Adapting Psycholinguistic Research for LLMs: Gender-inclusive Language in a Coreference Context

Marion Bartl[1,2] **Thomas Brendan Murphy**[1,3] **Susan Leavy**[1,2]
[1] Insight SFI Research Centre for Data Analytics
[2] School of Information and Communication Studies
[3] School of Mathematics and Statistics
University College Dublin

Correspondence: marion.bartl@insight-centre.org

Abstract

Gender-inclusive language is often used with the aim of ensuring that all individuals, regardless of gender, can be associated with certain concepts. While psycholinguistic studies have examined its effects in relation to human cognition, it remains unclear how Large Language Models (LLMs) process gender-inclusive language. Given that commercial LLMs are gaining an increasingly strong foothold in everyday applications, it is crucial to examine whether LLMs in fact interpret gender-inclusive language neutrally, because the language they generate has the potential to influence the language of their users. This study examines whether LLM-generated coreferent terms align with a given gender expression or reflect model biases. Adapting psycholinguistic methods from French to English and German, we find that in English, LLMs generally maintain the antecedent's gender but exhibit underlying masculine bias. In German, this bias is much stronger, overriding all tested gender-neutralization strategies.

1 Introduction

Over the last few decades, activism by feminist linguists has led to increased use of gender-neutral or gender-fair wording, especially in grammatical gender languages such as French or German (Usinger and Müller, 2024; Burnett and Pozniak, 2021). The aim of these forms is to alleviate masculine-default bias and establish representation for people with non-binary gender identities (Freed, 2020). Psycholinguistic studies have shown that gender-neutral alternatives can increase the visibility of women and non-binary people (Tibblin et al., 2023; Fatfouta and Sczesny, 2023).

As Large Language Models (LLMs) are embedded into everyday systems and are used as writing assistants and content creators, the language they generate can have an impact on equal treatment and linguistic representation of women and non-binary people[1]. However, despite the fact that gender bias in NLP has been examined from many different angles (Gupta et al., 2024a), gender-inclusive language in the context of LLMs has only begun to be investigated (Bartl and Leavy, 2024; Watson et al., 2025, a.o.). The processing of gender-inclusive vs. gendered language remains under-explored in English LLMs (Watson et al., 2023) and, to our knowledge, entirely unexamined in German LLMs. To address this, we compare the processing of gendered and gender-inclusive language in both English, a notional gender language, and German, a grammatical gender language.

We adapt a psycholinguistic study by Tibblin et al. (2023) to explore how the presence of masculine, feminine or neutral gender in one sentence influences (1) the likelihood of a reference to that gender in a subsequent sentence and (2) the gender mentioned in an LLM-generated completion. We find that while English LLMs generally keep antecedent and coreferent gender consistent, they are unlikely to use *they* as a singular pronoun and contain underlying masculine bias. The German LLM we tested showed a strong preference for masculine coreferents, regardless of the gender or gender-inclusive strategy used in the antecedent phrase. We also find evidence that German gender-inclusive language strategies increase the probability of feminine and neutral gender. This finding encourages us to believe that the use of gender-inclusive over generic masculine expressions in German LLMs has the potential to diversify gender representation.

Contributions This study translates psycholinguistic methodologies to LLMs, enabling com-

[1]Following Monro (2019), we use *non-binary* as "an umbrella term that includes those whose identity falls outside of or between male and female identities; as a person who can experience both male and female, at different times, or someone who does not experience or want to have a gender identity at all."

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 451–467

parisons between human and model reasoning. It introduces a novel approach to assessing whether gender-inclusive expressions promote gender-neutral interpretations within LLMs[2]. Additionally, it provides the first analysis of German gender-inclusive strategies in this context, showing that they partially achieve their intended effects by increasing associations with feminine and neutral gender, aligning with psycholinguistic findings.

2 Bias Statement

In this work, we define *gender bias* in an LLM as the tendency to assign higher likelihoods to gendered linguistic forms when referring to an entity that was initially introduced in a gender-neutral way. This behavior can result in *representational harms* (Blodgett et al., 2020): specifically, if masculine forms are used to refer to previously introduced gender-neutral nouns which describe a group or person of unknown gender, women and non-binary people are excluded from representation. Such linguistic erasure can reinforce their marginalization in society (Pauwels, 2003; Dev et al., 2021; Ovalle et al., 2023).

3 Background

The field of **feminist psycholinguistics** is concerned with evaluating human biases related to language. Studies have shown how masculine generics are in fact not interpreted generically (Noll et al., 2018), and that changing the language to be gender-inclusive also increases mental representation for women and non-binary people (Sato et al., 2025; Mirabella et al., 2024). The term *gender-inclusive language* describes linguistic strategies and neologisms to eliminate male-as-norm bias (*chairman→chairperson*) and emphasize alternative terms that do not reinforce a heteronormative, binary model of gender (*husband/wife→spouse*).

Large Language Models (LLMs) have also been shown to exhibit various social biases, including gender bias (Gupta et al., 2024b). However, few studies ha explored the **processing of gender-inclusive language within LLMs**. There are two main areas of investigation: gender-inclusive role nouns (*fire fighter*, *chairperson*, etc.) and gender-neutral pronouns such as singular *they*. The present research addresses both.

To investigate the processing of **gender-inclusive role nouns in LLMs**, Watson et al. (2023) adapted a psycholinguistic study on sentence acceptability judgments and social attitudes for BERT (Papineau et al., 2022; Devlin et al., 2019). They first calculated BERT's relative probability of a given masculine, feminine or neutral role noun (e.g. *fireman/firewoman/fire fighter*) within a sentence context. BERT's responses were then connected to the social attitudes of the human participants giving the same responses. The researchers found that BERT aligned most with people who had moderate to conservative views.

There are several studies examining **gender-neutral pronouns in LLMs**. Brandl et al. (2022) draw on psycholinguistic research into Swedish neopronouns and adapted an eye-tracking study for LLMs. They demonstrated that while humans do not have trouble processing neopronouns (Vergoossen et al., 2020), they are associated with greater processing difficulty in LLMs. Correspondingly, models also have lower pronoun fidelity for feminine and singular *they* pronouns, meaning that they are less likely to use them even if they were introduced alongside a corresponding entity (Gautam et al., 2024). When comparing an LLM's processing of singular *they* in a generic sense vs. referring to a specific person, models have less trouble with generic *they* (Baumler and Rudinger, 2022). In terms of social attitudes, BERT's likelihood to generate singular *they* resembled the judgments of participants with low to moderate acceptance of non-binary gender (Watson et al., 2023).

Psycholinguistic studies that were previously adapted for LLMs, including the research this paper is based on, often contain *anaphora*. Anaphora is defined "in a looser sense, [as] any relation in which something is understood in the light of what precedes it" (Matthews, 2014). The preceding term is the *antecedent*, while the referring term is the *coreferent*. The resolution of this relationship, finding the corresponding antecedent for a coreferent, is a large research field within NLP. **Coreference Resolution** (CR) is is relevant for downstream NLP tasks such as named entity recognition, summarization or question answering (Liu et al., 2023). CR systems have previously been shown to exhibit gender bias, relying on stereotypes instead of syntactic information or real-world gender distributions (Rudinger et al., 2018; Kotek et al., 2023).

