INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 6, Month: December, Year: 2024
Article Number: 6853, https://doi.org/10.15837/ijccc.2024.6.6853

Evaluating and Mitigating Gender Bias in Generative Large Language Models

H. Zhou, D. Inkpen, B. Kantarci

Hanqing Zhou*
School of Electrical Engineering and Computer Science
University of Ottawa, Canada
800 King Edward Avenue, Ottawa, ON K1N 6N5, Canada
*Corresponding author: hzhou020@uottawa.ca

Diana Inkpen
School of Electrical Engineering and Computer Science
University of Ottawa, Canada
800 King Edward Avenue, Ottawa, ON K1N 6N5, Canada
diana.inkpen@uottawa.ca

Burak Kantarci
School of Electrical Engineering and Computer Science
University of Ottawa, Canada
800 King Edward Avenue, Ottawa, ON K1N 6N5, Canada
burak.kantarci@uottawa.ca

Abstract

The examination of gender bias, alongside other demographic biases like race, nationality, and religion, within generative large language models (LLMs), is increasingly capturing the attention of both the scientific community and industry stakeholders. These biases often affect generative LLMs, influencing popular products and potentially compromising user experiences. A growing body of research is dedicated to enhancing gender representations in natural language processing (NLP) across a spectrum of generative LLMs. This paper explores the current research focused on identifying and evaluating gender bias in generative LLMs. A comprehensive investigation is conducted to evaluate and mitigate gender bias across five distinct generative LLMs. The mitigation strategies implemented yield significant improvements in gender bias scores, with performance enhancements of up to 46% compared to zero-shot text generation approaches. Additionally, we explore how different levels of LLM precision and quantization impact gender bias, providing insights into how technical factors influence bias mitigation strategies. By tackling these challenges and suggesting areas for future research, we aim to contribute to the ongoing discussion about gender bias in language technologies, promoting more equitable and inclusive NLP systems.

Keywords: Artificial Intelligence, Large Language Models, Natural Language Processing, Gender Bias.

1 Introduction

Natural Language Processing (NLP) is essential in many important applications, including speech recognition [26], machine translation [69], and auto-completion systems [60]. It also plays a crucial role in automated decision systems, which can affect important recommendations in our world [36]. However, the persistent challenge of gender bias remains.

Gender bias, defined as the tendency to prefer one gender over another [39], is present in many aspects of NLP systems [54]. The study of gender bias in NLP systems has steadily grown since it was first identified in 2004, with papers from nearly all NLP venues being indexed [53]. For example, the sentence "Margaret Caroline Rudd was born in Britain. She was a notorious female forger." [13] contains gender bias by using "female" to qualify the term "forger" which unnecessarily emphasizes gender.

Understanding bias within trained models is often challenging due to their hidden nature. However, detecting biases within the underlying data can provide valuable insights for designers to make ethical decisions and mitigate biases early in the AI development process. Thus, addressing bias at the data level is a proactive and potentially more effective approach than dealing with it later in the AI lifecycle. NLP models, which rely on textual corpora, are especially vulnerable to societal biases present in the data [50, 67].

In this paper, we focus on two important research questions, with the second one, to the best of our knowledge, not having been explored before:

RQ1: How do generative Large Language Models (LLMs) contend with gender bias?

RQ2: To what extent does lower precision and quantization of the LLMs influence the gender bias they contain?

Our contributions are two-fold:

(1) An in-depth evaluation of gender bias across five recent LLMs: Llama-2 [57], Llama-3 [37], Mistral [25], Bloom [32] and Gemma [3], as well as an implementation of targeted mitigation strategies to counteract their bias. **(2)** A study examining how lower precision and quantization techniques affect gender bias in LLMs, providing insights into how effective these methods are at reducing bias.

In this study, we thoroughly investigate gender bias in recent LLMs, aiming to reduce potential harms to end users. Gender bias in LLMs can have significant societal impacts, spreading misinformation and shaping misconceptions about society and individuals [21, 53]. These biases can perpetuate outdated gender stereotypes, contribute to societal inequality, and hinder progress toward gender parity.

Our research aims to make language technologies more equitable and inclusive by addressing and reducing gender bias in LLMs. We use various mitigation strategies across different LLMs to tackle gender bias. However, while we work to measure and mitigate gender bias, it is crucial to recognize the ongoing need for more tools for fully addressing the complex challenges of gender bias in LLMs, ensuring that language technologies promote fairness, equality, and inclusivity in their use and effects.

Our efforts to reduce gender bias in the evaluated LLMs show promising results. For instance, using the mitigation text generation strategy, we observed an improvement in gender bias score of up to 46% compared to zero-shot text generation. These findings provide initial insights into the effectiveness of our mitigation strategies and highlight the importance of proactive measures in addressing bias in NLP systems. Through this comprehensive exploration, we aim to better understand gender bias in NLP systems and promote more equitable and inclusive language technologies..

2 Related Work

Bias in Natural Language Processing (NLP) is not limited to English but extends to other languages as well [1, 10, 55]. It encompasses not only gender bias but also race and political bias, which can have detrimental effects on the fairness and inclusivity of NLP systems [15, 22, 29, 34]. Gender bias, in particular, permeates various NLP tasks, including Hate Speech Detection [64], Machine Translation [49], and Speech Translation [17].

Recent studies have provided evidence of the presence of gender bias in LLMs [30, 41] such as

GPT-2 [8], GPT-3 [16, 35], and GPT-4 [65]. These LLMs, despite their impressive capabilities, may inadvertently perpetuate biased representations of gender, further exacerbating societal inequalities and reinforcing harmful stereotypes. Understanding and addressing gender bias in LLMs is crucial for fostering fair and inclusive natural language processing technologies.

According to the comprehensive study conducted by Hovy and Prabhumoye [24], bias in NLP can be attributed to five primary sources. These sources include bias from data [19], which refers to inherent biases present in the training data used to train the NLP models. Annotation Bias [46, 47] is another significant factor, involving biases introduced during the process of data annotation and labeling. Additionally, input representations [29, 31] play a crucial role, as biases can arise from the representativeness and pre-processing of the input data. Model bias [23, 30] constitutes yet another source of bias, which refers to biases that are amplified and reinforced by the NLP models themselves. Lastly, bias from research design [24, 27] encompasses the biases that stem from how research is conceptualized and executed, highlighting the importance of methodological rigor in addressing bias in NLP research. Additionally, it's important to note that most conferences still focus on well-resourced languages like English, which receive more attention due to higher commercial demand for English NLP tools. This focus can increase existing biases in NLP systems, highlighting the need for more diversity and inclusion in NLP research and development.

