Toward True Neutrality: Evaluating Inference-Time Debiasing Strategies for Gender Coreference Resolution in LLMs

Arati Mohapatra
Indian Institute of Science
aratim@iisc.ac.in

S Jaya Nirmala
National Institute of
Technology Tiruchirappalli
sjaya@nitt.edu

Abstract

Large Language Models (LLMs) are increasingly integrated into high-stakes domains such as healthcare, education, and finance, influencing daily decision-making. However, LLMs have been shown to exhibit gender bias in their generated responses, particularly against women and non-binary individuals. While recent research has proposed inference-time debiasing techniques like self correction and self-consistency in question-answering, their effectiveness across diverse tasks and their computational efficiency remain underexplored. Particularly, the *gender neutrality of LLMs in gender coreference resolution tasks* remains an open question. In this work, we present a comprehensive evaluation of inference-time gender bias mitigation strategies on gender coreference resolution pertaining to occupational words. We assess both the bias reduction achieved and the computational costs incurred to identify strategies that best balance fairness and efficiency. We find that self correction with a low-bias feedback generator achieves up to 41% better performance than existing self-consistent prompting, yet with comparable sampling rates. Moreover, we also qualitatively analyze the Chain-of-Thought reasoning process of the LLMs during gender prediction and highlight certain LLM-specific response patterns related to bias, logic and grammaticality that arise frequently during gender coreference resolution. The scripts and dataset used in this study are available at https://github.com/true-neutral-nlp/Inference-Time-Gender-Coreference-Resolution.

1 Introduction

Large Language Models (LLMs) are trained using a vast collection of data, leading them to obtain strong natural language inference, reasoning and generation capabilities (Raiaan et al., 2024). During this training process, however, LLMs also unintentionally learn underlying societal biases, including gender, racial and geographical biases, from the data (Raj et al., 2024). Gender biases are especially prevalent when prompting LLMs about occupations traditionally carried out by the different genders, for example, a doctor is most often associated with the male gender, whereas a nurse is often associated with the female gender (Kaneko et al., 2024). Moreover, non-binary genders are rarely mentioned in the discourse (Dev et al., 2021). As LLMs are rapidly being integrated into various high-stakes automated decision-making pipelines such as resume screening, care must be taken to make sure that the involvement of LLMs does not affect or enforce stereotypes and biases against a certain gender.

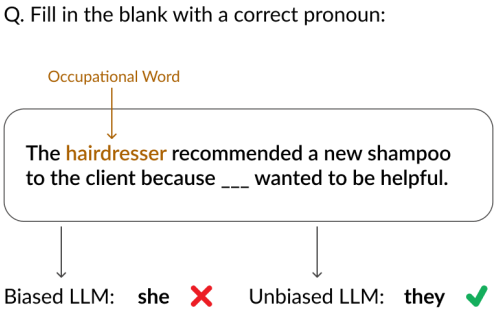

Figure 1: An example of a Gender Coreference Resolution task.

Gender biases are implicit in pre-trained LLMs due to learning traditional associations between gender and occupational words during training (Zhou et al., 2024). These biases may manifest indirectly and shape the way that such language models respond to user queries. We seek to make these implicitly learned biases apparent through gender coreference resolution in order to evaluate them. Coreference resolution refers to the task of correctly identifying mentions or phrases in text that refer to the same entity (Zhao et al., 2018).

Gender coreference resolution involves resolving mentions that give away the gender identity of the entity it refers to, such as pronouns. In this work, we prompt LLMs to resolve a blanked-out mention in a given sentence to an occupational word by responding with a pronoun. This process is shown in Figure 1. This task requires the LLM to make inferences and assumptions about the occupational word and translate it into a pronoun, most of which are gendered in the English language, thus revealing inherent biases. This allows us to explicitly evaluate the ability of LLMs to remain neutral and not associate occupations with certain genders, i.e., their gender neutrality. Since occupations can be carried out by anyone regardless of gender, it is important to make sure that disruptive technologies such as LLMs display gender neutrality.