To evaluate CR systems for gender biases, challenge datasets based on the Winograd

[2]Code and data are openly available at https://github.com/marionbartl/GIL-coref-context

schema (Levesque et al., 2012) were developed (Rudinger et al., 2018; Zhao et al., 2018). These datasets contain instances in which a pronoun must be resolved to refer to one of two previously mentioned entities, such as in the sentence "The paramedic performed CPR on the passenger even though *she/he/they* knew it was too late." (Rudinger et al., 2018). While most challenge datasets contain a single sentence, and assess the resolution of singular pronouns, this research focuses on coreference between two different sentences in both singular and plural.

In **German**, the issue of gender-inclusive language is more intricate than in English. German marks nouns, articles and adjectives for masculine, feminine or neuter gender, traditionally using masculine forms as the generic. Similar to English, masculine generics have a predominantly masculine interpretation (Fatfouta and Sczesny, 2023), which is also reflected in NLP models trained on German text (Schmitz et al., 2023). To increase women's visibility and/or take gender out of the equation, feminist scholars pushed for linguistic strategies to make role nouns more inclusive (Sczesny et al., 2016; Dick et al., 2024). In NLP, there have been efforts to automate the integration of these strategies into text (Amrhein et al., 2023), as well as research on gender-neutral machine translation into German (Lardelli et al., 2024b,a). However, it is unclear how German gender-fair language is processed by an LLM and we aim to provide some initial answers to this issue in this paper.

4 Methodology

In order to uncover how LLMs process gender-inclusive in contrast to gendered language, we adapted Tibblin et al.'s (2023) study design of sentence pairs containing antecedent and coreferent phrases (§4.1). We used several LLMs (§4.2) for our experiments on measuring the probability of specific gendered or gender-neutral terms (§4.3) and analyzing the gender contained in model generations (§4.4).

4.1 Dataset Creation

We adapted a study design with 44 sentence pairs by Tibblin et al. (2023). The French sentences in this study design were translated into English and German using ChatGPT and manually verified. Each instance in the dataset contains two subsequent phrases. Phrase 1 contains an *antecedent*, a plural noun phrase that is either gendered (*kings*, *au pair girls*) or gender-neutral (*oenologists*, *volunteers*). Phrase 2 contains as the *coreferent* the noun *men* or *women*. The content of the phrases can be coherent (1a) or incoherent (1b).

(1) a. *The **midwives** were entering the hospital. Given the good weather, some of the (**women**|**men**) were not wearing jackets.*

b. *The **referees** were watching the match in the rain. Because of the good weather, most of the **men** were wearing shorts.*

Using the 11 incoherent instances (cf. 1b) vs. taking them out had little impact on the outcome of our initial experiments, we therefore retained all 44 instances for experiments measuring coreferent probability. Translating the data into English did not always retain the original gender of the antecedent (*Hôtesses de l'air*$_{fem}$ – *flight attendants*$_{neut}$). The original data moreover contained imbalanced numbers of gendered/gender-neutral antecedents, which was undesirable for our analysis. We therefore decided to use the data as templates. A template consists of two phrases, the first one with a placeholder for an antecedent, the second with a placeholder for a coreferent.

4.1.1 Data for Measuring Coreferent Probability

English Our final English dataset comprises 13,464 instances for the plural (PL) condition and 14,652 instances for the singular (SG) condition. The PL dataset includes 34 antecedent triplets, each paired with three coreferent nouns—*men*, *women*, and *people*—across 44 templates. The SG dataset consists of 37 antecedent triplets, each paired with the pronouns *he*, *she*, and *they*, across 44 templates. To collect the English antecedents, we utilized gendered terms and their neutral replacements from Bartl and Leavy (2024), selecting terms that shared the same neutral equivalent for both masculine and feminine forms (e.g. *swordswoman–swordsman–fencer*). Any triplets that were semantically implausible within our template context (e.g., *humankinds*) were manually excluded. This resulted in 34 verified triplets for the PL condition and 37 for the SG condition.

German The final German dataset comprises 10,560 instances, constructed from 10 antecedents, each having eight gender-inclusive variations,

lang.	number	phrase 1	phrase 2
EN	PL	The (*sportsmen* | *sportswomen* | *athletes*) were waiting on the steps.	It was obvious that some of the (*men* | *women* | *people*) were in a really good mood.
	SG	The (*sportsman* | *sportswoman* | *athlete*) was waiting on the steps.	It was obvious that (*he* | *she* | *they*) (was | were) in a really good mood.
DE	PL	Die (*Tierärzte* | *Tierärztinnen* | *Tierärztinnen und Tierärzte* | *Tierärzte und Tierärztinnen* | *TierärztInnen* | *Tierärzt*innen* | *Tierärzt:innen* | *Tierärzt_innen*) warteten auf den Stufen.	Es war offensichtlich, dass einige (*Männer* | *Frauen* | *Leute*) wirklich guter Laune waren.

Table 1: Examples of antecedent and coreferent combinations for English and German experiments. The templates for English and German are the same, the German antecedents translate to *veterinarian*.

#	strategy	DE example	EN translation
1	masculine	*Akademiker*	$\text{academics}_{\text{masc}}$
2	feminine	*Akademikerinnen*	$\text{academics}_{\text{fem}}$
3	coordinated (masc. first)	*Akademiker und Akademikerinnen*	$\text{academics}_{\text{masc}}$ and $\text{academics}_{\text{fem}}$
4	coordinated (fem. first)	*Akademikerinnen und Akademiker*	$\text{academics}_{\text{fem}}$ and $\text{academics}_{\text{masc}}$
5	capital I	*AkademikerInnen*	$\text{academics}_{\text{mascFem}}$
6	colon	*Akademiker:innen*	$\text{academics}_{\text{masc:fem}}$
7	asterisk	*Akademiker*innen*	$\text{academics}_{\text{masc*fem}}$
8	underscore	*Akademiker_innen*	$\text{academics}_{\text{masc\_fem}}$

Table 2: Examples of different strategies for gender-inclusive language in German.

paired with three coreferent nouns (*Männer* '*men*', *Frauen* '*women*', and *Personen* '*persons*') across 44 templates. To ensure a truly gender-neutral antecedent noun phrase, we maintained coreferent pairs in the plural form, as the German singular inherently marks gender through its article. Instead of translating the English triplets we used professions from the French data to avoid data expansion, given that each antecedent in English had only three variations, whereas German antecedents had eight (Table 5 in Appendix A). The German gender-inclusive strategies used are outlined in Table 2: we include masculine and feminine forms for reference (strategies 1 and 2), as well as strategies that express both masculine and feminine gender (strategies 3–5) or incorporate non-binary genders (strategies 6–8). The latter use characters such as the gender star (*), colon (:), or underscore (_; Dick et al., 2024).