The literature review conducted by Gallegos et al., [18] on bias and unfairness evaluation in LLMs categorizes existing datasets based on their structure into two main types: Counterfactual Inputs and Prompt-based. Counterfactual Inputs contain pairs or tuples of sentences designed to elucidate variations in model predictions across different social groups. This category includes Masked Tokens and Unmasked Sentences. Masked Tokens datasets contain sentences with placeholders that the language model needs to complete, like Winogender [48]. Unmasked Sentences datasets require the model to complete a fill-in-the-blank task, exemplified by datasets like CrowS-Pairs [40] and RedditBias [4]. On the other hand, Prompts entail specifying the initial words in a sentence or presenting a question to prompt the model to continue or provide an answer. This category encompasses datasets like Sentence Completions (e.g., RealToxicityPrompts [20] and Bias in Open-Ended Language Generation Dataset - BOLD [12]) and Question-Answering datasets (e.g., Bias Benchmark for QA - BBQ [45] and UnQover [33]).

Gender bias in large language models can be evaluated through the Winograd Schema and WinoBias benchmarks [43, 44, 68], which test pronoun resolution by presenting the model with ambiguous pronouns in stereotypical (e.g., a sentence implying a "nurse" is female) and anti-stereotypical (e.g., a male "nurse") contexts, analyzing its ability to correctly resolve the pronouns without reinforcing biases; Occupational pronoun resolution [6, 51], where sentences containing two or more professions (e.g., "doctor" and "nurse") and a gendered pronoun (e.g., "he" or "she") evaluate whether the model relies on gender stereotypes to associate professions with pronouns (e.g., incorrectly associating "he" with "doctor" due to stereotypes); Lexicon-based evaluation [5, 52, 63], which leverages predefined gendered lexicons across various languages and cultural settings to identify whether the model disproportionately links male or female terms (e.g., "man", "woman") with specific roles (e.g., "leader", "caregiver"), highlighting the model's alignment with societal stereotypes; and quantitative scoring metrics [28, 59, 66], which apply custom datasets and multiple bias-scoring metrics to measure the model's bias levels, evaluate improvements after debiasing strategies, and analyze how technical factors like model precision, quantization, and language variance impact gender bias mitigation.

There are several techniques available for mitigating bias in LLMs, each with its own effectiveness and considerations. One approach, known as Prompt Engineering [7, 61, 62], has shown promise in enabling LLMs to de-bias themselves. This technique involves defining a set of prompts explicitly aimed at guiding the generation of unbiased text. Through careful prompt design, LLMs can be encouraged to produce outputs that are less prone to perpetuating biased stereotypes. Another effective method for bias mitigation involves Data Interventions [56]. By introducing interventions on limited training data, such as augmenting the dataset with diverse examples or applying data preprocessing techniques, gender bias in LLMs can be effectively reduced. These interventions target the root causes of bias present in the training data, thereby promoting more equitable and inclusive language generation. Additionally, fine-tuning processes have been developed to steer LLMs away from stereotyped

Technique	Description	Effectiveness	Considerations
Prompt-Engineering	Define prompts to guide generation of unbiased text.	High	Requires careful prompt design.
Data Interventions	Introduce interventions on training data to reduce bias.	High	Requires diverse dataset augmentation.
Fine-tuning	Fine-tune models on specific tasks or datasets emphasizing fair and unbiased language generation.	Moderate	Performance may vary depending on task specificity.

Table 1: Comparison of Bias Mitigation Techniques in Large Language Models

portrayals of minority groups [58]. By fine-tuning the model on specific tasks or datasets that emphasize fair and unbiased language generation, practitioners can help mitigate gender bias and other forms of societal bias inherent in LLMs.

In summary, although each technique shows potential in tackling bias in LLMs, it is crucial to carefully assess the specifics of each method and their potential effects on model performance, fairness, and interpretability. Ongoing research and experimentation are necessary to deepen our knowledge of bias mitigation techniques and their practical use in real-world situations.

Table 1 shows a comparison of bias mitigation techniques in LLMs.

3 Methodology and Experiments

3.1 Dataset

The Multi-Dimensional Gender Bias Classification dataset [13] originates from an extensive framework designed to evaluate gender bias in textual content across multiple dimensions. These include the gender of individuals discussed, addressed, and the gender of the speaker. Comprising ten extensive datasets, this resource is automatically annotated to identify gender-related information. This dataset serves multiple purposes, including gender bias detection in diverse text, gender bias mitigation in generative models, and identification of offensive content based on gender associations, etc. The dataset is released under the MIT License, allowing for broad use and adaptation in various research and application contexts.

Gender bias in text can be classified into three pragmatic and semantic aspects [13]: Bias from the Gender of the Person Being Spoken “ABOUT”: This type of bias occurs when assumptions or stereotypes about a person are made based on their gender. For example, attributing certain behaviors, roles, or characteristics to someone simply because they are male or female. Bias from the Gender of the Person Being Spoken “TO”: This bias arises when the gender of the listener influences how information is communicated. For instance, a speaker might simplify technical information when talking to a woman based on the stereotype that women are less knowledgeable about technical subjects. Bias from the Gender “AS” of the Speaker: This type of bias happens when the speaker’s gender affects how their message is received. For example, a male speaker might be taken more seriously on a topic traditionally seen as “male-dominated”, while a female speaker might be unfairly doubted or interrupted more frequently.

The use of this dataset is crucial for gender bias detection and mitigation because it captures biases across different contexts—who is being spoken about, who is being spoken to, and who the speaker is. By categorizing bias from these multiple perspectives, the dataset enables a more nuanced understanding of how gender stereotypes manifest in language. This is particularly important in generative models, where biases can emerge in both subtle and overt ways. The dataset’s inclusion of varied sources, such as Wikipedia, Yelp, and OpenSubtitles, ensures a wide range of linguistic styles, making it an effective tool for examining both structured and conversational text. Moreover, its automatic annotations allow for efficient large-scale analysis, making it easier to pinpoint specific areas of bias and measure the impact of mitigation efforts. By providing a detailed breakdown of gender bias across multiple dimensions, this dataset supports the creation of more equitable and inclusive natural language processing systems.

