GeNRe: a French Gender-Neutral Rewriting System Using Collective Nouns

Anonymous ACL submission

Abstract

A significant portion of the textual data used in the field of Natural Language Processing (NLP) exhibits gender biases, particularly due to the use of masculine generics (masculine words that are supposed to refer to mixed groups of men and women), which can perpetuate and amplify stereotypes. Gender rewriting, a NLP task that involves automatically detecting and replacing gendered forms with neutral or opposite forms (e.g., from masculine to feminine), can be employed to mitigate these biases. While such systems have been developed in a number of languages (English, Arabic, Portuguese, German, French), automatic use of gender neutralization techniques (as opposed to inclusive or gender-switching techniques) has only been studied for English. This paper presents GeNRe, the very first French gender-neutral rewriting system using collective nouns, which are gender-fixed in French. We introduce a rule-based system (RBS) tailored for the French language alongside two fine-tuned large language models trained on data generated by our RBS. We also explore the use of instruction models to enhance the performance of our other systems and find that Claude 3 Opus combined with our dictionary achieves results close to our RBS. Through this contribution, we hope to promote the advancement of gender bias mitigation techniques in NLP for French.

1 Introduction

Since the 1970s, several psycholinguistic studies have focused on how language influences thoughts (Berlin and Kay, 1969; Kay and McDaniel, 1978). Further studies examining gender in language showed that it could lead to cognitive biases (Jacobson and Insko, 1985; Sczesny et al., 2016), particularly when it comes to the use of masculine generics (MG), that is masculine words that are supposed to refer to mixed groups of men and women (Braun et al., 2005; Richy and Burnett, 2021; Gygax et al., 2008, 2019). For example, Stahlberg et al. (2001) showed that when asked to name a celebrity in a certain field in German, respondents were more likely to give the name of a man when a masculine generic was used in the question.

Gender bias in natural language processing (NLP) models is a critical issue that can lead to biased predictions and the amplification of bias in the training data (Ducel et al., 2024; Lu et al., 2020; Stanczak and Augenstein, 2021; Kotek et al., 2023). This problem is particularly relevant for machine translation systems, which are highly susceptible to gender bias when translating between languages with different grammatical gender systems (Savoldi et al., 2021; Vanmassenhove, 2024). Data augmentation, which involves balancing the amount of data for all genders in a specific language, has been proposed as a potential solution to debias NLP systems (Zhao et al., 2018). This led to the development of an NLP task known as "gender rewriting," whose goal is to automatically propose alternatives to sentences containing MG.

As of yet, automatic gender neutralization techniques have only been developed in English (Vanmassenhove et al., 2021; Sun et al., 2021). Thus, we develop a French gender-neutral rewriting system using human collective nouns (CN), defined by Lecolle (2019) as "nouns referring to entities comprised of groups of individuals."[1] CNs have been widely discussed in the literature, especially when it comes to French (Flaux, 1999; Lammert, 2010; Lammert and Lecolle, 2014; Lecolle, 2019). Since, in French, this type of noun has a gender which does not depend upon the referent's,[2] it is an effective way of achieving gender neutralization. This gender-neutral rewriting system, GeNRe (**Ge**nder-**N**eutral **Re**writing System Using French Collective

[1] In French: « nom désignant une entité composée d'un ensemble d'individus humains. »

[2] For instance, "la police" ("police") refers to both policemen and policewomen.

Nouns), is the very first gender-neutral rewriting system for French[3] and could foster the development of other types of gender rewriting systems for that language in the future.