4.1.2 Data for Coreferent Generation

In the second set of experiments, we used the models to generate the continuation of Phrase 2 instead of measuring the probability of specific coreferents. The final dataset for coreferent generation comprised 630 instances for English and 160 instances for German. We worked with heavily reduced datasets to minimize annotation workloads and reduce variability in the generations. The English dataset was reduced by using the 33 templates with coherent phrases (Example (1a)) and selecting a reduced set of seven high-frequency plural triplets (Table 3). For German, we used the same ten antecedent terms in eight gender variations (§4.1.1) with 2 coherent templates.

4.2 Models

We used six English and one German LLM in the experiments (Table 4 in Appendix A). The models were selected to enable comparison between model sizes and performances. For the English experiments we used GPT-2 (Radford et al., 2019) as a baseline, allowing for comparability due to its widespread use in prior research. We also tested an adaptation of GPT-2 by Bartl and Leavy (2024), which was fine-tuned with gender-neutral data in order to mitigate gender stereotyping in the model. This model is particularly relevant because our experiments assess how gender-neutral language is processed by LLMs. It can therefore provide insights into how a model that has seen additional gender-neutral language would process gender-neutral language differently. We also tested the 1B, 7B and 13B models from the OLMo suite (Groeneveld et al., 2024a), which are fully open-source, improving transparency for the research community. The different sizes allow us to show the impact of model size on the processing of gendered language. Qwen2.5 (32B) (Yang et al., 2024) was included as our largest model and the best performing pre-trained single-model LLM on the huggingface OpenLLM Leaderboard[3] at the time of experimentation (December 2024) within the hardware limitations of our institution.

[3] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard/

4.3 Measuring Coreferent Probability

We used the LLMs to predict the joint two phrases up to the coreferent (*men/women/people*), and then obtained the log probability of the coreferent ($\log(p)$) from the probability distribution over the vocabulary. For split coreferents, we took the probability of the first component token. Averaging the probabilities of all component tokens would have inflated probabilities, as each component serves as a strong predictor for the subsequent token.

4.4 Coreferent Generation and Annotation

We used the models to generate eight tokens for English and ten for German.The generated continuations were then annotated for gender of the entity mentioned, and whether the mentioned entity was a coreferent of the antecedent in the first sentence.

English Three annotators were recruited out of a pool of PhD researchers at our institution. Two were native and one was a fluent English speaker. All annotators were paid €60 for 630 items of annotation, each with two labels per item (`gender` and `coreference`). The annotation guidelines can be found in Figure 4 in the Appendix.

Fleiss' kappa was calculated to assess inter-annotator agreement. For the `gender` labels, the annotations showed $\kappa = 0.757$. For the `coreference` labels, the annotators reached a slightly lower score of $\kappa = 0.671$. This is not surprising given that coreference labeling might have been complicated by mentions of several entities or ambiguous phrasing, among others. However, both of these scores are in the range of "substantial agreement", according to Landis and Koch (1977). We then calculated the final `gender` and `coreference` labels based on the majority label. Instances for which all three annotators provided different labels were labeled as `NULL`. There were 22 `NULL` labels for `gender` and eight `NULL` labels for the presence of `coreference`.

German (pilot) Due to the lack of German-speaking annotators one of the authors, a linguist and native speaker of German, annotated the German sentence completions in a pilot experiment. Each completion was annotated for mentioned gender and presence of a coreferent to the antecedent.

5 Results

This section lays out the results for our experiments on coreferent probability and coreferent generation. For each of these, we will first present the English and then the German results.

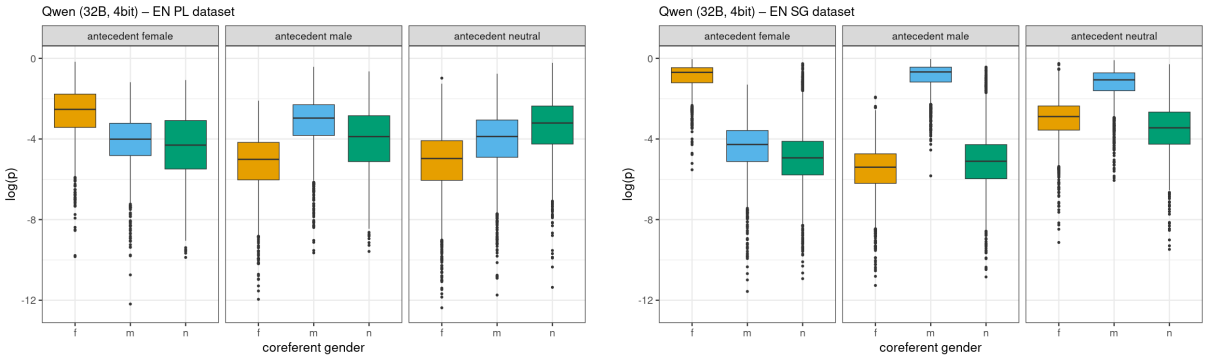
5.1 Coreferent Probability

English For our English results, we provide illustrations for and discuss Qwen-2.5 in detail, as it is the largest and best performing model of those we evaluated. Its results would therefore mirror most closely state-of-the-art models. However, the results for all English models (except the fine-tuned model) follow similar patterns. We provide results and illustrations for the other models, such as the OLMo suite (Figure 5), and the fine-tuned GPT-2 (Figure 6) in Appendix B.

We performed a two-way ANOVA on the coreferent probabilities produced by Qwen-2.5 (and all other models, cf. Table 6 in the Appendix), testing the effect of antecedent and coreferent gender on the probability of the coreferent. Effect sizes were labeled following Field et al.'s (2012) recommendations. The ANOVA showed that in the PL setting, the main effect of antecedent gender is statistically significant and small ($F(2, 13455) = 138.59$, $p < .001$; $\eta^2 = 0.02$, 95% CI [0.02, 1.00]), which also applied to the main effect of coreferent gender ($F(2, 13455) = 178.33$, $p < .001$; $\eta^2 = 0.03$, 95% CI [0.02, 1.00]). The interaction between antecedent and coreferent gender is statistically significant and large ($F(4, 13455) = 809.94$, $p < .001$; $\eta^2 = 0.19$, 95% CI [0.18, 1.00]). This indicates that in the coreference constructions we are investigating, the probability of the coreferent is most influenced by the correspondence between antecedent and coreferent gender.

Figure 1 illustrates the distribution of coreferent probability for the English Qwen-2.5 model in both PL and SG setting. In the PL setting, the model behaves as expected, producing the highest coreferent probability when antecedent gender and coreferent gender correspond (e.g. *The **bowmen** were going down the street. Some of the **men** were in a good mood.*). However, for feminine antecedents, masculine coreferents have the second highest probability, indicating masculine bias in the model. The Tukey post-hoc test showed a 21% lower probability for neutral than masculine coreferents following feminine antecedents (F:N/F:M[4] = $e^{-0.236} \approx 0.79$, $p < .001$). This masculine bias is also evident for neutral antecedents. Here, the Tukey post-hoc test showed a probability that was three times higher

[4]This notation indicates antecedent gender before and coreferent gender after the colon.