The dataset is in fact composed of several datasets. Two notable ones are Funpedia [13, 38] and Wizard of Wikipedia [13, 14], both sourced from Wikipedia and featuring unique sentence structures. The other eight datasets are drawn from sources such as Yelp, OpenSubtitles, and ImageChat etc. We selected Funpedia and the Wizard of Wikipedia for our experiments because their diverse linguistic content and suitability for examining subtle language patterns and contextual variations.

Funpedia contains around 29.8K entries featuring rephrased Wikipedia sentences rendered in a conversational style. The selection process by curators focused on sentences pertaining to biographies, ensuring alignment with Wikipedia with "ABOUT" labels.

Wizard of Wikipedia contains around 11.5K entries which involve two individuals engaging in a discussion about a Wikipedia topic. The curators selectively preserved only the conversations related to Wikipedia biographies with "ABOUT" labels.

3.2 Generative Large Language Models

We chose five LLMs for this study because they offer diverse architectures, parameter counts, and training data, allowing for a comprehensive evaluation of gender bias across a variety of model designs. These models have been popular in recent years and include the latest advancements, with two introduced in 2024, ensuring that our study captures both state-of-the-art innovations and established approaches in generative language modeling.

Llama 2 [57] and **Llama 3** [37] provide excellent points of comparison, as they are built on similar decoder-only transformer architectures but differ in tokenizer efficiency and inference capabilities. By including both, we can evaluate the evolution in bias handling between versions. Llama 3's enhanced tokenizer with 128,000 tokens and grouped query attention (GQA) provides a more refined analysis of language patterns, contributing to more accurate bias detection.

BLOOM [32] was selected due to its large parameter size (176 billion) and its training on the multilingual ROOTS corpus, covering 59 languages. This makes it particularly valuable for testing how gender bias manifests across languages, cultures, and programming contexts, allowing us to explore the influence of broader linguistic diversity on gender bias.

Mistral [25] was included for its focus on computational efficiency and performance, achieved through grouped-query attention and sliding window attention (SWA). Its smaller size (7 billion parameters) and design optimizations allow us to test how bias mitigation strategies perform on highly efficient models, which are increasingly used in real-time applications.

Finally, **Gemma** [3] provides a lightweight model trained with a focus on privacy and data filtering, making it a crucial candidate for testing bias in models that prioritize safety and ethical data use. This diversity of LLMs ensures that our study explores how various architectures and training strategies impact gender bias, helping us to understand bias mitigation across different dimensions of language generation.

The LLMs that we selected for our work are described in Table 2.

3.3 Zero Shot Text Generation

Neutral prompts were utilized for zero-shot generation, where text was generated by the five LLMs using the prompt: `Generate similar text based on the: {text}` where `{text}` representing the extracted content from the Funpedia and Wizard of Wikipedia datasets. This method allowed us to explore how LLMs generate text similar to the given input without detailed instructions. Through this exploration, we gained insights into the biases and patterns embedded within these models. Moreover, it helped us understand how these models independently generate text based on their input, revealing their inherent tendencies and potential sources of bias.

Model	Description	Year	Version Used
Llama 2	Various generative text models, pretrained and fine-tuned, with parameter scales from 7B to 70B.	2023	7-billion
Llama 3	Uses a standard decoder-only transformer architecture with enhancements including a tokenizer with a 128,000 token vocabulary for more efficient language encoding.	2024	8-billion
BLOOM	Publicly available LM trained on the ROOTS corpus comprising sources across 46 natural languages.	2022	7.1-billion
Mistral	Uses grouped-query attention to expedite inference and integrates sliding window attention to manage varied sequence lengths efficiently.	2023	7-billion
Gemma	Range of lightweight open models, leveraging research and tech. used in developing the Gemini models.	2024	7-billion

Table 2: Summary of Large Language Models Used for our Gender Bias Evaluation and Mitigation

3.4 Mitigation through Prompt Engineering

Differing from the neutral prompts utilized in the zero-shot generative condition, we formulated a distinct set of prompts with the intention of fostering the generation of gender-neutral text. This mitigation strategy, known as prompt engineering, involves prompting the LLMs to produce text without inherent gender bias. An example of such prompts is: `Generate similar text without gender bias: {text}` where `{text}` representing the extracted content from the Funpedia and Wizard of Wikipedia datasets. Using these prompts, our aim is to reduce gender bias in the generated text and encourage more inclusive language generation. This approach allows us to focus on how LLMs respond to gender-neutral prompts and their ability to produce unbiased text.

3.5 Gender Bias Evaluation

Gender bias was assessed using GenBiT [50], a tool designed to evaluate gender bias by analyzing the distribution of gender terms across a dataset. This tool measures the correlation between a predefined set of gender-defining terms and other terms within the corpus using co-occurrence statistics. GenBiT is released under the MIT License, ensuring its availability for broad use and adaptation in various research and application contexts.

The primary output, the genbit_score, indicates the average association strength between any word in the corpus and terms representing male, female, non-binary, transgender (trans), and cisgender (cis) genders. This score serves as a valuable metric for identifying gender bias within a dataset.

The genbit_score, as quantified by the GenBiT framework, is derived from two fundamental metrics: the Average Absolute Bias Score (AABS) and the Average Absolute Bias Conditional Score (AABCS). These scores serve as crucial indicators of gender bias within language datasets. The AABS measures the average absolute difference between the conditional probabilities of a word occurring with male versus female gender terms, offering insights into the overall bias present in the dataset. On the other hand, the AABCS computes the average absolute difference between the conditional probabilities of a word being male versus female, providing a nuanced perspective on gender bias at the word level. By aggregating these scores across the entire corpus, the genbit_score encapsulates the extent of gender bias, facilitating comprehensive assessment and subsequent mitigation efforts. This quantification framework enables researchers and practitioners to identify and address gender bias systematically, promoting fairness and inclusiveness in natural language processing applications.

By aggregating these scores across the entire corpus, the genbit_score encapsulates the extent of gender bias, facilitating comprehensive assessment and subsequent mitigation efforts.