2 The Task of Gender Rewriting

While Alhafni et al. (2022b) were the first to define this task as "gender rewriting," similar efforts had already been pursued for Arabic (Habash et al., 2019), German (Pomerenke, 2022), and English (Sun et al., 2021). Alhafni et al. (2022b) define this task as: "generating alternatives of a given Arabic sentence to match different target user gender contexts." (2). While this definition works well for the work by Alhafni et al. (2022b), as they focus specifically on Arabic and create a system to switch between the masculine gender and the feminine gender, it is not universally applicable. Indeed, among the aforementioned works, several approaches to gender rewriting have been explored: Habash et al. (2019) and Alhafni et al. (2022a) developed a system to transform Arabic sentences with masculine words into sentences with feminine equivalents, and vice versa. The system created by Pomerenke (2022) provides inclusive suggestions for input sentences in German and has led to the publication of an online resource letting the user choose the type of inclusive transformation to apply. More recently, Veloso et al. (2023) also developed an inclusive gender-rewriting system for Portuguese, and Lerner and Grouin (2024) for French. Finally, Sun et al. (2021), Vanmassenhove et al. (2021) and He et al. (2021) created systems to neutralize gender in an English input sentence, but no such system exists for French. As part of this work and in order to accomodate a larger amount of languages and transformation types, we reframe the initial task definition given by Alhafni et al. (2022b) as "generating one or more alternative sentences that either neutralize gender, adopt inclusive forms, or switch to a different gender".

3 Gender in French

In French, nouns (N) are classified as either masculine or feminine, and the gender of a noun influences the form of determiners (D), adjectives (A) and past participle verbs (V) that are syntactically related. Similarly, coreferent pronouns (P), that is pronouns that are used to refer to something which has already been mentioned previously, also feature the same gender. Examples 1 (masculine) and 2 (feminine) highlight the syntactic differences that arise when using either a masculine unanimate noun ("courrier", *mail*) and a feminine unanimate noun ("lettre", *letter*).

D N A V
(1) **Le** courrier **recommandé** a été **écrit** récemment. **Il** est **adressé** à son mari.
P V
(The registered letter [m.] has been recently written. It [m.] is adressed to her husband.)

D N A V
(2) **La** lettre **recommandée** a été **écrite** récemment. **Elle** est **adressée** à son mari.
P V
(The registered letter [f.] has been recently written. It [f.] is adressed to her husband.)

The gender of human role nouns reflects the sociological gender of the referent (for instance, "danseuse" refers to a female dancer), while gender of nouns referring to unanimated beings is arbitrary (Watbled, 2012).

The masculine gender for human nouns is considered to be the "default" gender in French, and can be used in a non-specific context (in the singular form, as in Example 3[4]) or to refer to groups of people composed of both men and women (in the plural form, as in Example 4).

(3) **Un professeur** doit savoir faire preuve d'autorité.
(A professor [m.] has to know how to show authority.)

(4) Les filles et les garçons sont **partis**.
(The girls and the boys **left** [m.].)

However, the use of masculine as the default gender can lead to both gender biases and invisibilizing women. While this also applies to other languages featuring a semantic grammatical gender system that classifies human nouns or pronouns based on real-world distinctions, as it has been demonstrated in studies conducted in German (Stahlberg et al., 2001) and English (Jacobson and Insko, 1985), when it comes to French in particular, a 2017 survey conducted by Harris Interactive (2017) following a methodology close to that of Stahlberg et al. (2001), found that when the masculine generic form is used, respondents tend to think of men. Similarly, according to Gabriel et al. (2018), masculine generic human nouns are more likely to be associated with male referents, and

[3] Code and data are made publicly available on GitHub, under license CC BY-SA 4.0 `https://github.com/REDACTED`

[4] In this example, "professeur" is considered as a masculne generics insofar as it does not refer to one specific male individual, but to any individual serving as "professor".

specifically highlighting the generic nature of MG does not have an effect on the biased perception of survey participants (Gygax et al., 2012).

Consequently, two main types of writing techniques can be used to avoid the use of MG: visibilization techniques and neutralization techniques.

Visibilization techniques seek to highlight the feminine ending of words by separating the masculine ending from the feminine one through the use of specific symbols (asterisk, interpunct: *professeur·e*, as in Example 5) or by affecting the feminine ending directly (using capital or bold letters). Neutralization techniques, on the other hand, mainly revolve around different types of words: epicene words, that is words whose form is the same for masculine and feminine, whether they may have a generic (e.g. "personne", *person*, as in Example 6) meaning or a specific (e.g. "spécialiste", *specialist*) one, or words that refer to groups of people, such as CNs (e.g. "lectorat", *readership*), these having a fixed gender which is not associated with the gender of the people within that group.