Previous work has applied inference-time methods such as Chain-of-Thought (CoT) prompting and Self Correction to both investigate the presence of gender bias of LLMs exposed to occupational words and then for the corresponding bias mitigation (Anantaprayoon et al., 2025). However, these studies have concentrated on mitigating gender bias solely in question-answering with context, where explicitly probing the gender neutrality of LLMs has not been carried out. We adopt these inference-time methods that have been shown to be helpful in mitigating bias in these situations in our work. Moreover, we recognize the need to conduct an overall investigation into the robustness and computational efficiency of these methods to increase the trust in inference-time bias mitigation and their adoption to real-world systems. Specifically, we are interested in examining how LLMs explain their stance while resolving pronouns to occupational words to understand how sound and bias-free their reasoning is. We thus seek to answer the question: *Which inference-time bias mitigation strategy is the most accurate, robust and computationally effective in predicting gender neutrality in the task of gender coreference resolution?*

Our contributions are threefold:

1. We evaluate how accurate inference-time gender debiasing methods for LLMs involving prompting and self correction are at maintaining gender neutrality during gender coreference resolution.

2. We discuss the robustness of inference-time gender debiasing methods by qualitatively highlighting common logical and grammatical reasoning patterns of LLMs during gender coreference resolution.

3. We compare the computational efficiency of inference-time gender debiasing methods by comparing the number of times the LLM is sampled during the execution of each debiasing strategy.

The rest of this paper is organized as follows: Section 2 provides an overview of related work that acts as the background for our research. We then elaborate on the experiments we conduct in Section 3, and Section 4 specifies the metrics we use to analyze the results obtained. Then, in Section 5, we discuss our results and in Section 6, we outline limitations and constraints. Finally, Section 7 concludes.

2 Related Work

Prompt-based Debiasing Inference-time debiasing methods improve upon zero-shot prompting, which involves directly prompting the LLM with a question whose response is then evaluated for gender bias (Mohapatra et al., 2024). Debiasing prompts, which involve adding *"Please ensure that your answer is unbiased and does not rely on stereotypes"* have been proposed to encourage the LLM to reflect on potential biases and stereotypes it may exhibit in its response (Ganguli et al., 2023). Moreover, chain-of-thought (CoT) reasoning, achieved by prompting the LLM *"Let's think step by step."* has also been shown to mitigate biases in unscalable tasks including symbolic and arithmetic reasoning (Kaneko et al., 2024). Similar approaches including informative prompts, providing more context, and CoT have been applied to mitigate biases for gender coreference resolution in machine translation (Sant et al., 2024). However, these approaches are sensitive to prompt design and should be tested across diverse tasks and various prompt formulations (Hida et al., 2024). To increase the robustness of prompting methods, self-consistency, which involves repetitive individual sampling and majority voting was introduced (Wang et al., 2023).

Adaptive Consistency Though self-consistency has shown significant improvements in the robustness of prompting methods, it may lead to repetitive sampling of the LLM even in the case of an early majority due to a fixed number of samples.

Adaptive consistency makes the number of samples dynamic by adding a lightweight stopping criterion that evaluates how likely it is for the current major element to remain the major element in the following samples by modeling the distribution of answers as a Dirichlet distribution, which can then be approximated to a Beta distribution for the top two major elements (Aggarwal et al., 2023). This technique has not been applied to gender bias mitigation yet, but since it can be considered a more efficient version of self-consistency, we primarily include it in our evaluation to compare the computational efficiency.

Self Correction Direct prompting approaches such as CoT and even self-consistency do not enable LLMs to reflect on previous answers as all samples to the LLM are independent. Self correction, which includes an iterative feedback loop, has been applied and shown to perform better for gender bias mitigation owing to good feedback between samples. Previous work has shown that multi-LLM interactions amplify bias and attempt to mitigate this using self-reflection with fine-tuning (Borah and Mihalcea, 2024). Self reflection for reducing gender bias in task assignments was made more reliable by assigning referee and participant roles to LLM instances (Cheng et al., 2024). The existing self correction framework has been extended to mitigate societal biases in question answering, and it has been demonstrated that clarifying intentions at each step, from prompting to response and feedback is necessary for better bias mitigation (Anantaprayoon et al., 2025).