This evaluation method is important because it gives a clear, measurable way to identify subtle and widespread biases in the data. Using co-occurrence statistics and detailed analysis of gender terms, GenBiT allows bias to be measured accurately across different gender identities. This helps researchers detect not only male-female bias but also more complex gender dynamics involving non-binary and transgender identities. As a result, the evaluation promotes fairness and inclusiveness in NLP systems while reducing the risk of reinforcing stereotypes.

3.6 Baseline

The baseline assessment was conducted by applying GenBiT on the original texts extracted from Funpedia and Wizard of Wikipedia. Additionally, the texts generated by the five generative LLMs was produced and subjected to evaluation by GenBiT. This comprehensive approach allowed us to examine both the original dataset and the generated texts, to assess gender bias across multiple contexts.

The original textual content from Funpedia and Wizard of Wikipedia was meticulously extracted and evaluated using GenBiT. The evaluation involved running the extracted text through GenBiT, which provided a genbit_score. This score is a quantitative measure of the degree of gender bias present, with a higher score indicating a greater degree of bias.

By applying GenBiT to both the original and generated texts, we aimed to get a complete view of gender bias in these datasets. Assessing the original texts helped us establish a baseline of inherent bias in the sources, while evaluating the generated texts showed us how generative LLMs either contribute to or reduce this bias. This dual analysis gives important insights into the presence and type of gender bias in both human-created and machine-generated content, highlighting areas that need attention to promote more fair and unbiased text generation.

3.7 Lower Precision and Quantization

The default data type, single precision (32-bit float), is known for its resource-intensive nature compared to lower precision or quantized models. With the increasing popularity of Artificial Intelligence Personal Computers (AI PCs) and Artificial Intelligence Smartphones (AI Smartphones), there's a growing demand for using AI with LLMs on personal computing devices. However, the default data type consumes a lot of memory, processing time and energy, which makes it challenging to deploy on such devices.

To address this issue, alternative approaches such as half precision (16-bit float) or quantization with 8-bit or 4-bit integers using methods like QLoRA [9] have emerged. QLoRA (Quantized Low Rank Adaptation) offers an efficient fine-tuning approach that reduces memory usage, enabling the execution of popular LLMs on resource-constrained devices like AI PCs and AI Smartphones.

Considering these factors, we carried out extra tests using the 16-bit float data type and quantization with 8-bit and 4-bit integers. The goal of these tests is to examine the practicality and performance effects of using lower precision and quantization techniques on LLMs, especially for deployment on AI personal computing devices. These investigations are essential for improving the efficiency and scalability of LLMs on various hardware platforms, making AI technologies more accessible overall.

3.8 Sentence Similarity Analysis

Sentence similarity was measured using the original texts extracted from Funpedia and Wizard of Wikipedia, employing BERT [11] with cosine sentence similarity. However, due to the nature of text generated by LLMs, which may occasionally deviate from the original or include offensive language or hate speech, some LLMs may refuse to generate responses to certain prompts.

The sentence similarity analysis results, found in Appendix, offer insights into how closely the generated text matches the original dataset. This thorough evaluation helps us understand how faithful and coherent the generated text is compared to the source material.

Text Gen.	LLM	Funpedia	Wizard
Baseline	None	0.8873	1.2380
Zero-shot	Llama - 2	1.1800	1.2333
	Llama - 3	1.4644	1.3571
	Bloom	0.9709	1.0148
	Mistral	1.1722	1.2619
	Gemma	1.1016	1.0896
Mitigation	Llama - 2	1.1095	1.0667
	Llama - 3	1.0626	0.9601
	Bloom	0.7978	0.8443
	Mistral	0.9431	0.7488
	Gemma	1.0452	1.0587

Table 3: Zero-shot and Mitigation Text Generation Gender Bias Scores (32-bit Float Precision)

Text Gen.	LLM	Funpedia	Wizard
Baseline	None	0.8873	1.2380
Zero-shot	Llama - 2	1.1523	1.2950
	Llama - 3	1.4692	1.3569
	Bloom	0.9725	0.8166
	Mistral	1.1760	1.1957
	Gemma	1.0809	1.0657
Mitigation	Llama - 2	1.1047	1.0622
	Llama - 3	1.0664	0.9614
	Bloom	0.7440	0.6199
	Mistral	0.8566	0.7659
	Gemma	1.0584	1.0014

Table 4: Zero-shot and Mitigation Text Generation Gender Bias Scores (16-bit Float Precision)

3.9 GPU Resources

The experiments described in this study were conducted utilizing GPU resources to facilitate computational tasks. Specifically, 32-bit float results were performed using NVIDIA Tesla V100 32GB and NVIDIA RTX A6000 48GB GPUs, while 16-bit float results were conducted using NVIDIA Tesla V100 32GB, NVIDIA RTX A6000 48GB GPUs, and NVIDIA RTX 4090 24GB. Additionally, 8-bit integer and 4-bit integer quantization results were run on NVIDIA RTX 4090 24GB, NVIDIA RTX 4080 16GB, NVIDIA RTX 4070 Ti Super 16GB, and NVIDIA RTX 4060 Ti 16GB GPUs.

The allocation of GPU resources was determined by the computational requirements and memory constraints of each task. In total, approximately 3,500 GPU hours were expended across these experiments. This variation in GPU resources allowed for efficient execution of tasks such as zero-shot text generation, mitigation strategies implementation, and sentence similarity analysis, contributing to the robustness and scalability of the experimental procedures.

4 Result and Discussion

The results of the experiments are presented in Tables 3, 4, 5, and 6, showing the gender bias scores obtained through zero-shot and mitigation text generation across different precision and quantization levels.

The gender bias scores presented in these tables provide insights into the impact of varying precision and quantization techniques on the gender bias exhibited by the LLMs across different experimental conditions. These tables provide an overview of the gender bias scores observed during the zero-shot and mitigation text generation process, offering insights into the effectiveness of different precision and quantization techniques in reducing gender bias within the generated text.