(5) **Un·e professeur·e** doit savoir faire preuve d'autorité.
(A [m./f.] professor [m./f.] has to know how to show authority.)

(6) **Une personne professeure** doit savoir faire preuve d'autorité.
(A person teaching has to know how to show authority.)

Given the impact of inclusive formulations on mitigating gender biases, developing a system capable of automatically rewriting text to reduce the prevalence of MG could be a valuable tool for data augmentation. Notably, to the best of our knowledge, no such system has been developed specifically for the French language, making this a pioneering effort in the field. By focusing on gender neutralization, our work aims to fill this gap and explore the potential of CNs and epicene words in promoting more inclusive language. We specifically chose to focus on gender neutralization due to it being a less explored issue in research comparatively to visibilization techniques. While works on gender neutralization and its application to NLP tools exist in Italian (Piergentili et al., 2023) and German (Lardelli and Gromann, 2023), no such efforts have been pursued for French. By specifically focusing on the use of CNs for gender neutralization, we aim to see how effective they can be as their usage in everyday language is still restricted to a few words and their full potential has not yet been explored.

Moreover, by focusing on gender neutralization, our work targets a writing technique that, compared to visibilization, tends to be less contentious among native French speakers, as it does not alter the spelling of existing words nor does it introduce non-standard or new punctuation marks to separate the feminine suffix from the base word form (Burnett and Pozniak, 2021). Finally, gender-neutralization challenges the binary male/female gender dichotomy and is better adapted for people whose gender falls outside of the traditional categories.

4 Methodology

To build our automatic gender-neutralization system, we propose three different approaches: a rule-based approach, a model fine-tuning approach, and an instruction model approach. To build the resources used for these systems, we first create a dictionary of French CNs and their member noun counterparts, which we describe in Section 4.1. In Section 4.2, we then give details about the datasets that we extracted sentences from for the development of our rule-based system, large language model (LLM) fine-tuning and evaluation. Finally, in Section 4.3, we explain our experimental design with the aforementioned model types. While our work focuses specifically on French, the methodology presented below is applicable to any language which can use collective nouns as a gender-neutralizing technique (e.g., Spanish) given a dictionary of human-member nouns. When it comes to syntactic changes, especially considering gender and number, those would be very similar in other romance languages such as Spanish, Italian or Portuguese. As a result, the amount of work needed to adapt our methodology to these languages specifically would be much lower compared to syntactically or morphologically more complex languages.

4.1 Dictionary

First, we manually created a dictionary with French CNs and their member noun counterparts. Three approaches were used to fill this dictionary: literature review, manual collecting and semi-automatic collecting.

Literature review. French CNs have been extensively studied in the linguistic literature. We drew on the list of 138 CNs by Lecolle (2019), the most exhaustive list of French CNs to our knowledge, which provided a comprehensive starting point for our dictionary. Some nouns were excluded from our dictionary due to their polysemy or restrictive semantics. For example, the CN "troupe" has multiple meanings (*troop*, *troupe*), and its use would require specifying the associated subdomain or group members to avoid confusion. Similarly, the semantics of nouns like "trio" have a too restrictive semantics, only applying to groups of exactly three people. After careful selection, we retained 105 entries from Lecolle's list.

Manual collecting. We empirically collected CNs from media and Internet sources over an extended period. This approach allowed us to identify nouns not presented in the literature on CNs, providing a complementary perspective to the literature review. With this approach, we added 46 entries to our dictionary.

Semi-automatic collecting. We scraped the French version of Wiktionary[5] to retrieve CNs with the suffix "-phonie", which refer to speakers of a language (e.g. "anglophonie", *English-speaking world*). We developed a Python script to generate equivalent CNs by replacing the suffix "-phonie" with "-phone" (e.g. "anglophone"). This approach enabled us to efficiently collect a set of nouns that follow a specific pattern, adding 164 entries, manually checked.

