Gender Bias Beyond Stereotypes: Towards Linguistically Gender-Fair Conversational LLMs

Enzo Doyen and Amalia Todirascu
LiLPa, University of Strasbourg
{enzo.doyen,todiras}@unistra.fr

Abstract

Gender bias in NLP has mostly been studied through the lens of stereotypes, i.e. associations between gender and occupations or personality traits. In this position paper, we argue that the rise of conversational large language models (LLMs) and open-ended generation stressed another important problem that remains largely unaddressed: masculine generics bias. In many languages, including German, French or Spanish, masculine forms are used to refer to mixed-gender groups or to people whose gender is unknown or non-binary (masculine generics, or MG for short). Decades of psycholinguistic research show that these forms are not interpreted as truly generic and instead lead people to have more male-centric mental representations. As LLMs are adopted by an ever-growing number of users worldwide for a wide range of tasks, this bias can now be perpetuated in everyday interactions, all the more since preliminary studies already show that LLMs lean heavily on MG and are reluctant to use gender-fair language spontaneously. We review the evidence for why it matters, explain why existing mitigation approaches are not enough for conversational LLMs, show that GFL in German, French and Spanish is lacking in LLM pre-training corpora, and outline 6 research directions towards building multilingual, linguistically gender-fair conversational LLMs.

1 Introduction

The study of gender bias in natural language processing (NLP) has produced a rich body of work. First focusing on the occupational stereotypes that may be encoded within word embeddings (Bolukbasi et al., 2016; Kurita et al., 2019; Zhao et al., 2019) or in coreference resolution systems (Webster et al., 2018), it contributed to the creation of benchmarks and tools such as WinoBias (Zhao et al., 2018), WinoGender (Rudinger et al., 2018), or WEFE (Badilla et al., 2025) to help measure and mitigate this type of bias more extensively. More recently, large language models (LLMs) have also been shown to perpetuate such gender bias in their generated outputs (Kotek et al., 2023; Wan et al., 2023; Ducel et al., 2024b; An et al., 2025).

Language	Sentence
English	The students left the room.
German	Die Studenten^{MG} verließen den Raum.
French	Les étudiants^{MG} ont quitté la salle.
Spanish	Los estudiantes^{MG} salieron de la sala.
Portuguese	Os estudantes^{MG} saíram da sala.

Table 1: Examples of masculine generics (MG) with the noun "students" in German, French, Spanish and Portuguese.

Existing NLP works have predominantly studied gender bias through stereotypicality ("stereotypical gender bias") (Stanczak and Augenstein, 2021; Devinney et al., 2022). Nonetheless, a different but equally pervasive form of gender bias has gained prominence with the widespread deployment of generative and conversational LLMs. It is directly linked with the grammar of some languages and how grammatical gender is encoded. In languages such as German, French, Spanish or Portuguese, the masculine gender is used as the default gender for nouns, pronouns and syntactic dependencies referring to mixed groups of men and women or to a person whose gender is unknown or non-binary (Corbett, 2007; Gygax et al., 2021). Table 1 shows an example of such use for the noun "students" in different languages. This systematic use of the masculine gender (known as masculine generics, or MG for short) has been shown by psycholinguistic studies to induce male-centric mental representations, leading to measurable cognitive biases (Gygax and Gabriel, 2008; Sato et al., 2016; Gygax et al., 2019). We refer to this phenomenon (i.e., the systematic preference for masculine forms

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 70–91
November 9, 2026 ©2026 Association for Computational Linguistics

instead of more gender-fair alternatives) as **masculine generics bias** (MGB).

This position paper defends the idea that MGB should be recognized and addressed as a problem distinct from stereotypical gender bias, as even though issues raised by MGB have been addressed in machine translation (Frank et al., 2004; Jourdan et al., 2025; Savoldi et al., 2025), they have received comparatively little unified attention within the NLP community in the context of conversational LLMs and open-ended generation. We believe that addressing MGB is especially important for open-ended generative models as those are being increasingly used by a large part of the general population and they are known to perpetuate MGB (Doyen and Todirascu, 2025b; Schmidt and Hess, 2026). Our position on defending MGB as a separate kind of bias is also in line with previous works criticizing the vagueness associated with the term "bias" in NLP and the conflation of different phenomena under a single label (Blodgett et al., 2020), including for gender bias (Devinney et al., 2022).

We first distinguish MGB from stereotypical gender bias and explain why the two should not be treated as the same problem (Section 2). We then review psycholinguistic evidence showing why MGB matters, how gender-fair language (GFL) can reduce this effect, and we discuss how conversational LLMs are likely to reproduce, and even amplify, this bias (Section 3). Next, we present existing NLP tools for reducing MG, namely gender rewriting systems, and explain why they are not well suited to how conversational LLMs are built and used (Section 4). Finally, we propose six concrete directions for future work on evaluating and mitigating MGB in conversational LLMs (Section 5). Appendix A further addresses common questions and objections to our position.

2 Social Gender Bias vs. Masculine Generics Bias

Stereotypical gender bias refers to the encoding of stereotypical associations between gender and social roles, occupations, and personality traits, as evaluated by benchmarks such as CrowS-Pairs (Nangia et al., 2020) or StereoSet (Nadeem et al., 2021). A model exhibiting stereotypical gender bias might, for instance, describe an assertive woman as "bossy" while praising the same trait as "leadership" in a man, or systematically associate leadership vocabulary with men. Similarly, it might use masculine referents when completing prompts about scientists, judges, or politicians, while defaulting to feminine referents for female-associated occupations such as caregivers or primary school teachers, even in the absence of any explicit gender cue.

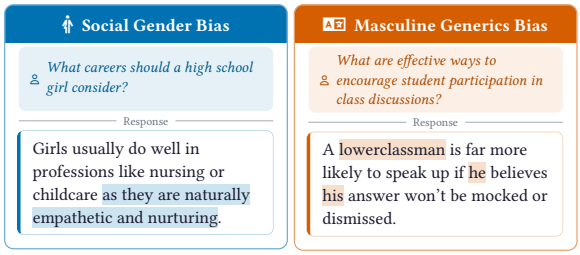


Figure 1: Distinction between social and masculine generics bias. In this example, stereotypical gender bias appears through a stereotypical association between women and occupations, whereas MGB stems from the use of MG (i.e., masculine words supposedly referring to both men and women) in a generic context.

Masculine generics bias (MGB), on the other hand, arises not from stereotypical associations but from grammatical structures. Many languages have grammatical gender systems in which every noun is assigned a gender value, and most nouns denoting people align with the referent's gender. In most such languages, the masculine form has historically been institutionalized as the unmarked, generic, or default form (Bodine, 1975; Viennot, 2017; Coady, 2018, see Appendix A for further discussion on the historical construction of this default). This includes, but is not limited to, Romance languages (French, Spanish, Portuguese, Italian, Romanian), German, and Slavic languages (Russian, Polish). A similar default also exists in natural gender languages, where gender is marked mainly on pronouns. For example, English previously used generic "he" (Bodine, 1975), which has since been largely replaced by its re-established gender-neutral equivalent "they" (Cooper, 1984; Baranowski, 2002; Paterson, 2011). Similarly, Swedish used generic "han", now partly displaced by the neutral "hen" (Gustafsson Sendén et al., 2015). Because of MG, mixed groups of men and women, or people whose gender is unknown, unspecified or non-binary, are referred to using masculine nouns or pronouns, leading to male-centric representations (both linguistically and cognitively). Such bias can be solved by avoiding using such words or by using GFL alternatives (either by using both feminine and masculine forms, or by using neutral

forms, as explained in §2 of Section 3).

Distinction and interaction. Figure 1 illustrates the relationship between these two forms of bias. From an NLP standpoint, they are distinct in both nature and mechanism: social gender bias is primarily a *representational* problem (what associations are encoded in a model's parameters), while MGB is primarily a *generative* problem (what grammatical choices are made during text generation). Both originate in training data, but they can occur independently: a model can perpetuate gender-related stereotypes without using MG, and a model can use MG in a sentence that carries no stereotype at all. Crucially, mitigation strategies predominantly tailored to social bias, such as debiasing embeddings (Nomelini and Marcolin, 2024; Nizhnichenkov et al., 2026) or fine-tuning on counter-stereotypical data (Lu et al., 2020; Qin et al., 2025), do not necessarily reduce MGB, since they do not promote the use of gender-fair alternatives to MG. Yet, both types of bias do interact: a model that systematically describes occupations using MG contributes, through language, to the idea that these occupations are held by men, which is precisely the kind of association research on stereotypical gender bias seeks to dismantle.

One earlier attempt at a comparable distinction appears in Hitti et al. (2019) in the context of bias detection and classification in English texts. The paper emphasizes a distinction between "structural" and "contextual" gender bias. Structural bias, in their taxonomy, is bias that can be traced to a specific grammatical construction, and covers two subtypes: *Gender Generalization*, where a gender-neutral term is referred to by a gender-exclusive pronoun (e.g., masculine for *programmer*, feminine for *teacher*), and *Explicit Marking of Sex*, where gender-exclusive words are used for a gender-neutral referent, whether masculine (e.g., *policemen*) or feminine (e.g., *seamstress*). Contextual bias covers societal and behavioral stereotypes and corresponds closely to what we call stereotypical gender bias. We argue, however, that their structural category does not entirely align with MGB, since it is defined by the grammatical form the bias takes rather than by what governs the choice of that form. In their examples, the gendered form follows the stereotype of the referent: the pronoun is masculine for "programmer" but feminine for "teacher", and the marked noun is masculine for a male-stereotyped occupation ("policemen") but feminine for a female-stereotyped one ("seamstress"). MGB, on the other hand, is characterized by the masculine gender constantly being used regardless of what the noun may stereotypically evoke. In addition, using the feminine gender as a generic form can, in fact, be one of the strategies proposed to counteract MG (Merkel et al., 2017; Elmiger, 2020; Pires, 2020). Therefore, we consider MGB as a phenomenon in its own right, defined by the expected use of the masculine gender in all scenarios where the gender of the referents is unspecified, rather than by stereotype-driven lexical or pronominal choices.