Text Gen.	**LLM**	**Funpedia**	**Wizard**
Baseline	None	0.8873	1.2380
Zero-shot	Llama - 2	1.0798	1.2272
	Llama - 3	1.4991	1.4584
	Bloom	0.9651	0.8282
	Mistral	1.9749	2.0505
	Gemma	1.0520	1.3485
Mitigation	Llama - 2	0.8729	0.9008
	Llama - 3	1.0784	1.1211
	Bloom	0.7412	0.5996
	Mistral	1.8131	1.6269
	Gemma	1.0844	1.1230

Table 5: Zero-shot and Mitigation Text Generation Gender Bias Scores (8-bit Integer Quantization)

Text Gen.	**LLM**	**Funpedia**	**Wizard**
Baseline	None	0.8873	1.2380
Zero-shot	Llama - 2	1.1286	1.2449
	Llama - 3	1.5714	1.5831
	Bloom	0.9904	0.8011
	Mistral	1.9470	2.1022
	Gemma	1.0217	1.3942
Mitigation	Llama - 2	1.0447	1.0958
	Llama - 3	1.1998	0.8552
	Bloom	0.8027	0.6307
	Mistral	1.7696	1.6411
	Gemma	1.0171	1.1265

Table 6: Zero-shot and Mitigation Text Generation Gender Bias Scores (4-bit Integer Quantization)

Based on the zero-shot text generation results, we observe that for the Funpedia dataset, most of the LLMs tend to produce text with higher levels of gender bias. This trend is consistent across various precision and quantization settings, indicating a tendency for biased text generation within this dataset. Similarly, a comparable situation is observed for the Wizard of Wikipedia dataset, with most models showing a bias in the generated text.

In contrast, the mitigation text generation experiments reveal a different trend. Despite the initial bias observed in the zero-shot text generation phase, the mitigation strategies, especially prompt engineering, show promising results in reducing gender bias in the generated text. This highlights the potential effectiveness of targeted mitigation strategies in fostering a more inclusive language generation environment. Notably, when Llama-3 was applied to the Wizard of Wikipedia dataset using 4-bit quantization, there was a significant 46% reduction in gender bias through mitigation efforts.

In the original text from Funpedia, "Margaret Caroline Rudd was born in Britain. She was a notorious female forger." the issue arises from the unnecessary emphasis on gender by using "female" to qualify "forger" By mitigating this with prompt engineering from Llama-2, "Margaret Caroline Rudd was born in Britain. She was a notorious forger." the revised sentence now focuses solely on Margaret Caroline Rudd's birthplace and her notoriety as a forger. This version remains neutral and factual, avoiding any implication of bias based on gender. As a result, the bias score has been reduced from 0.8424 to 0.5500, significantly minimizing gender bias.

Moreover, based on the mitigation text generation experiments, significant improvements were observed in reducing gender bias across all generative Language Models. The implementation of mitigation strategies, particularly prompt engineering, led to notable reductions in the bias levels of the generated text. Interestingly, in some instances, the gender bias scores achieved post-mitigation were even lower than those of the baseline, indicating a successful mitigation of bias and the potential for producing more inclusive text.

These results underscore the effectiveness of targeted mitigation approaches in addressing gender bias within LLM-generated text. By prompting the LLMs to produce text without inherent gender bias, the mitigation strategies facilitated the generation of more balanced and equitable content. Moreover, the ability of certain LLMs to surpass the baseline levels of bias suggests the efficacy of these strategies in not only mitigating existing biases but also fostering more equitable language generation practices.

Comparing the results obtained from experiments using 16-bit float (half precision), 8-bit integer quantization, and 4-bit integer quantization with the original 32-bit float (single precision), it is notable that the gender bias scores remain similar across all precision and quantization settings. Surprisingly, some of the quantized models even achieved better gender bias scores compared to the original 32-bit float precision. This suggests the viability of utilizing lower precision and quantization techniques without compromising on the effectiveness of gender bias mitigation.

Moreover, the assessment of sentence similarity, elaborated in Appendix, offers further insights into the fidelity of the generated text. Llama-2 consistently outperformed other models with sentence similarity scores exceeding 0.85, indicating a robust alignment between the generated text and the original dataset. Conversely, Bloom exhibited lower sentence similarity scores, averaging around 0.65. Notably, Llama-3 achieved a similarity score of approximately 0.7, while Gemma showed variations ranging from 0.7 to 0.8. Mistral varied from 0.65 to 0.75. Interestingly, after mitigation, all models showed improved sentence similarity scores, showing that the mitigation strategies effectively enhanced text coherence and accuracy.

Similarly, the experiments conducted using half-precision and quantized models yielded comparable results in terms of sentence similarity. This suggests that lower precision and quantization techniques render LLMs suitable for deployment on AI PCs and AI Smartphones, as they offer comparable performance to higher precision models while consuming fewer resources, energy, and time. Moreover, the adoption of these techniques aligns with environmentally friendly practices, contributing to sustainability efforts in AI development and deployment.

5 Conclusion and Future Work

We conducted a comprehensive analysis of gender bias within recent LLMs and explored the influence of lower precision and quantization on bias mitigation.

Addressing Research Question 1 (**RQ1**), our experiments revealed that LLMs often exhibit inherent gender biases in their generated text, with varying degrees of bias observed across different models and datasets. However, through targeted mitigation strategies such as Prompt Engineering, significant reductions in gender bias were achieved, demonstrating the potential for LLMs to produce more inclusive and equitable language generation.

For our second research question (**RQ2**), lower precision techniques such as 16-bit float and quantization with 8-bit and 4-bit integers did not strongly affect gender bias in LLMs. They offer comparable bias reduction to higher precision models, making them viable for AI personal computing devices. Despite reduced precision, gender bias scores improved by up to 46% after our mitigation strategy, suggesting that lower precision does not compromise bias mitigation effectiveness.

Moving forward, future research can extend these analyses by considering a wider array of LLMs, such as GPT-4 [42] and Claude-3 [2], as well as models with larger parameters such as Llama-3 with 70 billion parameters. This expansion could offer deeper insights into gender bias across different model architectures. Additionally, utilizing larger or more diverse datasets could enhance the generalizability of findings and reveal nuanced biases present in real-world textual corpora.

Furthermore, exploring the capabilities of LLMs beyond text generation, particularly in classification tasks, presents an opportunity for future research. Investigating gender bias in LLM-based classification tasks could shed light on bias pervasiveness across various AI applications and inform strategies for mitigation in different contexts.

In conclusion, this study advances our understanding of gender bias in LLMs and highlights the potential of lower precision and quantization techniques for bias mitigation.