In total, our dictionary thus contains 315 entries. Table 1 contains a few examples of entries in our dictionary.

4.2 Datasets

Using our dictionary, we searched for occurrences of masculine plural member nouns in a French Wikipedia dataset with 1.58 million texts (graelo, 2023)[6]. We extracted 292,076 sentences containing such nouns. In addition, we also extracted French sentences from the Europarl EN-FR corpus (Koehn, 2005), a corpus created from the proceedings of the European Parliament and available in 21 languages, including English and French. This corpus was filtered to include French sentences only, and 106,878

[5] https://fr.wiktionary.org/wiki/

[6] Dataset made available here: https://huggingface.co/datasets/graelo/wikipedia. License: CC-BY-SA-3.0

Collective noun	Member noun (masc. plural)
académie (*academy*)	académiciens (*academicians*)
armée (*army*)	soldats (*soldiers*)
milice (*militia*)	miliciens (*militiamen/women*)
artillerie (*artillery*)	artilleurs (*artillerists*)
auditoire (*listenership*)	auditeurs (*listeners*)
ballet (*ballet*)	danseurs (*dancers*)
police (*police*)	policiers (*police officers*)

Table 1: Collective noun-member noun dictionary overview

additional sentences were extracted for model fine-tuning and evaluation (total 398,954). Both of these corpora are made available for research purposes.

For the rule-based system specifically, tags were automatically added at the beginning and at the end of each member phrase in the extracted sentences, with the ID of the entry in the dictionary. This was done because member nouns may have several CN counterparts, leading to several different sentences being generated in addition to the main one. For instance, the member noun "soldats" (*soldiers*) could well be replaced with CNs "armée" (*army*) "bataillon" (*battalion*), "infanterie" (*infantry*) or "régiment" (*regiment*). As we used data generated by our rule-based system for model fine-tuning (see Section 4.3.2), this was especially useful to generate all the possible variations of the input sentence, and thus increase the number of examples the models were trained on. Moreover, the use of tags also helps ensure the member nouns to be replaced in the input sentence, as only those that are between tags will be taken into account. Example 7 shows how these tags are used.

(7) Un historique permet de lister <n-126>les auteurs</n> et de consulter les modifications successives de l'article par <n-68>ses rédacteurs</n>.
(A history allows one to list <n-126>the authors</n> and view successive modifications to the article by <n-68>its editors</n>.)

Finally, we created a corpus-specific evaluation dataset comprised of 250 sentences from each corpus (total 500), and we manually gender-neutralized each sentence to have gold sentences.

4.3 Models

In this section, we present three different model types for gender-neutral rewriting: a rule-based model, two fine-tuned language models, and an instruct-based language model. Each model takes a different approach to the task, allowing us to compare their performance.

4.3.1 Rule-based model

We developed a rule-based system (RBS) to automatically apply the correct syntactic rules when converting a member noun into a CN, which leads to number and gender changes in the sentence.

The RBS consists of two main components: a syntactic dependency detection component and a generation component.

The dependency detection component primarily relies on spaCy (Montani et al., 2024) with the fr_core_news_sm pipeline as well as a set of rules to detect the words that are syntactically related to the member noun that needs to be replaced.

The generation component replaces each member noun in the sentence with its CN counterpart found in the dictionary, adjusting the determiner, handling elision, and reinflecting the detected dependencies using *inflecteur* (Chuttarsing, 2021), a Python module leveraging the Delaf French morphological dictionary[7] and *french-camembert-postag-model*[8], a CamemBERT-based (Martin et al., 2020) part of speech (POS) tagging model for French. Our RBS also makes additional replacements for past participles and object pronouns as these are not always being well handled by the *inflecteur* Python module. If no member nouns are detected in the sentence, the original sentence will be returned instead as it is already considered gender-neutral. Figure 1 shows an overview of the rule-based model pipeline.