(a) plural

(b) singular

Figure 1: Distribution of $\log(p)$ of coreferent gender by antecedent gender

for masculine than feminine coreferents following neutral antecedents (N:M/N:F = $e^{1.107} \approx 3.03$, $p < .001$).

The SG setting (Figure 1b) is similar to the PL in that matching antecedent and coreferent gender result in the highest probability for masculine and feminine coreferents, for which we used the pronouns *he* and *she*, respectively. Similar to the PL, *he* as a coreferent had a 31% higher probability than the neutral coreferent *they* for a feminine antecedent (Tukey post-hoc test: F:N/F:M = $e^{-0.37} \approx 0.69$, $p < .001$), pointing either to masculine bias in the model, or the possibility that singular *they* is not well-recognized or accepted by the LLM. This phenomenon can also be observed for neutral antecedents, after which the masculine coreferent *he* has the highest probability, followed by *she* and singular *they*. In fact, the Tukey post-hoc test showed that masculine coreferents following a neutral antecedent had an 88% higher probability than neutral coreferents (N:N/N:M = $e^{-2.16} \approx 0.12$, $p < .001$). This result shows that the pronoun *they* is not fully accepted by the model as a singular pronoun.

German The effects of antecedent gender, coreferent gender, and their interaction on the probability of the coreferent as predicted by Leo Mistral 7B was tested with a two-way ANOVA, as with the English models. Effect sizes were labeled following Field et al.'s (2012) recommendations. The main effect of antecedent gender for the German model is statistically significant and small ($F(7, 10536) = 42.74, p < .001; \eta^2 = 0.03$, 95% CI [0.02, 1.00]), and the main effect of coreferent gender is statistically significant and large ($F(2, 10536) = 2601.35$, $p < .001$; $\eta^2 = 0.33$, 95% CI [0.32, 1.00]). The interaction between antecedent and coreferent gender is statistically significant and small ($F(14, 10536) = 36.63, p < .001$; $\eta^2 = 0.05$, 95% CI [0.04, 1.00]).

In the German ANOVA, contrary to the English results, coreferent gender is the biggest predictor for coreferent probability and not the interaction term. These results become more clear when looking at the probability distributions in Figure 2: the masculine continuation *Männer* 'men' always shows a much higher probability than *Frauen* 'women' and *Personen* 'persons'. Therefore, the ANOVA results show coreferent gender to be more predictive than the interaction term.

It can also be seen in Figure 2 that all German gender-inclusive language strategies lead to an increase in the probability of feminine and gender-neutral coreferents. In the ANOVA results, this finding is supported by the small interaction between antecedent and coreferent gender. The highest probability for the feminine coreferent can be seen with a feminine antecedent, which is somewhat expected. The second highest probability of a feminine coreferent is brought about by the asterisk strategy, which could be due the feminine PL suffix *-innen* contained in this strategy. However, the capital-I, colon and underscore strategies also contain *-innen*. Feminine coreferents generally have the second-highest probability for all gender-inclusive language strategies we tested, meaning that neither strategy favors the generation of *Personen* 'persons' as a gender-neutral coreferent.

5.2 Coreferent Generation

English As discussed in Section 4.4, we used majority voting over our three annotation labels to generate the final labels. Out of 630 sentence completions, 396 (62.86%) were labeled as containing a

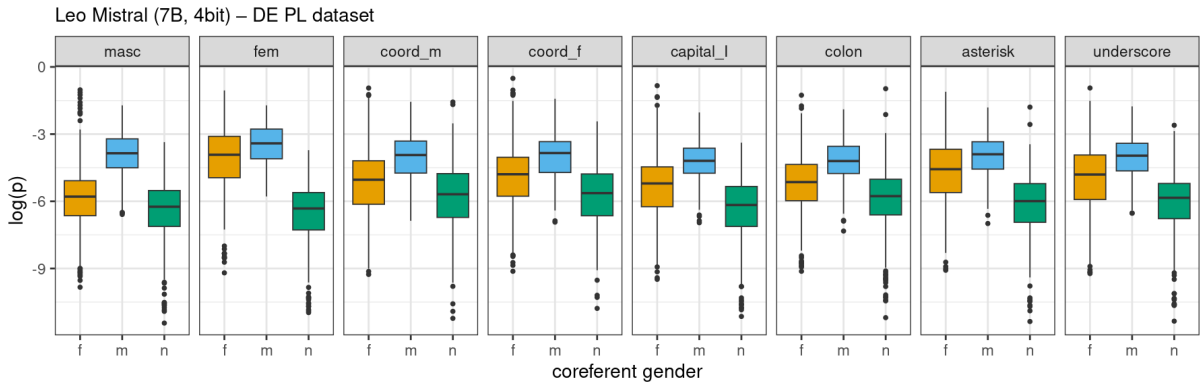


Figure 2: Effect of different gender-inclusive language strategies on coreferent gender probability

coreferent of the antecedent, 226 (35.87%) were labeled as not containing a coreferent, and 8 (1.27%) instances were inconclusive (labeled NULL).

We ran χ^2 tests of independence for both the coreference and no-coreference groups, which were statistically significant ($p < .001$). Effect sizes were labeled following Funder and Ozer's (2019) recommendations. In the coreference group, the effect of antecedent gender is very large, ($\chi^2 = 739.57$, $p < .001$; Adjusted Cramer's v = 0.96, 95% CI [0.90, 1.00]). In the no coreference group, the effect of antecedent gender is medium ($\chi^2 = 40.12$, $p < .001$; Adjusted Cramer's v = 0.28, 95% CI [0.16, 1.00]).

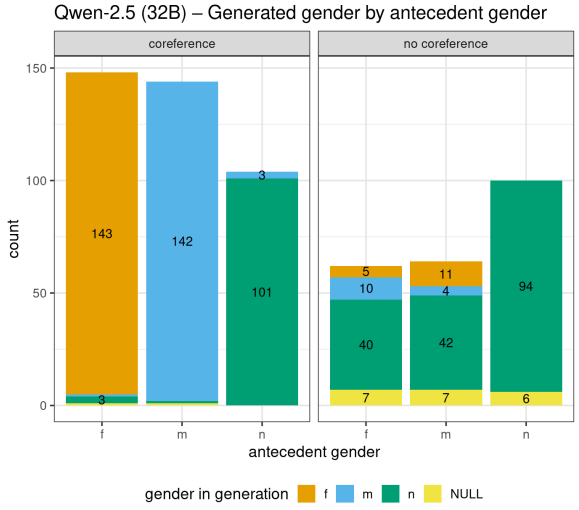


Figure 3: Gender mentioned in the sentence continuation, split by whether or not the generation contains a coreferent of the antecedent

The distribution of coreferent genders based on antecedent gender and divided by whether or not the continuation contains coreference is illustrated in Figure 3. Figure 3 shows that if the model generates a coreferent, the coreferent gender follows the antecedent gender with an overwhelming majority. However, the model generates a coreferent less often when the antecedent is neutral than when it is masculine or feminine. In cases where the continuation does not contain a coreferent of the antecedent, neutral entities are generated most often. There are also some generations of feminine gender following a masculine antecedent, and vice versa. This is likely due to prevalence of couplets such as *husband/wife*. Thus, when Phrase 1 mentions *husbands*, Phrase 2 is likely to mention *wives*.