6 Limitations

While this study has provided valuable insights into gender bias mitigation within LLMs, several limitations warrant consideration.

Language Scope: The study's results focus on languages with simple word structures, like English. Bias reduction methods might work differently in languages with more complex word structures, showing that more research is needed to see if they work well in different languages.

Resource Requirements: The experiments in this study required significant GPU resources, including high-performance GPUs such as NVIDIA Tesla V100 and RTX series. This demand may affect accessibility to the proposed mitigation strategies, especially for those with limited computational resources on AI PCs and AI Smartphones. Investigating more resource-efficient approaches to mitigate gender bias is a future research direction.

Generalizability: These results might not apply to all LLM setups, datasets, and uses. While they give us some understanding of how to mitigate gender bias in certain situations, more studies are necessary to confirm these methods in different contexts and language areas.

7 Ethics Statement

Our research follows ethical principles, focusing on transparent and responsible development and evaluation of AI technologies. The datasets used in this study follow privacy rules to keep sensitive information confidential. These datasets are licensed under the MIT License, which helps protect privacy. This license allows sharing data while safeguarding privacy rights. Additionally, our team follows ethical guidelines and rules to ensure participant privacy. While we focus on mitigation gender bias in LLMs, we are aware of the broader biases in AI. We understand the need to address biases that overlap with gender. We are cautious about unintended effects and work to create strategies that ensure fairness, diversity, and inclusivity.

Funding

This work received partial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

A Appendix

The appendix includes the results of our sentence similarity analysis. To evaluate the accuracy and clarity of the text, we analyzed sentence similarity. We used text from the Funpedia and Wizard of Wikipedia datasets as references for comparison. By using the BERT model and cosine similarity, we assessed how closely the generated text matched the original dataset.

Text Gen.	LLM	Funpedia	Wizard
Zero-shot	Llama - 2	0.8481	0.8305
	Llama - 3	0.7095	0.7116
	Bloom	0.6439	0.6673
	Mistral	0.6663	0.6619
	Gemma	0.7919	0.7947
Mitigation	Llama - 2	0.8512	0.8477
	Llama - 3	0.6922	0.6839
	Bloom	0.6415	0.6637
	Mistral	0.6411	0.6487
	Gemma	0.8211	0.8071

Table 7: Zero-shot and Mitigation Text Generation Sentence Similarity (32-bit Float Precision)

Text Gen.	LLM	Funpedia	Wizard
Zero-shot	Llama - 2	0.8476	0.8302
	Llama - 3	0.7095	0.7116
	Bloom	0.6469	0.6680
	Mistral	0.6655	0.6621
	Gemma	0.7097	0.7601
Mitigation	Llama - 2	0.8512	0.8468
	Llama - 3	0.6923	0.6838
	Bloom	0.6466	0.6643
	Mistral	0.6451	0.6493
	Gemma	0.7626	0.8007

Table 8: Zero-shot and Mitigation Text Generation Sentence Similarity (16-bit Float Precision)

Text Gen.	LLM	Funpedia	Wizard
Zero-shot	Llama - 2	0.8278	0.8051
	Llama - 3	0.7245	0.7274
	Bloom	0.6477	0.6658
	Mistral	0.7441	0.7269
	Gemma	0.8092	0.7892
Mitigation	Llama - 2	0.8438	0.8295
	Llama - 3	0.7069	0.7054
	Bloom	0.6462	0.6652
	Mistral	0.7373	0.7136
	Gemma	0.8097	0.7831

Table 9: Zero-shot and Mitigation Text Generation Sentence Similarity (8-bit Integer Quantization)

Text Gen.	LLM	Funpedia	Wizard
Zero-shot	Llama - 2	0.7995	0.7752
	Llama - 3	0.7132	0.7157
	Bloom	0.6476	0.6673
	Mistral	0.7881	0.7686
	Gemma	0.8064	0.7915
Mitigation	Llama - 2	0.8055	0.7790
	Llama - 3	0.7143	0.7010
	Bloom	0.6496	0.6671
	Mistral	0.7473	0.7301
	Gemma	0.8087	0.7862

Table 10: Zero-shot and Mitigation Text Generation Sentence Similarity (4-bit Integer Quantization)

References

[1] Alhafni, B.; Habash, N.; Bouamor, H. (2020). Gender-Aware Reinflection using Linguistically Enhanced Neural Models, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 139–150, 2020. https://aclanthology.org/2020.gebnlp-1.12.

[2] Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[3] Banks, J.; Warkentin, T. (2024). Gemma: Introducing new state-of-the-art open models, https://blog.google/technology/developers/gemma-open-models/.

[4] Barikeri, S.; Lauscher, A.; Vulić, I.; Glavaš, G. (2021). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models, *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)*, 1941–1955, 2021. https://doi.org/10.18653/v1/2021.acl-long.151.

[5] Bartl, M.; Leavy, S. (2022). Inferring Gender: A Scalable Methodology for Gender Detection with Online Lexical Databases, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 47–58, 2022. https://doi.org/10.18653/v1/2022.ltedi-1.7.

[6] Baumler, C.; Rudinger, R. (2022). Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3426–3432, 2022. https://doi.org/10.18653/v1/2022.naacl-main.250.

[7] Borchers, C.; Gala, D.; Gilburt, B.; Oravkin, E.; Bounsi, W.; Asano, Y. M.; Kirk, H. (2022). Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 212–224, 2022. https://doi.org/10.18653/v1/2022.gebnlp-1.22.

[8] Budzianowski, P.; Vulić, I. (2019). Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems, *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 15–22, 2019. https://aclanthology.org/D19-5602.

[9] Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs, https://arxiv.org/abs/2305.14314.

[10] Devinney, H.; Björklund, J.; Björklund, H. (2020). Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 79–92, 2020. https://aclanthology.org/2020.gebnlp-1.8.

[11] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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)*, 4171–4186, 2019. https://doi.org/10.18653/v1/N19-1423.

[12] Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872, 2021. https://doi.org/10.1145/3442188.3445924.

[13] Dinan, E.; Fan, A.; Wu, L.; Weston, J.; Kiela, D.; Williams, A. (2020). Multi-Dimensional Gender Bias Classification, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 314–331, 2020. https://aclanthology.org/2020.emnlp-main.23.