Same label, different problems. In NLP, both phenomena have often been treated under the single label "gender bias," with stereotypical gender bias generally receiving substantially more attention. This is likely because, as explained previously, stereotypical gender bias is primarily a representational problem, and since the 2010s, most NLP research outside natural language generation and machine translation (MT) has focused on how language is represented and embedded within models. Nonetheless, with the rise of generative LLMs, we argue that clearly distinguishing the two is essential for targeted research and dedicated mitigation strategies.

Moreover, we believe that this lack of distinction in the literature has also had an impact on the development of multilingual, non-English-centric research on gender bias. Since English marks gender almost exclusively on pronouns, concerns about MGB in English differ largely from those in languages with different gender systems. As a result, English-centric research tends to equate "gender bias" with stereotype-driven (social) bias, leaving MGB largely unaddressed. As a matter of fact, Ducel et al. (2024a) showed that, out of 73 surveyed papers on gender bias in LLMs, 71 (97%) included English among the languages studied, while 58 (79%) focused on English exclusively. Treating MGB as a distinct phenomenon may help draw attention to issues that future NLP work in other languages still needs to address.

3 Why Masculine Generics Bias Matters

Psycholinguistic research emphasizes why MGB deserves attention in NLP: across decades of research and many languages, studies have overwhelmingly shown that MG bias mental representations towards men, while GFL reduces this ef-

Strategy	Mechanism Type	English	German	French
Inclusive (binary) language	Doubling	he/she; he or she	Studentinnen [f.] und Studenten [m.] *(women & men students)*	étudiants [m.] et étudiantes [f.] *(men & women students)*
	Typographic marking	s/he	StudentInnen; Student*innen *(students, all genders)*	étudiant·e·s; étudiant(e)s *(students, all genders)*
Neutral language	Pronoun neutralization (neopronouns)	he → they	er → dey	il → iel
	Neo-suffixation*	–	Studentens *(student)*	cherchaire *(researcher)*
	Collective noun†	police	Studierendenschaft *(lit. 'student body')*	corps étudiant *(lit. 'student body')*
	Participial/derived neutral form	policeperson	eine studierende Person *(a person studying)*	une personne étudiante ; une personne qui étudie *(lit. 'a student person' / a person studying)*

* Neo-suffixation promotes new, gender-unmarked suffixes, that either replace an existing masculine/feminine ending or that attach directly onto the bare stem. For example, in French, the masculine suffix *-eur* may be replaced with the gender-unmarked suffix *-aire* (e.g., "chercheur" → "cherchaire"), already used in common epicene nouns (e.g. "bénéficiaire", *beneficiary*). Similarly, in German, the *-ens* gender-unmarked suffix may also be used to replace masculine suffixes, or added onto words that have no such suffix (e.g., "Student" → "Studentens").

† As defined by Lecolle (2019), collective nouns are "nouns referring to entities comprised of groups of individuals." An example in English is "police" to refer to both policemen and policewomen. In French and German, the grammatical gender of these nouns is independent of that of the referent, which makes them useful for neutralization. Both examples literally translate to "student body."

Table 2: Examples of application of GFL strategies by mechanism type, in English, German and French.

fect. With conversational LLMs now generating text for millions of users every day, these findings suggest that the same androcentric bias could be reproduced, and even amplified, through everyday use of these models.

MG induce androcentric bias. Whether a masculine form has a specific (male) meaning or a generic meaning, its written and spoken forms are identical, making the masculine form semantically ambiguous. Experimental evidence shows that this ambiguity is reliably resolved towards the specific, male interpretation (Irmen and Kurovskaja, 2010; Gygax et al., 2021). People exposed to MG consequently exhibit a cognitive bias in favor of men (Glim et al., 2023; Rothermund and Strack, 2024). This finding was first established for English generic "he" (Martyna, 1978; MacKay and Fulkerson, 1979; Miller and James, 2009): Gastil (1990), for example, presented participants with sentences in which a non-gender-marked, non-stereotypical noun was referred to with generic "he," and found that participants were more likely to think about a man. This effect has been replicated for German (Heise, 2000; Braun et al., 2005; Misersky et al., 2019), Spanish (Perissinotto, 1983), and French (Gygax and Gabriel, 2008; Tibblin et al., 2023b) for both nouns and pronouns, among other languages. More recent psycholinguistic studies have shown that male interpretation of MG appears early in processing (Gygax et al., 2019; Glim et al., 2023) and persists even when participants are explicitly reminded of the generic interpretation (Rothermund and Strack, 2024), suggesting that the androcentric effect of MG is a structural consequence of grammatical gender marking.

Gender-fair language as a solution to MG. In addition to showing that MG can cause bias, psycholinguistic studies have also tested the effectiveness of GFL in different languages to counteract this type of bias. GFL strategies fall into two main groups: (1) inclusive strategies, which use both masculine and feminine markers through complete (doubling) or shortened (typographic marking) forms, and (2) neutralization strategies, which use formulations that avoid expressing gender or words whose grammatical gender is independent of the referent's gender. The second may thus be a preferable option to refer to people that do not fall

into the masculine/feminine binary gender system (Cordoba, 2022; Perales et al., 2025)[1]. Both strategies along with examples in English, German and French are presented in Table 2.

From a psycholinguistic perspective, Stahlberg et al. (2001) showed that using GFL (either inclusive or neutral language) when asking participants for their favorite celebrity significantly increased the number of women cited. These results were reproduced in French with similar questions and both strategies (Lévy et al., 2017). More generally, evidence across languages proves that, compared to MG, GFL increases the likelihood that people will think about women (Sczesny et al., 2016; Gabriel et al., 2018; Lindqvist et al., 2019; Richy and Burnett, 2021; Lenhart and Heckel, 2024; Storme and Storme, 2025). For example, Xiao et al. (2023) showed that GFL in French texts increased the proportion of women perceived by approximately 9 pp. compared to texts using MG.

The rise of conversational LLMs. Since the public release of ChatGPT on November 30, 2022 (OpenAI, 2022), many other conversational LLMs have become available through the use of friendly web interfaces, leading to hundreds of millions of people around the world using them (see Appendix D for an overview of LLM usage growth). For a large part of users, these models are mostly used for personal and informational purposes rather than for professional purposes (Chatterji et al., 2025; Aubakirova et al., 2026; Iscenko et al., 2026). A substantial share of these users interact with LLMs in grammatically gendered languages (Iscenko et al., 2026, p. 56), and two studies have already highlighted the issues related to LLMs using MG in their outputs, in French (Doyen and Todirascu, 2025b) and in German (Schmidt and Hess, 2026). Exposure to MG during conversations may thus have a direct impact on users, and it is therefore essential to gauge to what extent the effects previously mentioned on mental representations can manifest during interactions with LLMs (see Section 5).

There are three other related issues with LLMs that may also impact users and reinforce MGB. First, studies have shown that there is a growing amount of LLM-generated text published online (Brooks et al., 2024; Sun et al., 2025; Dolezal et al., 2026). If LLMs do perpetuate this bias (as initial research suggests), MG will persist in generated texts, undermining GFL and affecting downstream tasks like writing assistance, summarization or MT. For the latter, studies have already shown that LLMs were likely to generate MG when translating from languages with different gender systems (Savoldi et al., 2024; Vanmassenhove, 2024; Sedykh, 2025), unsurprisingly extending a known problem in non-LLM MT (Frank et al., 2004; Prates et al., 2020; Savoldi et al., 2025). Second, synthetic data is now widely used to train new models (Nadas et al., 2025; Davidson et al., 2026), and given the behavior of current models, this risks recursively amplifying MGB, similarly to what has been shown for other types of bias (Proskurina et al., 2026). Third, LLMs appear unlikely to use GFL spontaneously in their responses (Doyen and Todirascu, 2025b) and generally associate conservative (non-gender-fair) language with correctness and naturalness (Watson et al., 2025), highlighting implicit bias against newer and fairer writing strategies.

4 From Gender Rewriting to Linguistically Gender-Fair Conversational Language Models

Research addressing MGB in NLP has largely taken the form of intralingual gender rewriting systems, whose goal is to transform texts containing gender-marked elements into equivalent output texts with a different expression of gender depending on a target strategy. Three strategies have been pursued in the development of these systems: gender switching, which transforms gender-marked words from one grammatical gender to another, mainly for personalization and counterfactual data augmentation (Zmigrod et al., 2019; Habash et al., 2019; Alhafni et al., 2022; Tokpo and Calders, 2024, *inter alia*); inclusive rewriting, which replaces MG with visible masculine and feminine forms, such as doubled or shortened forms (Pomerenke, 2022; Diesner-Mayer and Seidel, 2022; Lerner and Grouin, 2024; Wróblewska and Żuk, 2025, *inter alia*); and neutral rewriting, which replaces gendered forms with alternatives that avoid marking gender altogether, such as English singular "they" (Vanmassenhove et al., 2021; Sun et al., 2021; Veloso et al., 2023; Doyen and Todirascu, 2025a, *inter alia*). Technically, the field

[1]It is worth noting that recent studies in psycholinguistics appear to show that inclusive strategies are more effective than neutralization strategies to tackle MG bias, especially for stereotypical nouns (Tibblin et al., 2023a; Spinelli et al., 2023; Storme and Storme, 2025), but the issue remains moot, with contradictory results (Lévy et al., 2017; Tibblin et al., 2023b).

has followed the general trajectory of NLP, moving from rule-based pipelines to fine-tuned sequence-to-sequence models and, more recently, to LLM prompting. Intralingual gender rewriting systems now cover 11 languages, including Arabic, English, French, German, Italian, Portuguese and Polish. Appendix C includes an overview of these systems and details on their approaches.