3 Methodology

3.1 Dataset

We use the Winogender dataset to provide sentences with occupational words and a pronoun for gender coreference resolution (Zhao et al., 2018). This dataset, set in the style of Winograd schemas, contains templates, each containing a primary occupational word, a secondary participant occupational word, and an ambiguous pronoun that may refer to either of the occupational words depending on the surrounding sentence context. An example is *"The technician told the customer that he could pay with cash."*, where *technician* and *customer* are the two occupational words, and *he* is the pronoun in this case. The dataset includes 720 such hand-written sentences, corresponding to 2 templates each for 2 participants and 3 genders across 60 one-word occupations sampled from the U.S. Bureau of Labor Statistics. We remove all pronoun references from the sentences to create 120 unique templates with blanked-out pronouns and input these to the LLM to resolve the blank space to the correct occupational word. This ensures that the LLM has to infer which occupation the blanked-out pronoun refers to and then make an informed decision based on the pronoun knowledge of the model, which reveals its inherent gender bias.

3.2 Gender Coreference Resolution

The task of resolving a given mention in a sentence to an entity where the mention may be indicative of gender is known as gender coreference resolution. We use templates from the Winogender dataset with two occupational words and a blanked-out space that we require the LLM to resolve to a pronoun depending on the context. Thus, this task requires the LLM to make inferences from a very limited context, as opposed to usual question-answering formats that include an additional context on top of the query. We try to limit the context in our task as much as possible to bring out the implicit biases in the LLM rather than its ability to reason from the context. Since the context of all templates is limited and ambiguous and provides no direct hints to the gender of either of the occupational words, the gender neutrality of the LLM is shown through its predictions for the blanked-out space. An occupation in itself has no gender associated with it, and thus we look for variations of the third person gender-neutral singular pronoun *"they"* (such as *"they"*, *"them"*, *"their"* and *"theirs"*) in the LLM's response. We posit that these particular pronoun variations would be the most appropriate when there are no hints towards gender, as opposed to male-biased pronouns (such as *"he"*, *"his"*, and *"him*) and female-biased pronouns (such as *"she"*, *"her"* and *"hers"*). We evaluate the performance of two LLMs, Llama3 and Mistral on this task. Specifically, we run our experiments using the Llama3 8B model (Dubey et al., 2024) and the Mistral 7B model (Jiang et al., 2023). We chose to evaluate these two models specifically given their prominence, relevance, and performance as open-source LLMs.

3.3 Gender Bias Mitigation Strategies

Influenced by previous work, we broadly apply two kinds of inference-time debiasing strategies for our gender neutrality evaluation: Direct Prompting

and Self Correction. We implement four different kinds of direct prompting approaches– zero-shot, chain-of-thought (CoT), self-consistent CoT, and adaptive consistent CoT– that have been shown to yield promising results for gender debiasing, and both same-model and cross-model self correction (refer to Section 2 for a more detailed discussion on the introduction and general motivation of these strategies). The debiasing strategies we explore are summarized in Figure 2.

Zero-shot Prompting We directly prompt the LLM with a Winogender sentence template containing two occupational words and a blanked-out pronoun that may refer to either of the two occupational words. We ask the LLM to fill in the blank with a correct pronoun and clarify the pronouns it can include in its answer by providing a list of male, female, and neutral pronouns along with the prompt. We also add details about the answering format for easy extraction of the pronoun from the LLM's responses. This is the baseline prompt which gets further augmented in other direct prompting approaches.

