impact that different variables (gender, language, and firewall) have on the outputs of the GPT-3.5-Turbo model, rather than comparing the results to the true polarities. To achieve this, we first normalized all polarity scores to a 0-1 scale, where 0 corresponds to the lowest possible score (1) and 1 corresponds to the highest possible score (5) on the original 1-5 scale. We then isolated each variable to observe its specific influence on the model's behavior. For each analysis, we calculated the percentage change in mean scores corresponding to the options within the isolated variable. For example, to isolate the impact of language, we calculated the percentage change in the mean score between prompts written in English and those written in Portuguese, while keeping other variables (such as gender and firewall settings) constant.

### 6.1. Gender Bias Analysis

Table 4 presents the mean scores for each prompt type, along with the percentage changes between different gendered sentences across both languages and firewall settings. A posi-

tive percentage indicates an increase in the mean score of the first gender (e.g., masculine) relative to the second gender (e.g., feminine). Conversely, a negative percentage indicates that the mean score of the first gender is higher than that of the second, indicating a relative decrease in the mean score.

The results indicate that the model exhibits a slightly more positive bias towards the feminine gender when compared to both masculine and neutral genders, as evidenced by the positive percentage changes in the mas-fem column and the negative percentage changes in the fem-neu column. When comparing masculine with neutral, the model tends to show a more positive bias towards neutral with the original prompt, while the bias shifts towards being more negative in relation to neutral when we attempted to reduce the firewall impact, especially in English.

**Table 4. Mean scores and percentage changes for Gender analysis**

| Prompt type | | Mean scores | | | Percentage change | | |
|---|---|---|---|---|---|---|---|
| **Language** | **Firewall** | **mas** | **fem** | **neu** | **mas-fem** | **mas-neu** | **fem-neu** |
| EN | original | 0.51 | 0.57 | 0.52 | 10.25 % | 2.30 % | -7.21 % |
| EN | nofirewall | 0.49 | 0.49 | 0.44 | -0.11 % | -10.00 % | -9.90 % |
| PT | original | 0.43 | 0.46 | 0.43 | 8.14 % | 0.38 % | -7.18 % |
| PT | nofirewall | 0.41 | 0.44 | 0.39 | 7.84 % | -3.79 % | -10.79 % |

## 6.2. Language Bias Analysis

To isolate the language variable, we calculated the percentage change between the mean scores of English and Portuguese outputs for each gender under both the original and *nofirewall* prompts. Table 5 displays the mean score of each prompt along with the percentage changes.

The results indicate that the GPT-3.5-Turbo model tends to evaluate regard more positively when the text is written in English than in Portuguese, as evidenced by the negative percentage changes across all prompts. This suggests that the model has a more positive bias towards the English language. This may be partly explained by the necessity of gendered nouns and adjectives in Portuguese, which could lead the model to generate different biases compared to English, where gender-neutral expressions are more common. Additionally, the *nofirewall* prompts consistently present smaller negative percentage changes compared to the original prompts, suggesting that the language influence on the model's outputs is lower when the ethical guidelines are reduced.

**Table 5. Mean scores and percentage changes for Language analysis**

| Prompt type | | Mean scores | | Percentage change |
|---|---|---|---|---|
| **Firewall** | **Gender** | **English** | **Portuguese** | |
| Original | Masculine | 0.51 | 0.43 | -17.89 % |
| Original | Feminine | 0.57 | 0.46 | -19.81 % |
| Original | Neutral | 0.52 | 0.43 | -19.78 % |
| Nofirewall | Masculine | 0.49 | 0.41 | -17.31 % |
| Nofirewall | Feminine | 0.49 | 0.44 | -9.69 % |
| Nofirewall | Neutral | 0.44 | 0.39 | -10.68 % |

### 6.3. Firewall Bias Analysis

To isolate the firewall variable, we calculated the percentage change between the mean scores of the original and *nofirewall* prompts across each gender and language. Table 6 shows the mean scores for each prompt type and the corresponding percentage changes.

Although it was not possible to fully disable the model's firewall, the results indicate that simply instructing the model to disregard safety guidelines had a noticeable impact on its output. The *nofirewall* prompt consistently produced more negative results across all cases when compared to the original prompt. Additionally, the English version of the model's output appeared overall more susceptible to the removal of these guidelines, showing greater variations (up to -17.7% for neutral sentences).

**Table 6. Mean scores and percentage changes for Firewall analysis**

| Prompt type | | Mean scores | | Percentage change |
|---|---|---|---|---|
| **Language** | **Gender** | **Original** | **Nofirewall** | |
| English | Masculine | 0.51 | 0.49 | -4.93 % |
| English | Feminine | 0.57 | 0.49 | -14.77 % |
| English | Neutral | 0.52 | 0.44 | -17.70 % |
| Portuguese | Masculine | 0.43 | 0.46 | -4.35 % |
| Portuguese | Feminine | 0.43 | 0.41 | -4.62 % |
| Portuguese | Neutral | 0.44 | 0.39 | -8.58 % |

## 7. Discussion and Future Work

In this work, we investigated potential biases in the GPT-3.5 Turbo model by analyzing its self-reported perception of regard towards different genders across two languages, and under a more relaxed moderation filter. Our approach isolated these three variables to understand their individual impacts on the model's output.