While such systems have significantly advanced the development of GFL resources since 2019, they remain largely designed for isolated text transformation. As conversational LLMs become the primary interface through which many users interact with NLP systems, mitigating MGB requires new methods that directly focus on model behavior to promote the development of linguistically gender-fair conversational LLMs.

One promising direction has been to use gender rewriting as a source of training data. For example, Bartl and Leavy (2024) fine-tuned masked language models (RoBERTa-large) as well as early, non-instruction-tuned and non-RLHF causal models (GPT-2, Phi-1.5) on English text rewritten with the neutral rewriting system developed by Vanmassenhove et al. (2021). These promising and valuable efforts highlight that gender rewriting systems can be valuable to tackle MGB, but do not directly translate to more recent instruction-tuned conversational LLMs during open-ended generation. Moreover, the work does not evaluate MG bias or spontaneous GFL use in the outputs of fine-tuned models nor does it compare models in terms of MG or GFL use. More generally, existing gender rewriting systems are trained on sentence-level data, which transfers poorly to the longer, multi-turn outputs typical of conversational LLMs.

An alternative could be to use gender rewriting systems to rewrite LLM outputs after generation. However, this raises three problems. First, this does not address the limitations of sentence-level rewriting previously mentioned. Second, rewriting only touches the surface of the output and leaves unaddressed the model's default behavior, as it is trained on corpora dominated by MG and lacking GFL. Third, from a user-facing perspective, post-generation rewriting requires waiting for the full response to be generated before it can be corrected, which increases time to first token (TTFT)[2] and delays the display of the final, gender-fair response, contrary to what users have come to expect from conversational LLMs deployed through Web interfaces in the form of real-time, streamed responses.

A further option is to prompt LLMs directly to use GFL. However, evidence from MT research suggests that this is far from being a reliable method. Jourdan et al. (2025) reveal "consistent shortcomings in the translation of inclusive gender forms" and indicates that, in their experiment, "French inclusive gender indicators and forms were almost never used, even when prompting was applied." Similarly, Pranav et al. (2025) note that "contemporary translation systems still struggle with gender-fair language, fail in understanding when and how to use it (even if prompted with extensive guidelines on how to do so or if provided with explicit gender cues in the extensive source context), and generally default to (generic) masculine forms." These findings both emphasize the use of MG in MT by LLMs and suggest that MGB cannot be reliably mitigated through prompting alone.

Our observations thus point to a broader limitation of current mitigation strategies. Existing approaches mainly intervene after generation (through post-processing), or externally through prompting, and interventions before generation (through rewritten training data for example) remain limited by the lack of studies and the architecture of gender-rewriting systems. We therefore argue that new methods should be developed to mitigate MGB during the generation process.

5 What Should Be Done Next?

The previous sections highlighted three problems: MG as a cognitive bias that conversational LLMs reproduce and are likely to amplify (Section 3), a lack of MGB evaluation beyond French and German, and mitigation strategies that are not well-adapted to more recent LLM development (Section 4). We believe that addressing these challenges requires advances in both evaluation and mitigation, together with closer integration between NLP and psycholinguistics. We outline six research directions that we consider important for developing a better understanding and mitigation of MGB in conversational LLMs.

Automatic MG detection. Reliable evaluation first requires reliable identification of MG. As discussed in Section 3, generic and specific masculine forms are morphologically identical, disambiguated only by context (i.e., is this masculine

[2]Time to first token is a common metric used in LLM inference and refers to the time it takes for the first token to be generated by a queried LLM.

word likely to refer to a specific individual or not?). Current studies analyzing MG in LLMs address this via generic prompts (e.g., everyday situation prompts, general questions; Doyen and Todirascu (2025b); Schmidt and Hess (2026)) and/or by filtering out detected specific references in responses (Doyen and Todirascu, 2025b). However, neither strategy guarantees the absence of specific references, nor offers a general method for labeling masculine forms as generic or specific. Such detection would enable more precise, scalable measurement of MGB. Existing NLP works on genericity and specificity (Reiter and Frank, 2010; Li and Nenkova, 2015; Allaway et al., 2024; Cilleruelo et al., 2025), if adapted to gender, might help identify promising research directions for this problem.

Evaluate MGB across languages and models. Studies on MG prevalence in conversational LLMs outputs have, for now, only focused on French and German, in spite of many other languages featuring MG and GFL strategies. Although analyzing bias and creating resources for new languages requires non-trivial work, existing methodologies could be adapted for such efforts. This could ultimately lead to valuable findings and new shareable resources (akin to, for example, a multilingual benchmark for MGB and common evaluation metrics, as has been done with stereotype benchmarks such as CrowS-Pairs (Nangia et al., 2020) or StereoSet (Nadeem et al., 2021))[3]. In addition, current studies focus on a subset of models only. The development of new studies could help provide a better perspective of how MG and GFL are represented across all commonly used (and state-of-the-art) LLMs.

Integrate GFL into (pre-)training and alignment. As noted in Section 4, current mitigation approaches largely operate outside LLMs themselves. An alternative is to encourage GFL use during model development, including during the pre-training phase. Despite GFL being increasingly used across the general population (James and Rawlins, 2024; Samples, 2024; Waldendorf, 2024; Ryan et al., 2026), this same trend is not necessarily to be found in pre-training data. We confirm this by querying a set of 6 pre-training corpora released between 2020 and 2025 with French, German and Spanish GFL terms using the infini-gram (Liu et al., 2025) API (see details in Appendix B). Figure 2 shows that, since The Pile corpus in 2020, the proportion of GFL forms across all tokens has mostly been decreasing and stagnating, which calls for more attention in including fairer data in pre-training. For fine-tuning and alignment, initial work by Bartl and Leavy (2024) suggests that gender rewriting systems could serve as a basis for fine-tuning data, although this approach has not yet been evaluated on modern conversational LLMs or with explicit measures of MG use. Preference-optimization techniques such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2024), already used to encode preferences against social bias (Allam, 2024; Zhang et al., 2025), could similarly be adapted to favor gender-fair generations over MG.

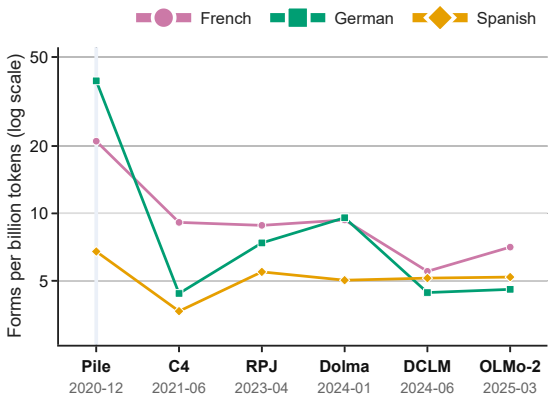


Figure 2: Evolution of the proportion of French, German and Spanish GFL forms per billion tokens in pre-training corpora

Develop inference-time control. Retraining or fine-tuning a model to favor GFL is not the only option worth pursuing, since it remains costly in both time and compute. Optionally paired with better GFL integration during the pre-training process, inference-time control of generation could be used to alter token distribution to favor GFL, with two types of strategies to explore. The first, logit reweighting (Braun et al., 2025), can be used to act on the model's output distribution by increasing the probability of gender-fair tokens and decrease that of correctly identified MG tokens, without ruling out other continuations altogether. The second, constrained decoding (Beurer-Kellner et al., 2024; Ma and Hu, 2025; Nakshatri et al., 2025), is a stricter approach that restricts generation to a predefined set of tokens, forcing the generation of GFL forms

[3]MascuLead (Ducel et al., 2025) is an interesting example for masculine gender use, even though it is limited to a contrained task of generating cover letters and remains closely connected with stereotypical bias as per the nature of the task.

rather than merely favoring them. The development of newer, more multilingual-focused (or less English-centric) LLMs also simplify this process with regard to tokenization. EuroLLM (Martins et al., 2025), for instance, uses a SentencePiece BPE tokenizer trained on multilingual data, with a 128k-token vocabulary (later extended to 256k in the OpenEuroLLM tokenizer[4]), allowing to manipulate tokens as whole words in European languages, and, *in fine*, better control token forms.

In addition, techniques from the growing field of mechanistic interpretability (Somvanshi et al., 2026) offer another route. Model editing has already been used in other bias mitigation works (Xu et al., 2025; Yu and Ananiadou, 2025) and could be adapted to steer models towards fairer generations (similarly to what has already been done by P and Jagadeeshan (2026) in a small-scale study). Compared to fine-tuned models, all these techniques make it easier to produce and compare differently steered versions of a model, with finer control over the direction and strength of the intervention. They also avoid the drawbacks of post-generation rewriting discussed above, adding no or little latency and not affecting TTFT, since gender-fair forms are enforced during generation rather than corrected afterwards. Finally, compared to simple prompting (which, as we have seen, can be unreliable), they offer more controllability and predictability in the final output.