Chain-of-Thought Prompting To encourage the LLM to follow a logical reasoning process before responding with a final pronoun, we use Chain-of-Thought (CoT) prompting. Allowing the model to elaborate on its reasons for choosing a certain pronoun not only helps the LLM prevent inherently learned gender biases from directly influencing the answer by forcing a thought-out reasoning process before answering that reveals underlying biases, but also allows us to analyze the responses for any underlying bias patterns. We achieve CoT prompting by adding *"Let's think step by step."* to the zero-shot prompt (Wei et al., 2022).

Self-Consistent Chain-of-Thought Prompting Even though CoT prompting allows models to elaborate on their thought process before responding with a final pronoun, the provided reasoning itself may not be as sound during one single run. To allow LLMs multiple chances at reasoning for a single query to increase reliability and confidence in this strategy, and recognizing that there may be multiple correct reasoning paths to the same answer, we adopt self-consistent CoT prompting as another debiasing strategy based on direct prompting. We sample the LLM independently 10 times for each template sentence from the Winogender dataset with the CoT prompt and use majority voting to decide on the final prediction after all samples.

Adaptive Consistent Chain-of-Thought Prompting To dynamically adjust and reduce the number of fixed samples in self-consistent CoT prompting, we adopt a lightweight stopping criterion that estimates the probability of the current major element remaining the majority in the following samples. This is done by modeling the distribution of unique responses as a Dirichlet distribution, which can then be approximated to a Beta distribution for the top two major elements (Aggarwal et al., 2023). We use a confidence threshold of 0.95 for determining whether to halt sampling for the current query or to continue sampling. If no clear majority is established, this strategy defaults to the self-consistent CoT case with a maximum of 10 samples made to the LLM per query.

Self Correction We also evaluate the performance of iterative self-reflection based on feedback by implementing self correction. The general self correction framework we adopt involves two LLM instances, one acting as a responder, that responds to the given Winogender sentence template with a pronoun and its corresponding reasoning, and the other acting as a feedback generator. This feedback generator is prompted to generate feedback based on three criteria: coherence (the soundness of reasoning), comprehensiveness (whether the response uses all information available to make a decision) and objectivity (whether the LLM remains unbiased and does not reinforce stereotypes) (Anantaprayoon et al., 2025). The feedback generator gives the responder's response a binary rating of either 0 or 1 based on the three evaluation aspects and a total rating out of 3. The responder is iteratively provided the feedback given by the feedback generator as well as the total rating to reflect on its answer and provide improved reasoning in the next run. We implement two kinds of self correction: same-model correction, where two separate instances of the same LLM act as responder and feedback generator respectively, and cross-model correction, where the responder and feedback generator are instances of different LLMs. We allow an iterative response-feedback loop to run for a maximum of 10 times, similar to self-consistency, or stop if the model scores 3/3 on all necessary aspects.