German (pilot) The results for the pilot experiments on German coreferent generation are illustrated in Figure 7 in Appendix B. The data are divided into instances where a coreferent noun was generated vs. when there was not. Out of the 160 instances labeled, 100 (62.5%) contained a coreferent, and 60 (37.5%) did not. These proportions of generations with and without the coreferent mirror those obtained for English (§5.2).

The Pearson's χ^2 test of independence between antecedent gender and generated coreferent gender suggests that the effect is statistically significant, and very large for the group in which a coreferent was generated ($\chi^2 = 171.79$, $p < .001$; Adjusted Cramer's v = 0.72, 95% CI [0.56, 1.00]). For the group in which no coreferent was generated, the χ^2 test also showed a statistically significant and very large effect ($\chi^2 = 70.88$, $p < .001$; Adjusted Cramer's v = 0.54, 95% CI [0.20, 1.00]).

Figure 7 shows that similar to the English results (Figure 3), masculine and feminine coreferents are mostly generated when the antecedent is masculine or feminine. However, feminine antecedents seem to be a clearer predictor for feminine coreferents, while there are some instances in which a neutral

coreferent is generated following a masculine antecedent. Moreover, gender-inclusive antecedents generally invoke gender-neutral coreferents (Figure 7), which is the intention of using these strategies. One specific case is that coordinated masculine and feminine forms (Table 2, #3 & #4) of the antecedents invoke coordinated coreferents, indicating a model tendency to keep using the same gender form in Phrase 2 that it has seen in Phrase 1.

6 Discussion

Both experiments on measuring coreferent probability and generation of coreferents demonstrated that generally, models tend to match coreferent gender to the antecedent gender. However, there are several caveats to this observation. For English models, whether or not the gender of the coreferent aligns with the antecedent depends on whether the sentences are singular or plural. Our English coreferent probability experiments in the singular setting (Figure 1b) showed that when the antecedent is neutral, the masculine pronoun *he* has the highest probability instead of *they*, meaning that models struggle to interpret the pronoun *they* as a singular pronoun. This finding was also reported by Gautam et al. (2024). In language generation applications, this might contribute to the erasure of people of non-binary gender who use *they/them* pronouns, as well as reinforce male-as-norm biases when people of unknown gender are referenced with masculine pronouns (Cao and Daumé, 2021).

Furthermore, in the English plural experiments the most probable coreferent gender generally follows the gender of the antecedent. However, the second- and third-highest gender probabilities paint a more nuanced picture (Figure 1). For both feminine and neutral antecedents, masculine coreferents are second-most likely. This illustrates bias, because an equitable model would display similar probabilities for feminine and masculine coreferents given a gender-neutral antecedent. For feminine antecedents, it would also assign higher probabilities to neutral over masculine coreferents. Thus, while the model prioritizes gendered context clues — a desirable behavior — it still exhibits an underlying masculine default bias.

This masculine bias was not just underlying but clearly visible in our German experiments. Measuring the probability of specific coreferents showed that *Männer* '*men*' always had a higher probability than either the feminine coreferent *Frauen* '*women*' or neutral coreferent *Personen* '*persons*'. This important finding shows that gender bias in the model outweighs information it received in the prompt, which might lead to a reinforcement of male-as-norm bias through a likely prevalence of masculine terms in the output. It is important to note, however, that the coreferent generation experiments for German did not show masculine bias to the same extent as the coreferent probability experiments. This might have been due to the model often simply repeating the antecedent phrase in the generations. In our coreferent probability experiments coreferent terms differed from the antecedent phrases.

One encouraging finding from the German experiments is that, despite masculine gender having the highest probability, gender-inclusive strategies help increase the probability of feminine and neutral coreferents. This supports one of the aims of using gender-inclusive language: to allow equal association of all genders with respective terms. Our findings clearly illustrate that the model we used does not show this equal association, however, it is promising that the use of gender-fair language can increase the probability of an association with gender-neutral and feminine terms. This finding mirrors the result of psycholinguistic studies into the effects of gender-inclusive language on humans (Tibblin et al., 2023; Sczesny et al., 2016).

7 Conclusion

This research adapted Tibblin et al.'s (2023)'s psycholinguistic experiments on the effects on gender-fair language on anaphora resolution to the domain of LLMs. We investigated how the use of gendered or gender-inclusive language within one sentence influences the generation of language in consecutive sentences. Our findings indicate that while English LLMs are likely to continue to use the gender of a mentioned entity in a subsequent sentence, there is an underlying prevalence for masculine gender. For German, this bias appears more pronounced, with masculine gender always having the highest probability in spite of feminine or neutral gender information in the previous sentence. However, with reference to Tibblin et al.'s (2023) findings, gender-inclusive language strategies in German also increase the probability of feminine and gender-neutral referents. This research therefore supports the value of using gender-inclusive language in an LLM context, especially in underrepresented languages like German.

8 Limitations

There are several limitations to our work. Firstly, we conducted **pilot experiments for German** coreferent generation due to a lack of annotators. The annotations for a small set of instances (160 sentence pairs, based on two out of 44 templates) were provided by one of the authors, who is a German native speaker and trained linguist. The reliance on a single annotator may introduce bias, however, the smaller sample size compared to English reduces the risk of variation. Moreover, 23% of the coreferent generations simply repeated the antecedent gender, supporting consistent gender assignment. Future work will address this issue by involving multiple annotators and expanding the number of templates and instances.

Secondly, the **types of models** covered mainly included smaller LLMs (1.5–32 billion parameters) due to hardware restrictions at our institution. In contrast, recently released DeepSeek-V3, contains a total of 671B parameters (DeepSeek-AI, 2024). Future research is needed to determine whether our findings hold for these larger models.

A third limitation is the **number of coreferents** tested. While we varied the antecedents, we used the same coreferents (PL: *women* (DE: *Frauen*), *men* (DE: *Männer*), *people* (DE: *Personen*); SG: *she*, *he*, *they*). This was done to follow the original setup by Tibblin et al. (2023). However, in LLMs it would also have been possible to measure the probability of several coreferent candidates. Still, our coreferent generation experiments partially alleviate this bias because they are based on the tokens with the highest probability.

Finally, we showed how LLMs handle gender-inclusive expressions from one sentence to another. However, LLMs often handle **longer contexts and exchanges**. Therefore, future research should be conducted in a setting with a longer context.