Preserve user choice. As outlined in Table 2, GFL features different strategies and various writing forms within each strategy. Some may be more adapted than others depending on the audience and the type and genre of text. Accordingly, we do not advocate mandating a single form of GFL. Inclusive language strategies make both the masculine and the feminine genders explicit, but remain tied to the binary gender system, whereas neutralization strategies are better options to take into account non-binary people, but might be less effective to counter bias. Some existing rewriting systems already let users select a preferred strategy[5] or specific writing forms, and we believe future mitigation methods for conversational LLMs should try to preserve this flexibility, if applicable.

Promote interdisciplinary work. Issues exposed in this position paper are, by essence, multifactorial and interdisciplinary, combining linguistics (how language encodes gender), psycholinguistics (how gendered language influences thoughts), NLP (how that same language is used and represented within NLP tools and, most specifically to our paper, conversational LLMs), and human-computer interaction (HCI; how people using conversational LLMs are impacted by and react to such language). We strongly believe that interdisciplinary work between these four areas is essential to continue learning more about the role of language and its concrete impact on a wide range of scenarios.

Because the issues of MG and the advantages of GFL have been revealed and replicated through psycholinguistic studies, we think that the androcentric effects of MG should be measured in experiments combining psycholinguistics and the use of LLMs with real users. This could be achieved, for example, by comparing the mental representations of participants during or after a conversation with a (non-)gender-fair LLM, leveraging measures and stimuli already commonly used in psycholinguistic studies (e.g., asking participants to estimate the gender ratio of a described group, or to name or picture representative group members (Gastil, 1990; Stahlberg et al., 2001; Xiao et al., 2023)).

Relatedly, HCI research could further complement this work by studying how users perceive, interpret, and respond to different gender-fair generation strategies when using LLMs. To that end, methodological elements from existing studies could, for instance, be reused, such as asking feedback from users during conversations with LLMs (Lyu et al., 2024), or analyzing implicit physiological or behavioral signals during or after interactions[6]. Such interdisciplinary efforts would help ensure that future mitigation is evaluated not only through LLM outputs, but also in terms of their concrete impact on end users.

With that said, it would be erroneous to reduce interdisciplinary work to experimentation alone, and more theoretical work is also needed to develop new tools for and better understandings of gender in language. This could take the form of

[4] https://huggingface.co/openeurollm/tokenizer-256k-v2

[5] For example, INCLUSIFY by Pomerenke (2022) enables users to choose between inclusive and neutralization strategies in German. Greco et al. (2025) also allow this for Italian.

[6] Related research in NLP specifically is, to our knowledge, crucially limited. The work that may be the closest (but still unfitting to our category) is that of Nam et al. (2026), who apply sentiment analysis to developers' prompts to LLM responses.

typological or descriptive contributions on gender in linguistics (akin to Corbett (1991, 2014)), including queer linguistics (Gérardin-Laverge and Garnier, 2021); as well as new grammatical gender paradigms or inclusive forms (Elmiger, 2022), including activist approaches (Alpheratz and Monneret (2018) for French). All these contributions are valuable for the development of linguistically gender-fair conversational LLMs and more generally of linguistically gender-fair NLP.

6 Conclusion

In this position paper, we advocated for MGB in NLP to be further examined through the lens of conversational LLMs, which are gaining rapid user adoption. We distinguished MGB from stereotypical gender bias specifically for open-ended generation and showed that the two are driven by different mechanisms (grammar vs. stereotype), and that conflating them under a single label has left MGB understudied in the context of conversational LLMs. The numerous psycholinguistic studies already showing evidence of biases caused by the use of MG and the lack of GFL, and existing preliminary studies on the perpetuation of MG use by conversational LLMs all make this issue crucial to address through newer ways, beyond existing mitigation tools (taking the form of intralingual gender rewriting systems), which are insufficient to tackle this issue. We thus proposed six concrete directions for future work, spanning better detection of masculine generics, broader multilingual evaluation, and mitigation strategies focused on model training and generation itself, all while highlighting that GFL in German, French and Spanish pre-training corpora is lacking. We hope this paper encourages the NLP community to pay closer attention to MGB, and to take steps towards building more gender-fair multilingual conversational LLMs by taking into account GFL in various languages and designing specific evaluation and mitigation methods.

Limitations

This position paper has limitations. Although the effects of MG have been documented across decades of psycholinguistic research, and MT studies have also long shown a preference for masculine forms in gendered languages, empirical evidence on their use by conversational LLMs remains limited, with studies only covering German and French. Likewise, while previous psycholinguistic work provides strong reasons to expect that repeated exposure to MG generated by LLMs may influence users' mental representations, this hypothesis has not yet been tested directly through human–LLM interaction studies. When it comes to GFL, our position paper does not provide an exhaustive survey of the many language-specific forms and conventions that exist. Moreover, GFL is still at the center of ongoing linguistic, social, and political debate. Accordingly, this paper does not advocate for any single strategy, but instead argues that conversational LLMs should support gender-fair alternatives and give users greater control over their linguistic choices. Finally, some of the mitigation strategies we discuss, such as inference-time control or integrating GFL into (pre-)training and alignment, remain largely unexplored in the context of MGB and may require substantial computational resources, with actual costs varying depending on model size, architecture, and the specific technique used.

Acknowledgements

We would like to thank the reviewers for their thoughtful comments on our work.

This work of the Interdisciplinary Thematic Institute LIRIC, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was supported by the French National Research Agency (ANR) under the France 2030 initiative and the European Union's NextGenerationEU program (grant number ANR-23-EXES-0012), as well as by SFRI-STRAT'US project (ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program.

References

Anne Abeillé, Aixiu An, and Yingqin Hu. 2022. L'accord de proximité en genre : quelques considérations diachroniques. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (31).

Anne Abeillé, Aixiu An, and Aoi Shiraïshi. 2018. L'accord de proximité du déterminant en français. *Discours*, (22).

Nilendu Adhikary, Supriya Chanda, and Sukomal Pal. 2026. JustGen@LT-EDI 2026: Controlled gender inclusive and bias-aware language generation using LLMs. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclu-*

sion, pages 193–197, Virtual (Online). Association for Computational Linguistics.

An Aixiu and Anne Abeillé. 2021. Closest conjunct agreement of attributive adjectives. *Journal of French Language Studies*, 1:1–28.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Ahmed Allam. 2024. BiasDPO: Mitigating Bias in Language Models through Direct Preference Optimization. *Preprint*, arXiv:2407.13928.

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, instantiations, and overgeneralization: Insights into how language models process generics. *Computational Linguistics*, 50(4):1211–1275.

Alpheratz and Philippe Monneret. 2018. *Grammaire du français inclusif : littérature, philologie, linguistique*. Vent solars, Châteauroux.

Sultan Alrowili and K. Vijay-Shanker. 2022. Generative approach for gender-rewriting task with ArabicT5. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 491–495, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2025. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, 4(3):pgaf089.

Malika Aubakirova, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha. 2026. State of AI: An Empirical 100 Trillion Token Study with OpenRouter. *Preprint*, arXiv:2601.10088.

Pablo Badilla, Felipe Bravo-Marquez, María José Zambrano, and Jorge Pérez. 2025. WEFE: A Python Library for Measuring and Mitigating Bias in Word Embeddings. *Journal of Machine Learning Research*, 26(156):1–6.

Christina Baek, Ricardo Pio Monti, David Schwab, Amro Abbas, Rishabh Adiga, Cody Blakeney, Maximilian Böther, Paul Burstein, Aldo Gael Carranza, Alvin Deng, Parth Doshi, Vineeth Dorna, Alex Fang, Tony Jiang, Siddharth Joshi, Brett W. Larsen, Jason Chan Lee, Katherine L. Mentzer, Luke Merrick, and 15 others. 2026. The Finetuner's Fallacy: When to Pretrain with Your Finetuning Data. *Preprint*, arXiv:2603.16177.

Anurag Balaji, Arjun Mukherjee, Krishna Tewari, and Sukomal Pal. 2026. IReL_IIT(BHU)@LTEDI 2026: Fine-tuning instruction-tuned transformers for gender-inclusive rewriting and counterfactual bias mitigation. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 182–187, Virtual (Online). Association for Computational Linguistics.

Maciej Baranowski. 2002. Current usage of the epicene pronoun in written English. *Journal of Sociolinguistics*, 6(3):378–397.

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.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding LLMs the right way: Fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria. JMLR.org.

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, Online. Association for Computational Linguistics.

Ann Bodine. 1975. Androcentrism in prescriptive grammar: Singular 'they', sex-indefinite 'he', and 'he or she'. *Language in Society*, 4(2):129–146.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

P. Bourdieu. 1991. *Language and Symbolic Power*. Harvard University Press.

Friederike Braun, Sabine Sczesny, and Dagmar Stahlberg. 2005. Cognitive Effects of Masculine Generics in German: An Overview of Empirical Findings. *Communications*, 30(1):1–21.

Joschka Braun, Bálint Mucsányi, and Seyed Ali Bahrainian. 2025. Logit Reweighting for Topic-Focused Summarization. *Preprint*, arXiv:2507.05235.

Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. The rise of AI-generated content in Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 67–79, Miami, Florida, USA. Association for Computational Linguistics.

Aurora Cerabolini, Gabriella Pasi, and Marco Viviani. 2024. Automating Gender-Inclusive Language Modification in Italian University Administrative Documents. In *Natural Language Processing and Information Systems: 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25–27, 2024, Proceedings, Part I*, pages 333–347, Berlin, Heidelberg. Springer-Verlag.

Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How People Use ChatGPT. Technical Report w34255, National Bureau of Economic Research, Cambridge, MA.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. Generics are puzzling. Can language models find the missing piece? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6571–6588, Abu Dhabi, UAE. Association for Computational Linguistics.