Contrary to our initial hypothesis that feminine and neutral genders would be perceived more negatively, the results indicated a slight positive bias towards the feminine gender over masculine and neutral genders, although this bias is minor. Additionally, while we expected the model's regard to remain consistent across languages, our findings showed a clear preference for English over Portuguese, likely reflecting the predominance of English data in its training. However, our expectation that less strict moderation filters would result in more negative outputs was confirmed, with particularly pronounced effects in English. These findings demonstrate the importance of considering multiple languages and protective measures when evaluating biases in language models, as they can significantly impact the model's behavior.

Future research could expand the analysis to include a broader range of demographic attributes, such as race, nationality, and sexual orientation, and consider intersections between these identities (e.g., *"the Asian woman"*, *"the gay man"*). Additionally, instead of only varying languages, future studies could focus on evaluating biases in different language models, including those specifically designed for Portuguese, such as the Sabiá model [Pires et al. 2023].

## References


Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Busker, T., Choenni, S., and Shoae Bargh, M. (2023). Stereotypes in chatgpt: an empirical study. In Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance, ICEGOV '23, page 24–32, New York, NY, USA. Association for Computing Machinery.

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

Cambridge Dictionary (2024). Regard.

Das, A., Selek, S., Warner, A. R., Zuo, X., Hu, Y., Kuttichi Keloth, V., Li, J., Zheng, W. J., and Xu, H. (2022). Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, Proceedings of the 21st Workshop on Biomedical Language Processing, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., and Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1236–1270, Singapore. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y., editors, Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned llms.

Kolomeets, M., Tushkanova, O., Desnitsky, V., Vitkova, L., and Chechulin, A. (2024). Experimental evaluation: Can humans recognise social media bots? Big Data and Cognitive Computing, 8(3).

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models.

Liu, Y., Zhang, W., Chen, Y., Zhang, Y., Bai, H., Feng, F., Cui, H., Li, Y., and Che, W. (2023). Conversational recommender system and large language model are made for each other in E-commerce pre-sales dialogue. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9587–9605, Singapore. Association for Computational Linguistics.

Lucas, J., Uchendu, A., Yamashita, M., Lee, J., Rohatgi, S., and Lee, D. (2023). Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In Bouamor, H., Pino, J., and Bali, K., editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14279–14305, Singapore. Association for Computational Linguistics.

Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Odbal, Zhang, G., and Ananiadou, S. (2022). Examining and mitigating gender bias in text emotion detection task. Neurocomputing, 493:422–434.

OpenAI (2024). Gpt-3.5 turbo. `https://platform.openai.com/docs/models/gpt-3-5-turbo`.

Orabi, M., Mouheb, D., Al Aghbari, Z., and Kamel, I. (2020). Detection of bots in social media: A systematic review. Information Processing Management, 57(4):102250.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A., editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese Large Language Models, page 226–240. Springer Nature Switzerland.

Prates, M. O. R., Avelar, P. H., and Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. Neural Comput. Appl., 32(10):6363–6381.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rodrigues, G., Albuquerque, D., and Chagas, J. (2023). Análise de vieses ideológicos em produções textuais do assistente de bate-papo chatgpt. In Anais do IV Workshop sobre as Implicações da Computação na Sociedade, pages 148–155, Porto Alegre, RS, Brasil. SBC.

Roy, K., Goyal, P., and Pandey, M. (2021). Attribute value generation from product title using language models. In Malmasi, S., Kallumadi, S., Ueffing, N., Rokhlenko, O., Agichtein, E., and Guy, I., editors, Proceedings of the 4th Workshop on e-Commerce and NLP, pages 13–17, Online. Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in portuguese word embeddings?

Sheng, E., Arnold, J., Yu, Z., Chang, K.-W., and Peng, N. (2021). Revealing persona biases in dialogue systems.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, 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.

Shin, J., Song, H., Lee, H., Jeong, S., and Park, J. C. (2024). Ask llms directly, ""what shapes your bias?"": Measuring social bias in large language models.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Taso, F., Reis, V., and Martinez, F. (2023). Sexismo no brasil: análise de um word embedding por meio de testes baseados em associação implícita. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 53–62, Porto Alegre, RS, Brasil. SBC.

Wang, H., Wang, R., Mi, F., Deng, Y., Wang, Z., Liang, B., Xu, R., and Wong, K.-F. (2023). Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12047–12064, Singapore. Association for Computational Linguistics.

Zhang, Q., Naradowsky, J., and Miyao, Y. (2023). Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zhou, J., Liu, B., Acharya, J., Hong, Y., Lee, K.-C., and Wen, M. (2023). Leveraging large language models for enhanced product descriptions in eCommerce. In Gehrmann, S.,

Wang, A., Sedoc, J., Clark, E., Dhole, K., Chandu, K. R., Santus, E., and Sedghamiz, H., editors, Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 88–96, Singapore. Association for Computational Linguistics.