References

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting Biased Models to De-bias Text: A Gender-Fair Rewriting Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

Connor Baumler and Rachel Rudinger. 2022. Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Heather Burnett and Céline Pozniak. 2021. Political dimensions of gender inclusive writing in Parisian universities. *Journal of Sociolinguistics*, 25(5):808–831.

Yang Trista Cao and Hal Daumé, III. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Anna-Katharina Dick, Matthias Drews, Valentin Pickard, and Victoria Pierz. 2024. GIL-GALaD: Gender Inclusive Language - German Auto-Assembled Large Database. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 7740–7745, Torino, Italia. ELRA and ICCL.

Ramzi Fatfouta and Sabine Sczesny. 2023. Unconscious Bias in Job Titles: Implicit Associations Between Four Different Linguistic Forms with Women and Men. *Sex Roles*, 89(11):774–785.

Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Sage publications.

Alice F. Freed. 2020. Women, Language and Public Discourse: Five decades of sexism and scrutiny. In *Innovations and Challenges: Women, Language and Sexism*. Routledge. Num Pages: 16.

David C. Funder and Daniel J. Ozer. 2019. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168. Publisher: SAGE Publications Inc.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024a. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024b. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024a. Sociodemographic Bias in Language Models: A Survey and Forward Path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024b. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, pages 12–24, New York, NY, USA. Association for Computing Machinery.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Publisher: International Biometric Society.

Manuel Lardelli, Timm Dill, Giuseppe Attanasio, and Anne Lauscher. 2024a. Sparks of fairness: Preliminary evidence of commercial machine translation as English-to-German gender-fair dictionaries. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 12–21, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Manuel Lardelli, Anne Lauscher, and Giuseppe Attanasio. 2024b. GeFMT: Gender-fair language in German machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 37–38, Sheffield, UK. European Association for Machine Translation (EAMT).

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

P. H. Matthews. 2014. *The concise Oxford dictionary of linguistics*, third;3rd; edition. Oxford University Press, Oxford.

Marta Mirabella, Claudia Mazzuca, Chiara De Livio, Bianca Di Giannantonio, Fau Rosati, Maric Martin Lorusso, Vittorio Lingiardi, Anna M. Borghi, and

Guido Giovanardi. 2024. The Role of Language in Nonbinary Identity Construction: Gender Words Matter. *Psychology of sexual orientation and gender diversity*. Publisher: Educational Publishing Foundation.

Surya Monro. 2019. Non-binary and genderqueer: An overview of the field. *International Journal of Transgenderism*, 20(2-3):126–131. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15532739.2018.1538841.

Jane Noll, Mark Lowry, and Judith Bryant. 2018. Changes Over Time in the Comprehension of He and They as Epicene Pronouns. *Journal of Psycholinguistic Research*, 47(5):1057–1068.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1246–1266, New York, NY, USA. Association for Computing Machinery.

Brandon Papineau, Rob Podesva, and Judith Degen. 2022. 'Sally the Congressperson': The Role of Individual Ideology on the Processing and Production of English Gender-Neutral Role Nouns. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Anne Pauwels. 2003. Linguistic Sexism and Feminist Linguistic Activism. In Janet Holmes and Miriam Meyerhoff, editors, *The Handbook of Language and Gender*, pages 550–570. Blackwell Publishing Ltd, Oxford, UK.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Sayaka Sato, Pascal Mark Gygax, Ute Gabriel, Jane Oakhill, and Lucie Escasain. 2025. Does Inclusive Language Increase the Visibility of Women, or Does It Simply Decrease the Visibility of Men? A Missing Piece of the Inclusive Language Jigsaw. *Collabra: Psychology*, 11(1):128470.

Dominic Schmitz, Viktoria Schneider, and Janina Esser. 2023. No genericity in sight: An exploration of the semantics of masculine generics in German. *Glossa Psycholinguistics*, 2(1).

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7(Journal Article):25–25. Place: Switzerland Publisher: Frontiers Research Foundation.

Julia Tibblin, Jonas Granfeldt, Joost van de Weijer, and Pascal Gygax. 2023. The male bias can be attenuated in reading: on the resolution of anaphoric expressions following gender-fair forms in French. *Glossa Psycholinguistics*, 2(1).

Johanna Usinger and Philipp Müller. 2024. Geschickt gendern - das Genderwörterbuch.

Hellen P. Vergoossen, Philip Pärnamets, Emma A. Renström, and Marie Gustafsson Sendén. 2020. Are New Gender-Neutral Pronouns Difficult to Process in Reading? The Case of Hen in SWEDISH. *Frontiers in Psychology*, 11. Publisher: Frontiers.

Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023. What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6790–6809, Toronto, Canada. Association for Computational Linguistics.

Julia Watson, Sophia S. Lee, Barend Beekhuizen, and Suzanne Stevenson. 2025. Do language models practice what they preach? examining language ideologies about gendered language reform encoded in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1201–1223, Abu Dhabi, UAE. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint*. ArXiv:2407.10671 [cs].

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, New Orleans, Louisiana. Association for Computational Linguistics.

A Data

number	neutral	feminine	masculine
PL	grandparents	grandmothers	grandfathers
	monarchs	queens	kings
	siblings	sisters	brothers
	parents-in-law	mothers-in-law	fathers-in-law
	parents	mothers	fathers
	children	daughters	sons
	spouses	wives	husbands

Table 3: High frequency English antecedents

lang.	model name	# parameters
EN	GPT2	1.5B
	GPT2 fine-tuned	1.5B
	OLMo	1B, 7B, 13B
	Qwen2.5	32B
DE	LeoLM Mistral[8]	7B

Table 4: Overview of LLMs used

[8] https://huggingface.co/jphme/em_german_leo_mistral

B Results

B.1 Model Size Comparison

Figure 5 shows the probability distributions for three OLMo models (Groeneveld et al., 2024b) of 1B, 7B and 13B parameters. Overall, the three models show similar distributions for all three antecedent genders that follow those discussed for the Qwen2.5 32B model (Figure 1): the highest probabilities are obtained when antecedent and coreferent gender match, and masculine gender has the second-highest probability for both neutral and feminine antecedent. The probabilities for masculine coreferents across all antecedents are highest for the smallest, 1B parameter model, which could indicate that masculine bias is highest for this model.

B.2 Models Fine-tuned with Gender-inclusive Language

Figure 6 presents the results for Bartl and Leavy's (2024) fine-tuned GPT-2 models. The models were fine-tuned for 3 epochs with an English corpus in which gendered terms were rewritten with gender-neutral variants and gendered singular pronouns (*he, she*) were replaced with singular *they*. The effects of pronoun replacement are clearly visible in the SG setting (Figure 6b): singular *they* has a much higher likelihood than other pronouns that even overrides gender information from the antecedent. This indicates that fine-tuning may serve as a method for enabling models to accept singular *they*, given that our findings demonstrate their difficulties with it (§6). However, the extent of replacement should likely be less comprehensive than in the experiments conducted by Bartl and Leavy's (2024). Further, in the PL setting, the probabilities resemble previously observed distributions for Qwen2.5 (Figure 1) and OLMo (Figure 5) for feminine and masculine antecedents. For neutral antecedents, however, masculine coreferents exhibit the highest probability, contrary to the intended effect of fine-tuning. We would have expected the fine-tuning process to enhance the likelihood of a neutral coreference and balance out associations between masculine and feminine coreferents. While fine-tuning with gender-neutral language might have been effective in reducing stereotyping (Bartl and Leavy, 2024), our results demonstrate that more fine-grained evaluation methods are necessary to comprehensively assess the effects.