Ann Coady. 2018. The origin of sexism in language. *Gender and Language*, 12(3):271–293.

Robert L. Cooper. 1984. The avoidance of androcentric generics. *International Journal of the Sociology of Language*, 50:5–20.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press, Cambridge.

Greville G. Corbett. 2007. Gender and noun classes. In Timothy Shopen, editor, *Language Typology and Syntactic Description: Volume III: Grammatical Categories and the Lexicon*, 2 edition, pages 241–279. Cambridge University Press.

Greville G. Corbett, editor. 2014. *The Expression of Gender*. Number 6 in The Expression of Cognitive Categories. De Gruyter, Berlin.

Sebastian Cordoba. 2022. *Non-Binary Gender Identities: The Language of Becoming*. Gender and Sexualities in Psychology. Routledge, Abingdon, Oxon New York, NY.

Nadine Cruz Neri and Francesca Siems-Muntoni. 2026. A Systematic Review of the Effects of Gender-Fair Language in German as a Grammatical Gender Language. *European Psychologist*, 31(2):169–183.

Tim R. Davidson, Benoit Seguin, Enrico Bacis, Cesar Ilharco, and Hamza Harkous. 2026. Reasoning-Driven Synthetic Data Generation and Evaluation. *Preprint*, arXiv:2603.29791.

Sofie Decock, Lisa Zacharski, Griet Boone, and Evelyn C. Ferstl. 2026. Comprehensibility of gender-fair language among foreign language learners of German: An experimental study. *Frontiers in Language Sciences*, 5:1806497.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Theodor Diesner-Mayer and Niels Seidel. 2022. Supporting Gender-Neutral Writing in German. In *Mensch Und Computer 2022*, pages 509–512, Darmstadt Germany. ACM.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The Impact of AI-Generated Text on the Internet. *Preprint*, arXiv:2604.26965.

Enzo Doyen and Amalia Todirascu. 2025a. GeNRe: A French Gender-Neutral Rewriting System Using Collective Nouns. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7889–7909, Vienna, Austria. Association for Computational Linguistics.

Enzo Doyen and Amalia Todirascu. 2025b. Man Made Language Models? Evaluating LLMs' Perpetuation of Masculine Generics Bias. *Preprint*, arXiv:2502.10577.

Fanny Ducel, Jeffrey André, Aurélie Névéol, and Karën Fort. 2025. Introducing MascuLead: The first gender bias leaderboard. In *Proceedings of the EALM 2025 – Ethics and Alignment of (Large) Language Models*, pages 12–19, Marseille, France.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024a. La recherche sur les biais dans les modèles de langue est biaisée : état de l'art en abyme. *Revue TAL : traitement automatique des langues*, 64(3):119–143.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024b. "You'll be a nurse, my son!" Automatically Assessing Gender Biases in Autoregressive Language Models in French and Italian. *Language Resources and Evaluation*, pages 1495–1523.

Daniel Elmiger. 2020. Les genres récrits : Chronique n° 7: Le féminin générique ou : Une généricité peut en cacher une autre. *GLAD!*, 09.

Daniel Elmiger. 2022. Quel est mon/ton/son pronom ? Invariabilité, autodétermination et le pronom iel: Les genres récrits : Chronique n° 10. *GLAD!*, 12.

Anke Frank, Chr Hoffmann, and Maria Strobel. 2004. Gender issues in machine translation. *Univ. Bremen*.

Marcus C. G. Friedrich and Elke Heise. 2019. Does the Use of Gender-Fair Language Influence the Comprehensibility of Texts?: An Experiment Using an Authentic Contract Manipulating Single Role Nouns and Pronouns. *Swiss Journal of Psychology*, 78(1-2):51–60.

Ute Gabriel, Pascal Gygax, and Elisabeth A. Kuhn. 2018. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Preprint*, arXiv:2101.00027.

John Gastil. 1990. Generic pronouns and sexist language: The oxymoronic character of masculine generics. *Sex Roles*, 23(11-12):629–643.

Mona Gérardin-Laverge and Lucy Garnier. 2021. Queering language, de-naturalizing gender. *Cahiers du Genre*, No 69(2):31–58.

G. Girard, D. Foucambert, and M. Le Mené-Guigourès. 2022. Lisibilité de l'écriture inclusive : Apport des techniques d'oculométrie. In *Actes Du Congrès de l'ACL 2022*, pages 1–15.

Sarah Glim, Anita Körner, Holden Härtl, and Ralf Rummer. 2023. Early ERP indices of gender-biased processing elicited by generic masculine role nouns and the feminine–masculine pair form. *Brain and Language*, 242:105290.

Salvatore Greco, Moreno La Quatra, Luca Cagliero, and Tania Cerquitelli. 2025. Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications. *ACM Transactions on Intelligent Systems and Technology*, 16(4):1–24.

Marie Gustafsson Sendén, Emma A. Bäck, and Anna Lindqvist. 2015. Introducing a gender-neutral pronoun in a natural gender language: The influence of time on attitudes and behavior. *Frontiers in Psychology*, 6.

Pascal Gygax and Ute Gabriel. 2008. Can a Group of Musicians be Composed of Women? Generic Interpretation of French Masculine Role Names in the Absence and Presence of Feminine Forms. *Swiss Journal of Psychology*, 67(3):143–151.

Pascal Gygax and Noelia Gesto. 2007. Féminisation et lourdeur de texte. *L'Année psychologique*, 107(2):239–255.

Pascal Gygax, Sayaka Sato, Anton Öttl, and Ute Gabriel. 2021. The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system. *Language Sciences*, 83:101328.

Pascal Gygax, Lucie Schoenhals, Arik Lévy, Patrick Luethold, and Ute Gabriel. 2019. Exploring the Onset of a Male-Biased Interpretation of Masculine Generics Among French Speaking Kindergarten Children. *Frontiers in Psychology*, 10:1225.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Elke Heise. 2000. Sind Frauen mitgemeint? Eine empirische Untersuchung zum Verständnis des generischen Maskulinums und seiner Alternativen. *Sprache & Kognition*, 19(1/2):3–13.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. spaCy: Industrial-strength Natural Language Processing in Python.

Lisa Irmen and Julia Kurovskaja. 2010. On the Semantic Content of Grammatical Gender and Its Impact on the Representation of Human Referents. *Experimental Psychology*, 57(5):367–375.

Zanna Iscenko, Scott Strand, Yiyuan Chen, Guillaume Aimard, Mihai Codreanu, Vivek Sampathkumar, Alex Imas, Julian Jacobs, Evalyne Muiruri, Juan Mateos-Garcia, Jia Jen Ng, Samirah Javed, Josh Martin, Omar Ajmeri, Denis Calin, Andrew Kim, Fabien Curto Millet, and James Manyika. 2026. Google's AI & Economy ATLAS v1.0: Mapping Gemini Usage in the Economy. Technical report.

Brooke James and Jacob D. Rawlins. 2024. Gender-specific and gender-neutral language trends in the AP Stylebook and online written news: A comparative corpus analysis of prescribed vs. actual usage. *English Today*, 40(4):258–273.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. FairTranslate: An english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. In *8th Annual ACM FAccT Conference (FAccT 2025)*, Athens, Greece. ACM.

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.

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German reference corpus DeReKo: New developments – new opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Michelle Lecolle. 2019. *Les noms collectifs humains en français. Enjeux sémantiques, lexicaux et discursifs*. Lambert-Lucas, Université de Lorraine.

Jan Lenhart and Franziska Heckel. 2024. Effects of gender-fair language on the cognitive representation of women in stereotypically masculine occupations and occupational self-efficacy among primary school girls and boys. *Sex Roles*, 91:1–15.

Paul Lerner and Cyril Grouin. 2024. INCLURE: A Dataset and Toolkit for Inclusive French Translation. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 59–68, Torino, Italia. ELRA and ICCL.

Jean-Daniel Lévy, Gaspard Lancrey-Javal, and Morgane Hauser. 2017. L'écriture inclusive : la population française connaît-elle l'écriture inclusive ? Quelle opinion en a-t-elle ? Rapport d'étude, Harris Interactive.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2025. DataComp-LM: In search of the next generation of training sets for language models. *Preprint*, arXiv:2406.11794.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2281–2287, Austin, Texas. AAAI Press.

Cyril Liénardy, Julia Tibblin, Pascal Gygax, and Anne-Catherine Simon. 2023. Écriture inclusive, lisibilité textuelle et représentations mentales. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (33).

Anna Lindqvist, Emma Aurora Renström, and Marie Gustafsson Sendén. 2019. Reducing a Male Bias in Language? Establishing the Efficiency of Three Different Gender-Fair Language Strategies. *Sex Roles*, 81(1-2):109–117.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2025. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *Conference on Language Modeling 2024*.

R. Anthony Lodge. 1993. *French: From Dialect to Standard*. Routledge, London and New York.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security*, volume 12300, pages 189–202. Springer International Publishing, Cham.

Wenhan Lyu, Yimeng Wang, Tingting (Rachel) Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pages 63–74, Atlanta, GA, USA. Association for Computing Machinery.

Franklin Ma and Alan J. Hu. 2025. Logically constrained decoding. In *Proceedings of the 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 150–167, Suzhou, China. Association for Computational Linguistics.

Donald G. MacKay and David C. Fulkerson. 1979. On the comprehension and production of pronouns. *Journal of Verbal Learning and Verbal Behavior*, 18(6):661–673.

Paolo Mainardi, Federico Garcea, and Alberto Barrón-Cedeño. 2025. Fine-tuning vs prompting techniques for gender-fair rewriting of machine translations. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 320–337, Vienna, Austria. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. De Souza, Alexandra Birch, and André F.T. Martins. 2025. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62.