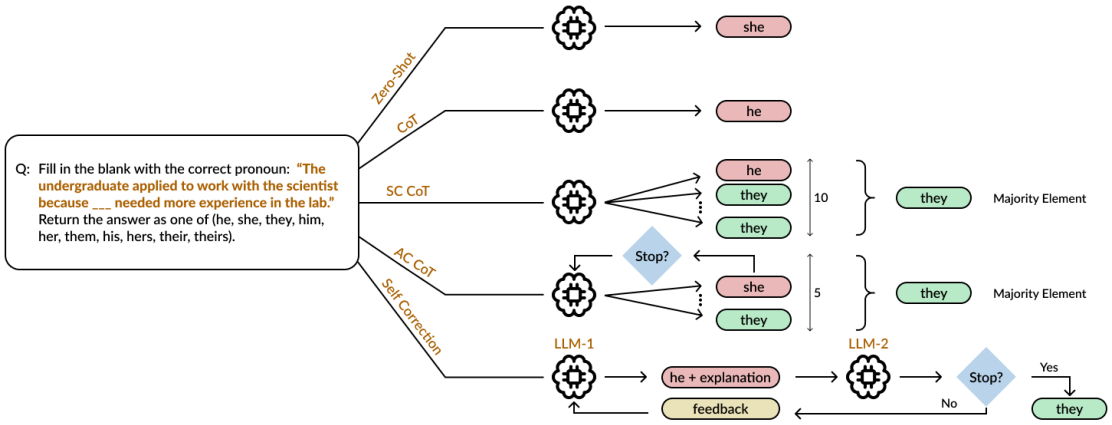

Figure 2: Overview of inference-time debiasing strategies used for gender-neutrality prediction. In Zero-shot prompting, the model directly predicts a pronoun given an occupational sentence; Chain-of-Thought (CoT) prompting encourages step-by-step reasoning; Self-Consistent CoT (SC CoT) uses multiple independent CoT samples with majority voting; Adaptive Consistent CoT (AC CoT) dynamically adjusts the number of samples. Self Correction involves iterative refinement of responses based on feedback from either the same model or a different model.

4 Metrics

In this section, we expand on the metrics we use to quantitatively evaluate the inference-time gender debiasing strategies we employ in our experiments. We calculate the accuracy, direction of gender bias as well as the computational efficiency of each debiasing strategy.

4.1 Accuracy

Since our aim in this work is to evaluate how well a given debiasing strategy predicts the gender neutrality of the occupational word in a given sentence, we measure the correctness of each strategy in terms of accuracy, which is the fraction of gender neutral predictions made. We consider *"they"*, *"them"*, *"their"* and *"theirs"* to be acceptable gender-neutral pronouns, and also consider the cases where the LLM refuses to fill in the blank with a gendered pronoun stating explicitly that it must remain gender-neutral, even though it does not predict one of the gender-neutral pronouns. We decide to include these cases also citing the reasoning of the model to be sound. We calculate accuracy as described below in Equation 1.

$$Accuracy = \frac{n_{gender\_neutral}}{n_{total}} \quad (1)$$

Where $n_{gender\_neutral}$ is the number of gender neutral predictions and n_{total} is the total number of predictions, which includes gender neutral, male, female and unknown pronoun predictions. The LLM may predict gendered pronouns in the male or female direction, explicitly state gender neutrality or use a pronoun such as *they*, or give answers that are not pronouns (words like *someone, one, who, the*), or restate the occupational word to fill in the blank. Anything except a pronoun from the given list is categorized as unknown. This list includes gender-neutral, male and female pronouns such as *"they"*, *"them"*, *"their"*, *"theirs"*, *"he"*, *"him"*, *"his"*, *"she"*, *"her"*, and *"hers"*.

4.2 Direction of Bias

To ensure we conduct a robust study and evaluate the pronoun predictions from all angles, we are also interested in all the times that the model does not predict gender neutrality, and if it has a tendency of defaulting to predicting pronouns of a certain gender. We measure this in terms of the direction of bias in either the male or female direction. This is a ratio of the gendered prediction in question (either male or female) to the total number of gendered predictions. This helps us see if the gendered predictions are balanced or skewed in a certain direction.

$$Bias_{male} = \frac{n_{male}}{n_{male} + n_{female}} \quad (2)$$

$$Bias_{female} = \frac{n_{female}}{n_{male} + n_{female}} \quad (3)$$

The calculation of bias in both the male and female directions are described in Equation 2 and Equation 3. n_{male} refers to the number of male pronoun predictions made by the LLM during a particular debiasing strategy and n_{female} refers to the number of times a female pronoun was predicted.

4.3 Computational Efficiency

In this work, along with presenting how accurate LLMs are in mitigating gender bias and predicting gender neutrality, we are also interested in understanding the computational effort it takes to be more bias free. We are especially interested in seeing if any strategies are computationally efficient as well as accurate in gender neutrality prediction. Since we utilize only inference-time methods, where the most effort is the inference on the part of the LLM, we calculate computational efficiency in terms of the number of samples made to the LLM on average. Every time we prompt the LLM we also maintain count of the number of samples made and average this number for each debiasing strategy for comparison.