Table 5: German antecedents

#	masculine	feminine	coordinated feminine first	coordinated masculine first	capital I	asterisk	colon	underscore	EN translation
1	Eigentümer	Eigentümerinnen	Eigentümerinnen und Eigentümer	Eigentümer und Eigentümerinnen	EigentümerInnen	Eigentümer*innen	Eigentümer:innen	Eigentümer_innen	owners
2	Allergologen	Allergologinnen	Allergologinnen und Allergologen	Allergologen und Allergologinnen	AllergologInnen	Allergolog*innen	Allergolog:innen	Allergolog_innen	allergists
3	Choreographen	Choreographinnen	Choreographinnen und Choreographen	Choreographen und Chore-ographinnen	ChoreographInnen	Choreograph*innen	Choreograph:innen	Choreograph_innen	choreographers
4	Beamte	Beamtinnen	Beamtinnen und Beamte	Beamte und Beamtinnen	BeamtInnen	Beamt*innen	Beamt:innen	Beamt_innen	civil servants
5	Radfahrer	Radfahrerinnen	Radfahrerinnen und Radfahrer	Radfahrer und Radfahrerinnen	RadfahrerInnen	Radfahrer*innen	Radfahrer:innen	Radfahrer_innen	cyclists
6	Akademiker	Akademikerinnen	Akademikerinnen und Akademiker	Akademiker und Akademikerinnen	AkademikerInnen	Akademiker*innen	Akademiker:innen	Akademiker_innen	academics
7	Önologen	Önologinnen	Önologinnen und Önologen	Önologen und Önologinnen	ÖnologInnen	Önolog*innen	Önolog:innen	Önolog_innen	oenologists
8	Schiedsrichter	Schiedsrichterinnen	Schiedsrichterinnen und Schiedsrichter	Schiedsrichter und Schiedsrich-terinnen	SchiedsrichterInnen	Schiedsrichter*innen	Schiedsrichter:innen	Schiedsrichter_innen	referees
9	Tierärzte	Tierärztinnen	Tierärztinnen und Tierärzte	Tierärzte und Tierärztinnen	TierärztInnen	Tierärzt*innen	Tierärzt:innen	Tierärzt_innen	veterinarians
10	Archäologen	Archäologinnen	Archäologinnen und Archäologen	Archäologen und Archäologinnen	ArchäologInnen	Archäolog*innen	Archäolog:innen	Archäolog_innen	archeologists

number	lang.	# obs.	LLM	quant.	$F_{\text{ante\_gender}}$	$F_{\text{coref\_gender}}$	$F_{\text{interaction}}$
PL	EN	13464	GPT-2	32bit	481.6	720.2	1629.7
			GPT-2-finetuned	32bit	119.8	3432.9	983.5
			OLMo 1B	4bit	184.3	799.1	1011.8
			OLMo 7B	4bit	67.3	142.8	720
			OLMo 13B	4bit	297.8	710.4	622.8
			Qwen 32B	4bit	138.6	178.3	809.9
	DE	9240	EM Leo Mistral 7B	4bit	42.74	2601.35	36.63
SG	EN	14652	GPT-2	32bit	876.6	7885.6	6336.3
			GPT-2-finetuned	32bit	111.9	44001.9	6835.5
			OLMo 1B	4bit	342.8	3998.4	4171.4
			OLMo 7B	4bit	706.3	2816.8	5509.6
			OLMo 13B	4bit	592.9	3212.2	3703.3
			Qwen 32B	4bit	1231	3866	4626

Table 6: ANOVA effect sizes for antecedent gender, coreferent gender and interaction for all LLMs tested. All effects significant with $p < .001$. **quant.** = model quantization.

B.3 German Coreferent Generation

Figure 7 visualizes how the different gender-inclusive strategies influence the gender mentioned in the generations. We differentiated by whether the model generation referred back to the antecedent (62.5%, left panel) or not (37.5%, right panel). What both conditions have in common is that in most cases gender-neutral antecedents effect a gender-neutral coreferent. For the `no coreference` group, indeed all coreferents are neutral. These results suggest that LLMs are likely to maintain gender-inclusive language when prompted with these forms. In fact, there were many instances in which the model simply repeated the antecedent phrase. This is why Figure 5.2 contains the additional coreferent category `masc_fem` to capture instances in which the model generated coordinated forms (Table 2, strategies 3&4). These were indeed only generated when prompted with a coordinated form.

For instances where the antecedent expressed only a single gender (masculine or feminine), Figure 7 shows the majority of masculine coreferents and all of the feminine coreferents corresponding with the respective antecedent. These results indicate that feminine gender in the antecedent is a very strong signal for future generations. The presence of some neutral coreferents for masculine antecedents suggests that masculine gender can sometimes have a generic interpretation. However, in most cases masculine gender has a masculine association.

The German coreferent generation results suggest that generated coreferents generally align with the antecedent gender in the prompt, indicating that gender-inclusive language can encourage gender-neutral generations. However, this contrasts with our coreferent probability experiments (Section 5.1), which revealed strong masculine biases. This suggests that German models rely on repetition rather than a genuinely gender-neutral interpretation.

Annotation Guidelines

The dataset you were given contains sentences that a language model was asked to complete.

Your task is to annotate each sentence based on the following questions:

1. **What gender does the human entity mentioned in the continuation have?**
   Possible answers are *masculine* (m), *feminine* (f) and *neutral* (n).
   If the continuation does not mention a human entity, please annotate the item as *NA*.
2. **Is it possible that the person/people in the continuation refers back to the person/people mentioned in the first sentence?**
   Possible answers are *YES* or *NO*
   If you are unsure, you can annotate the item with *UNSURE*

Examples:

Sentences	**Continuation**	**gender mentioned**	**refers back?**
The daughters were waiting near the barrier. After a while, most of the	men had been drinking.	m	no
The sons were waiting near the barrier. After a while, most of the	wolves had been	NA	no
The spouses were waiting near the barrier. After a while, most of the	women were sweating. The	f	yes