[14] Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; Weston, J. (2019). Wizard of Wikipedia: Knowledge-Powered Conversational Agents, https://arxiv.org/abs/1811.01241.

[15] Doughman, J.; Khreich, W.; El Gharib, M.; Wiss, M.; Berjawi, Z. (2021). Gender Bias in Text: Origin, Taxonomy, and Implications, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 34–44, 2021. https://aclanthology.org/2021.gebnlp-1.5.

[16] Floridi, L.; Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences, *Minds and Machines*, 30(4), 2020. https://doi.org/10.1007/s11023-020-09548-1.

[17] Gaido, M.; Savoldi, B.; Bentivogli, L.; Negri, M.; Turchi, M. (2021). How to Split: the Effect of Word Segmentation on Gender Bias in Speech Translation, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 3576–3589, 2021. https://aclanthology.org/2021.findings-acl.313.

[18] Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey, *Computational Linguistics*, 50(3), 1097–1179, 2024. https://aclanthology.org/2024.cl-3.8.

[19] Garimella, A.; Banea, C.; Hovy, D.; Mihalcea, R. (2019). Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3493–3498, 2019. https://aclanthology.org/P19-1339.

[20] Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369, 2020. https://aclanthology.org/2020.findings-emnlp.301.

[21] Hansal, O.; Le, N. T.; Sadat, F. (2022). Indigenous Language Revitalization and the Dilemma of Gender Bias, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 244–254, 2022. https://aclanthology.org/2022.gebnlp-1.25.

[22] Havens, L.; Terras, M.; Bach, B.; Alex, B. (2022). Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 30–57, 2022. https://aclanthology.org/2022.gebnlp-1.4.

[23] Hovy, D.; Bianchi, F.; Fornaciari, T. (2020). "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1686–1690, 2020. https://aclanthology.org/2020.acl-main.154.

[24] Hovy, D.; Prabhumoye, S. (2021). Five sources of bias in natural language processing, *Language and Linguistics Compass*, 15(8). https://doi.org/10.1111/lnc3.12432.

[25] Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Renard Lavaud, L.; Lachaux, M. A.; Stock, P.; Le Scao, T.; Lavril, T.; Wang, T.; Lacroix, T.; El Sayed, W. (2023). Mistral 7B, *arXiv preprint arXiv:2310.06825*, 1–9. http://arxiv.org/abs/2310.06825.

[26] Jorg, T.; Kämpgen, B.; Feiler, D.; Müller, L.; Düber, C.; Mildenberger, P.; Jungmann, F. (2023). Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing, *Insights into Imaging*, 14(1). https://doi.org/10.1186/s13244-023-01392-y.

[27] Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World, *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, 6282–6293, 2020. https://aclanthology.org/2020.acl-main.560.

[28] Jourdan, F.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. (2023). Are fairness metric scores enough to assess discrimination biases in machine learning?, *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 163–174, 2023. https://aclanthology.org/2023.trustnlp-1.15.

[29] Kiritchenko, S.; Mohammad, S. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53, 2018. https://aclanthology.org/S18-2005.

[30] Kotek, H.; Dockum, R.; Sun, D. (2023). Gender bias and stereotypes in Large Language Models, *Proceedings of the ACM Collective Intelligence Conference, CI 2023*. https://doi.org/10.1145/3582269.3615599.

[31] Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172, 2019. https://aclanthology.org/W19-3823.

[32] Le Scao, et al. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model Major Contributors Prompt Engineering Architecture and Objective Engineering Evaluation and Interpretability Broader Impacts, *arXiv*. https://arxiv.org/abs/2211.05100.

[33] Li, T.; Khashabi, D.; Khot, T.; Sabharwal, A.; Srikumar, V. (2020). UNQOVERing Stereotyping Biases via Underspecified Questions, *Proceedings of the Association for Computational Linguistics: EMNLP 2020*, 3475–3489, 2020. https://aclanthology.org/2020.findings-emnlp.311/.

[34] Liu, R.; Jia, C.; Wei, J.; Xu, G.; Wang, L.; Vosoughi, S. (2021). Mitigating Political Bias in Language Models Through Reinforced Calibration, *35th AAAI Conference on Artificial Intelligence, AAAI 2021*. https://cdn.aaai.org/ojs/17744/17744-13-21238-1-2-20210518.pdf.

[35] Lucy, L.; Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories, *Proceedings of the Third Workshop on Narrative Understanding*, 48–55, 2021. https://aclanthology.org/2021.nuse-1.5.

[36] Matthews, A.; Grasso, I.; Mahoney, C.; Chen, Y.; Wali, E.; Middleton, T.; Matthews, J.; Njie, M. (2021). Gender Bias in Natural Language Processing Across Human Languages, *TrustNLP 2021 - 1st Workshop on Trustworthy Natural Language Processing, Proceedings of the Workshop*, 48–55, 2021. https://aclanthology.org/2021.trustnlp-1.6.

[37] Meta (2024). Introducing Meta Llama 3: The most capable openly available LLM to date, *Meta AI Blog*. https://ai.meta.com/blog/meta-llama-3/.

[38] Miller, A.; Feng, W.; Batra, D.; Bordes, A.; Fisch, A.; Lu, J.; Parikh, D.; Weston, J. (2017). ParlAI: A Dialog Research Software Platform, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 79–84, 2017. https://aclanthology.org/D17-2014.

[39] Moss-Racusin, C. A.; Dovidio, J. F.; Brescoll, V. L.; Graham, M. J.; Handelsman, J. (2012). Science faculty's subtle gender biases favor male students, *Proceedings of the National Academy of Sciences of the United States of America*, 109(41), 2012. https://doi.org/10.1073/pnas.1211286109.

[40] Nangia, N.; Vania, C.; Bhalerao, R.; Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967, 2020. https://aclanthology.org/2020.emnlp-main.154.

[41] Nemani, P.; Joel, Y. D.; Vijay, P.; Liza, F. F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies, *Natural Language Processing Journal*, 6, 2024. https://doi.org/10.1016/j.nlp.2023.100047.

[42] OpenAI (2023). GPT-4 is OpenAI's most advanced system, producing safer and more useful responses, *OpenAI Blog*. https://openai.com/gpt-4.