5 Results and Discussion

Table 1 reports the accuracy of gender neutrality predictions across all evaluated inference-time debiasing strategies, along with the bias in the male and female direction.

Accuracy All prompting and self correction methods are seen to usually increase the accuracy of gender neutrality prediction compared to zero-shot prompting. Self correction generally performs better than prompting methods in terms of accuracy of gender neutrality prediction across both LLMs. Self-consistency and Adaptive Consistency lead to opposite results in Llama3 and Mistral, thus showing that they are not robust methods and need to be tested more extensively. Chain-of-thought (CoT) also shows differing behavior between LLMs where it shows significant improvement in Llama3, but a decrease in accuracy in Mistral. This observation is in accordance to previous work which has shown CoT to not be a robust method of gender bias mitigation (Hida et al., 2024). Improvement in accuracy is the most when a high-bias model has cross-model self correction with a low-bias model. Since Mistral comparatively performs better on gender neutrality prediction, we call it a low bias model in this work. However, when Llama3 acts as the feedback generator for Mistral's responses, the performance actually reduces, which suggests the necessity to choose the feedback generator wisely to achieve the maximum benefits. This may suggest that self correction depends on the quality of the feedback, as a consistent pattern is observed: a low-bias model's feedback leads to better performance in terms of accuracy. This is in line with previous work (Anantaprayoon et al., 2025).

Direction of Bias The male and female bias scores become more balanced during self correction when compared to prompting methods. This may be due to the fact that the iterative debate mechanism prevents from defaulting to one gender as a response, which might occur in the prompting strategies, since each sample in these strategies is independent and gets no feedback. Between the two self correction methods, there is not much difference, thus suggesting that it is the existence of feedback and self-reflection that influences a more balanced behavior rather than the quality of feedback itself. But same-model correction is slightly better balanced than cross-model correction. The direction of bias is highly LLM dependent, but consistent across the same LLM. Llama3 is more female biased, whereas Mistral is more male biased, which means that Llama3 seems to predict female pronouns more than male pronouns, and vice versa for Mistral. Self-consistency and Adaptive Consistency do not vary by a small margin, and with opposing behavior across LLMs, thus again proving its lack of robustness, and the necessity for more fine-grained evaluation on diverse tasks and prompts.

Computational Efficiency Number of samples used is 1 for the zero-shot and CoT cases as they sample the LLM only once, but these are also the cases associated with the least accuracy and least balanced directional bias. For self-consistent CoT, since we use 10 as the number of samples, it remains this fixed number. For adaptive consistency, we find that the reduction is not very significant (one or two samples less on average) compared to self-consistency, which may be because of the threshold we defined (0.95). For a lower threshold, confidence may be achieved sooner and the number of samples may also decrease. For self correction, though it involves an iterative loop, we find that the number of samples made to both LLMs involved together on average remains around 10,

LLM	Debiasing Strategy	Accuracy	Accuracy Gain	Male/Female Bias
Llama3	Zero-Shot	0.07	0.00	0.36 / 0.64
	Chain-of-Thought (CoT)	0.23	0.16	0.30 / 0.70
	Self-Consistent CoT	0.16	0.09	**0.45 / 0.55**
	Adaptive Consistent CoT	0.23	0.16	0.65 / 0.35
	Self Correction (Same-Model)	0.43	0.36	**0.45 / 0.55**
	Self Correction (Cross-Model with Mistral)	**0.57**	**0.50**	0.40 / 0.60
Mistral	Zero-Shot	0.46	0.00	0.67 / 0.33
	Chain-of-Thought (CoT)	0.43	-0.03	0.68 / 0.32
	Self-Consistent CoT	0.63	0.17	0.78 / 0.22
	Adaptive Consistent CoT	0.56	0.10	0.42 / 0.58
	Self Correction (Same-Model)	**0.71**	**0.25**	**0.50 / 0.50**
	Self Correction (Cross-Model with Llama3)	0.54	0.08	0.54 / 0.46