Figure 4: Annotation guidelines given to annotators for English data

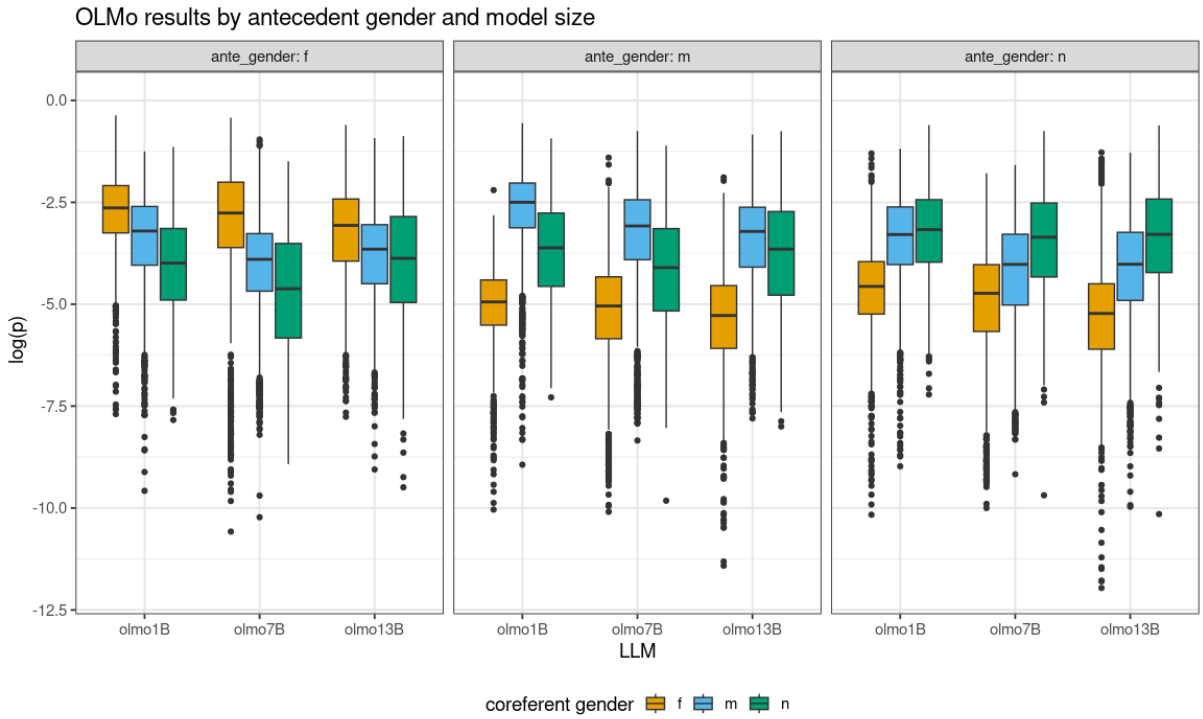
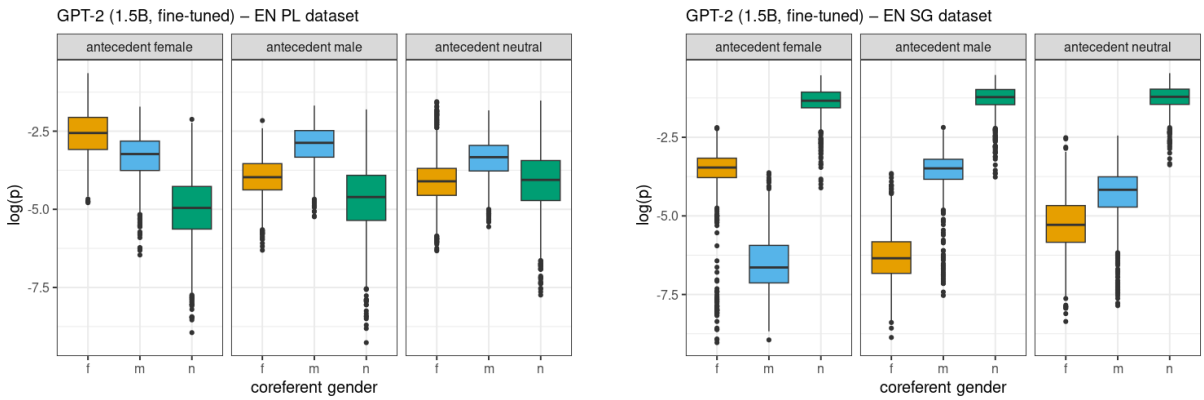


Figure 5: Coreferent probabilities for three OLMo model sizes for feminine, masculine and neutral antecedent gender



(a) plural

(b) singular

Figure 6: Distribution of $\log(p)$ of coreferent gender by antecedent gender in the PL and SG setting

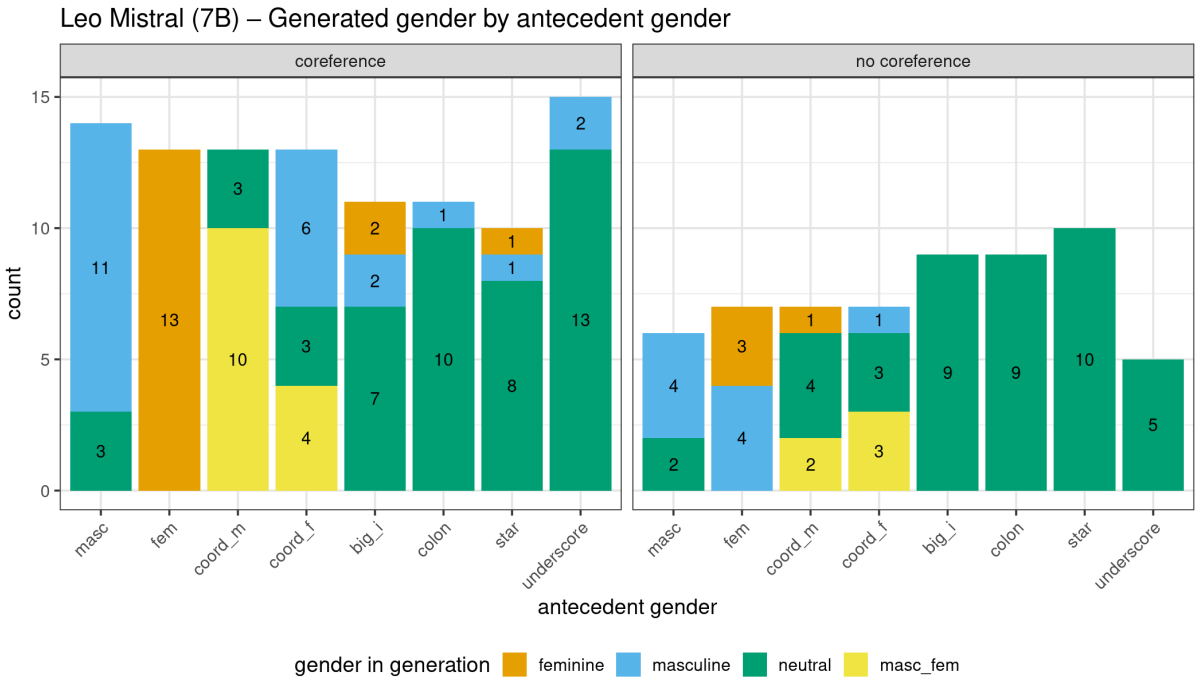


Figure 7: Generated gender for German model, divided by whether or not the continuation contains a coreferent of the antecedent