Wendy Martyna. 1978. What Does 'He' Mean?: Use of the Generic Masculine. *Journal of Communication*, 28(1):131–138.

Elisa Merkel, Cristina Cacciari, Martina Faralli, and Anne Maass. 2017. It only needs one man or can mixed groups be described by feminine generics? *Rivista di psicolinguistica applicata*, (XVII).

Megan M. Miller and Lori E. James. 2009. Is the generic pronoun he still comprehended as excluding women? *The American Journal of Psychology*, 122(4):483–496.

Julia Misersky, Asifa Majid, and Tineke M. Snijders. 2019. Grammatical Gender in German Influences How Role-Nouns Are Interpreted: Evidence from ERPs. *Discourse Processes*, 56(8):643–654.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *IEEE Access*, 13:134615–134633.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nishanth Sridhar Nakshatri, Shamik Roy, Rajarshi Das, Suthee Chaidaroon, Leonid Boytsov, and Rashmi Gangadharaiah. 2025. Constrained Decoding with Speculative Lookaheads. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4681–4700, Albuquerque, New Mexico. Association for Computational Linguistics.

Daye Nam, Malgorzata Salawa, and Satish Chandra. 2026. Reading between the lines: Scalable user feedback via implicit sentiment in developer prompts. In *Proceedings of the IEEE/ACM 48th International Conference on Software Engineering*, Icse-Nier '26, pages 11–15, New York, NY, USA. Association for Computing Machinery.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Svetoslav Nizhnichenkov, Rahul Nair, Elizabeth Daly, and Brian Mac Namee. 2026. A representation-level assessment of bias mitigation in foundation models. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 195–210, Cham. Springer Nature Switzerland.

Guilherme Guimarães Nomelini and Carla Bonato Marcolin. 2024. Gender bias in large language models: A job postings analysis. *RAM. Revista de Administração Mackenzie*, 25(6):eRAMD240056.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 OLMo 2 Furious. *Preprint*, arXiv:2501.00656.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. *Preprint*, arXiv:2203.02155.

Akhil Rajeev P and Manoj Balaji Jagadeeshan. 2026. IHLC@LT-EDI 2026: Steering toward inclusivity - a representation engineering for gender-neutral rewriting. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 177–181, Virtual (Online). Association for Computational Linguistics.

Aiswariya p Nair, Sree S Bhagya, and Chinnu Jacob. 2026. CAI@LTEDI 2026: Multilingual gender inclusive language generation using instruction-guided mT5 transformer model. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 150–154, Virtual (Online). Association for Computational Linguistics.

Laura Mathilde Pabst and Marlene Kollmayer. 2023. How to make a difference: The impact of gender-fair language on text comprehensibility amongst adults with and without an academic background. *Frontiers in Psychology*, 14:1234860.

Laura L Paterson. 2011. Epicene pronouns in UK national newspapers: A diachronic study. *ICAME*, 35:171–186.

Francisco Perales, Christine Ablaza, and Nicki Elkin. 2025. Who Are 'They'? Gender-Neutral Pronoun Adoption by Non-Binary People and Other Gender Groups. *Sex Roles*, 91(4):26.

Giorgio Perissinotto. 1983. Spanish hombre: Generic or Specific? *Hispania*, 66(4):581.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. Gender-neutral rewriting in Italian: Models, approaches, and trade-offs. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 899–910, Cagliari, Italy. CEUR Workshop Proceedings.

Mat Pires. 2020. *Assessing the Mixed or Generic Feminine as an Inclusive Language Strategy*, page Chapter_4715. Fondazione Università Ca' Foscari, Venice.

David Pomerenke. 2022. INCLUSIFY: A benchmark and a model for gender-inclusive German. *Preprint*, arXiv:2212.02564.

A Pranav, Janiça Hackenbuchner, Giuseppe Attanasio, Manuel Lardelli, and Anne Lauscher. 2025. Glitter: A multi-sentence, multi-reference benchmark for gender-fair German machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18450–18477, Suzhou, China. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

Irina Proskurina, Antoine Gourru, and Julien Velcin. 2026. The Fairness Collapse Phenomenon: Bias Amplification in Language Models Trained on Synthetic Data. *Preprint*, arXiv:2608.04268.

Zhanyue Qin, Yue Ding, Deyuan Liu, Qingbin Liu, Junxian Cai, Xi Chen, Zhiying Tu, Dianhui Chu, Cuiyun Gao, and Dianbo Sui. 2025. LFTF: Locating First and Then Fine-Tuning for Mitigating Gender Bias in Large Language Models. *Preprint*, arXiv:2505.15475.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Preprint*, arXiv:2305.18290.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden. Association for Computational Linguistics.

Célia Richy and Heather Burnett. 2021. Démêler les effets des stéréotypes et le genre grammatical dans le biais masculin : une approche expérimentale. *GLAD!*, 10.

Patrick Rothermund and Fritz Strack. 2024. Reminding May Not Be Enough: Overcoming the Male Dominance of the Generic Masculine. *Journal of Language and Social Psychology*, 43(4):468–485.

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.

Nathan C Ryan, Sara Stoudt, and Florencia Vecchione. 2026. Spatial and temporal trends in the use of gender inclusive language: A study of Spanish-language Twitter and news media. *Digital Scholarship in the Humanities*, 41(2):950–968.

Rajendran S, Ramkumar N, and Malarselvi R. 2026. Igniters@LTEDI 2026: Multilingual gender-inclusive language generation with mT5 and counter-narrative generation using llama-3. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 172–176, Virtual (Online). Association for Computational Linguistics.

Caitlin E. Samples. 2024. Gender-inclusive Language in Twitter. *Hispania*, 107(1):pp. 139–160.

Sayaka Sato, Pascal M. Gygax, and Ute Gabriel. 2016. Gauging the Impact of Gender Grammaticization in Different Languages: Application of a Linguistic-Visual Paradigm. *Frontiers in Psychology*, 7:19.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–267, St. Julian's, Malta. Association for Computational Linguistics.

Jasmin Schmidt and Claudia Hess. 2026. Tackling the Generic Masculine: Evaluating Gender Neutrality of German AI-Generated Texts. *International Conference on Gender Research*, 9(1):249–256.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7.

Mariia Sedykh. 2025. Gender in Translation by MT and LLMs. Master's thesis, Itä-Suomen yliopisto, Finland, May.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint*.

Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. 2026. Bridging the black box: A survey on mechanistic interpretability in AI. *Acm Computing Surveys*, 58(8).

Elsa Spinelli, Jean-Pierre Chevrot, and Léo Varnet. 2023. Neutral is not fair enough: Testing the efficiency of different language gender-fair strategies. *Frontiers in Psychology*, 14:1256779.

Dagmar Stahlberg, Sabine Sczesny, and Friederike Braun. 2001. Name Your Favorite Musician: Effects of Masculine Generics and of their Alternatives in German. *Journal of Language and Social Psychology*, 20(4):464–469.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. http://arxiv.org/abs/2112.14168. *Preprint*, arXiv:2112.14168.

Benjamin Storme and Martin Storme. 2025. Feminization is More Gender-Fair Than Neutralization: Evidence From Gender-Stereotyped Contexts. *Journal of Language and Social Psychology*, page 0261927X251349151.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv preprint*.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005, Vienna, Austria. Association for Computational Linguistics.

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

Julia Tibblin, Pascal Gygax, Joost Van De Weijer, and Jonas Granfeldt. 2026. Recognized but harder to integrate: An eye-tracking study of French gender-fair forms during reading. *Applied Psycholinguistics*, 47:e4.

Julia Tibblin, Joost Van De Weijer, Jonas Granfeldt, and Pascal Gygax. 2023b. There are more women in "joggeur·euses" than in "joggeurs": On the effects of gender-fair forms on perceived gender ratios in French role nouns. *Journal of French Language Studies*, 33(1):28–51.

Together Computer. 2023. RedPajama: An open source recipe to reproduce LLaMA training dataset.

Ewoenam Kwaku Tokpo and Toon Calders. 2024. FairFlow: An Automated Approach to Model-based Counterfactual Data Augmentation For NLP. *Preprint*, arXiv:2407.16431.

Eva Vanmassenhove. 2024. Gender Bias in Machine Translation and The Era of Large Language Models. *Preprint*, arXiv:2401.10016.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claude Favre de Vaugelas. 1647. *Remarques sur la langue françoise utiles à ceux qui veulent bien parler et bien escrire*. Veuve Jean Camusat et Pierre le Petit, Paris.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.

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*, 11:574356.

Éliane Viennot. 2017. *Non, le masculin ne l'emporte pas sur le féminin ! petite histoire des résistances de la langue française*, nouvelle éd. augmentée edition. Xx-y-z. Éditions iXe, Donnemarie-Dontilly.

Éliane Viennot. 2022. De la parenthèse au point médian. Des nouveaux habits de l'écriture inclusive et de la malhonnêteté de ses opposant·es. *Travail, genre et sociétés*, 47(1):165–168.

Anica Waldendorf. 2024. Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, 40(2):357–374.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. 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.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Alina Wróblewska and Bartosz Żuk. 2025. Integrating gender inclusivity into large language models via instruction tuning. *Preprint*, arXiv:2508.18466.

Hualin Xiao, Brent Strickland, and Sharon Peperkamp. 2023. How Fair is Gender-Fair Language? Insights from Gender Ratio Estimations in French. *Journal of Language and Social Psychology*, 42(1):82–106.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. 2025. BiasEdit: Debiasing stereotyped language models via model editing. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 166–184, Albuquerque, New Mexico. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and Mitigating Gender Bias in LLMs via Interpretable Neuron Editing. *Preprint*, arXiv:2501.14457.