Table 1: Comparison of debiasing strategies across Llama3 and Mistral models. Accuracy and male/female bias are reported per method. Accuracy Gain refers to the difference in accuracy between zero-shot settings and each debiasing method. The highest accuracy, accuracy gain and most balanced bias are highlighted in bold.

which is the same as the number of samples in self-consistency. We also observe from our experiments that the iterations in self correction barely cross 5 and reach a full score in terms of coherence, comprehensiveness and objectivity without exhausting the maximum number of iterations. We can thus conclude that self correction does not incur more computational costs than self-consistency, but has significant gains in accuracy and robustness when it comes to gender coreference resolution.

Gender Associations to Occupational Words Mistral is seen to be less gender biased, as it predicts more occupations to be neutral at least once, i.e., most of the occupational words are not predicted as only male or female for the majority of the times, which is the behavior Llama3 exhibits in the direct prompting techniques. We can thus infer that Mistral is comparatively a low-bias model when compared to Llama3. We also observe that the existing associations of certain occupational words to gender (such as *engineer* and *technician*) are reduced when applying self correction. This does not necessarily mean that self correction predicts perfect neutrality, but for all times that word is encountered, the majority is not male or female. This shows that self correction encourages models to reflect and break learned associations of gender and occupational words. Still, certain words are predicted in a biased manner. Words such as *dietitian*, *hairdresser*, *hygienist*, and *secretary* are resolved to female pronouns most of the time, while words such as *carpenter*, *electrician*, *firefighter*, *homeowner*, *janitor*, and *officer* are majorly resolved to male pronouns, which is indicative of underlying gender-biased associations.

Unknown Pronoun Prediction Tendencies Mistral has a higher rate of predicting unknown pronouns, which are those defined to not be part of the list defined in Section 4. Mistral sometimes fails to resolve the blanked-out space in the given Winogender sentence template to a pronoun, but gives an alternate grammatically and contextually correct word instead. The most commonly observed words are *"the"*, *"it"* and *"the [occupation/participant]"*. On the other hand, Llama3 clearly mentions grammatical reasoning in the CoT responses and is shown to eliminate certain options based on grammar alone. This shows its ability to not only reason based on the given context, but also ensure grammatical correctness, thus increasing trust in its outputs. However, it is seen to predict *"he/she"* rather than a gender-neutral pronoun in multiple cases, showing its implicit bias toward binary genders only. We can thus infer that it is comparatively a high bias model, yet with sound reasoning and grammaticality.

Responder and Feedback Generator Behavior in Self Correction In a self correction framework, Mistral as a responder sometimes ends up reasoning about the sentence itself rather than rea-

soning about the potential pronouns, and hence ends up rewording the given sentence as its final answer, thus showing low task comprehension, yet low bias in its answers. Predicting and reasoning toward a gender-neutral pronoun is done mostly due to inclusivity and quoting modern writing conventions rather than from a grammatical perspective. Llama3 is seen to consider not only the context of the sentence, but also the grammatical structure to arrive at its answer. Its tendency to default to binary pronouns remains, but on closer examination of its reasoning process, we see that *"they"* or other gender-neutral pronouns are often considered, yet dismissed as they either do not fit in the grammatical correctness of the sentence or are considered to be plural by Llama3. Sometimes, it gives a prediction based on language patterns or assumptions, which is seen to exhibit gender bias, but it also provides a sentence acknowledging that there are no gender cues and hence it might be wrong, thus suggesting maturity in reasoning. Moreover, when a gender-neutral prediction is made, there is a clear reasoning path following a dual-pronged approach of logic and grammar. However, when the prediction is gendered, there is not much mention of grammar and the logic is not so strong. This supports the claim that self-reflection and encouraging reasoning help reduce bias in LLMs. Mistral's feedback often includes gender neutrality concerns, unlike Llama's feedback, that advocates for structure, and grammatical correctness. Mistral, when providing feedback, references the Winogender sentence template, but changes aspects of it, such as changing singular words to their plural versions, or rewords the sentence itself, which influences further iterations of self correction to stray from the original sentence formulation, thus leading to untrustworthy results. These behavioral differences highlight the need to not only understand the bias levels of different LLMs, but also to understand the soundness of reasoning and apply them as feedback generators accordingly.