4.3.2 Fine-tuned models

Previous research on gender rewriting has focused on training neural models as well as fine-tuning large language models using data generated by RBS to improve task-specific performance. While some studies (Sun et al., 2021; Veloso et al., 2023) showed a decrease in performance compared to RBS, Vanmassenhove et al. (2021) found a notable

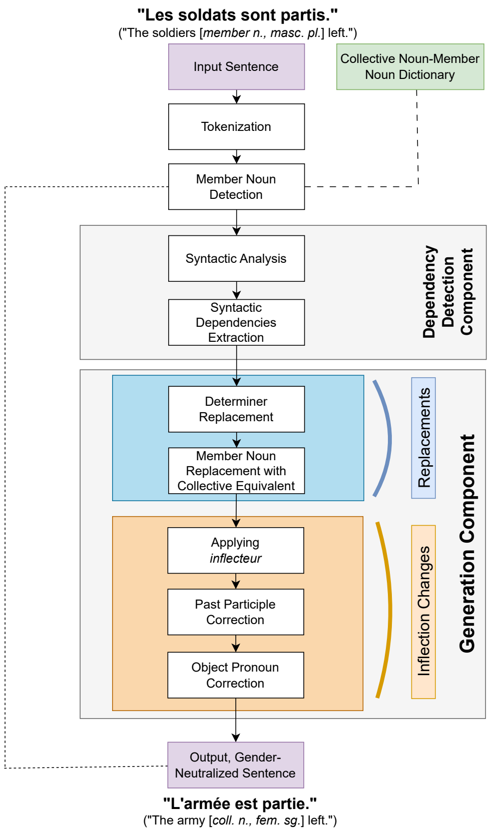


Figure 1: Rule-based model replacement pipeline overview

improvement of 0.27 in WER. We aim to investigate whether fine-tuning large language models can significantly improve the results of RBS, hypothesizing that the linguistic knowledge acquired by these models during training on large text corpora will help resolve errors in the training corpus and enhance results.

Two Seq2seq LLMs, t5-small (Raffel et al., 2020) and m2m100_418M (Fan et al., 2020), were selected for the experiments. Those models were chosen for their great text-to-text performance and their relatively small size, making the training process easier. Furthermore, as m2m100_418M had already been used by Veloso et al. (2023), we want to compare the results we can get for our specific task. Both models were fine-tuned using our two RBS-generated corpora (Wikipedia and Europarl) containing gender-neutralized and non-gender-neutralized sentence pairs. The training dataset for each model consisted of 60,000 sentence pairs per corpus, and the validation dataset had 6,000 (10%). Hyperparameters used for training are available in Appendix A.

[7] https://uclouvain.be/fr/instituts-recherche/ilc/cental/delaf-2-0.html

[8] https://huggingface.co/gilf/french-camembert-postag-model

Type	WER (↓)	BLEU (↑)	Cos. sim. (↑)
Baseline (unchanged)	13.35%	80.55	0.914
GeNRe-RBS	**3.40%**	93.43	**0.982**
GeNRe-T5	5.11%	90.68	0.968
GeNRe-M2M-100	5.40%	90.17	0.967
Claude 3 Opus-BASE	12.16%	82.98	0.925
Claude 3 Opus-DICT	3.75%	**93.64**	0.975
Claude 3 Opus-CORR	10.17%	85.13	0.95

Table 2: Results by model type. Bold indicates the best results overall.

4.3.3 Instruction model

The rapid development of LLMs and advances in NLP have demonstrated the ability to manipulate language models' behavior to predict text continuations and perform specific tasks without explicit training, leading to instruction models such as InstructGPT (Ouyang et al., 2022), or, more recently, Llama 3 (Grattafiori et al., 2024) or DeepSeek-V3 (DeepSeek-AI et al., 2024). This is primarily achieved through the use of "prompts" or instructions given to the language model (Liu et al., 2021). While some studies have briefly mentioned the potential of instruction models to reduce gender biases in automatically generated texts, and have occasionally experimented with such models,[9] no gender rewriting study has yet analyzed their capabilities for this specific task. We chose Claude 3 Opus `claude-3-opus-20240229` due to its best text generation performance at the time of the experiments (Anthropic, 2024) and its API being free to use during the period the experiments were conducted.[10]

To comprehensively evaluate the performance of Claude 3 Opus, we designed three distinct types of instructions to test its ability to generate gender-neutral texts. Corresponding prompts are available in Appendix B.