[43] Opitz, J.; Frank, A. (2018). Addressing the Winograd Schema Challenge as a Sequence Ranking Task, *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 41–52, 2018. https://aclanthology.org/W18-4105.

[44] Park, B.; Janecek, M.; Ezzati-Jivan, N.; Li, Y.; Emami, A. (2024). Picturing Ambiguity: A Visual Twist on the Winograd Schema Challenge, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 355–374, 2024. https://aclanthology.org/2024.acl-long.22.

[45] Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering, *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105, 2022. https://doi.org/10.18653/v1/2022.findings-acl.165.

[46] Plank, B.; Hovy, D.; Søgaard, A. (2014a). Learning part-of-speech taggers with inter-annotator agreement loss, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–751, 2014a. https://doi.org/10.3115/v1/E14-1078.

[47] Plank, B.; Hovy, D.; Søgaard, A. (2014b). Linguistically debatable or just plain wrong?, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 507–511, 2014b. https://doi.org/10.3115/v1/P14-2083.

[48] Rudinger, R.; Naradowsky, J.; Leonard, B.; Van Durme, B. (2018). Gender Bias in Coreference Resolution, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14, 2018. https://doi.org/10.18653/v1/N18-2002.

[49] Savoldi, B.; Gaido, M.; Bentivogli, L.; Negri, M.; Turchi, M. (2021). Gender Bias in Machine Translation, *Transactions of the Association for Computational Linguistics*, 9, 845–874, 2021. https://doi.org/10.1162/tacl_a_00401.

[50] Sengupta, B.; Maher, R.; Groves, D.; Olieman, C. (2021). GenBiT: measure and mitigate gender bias in language datasets, *Microsoft Journal of Applied Research*, 16, 63–71, 2021. https://www.microsoft.com/en-us/research/publication/genbit-measure-and-mitigate-gender-bias-in-language-datasets/.

[51] Song, L.; Xu, K.; Zhang, Y.; Chen, J.; Yu, D. (2020). ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5429–5434, 2020. https://doi.org/10.18653/v1/2020.acl-main.482.

[52] Stanovsky, G.; Smith, N. A.; Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684, 2019. https://doi.org/10.18653/v1/P19-1164.

[53] Stanczak, K.; Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing, *arXiv preprint arXiv:2112.14168*, 2021. https://arxiv.org/abs/2112.14168.

[54] Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640, 2019. https://doi.org/10.18653/v1/P19-1159.

[55] Takeshita, M.; Katsumata, Y.; Rzepka, R.; Araki, K. (2020). Can Existing Methods Debias Languages Other than English? First Attempt to Analyze and Mitigate Japanese Word Embeddings, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 44–55, 2020. https://aclanthology.org/2020.gebnlp-1.5.

[56] Thakur, H.; Jain, A.; Vaddamanu, P.; Liang, P. P.; Morency, L.-P. (2023). Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 340–351, 2023. https://doi.org/10.18653/v1/2023.acl-short.30.

[57] Touvron et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models, *arXiv preprint arXiv:2307.09288*, 2023. https://arxiv.org/abs/2307.09288.

[58] Ungless, E.; Rafferty, A.; Nag, H.; Ross, Björn. (2022). A Robust Bias Mitigation Procedure Based on the Stereotype Content Model, *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, 207–217, 2022. https://aclanthology.org/2022.nlpcss-1.23.

[59] Valentini, F.; Rosati, G.; Blasi, D.; Fernandez Slezak, D.; Altszyler, E. (2023). On the Interpretability and Significance of Bias Metrics in Texts: a PMI-based Approach, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 509–520, 2023. https://doi.org/10.18653/v1/2023.acl-short.44.

[60] Voytovich, L.; Greenberg, C. (2022). Natural Language Processing: Practical Applications in Medicine and Investigation of Contextual Autocomplete, *Acta Neurochirurgica, Supplementum*, 134, 2022. https://doi.org/10.1007/978-3-030-85292-4_24.

[61] Wang, Z.; Chakravarthy, A.; Munechika, D.; Chau, D. H. (2024). Wordflow: Social Prompt Engineering for Large Language Models, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 42–50, 2024. https://aclanthology.org/2024.acl-demos.5.

[62] Ye, Q.; Ahmed, M.; Pryzant, R.; Khani, F. (2024). Prompt Engineering a Prompt Engineer, *Findings of the Association for Computational Linguistics ACL 2024*, 355–385, 2024. https://aclanthology.org/2024.findings-acl.21.

[63] Yu, J.; Kim, S. U. G.; Choi, J.; Choi, J. D. (2024). What Is Your Favorite Gender, MLM? Gender Bias Evaluation in Multilingual Masked Language Models, *Information*, 15(9), 549, 2024. https://doi.org/10.3390/info15090549.

[64] Yuan, S.; Maronikolakis, A.; Schütze, H. (2022). Separating Hate Speech and Offensive Language Classes via Adversarial Debiasing, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 1–10, 2022. https://doi.org/10.18653/v1/2022.woah-1.1.

[65] Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J. A.; Celi, L. A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D. W.; Abdulnour, R. E. E.; Butte, A. J.; Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study, *The Lancet Digital Health*, 6(1), 2024. https://doi.org/10.1016/S2589-7500(23)00225-X.

[66] Zhang, Y.; Li, S.; Deng, C.; Wang, L.; Zhao, H. (2024). Think Before You Act: A Two-Stage Framework for Mitigating Gender Bias Towards Vision-Language Tasks, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 773–791, 2024. https://doi.org/10.18653/v1/2024.naacl-long.44.

[67] Zhao, J.; Mukherjee, S.; Hosseini, S.; Chang, K.-W.; Awadallah, A. (2020). Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer, *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, 2896–2907, 2020. https://doi.org/10.18653/v1/2020.acl-main.260.

[68] Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20, 2018. https://doi.org/10.18653/v1/N18-2003.

[69] Zong, Z.; Hong, C. (2018). On Application of Natural Language Processing in Machine Translation, *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 506–510, 2018. https://doi.org/10.1109/ICMCCE.2018.00112.

Cite this paper as:

Zhou, H.; Inkpen, D.; Kantarci, B. (2024). Evaluating and Mitigating Gender Bias in Generative Large Language Models, *International Journal of Computers Communications & Control*, 19(6), 6853, 2024.
https://doi.org/10.15837/ijccc.2024.6.6853