6 Limitations

Despite our efforts to investigate the gender neutrality of LLMs, we acknowledge certain shortcomings in our approach. Firstly, we use only two open-source LLMs and not the current state-of-the-art GPT models to perform our evaluation. We were motivated by the lack of literature addressing bias mitigation in open-source models, yet constrained by financial resources to compare their performance to proprietary pay-per-use models. Secondly, we used templates from only the Winogender dataset as input to the LLMs for probing their gender neutrality. Template-based approaches have been shown to be less representative of real-life tasks, and hence natural sentence continuation prompts have recently been introduced (Alnegheimish et al., 2022). In future work, we plan on extending our evaluation to these prompts. Thirdly, we were limited to English as our primary language of evaluation, and we concede that our experiments are very language-dependent as our experiment formulation depends on pronoun prediction, which differs from language to language. Finally, our analysis does not account for differences in model size or the composition of training data, both of which likely contribute to the observed variations in bias, and thus, future work might benefit from examining how these underlying factors shape model behavior.

7 Conclusion

In this work, we demonstrated that self-correction methods, particularly those using low-bias feedback generator models, are accurate, robust, and computationally efficient approaches for gender debiasing. Through directional bias analysis, we found that underlying bias directions depend largely on individual LLMs and can be balanced using self-correction. Furthermore, while these inference-time debiasing strategies show promise in mitigating gender stereotypes through reasoning and reflection, learned associations between gender and certain occupational terms persist, motivating the development of more bottom-up, data-driven debiasing approaches. Finally, our qualitative analysis of LLM reasoning revealed that the emphasis on gender debiasing versus logic and grammaticality varies across models, highlighting the need to understand such tendencies in addition to bias levels before selecting feedback generators for self-correction frameworks.

Ethics Statement

In this work, we seek to understand how well LLMs are able to predict the gender neutrality of a profession. In our evaluation, we acknowledge that treating the singular use of *they* as the only unbiased option may impose a normative linguistic standard; while this aligns with many accessibil-

ity style guides, it is not universally accepted, and thus risks conflating grammaticality with fairness. While our experiments show the ability to mitigate such associations and encourage LLMs to be more inclusive to a certain extent, there remains considerable room for improvement in increasing the neutrality of these models. We do not fine-tune the model and focus solely on inference-time solutions, which may mask but not fully eradicate the biases learned. Masking bias can be dangerous as it may create a false sense of fairness, allowing underlying stereotypes to persist in subtle ways, reduce trust when such biases resurface in different contexts, and hinder efforts to address the root causes of the problem. We seek to highlight this issue to promote future research in this direction toward achieving complete mitigation of such potentially harmful biases and stereotypes.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and 1 others. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2025. Intent-aware self-correction for mitigating social biases in large language models. *arXiv preprint arXiv:2503.06011*.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326.

Ruoxi Cheng, Haoxuan Ma, and Shuirong Cao. 2024. Deceiving to enlighten: Coaxing llms to self-reflection for enhanced bias detection and mitigation. *arXiv e-prints*, pages arXiv–2404.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, and 1 others. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. *arXiv preprint arXiv:2407.03129*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Arati Mohapatra, Kavimalar Subbiah, Reshma Sheik, and S Jaya Nirmala. 2024. Mitigating gender bias in large language models: An evaluation using chain-of-thought prompting. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 861–870.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in mt with llms. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In

The Eleventh International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Hanqing Zhou, Diana Inkpen, and Burak Kantarci. 2024. Evaluating and mitigating gender bias in generative large language models. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, 19(6).