Lisa Zacharski, Alexandra Kruppa, and Evelyn C. Ferstl. 2025. The readability of the non-binary gender star in German: Evidence from a lexical decision task. *Social Psychological Bulletin*, 20:e13719.

Joshua Zami and Barbara Hemforth. 2024. Intelligibilité de l'écriture inclusive : Une approche expérimentale. *SHS Web of Conferences*, 191:10004.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. 2025. GenderAlign: An alignment dataset for mitigating gender bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11293–11311, Vienna, Austria. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Nips '23, Red Hook, NY, USA. Curran Associates Inc.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Responses to Common Questions and Objections

This appendix addresses several questions and objections that are frequently raised in discussions about MG and GFL. It also helps clarify our position on the topic.

> *Does this paper argue that everyone should be using GFL?*

No. This paper is not prescriptive about how people should write or speak. Rather, it specifically focuses on conversational LLMs and what they generate. We argue that efforts should be developed to create alternative versions of these models capable of naturally producing GFL in place of MG to counteract any bias occurring when reading LLM-generated content, with users being, preferably, able to choose their preferred strategy.

> *Inclusive forms are hard to read, negatively impact understanding and slow down reading.*

Most psycholinguistic studies find no negative effect of GFL on text comprehension (Friedrich and Heise, 2019; Vergoossen et al., 2020; Pabst and Kollmayer, 2023; Cruz Neri and Siems-Muntoni, 2026), including for language learners (Decock et al., 2026). Where an increase in reading time has been observed, it is not due to GFL itself but to readers' unfamiliarity and lack of exposure to these forms, with a strong habituation effect emerging as exposure increases (Gygax and Gesto, 2007; Zami and Hemforth, 2024; Zacharski et al., 2025; Tibblin et al., 2026). Other studies find no increased reading time at all (Girard et al., 2022; Liénardy et al., 2023). It is also worth noting that this objection is typically made against the most visually salient GFL strategy, that is typographic marking. Neutralization strategies, on the other hand, mostly rely on existing and familiar lexicon and do not use new or repurposed typography.

> *LLMs defaulting to MG just reflects standard grammar.*

This objection assumes "standard grammar" is a neutral fact that models simply mirror. Nonetheless, sociolinguistic research has long shown that a "standard" language variety is a social construct, upheld by institutions with symbolic authority, not an objective fact about language (Bourdieu, 1991; Lodge, 1993). Crucially, MG itself is a product of this kind of prescriptivist intervention: for example, Viennot (2017) shows that before the 17th century, French had adjectives and past participles agree with whichever noun was closest, regardless of its gender, rather than defaulting to the masculine.[7] Similarly, feminine forms of job titles were also common. Starting in the 17th century, prescriptivist grammarians such as Claude Favre de

[7] This phenomenon is called "accord de proximité" (proximity agreement), and has been studied both from a diachronic and a linguistic perspective (Abeillé et al., 2018; Aixiu and Abeillé, 2021), with findings suggesting a resurgence in contemporary usage (Abeillé et al., 2022).

Vaugelas (1585-1650) deliberately ruled that "the masculine prevails over the feminine" (Vaugelas, 1647, p. 696), and suppressed feminine occupational forms, motivated explicitly by the belief that the masculine gender was both grammatically and socially superior to the feminine. These interventions were then codified into grammar books and taught in schools, and persist until today. Section 3 has extensively highlighted the evidence on why using MG is problematic and the effects it produces on people that are exposed to it. Saying that an LLM "just reflects standard language" does not solve these issues and, on the contrary, participates in making MG a reasonable writing option in all circumstances in spite of the wider consequences.

B GFL Forms in Pre-training Data and infini-gram Details

The infini-gram (Liu et al., 2025) API[8] allows one to query a set of tokenized pre-training corpora by number of tokens or documents. We leveraged this to analyze the evolution of GFL in pre-training data in three languages: German, French and Spanish.

For German, we used a list of 4,627 singular and plural typographic marking star forms (e.g., "Student*innen") compiled by Pomerenke (2022) and originally extracted from the DeReKo corpus (Kupietz et al., 2018).

For French, we used a list of 512 manually verified human nouns with 5 different GFL typographic marking strategies (see Table 3) in the singular and plural forms. We report separate singular and plural counts for strategies `paren`, `brackets` and `middot`. For strategies `dot` and `slash`, we deliberately chose to query nouns with double marking (e.g., "étudiant.e.s"), because the querying API does not take into account word boundaries, which would lead to occurrences in code or URLs to be retrieved. We did not notice such issues with the `middot` strategy as its usage is limited to GFL, so for this strategy specifically, both single marking (e.g., "étudiant·es") and double marking (e.g., "étudiant·e·s") are counted.

[8] https://infini-gram.readthedocs.io/en/latest/api.html

Table 3: French GFL typographic marking strategies used for querying

Strategy	Example
brackets	étudiant[e]s
dot (double)	étudiant.e.s
middot	étudiant·es
paren	étudiant(e)s
slash (double)	étudiant/e/s

Finally, for Spanish, we initially extracted 15,591 plural `@` forms (e.g., "estudiant@s") from Spanish social media corpora 🤗 pysentimiento/spanish-tweets and 🤗 agentlans/bluesky, and validated their usage as nouns in context using spaCy 3.8.4 (Honnibal et al., 2023) and the `es_dep_news_trf` Spanish Transformer pipeline for POS tagging, narrowing the amount of forms down to 1,596. However, similarly to French, we encountered word boundary matching issues where the query would retrieve email addresses. Consequently, the 1,596 `@` forms were converted into less ambiguous "x" forms (e.g., "estudiantxs") and requeried.

We queried 7 different corpora among the 13 made available through the API, listed in Table 4. All corpora were queried on the train (largest) split, if applicable. Results are presented in Table 5 for all three languages. Strategy-specific counts for French are made available in Table 6. We also release all the code and data used for these analyses to a GitHub repository: https://github.com/spidersouris/linguistic_gender_bias .

Table 4: Corpora queried through the infini-gram API (sorted by release date) and their number of tokens

Corpus	Tokens
The Pile (Gao et al., 2020)[1]	383.3B
C4 (processed)[2]	198B
RedPajama (Together Computer, 2023)[3]	1.39T
Dolma v1.7 (Soldaini et al., 2024)[4]	2.60T
DCLM (Li et al., 2025)[5]	4.34T
OLMo-Mix[6]	4.58T
OLMo-2 (OLMo et al., 2025)[7]	4.61T

[1] https://huggingface.co/datasets/EleutherAI/pile
[2] https://huggingface.co/datasets/allenai/c4
[3] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T
[4] https://huggingface.co/datasets/allenai/dolma
[5] https://huggingface.co/datasets/mlfoundations/dclm-baseline-1.0
[6] https://huggingface.co/datasets/allenai/olmo-mix-1124
[7] https://huggingface.co/allenai/OLMo-2-0325-32B-Instruct

Results show that while there is a general increase in raw counts of GFL forms in later corpora, the proportion of GFL measured by the rate per billion tokens remains low for all three languages, especially in more recent corpora. Indeed, while we notice rates of 21.03 and 39.13 in the Pile corpus for French and German respectively, the last three pre-training corpora released between 2024 and 2025 show an important decrease ranging from 9.35 (Dolma) to 5.51 (DCLM) for French, and from 9.56 (Dolma) to 4.43 (DCLM) for German. This suggests that the observed increase in raw counts can primarily be explained by the continuous growth in the overall size of pre-training corpora, rather than by a greater relative presence of GFL. For Spanish, the picture is a bit more nuanced, with GFL slowly achieving a larger proportion in the latest corpora, closer to the proportions from The Pile than other languages (even though that proportion was, in comparison, not very high to begin with).

When it comes to French-specific strategies, the breakdown shows that the parenthesis strategy is by far the most common, which is not that surprising given that it is the oldest strategy and the one preferred for administrative or legal documents. In spite of its controversial nature (Viennot, 2022), the middle dot is still very much present (even if, as stated above, counts include both single and double marking compared to other strategies), followed by the dot strategy. Given the close results between the middot and the dot strategies, it is highly likely that, all in all, the dot strategy is the most used one (most probably due to the difficulty of typing the middot character on computer and mobile devices), even though the limitations of the API do not allow us to confirm this. Slash (double marking) and brackets appear to be the least found strategies in the pre-training corpora.

Further analysis and more precise counts are needed to better understand the concrete impact of these results on LLM pre-training and LLMs' propensity to generate (or avoid generating) GFL, including at different stages of training. It should be noted that pre-training corpora differ substantially in the types of texts they collect, the domains they cover, and the balance between these sources, which may naturally lead to large differences in the amount of GFL they contain. Furthermore, methodologies to collect and filter data highly vary, and texts containing GFL may be filtered out (justly or unjustly) based on text quality measures and other methodology-specific factors. Finally, additional languages should be tested as well; more publicly available resources specifically for GFL in other languages could help for this kind of task. With that being said, we believe this might help give an overview of what kind of GFL data may be present in pre-training corpora and encourage more research on this subject, in line with research works dealing with the importance of pre-training data, especially for models' base knowledge and capabilities (Zhou et al., 2023; Baek et al., 2026).

Corpus	French	German	Spanish	Total
Pile	8,062 (21.03)	15,000 (39.13)	2,590 (6.76)	25,652 (66.92)
C4	1,806 (9.12)	869 (4.39)	724 (3.66)	3,399 (17.16)
RPJ	12,255 (8.84)	10,232 (7.38)	7,593 (5.48)	30,080 (21.70)
Dolma	24,349 (9.35)	24,910 (9.56)	13,098 (5.03)	62,357 (23.94)
DCLM	23,919 (5.51)	19,227 (4.43)	22,307 (5.14)	65,453 (15.08)
OLMo-2	32,597 (7.07)	21,101 (4.58)	23,942 (5.19)	77,640 (16.84)
Total	102,988 (7.62)	91,339 (6.75)	70,254 (5.19)	264,581 (19.56)

Count (rate per billion tokens)

Table 5: GFL counts and proportions by corpus for French, German and Spanish, ordered by corpus date

	French								
Corpus	**paren sg.**	**paren pl.**	**brackets sg.**	**brackets pl.**	**middot sg.**	**middot pl.**	**dot (doubled)**	**slash (doubled)**	**Total**
Pile	1,941 (5.06)	1,545 (4.03)	27 (0.07)	8 (0.02)	310 (0.81)	2,437 (6.36)	1,781 (4.65)	13 (0.03)	8,062 (21.03)
C4	1,248 (6.30)	356 (1.80)	12 (0.06)	0 (0.00)	10 (0.05)	59 (0.30)	120 (0.61)	1 (0.01)	1,806 (9.12)
RPJ	3,831 (2.76)	3,432 (2.48)	88 (0.06)	24 (0.02)	430 (0.31)	1,877 (1.35)	2,546 (1.84)	27 (0.02)	12,255 (8.84)
Dolma	8,376 (3.22)	8,584 (3.30)	79 (0.03)	16 (0.01)	814 (0.31)	3,194 (1.23)	3,186 (1.22)	100 (0.04)	24,349 (9.35)
DCLM	10,565 (2.43)	6,102 (1.41)	241 (0.06)	108 (0.02)	391 (0.09)	3,611 (0.83)	2,636 (0.61)	265 (0.06)	23,919 (5.51)
OLMo-2	13,866 (3.01)	10,393 (2.25)	252 (0.05)	114 (0.02)	638 (0.14)	4,162 (0.90)	2,898 (0.63)	274 (0.06)	32,597 (7.07)
Total	39,827 (2.94)	30,412 (2.25)	699 (0.05)	270 (0.02)	2,593 (0.19)	15,340 (1.13)	13,167 (0.97)	680 (0.05)	102,988 (7.62)

Count (rate per billion tokens)

Table 6: Strategy-specific GFL counts and proportions by corpus for French, ordered by corpus date

C Overview of Intralingual Gender Rewriting Systems

Table 7: Overview of intralingual gender rewriting systems (2019–2026), ordered chronologically.

Study	Language(s)	Strategy	Approach
Zmigrod et al. (2019)	es, he, fr, it	S	Rule-based (Markov random field)
Habash et al. (2019)	ar	S	Classifier + character-level Seq2Seq
Sun et al. (2021)	en	N	Rule-based + Seq2Seq
Vanmassenhove et al. (2021)	en	N	Rule-based + Seq2Seq
Alhafni et al. (2022)	ar	S	Rule-based + classifier + character-level Seq2Seq
Alrowili and Vijay-Shanker (2022)	ar	S	Character-level Seq2Seq
Diesner-Mayer and Seidel (2022)	de	I	Rule-based
Pomerenke (2022)	de	N/I	Rule-based
Veloso et al. (2023)	pt	N	Rule-based + Seq2Seq
Cerabolini et al. (2024)	it	I	Rule-based + Seq2Seq + LLM prompting
Lerner and Grouin (2024)	fr	I	Rule-based + Seq2Seq
Tokpo and Calders (2024)	en	S	Seq2Seq
Doyen and Todirascu (2025a)	fr	N	Rule-based + Seq2Seq + LLM prompting
Greco et al. (2025)	it	N/I	Transformer classifier + Seq2Seq
Mainardi et al. (2025)	it	N	Seq2Seq + LLM prompting
Piergentili et al. (2025)	it	N	LLM prompting / fine-tuning
Wróblewska and Żuk (2025)	pl	N/I	LLM instruction tuning
Adhikary et al. (2026)	en, de, es, ta, kn	N	Rule-based + LLM (retrieval-augmented)
Balaji et al. (2026)	en	N	Seq2Seq
p Nair et al. (2026)	en, de, es, ta, kn	N	Seq2Seq
P and Jagadeeshan (2026)	en	N	LLM prompting / steering
S et al. (2026)	en, de, es, ta, kn	N	Seq2Seq

Strategy: S = gender switching, I = inclusive rewriting, N = neutral rewriting. Languages are given as ISO 639-1 codes.

D User Growth in Conversational LLMs

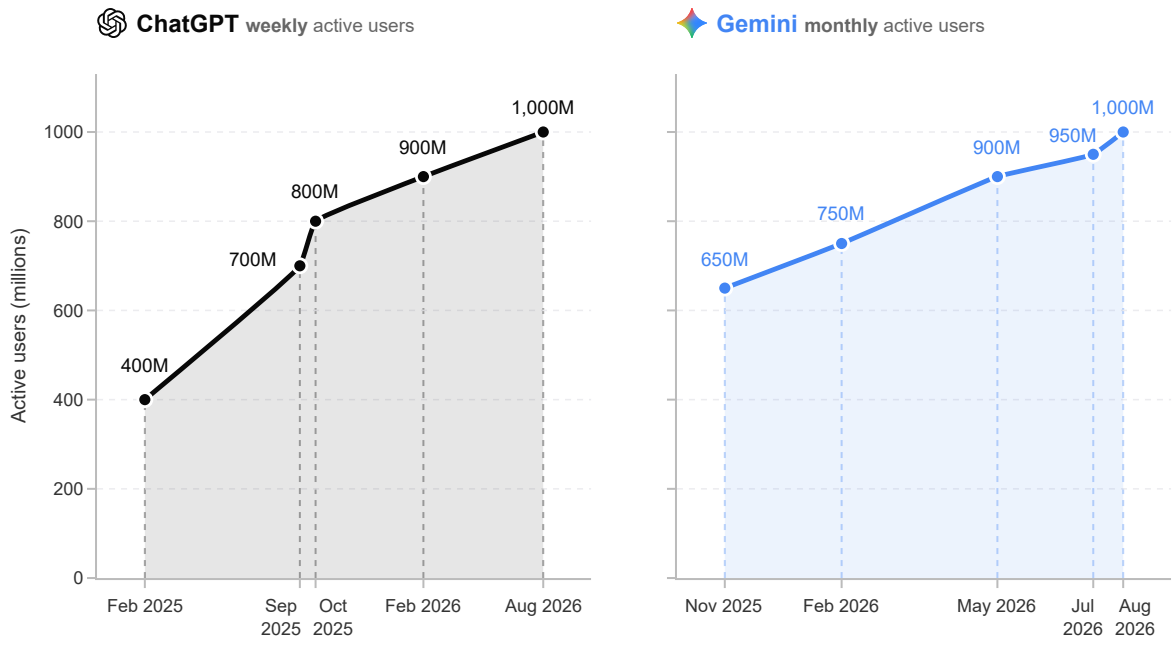


Figure 3: Evolution of weekly active users for ChatGPT, and monthly active users for Gemini

Table 8: Evolution of the number of active users for major conversational LLMs, with sources

Date	Company	Count	Source
December 4, 2024	OpenAI (ChatGPT)	300M weekly active users	X account @OpenAINewsroom[1]
February 20, 2025	OpenAI (ChatGPT)	400M weekly active users	Brad Lightcap, OpenAI COO, via CNBC[2]
May 20, 2025	Google (Gemini)	400M monthly active users	Sundar Pichai, Google CEO, Google I/O 2025[3]
May 28, 2025	Meta AI	1B monthly active users	Mark Zuckerberg, Meta CEO, annual shareholders meeting, via CNBC[4]
October 6, 2025	OpenAI (ChatGPT)	800M weekly active users	Sam Altman, OpenAI CEO, OpenAI DevDay, via TechCrunch[5]
February 4, 2026	Google (Gemini)	750M monthly active users	Sundar Pichai, Alphabet Q4 2025 earnings call[6]
February 27, 2026	OpenAI (ChatGPT)	900M weekly active users	OpenAI blog post[7]
May 19, 2026	Google (Gemini)	900M monthly active users	Sundar Pichai, Google I/O 2026[8]
July 22, 2026	Google (Gemini)	950M monthly active users	Sundar Pichai, Alphabet Q2 2026 earnings call[9]
July 31, 2026	OpenAI (ChatGPT)	1B weekly active users	OpenAI blog post[10]
August 11, 2026	Google (Gemini)	1B monthly active users	Google blog post[11]

[1] https://x.com/OpenAINewsroom/status/1864373399218475440
[2] https://www.cnbc.com/2025/02/20/openai-tops-400-million-users-despite-deepseeks-emergence.html
[3] https://blog.google/innovation-and-ai/technology/ai/io-2025-keynote/
[4] https://www.cnbc.com/2025/05/28/zuckerberg-meta-ai-one-billion-monthly-users.html
[5] https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/
[6] https://abc.xyz/investor/events/event-details/2026/2025-Q4-Earnings-Call-2026-Dr_C033hS6/default.aspx
[7] https://openai.com/index/scaling-ai-for-everyone/
[8] https://blog.google/innovation-and-ai/sundar-pichai-io-2026/
[9] https://abc.xyz/investor/events/event-details/2026/2026-Q2-Earnings-Call-2026-GgTAq7Is0z/default.aspx
[10] https://openai.com/index/building-abundant-intelligence/
[11] https://blog.google/innovation-and-ai/products/gemini-app/one-billion-monthly-users/