- The "BASE" instruction provides a basic task description, asking the model to make the sentence inclusive by replacing MG with their CN equivalents, without explicitly specifying the replacement word.
- The "DICT" instruction leverages our collective noun dictionary and asks the model to replace MG with their corresponding CNs, those being explicitly mentioned. There are two different versions for the "DICT" instruction: "DICT-SG", used when only one generic masculine noun with a matching CN was found in the sentence, and "DICT-PL", used when several generic masculine nouns with matching CNs were found.
- The "CORR" instruction takes sentences generated by our RBS as input and tasks the model with correcting potential errors, such as mismatches between verb and adjective numbers and genders.

5 Results

To evaluate the performance of our different rewriting models, we leverage two evaluation metrics commonly used for the task of gender rewriting: Word Error Rate (WER) and BLEU (Papineni et al., 2002). JiWER 3.0.3[11] and sacrebleu 2.4.2[12] Python packages were used with default parameters. We also provide cosine similarity, an additional metric not used for previous gender rewriting works.

Average results of each model on the two corpora are available in Table 2.

The RBS and Claude 3 Opus-DICT achieved the best results in our experiments, with the RBS achieving 3.40% WER and 0.982 cosine similarity, and Claude 3 Opus-DICT achieving 93.64 BLEU. The fine-tuned models also showed mostly promising results, even though lower than the RBS and Claude 3 Opus DICT (5.11% WER, 90.68 BLEU and 0.968 cosine similarity for T5; 5.40% WER, 90.17 BLEU and 0.967 cosine similarity for M2M-100). Comparing the two of them, they achieved

[9] For instance, Veloso et al. (2023) tried to make use of OpenAI's ChatGPT to generate gender-inclusive sentences in Portuguese, and suggested that the use of instruction models could prove useful to automatically create gender-inclusive datasets.

[10] For the announcement, see `https://www.anthropic.com/news/claude-3-family`; for API usage, see `https://docs.anthropic.com/en/docs/about-claude/models`

[11] `https://pypi.org/project/jiwer/`

[12] `https://pypi.org/project/sacrebleu/`

similar results, with the T5 model slightly outperforming M2M-100. However, both models showed a minor decrease in performance compared to the RBS. As a result, similarly to Veloso et al. (2023) and in contrast with the findings of Vanmassenhove et al. (2021), we do not find a significant improvement compared to our RBS following fine-tuning.

6 Discussion

We provide the distribution of errors made by the models in Figure 2. For GeNRe-RBS, GeNRe-T5 and GeNRe-M2M-100 models, errors were manually annotated for the Europarl corpus[13] and agreed upon by two annotators based on previously defined error type descriptions and hypothetical examples. Error types are divided into three main categories: POS (ADJ, DET, DET_COREF, PRON_COREF, VERB), text generation (CASE, GEN_FAILURE, SPECIAL_CHAR) and other (ELISION, MISID_NOUN, PUNCT, SEM, UNREPLACED). Error categories were first created by looking at the sentences generated by the RBS and the fine-tuned models, and were then applied for each sentence and each model (500 sentences × 3). Multiple error types may be applied to one sentence. We provide a description of each error type in Table 6. Text generation errors, labeled with (N) both in Figure 2 and Table 6, are strictly specific to the fine-tuned models.

Since the text generated by instruction models is much less deterministic than what can be found in sentences modified by the RBS or the fine-tuned models, we have leveraged GPT-4o mini (OpenAI, 2024) and in-context learning (Brown et al., 2020) to generate error labels specific to the instruction model outputs. This was done in three steps: first, we asked GPT-4o mini to compare golden sentences and Claude 3 Opus-generated sentences for each instruction (BASE, DICT, CORR), and generate a short explanation of the errors. Second, these explanations were used to query the model once again (with the same previous context and sentences), this time asking it to generate error labels. Previously generated error labels were automatically added to the prompt as they were generated and the model was instructed to reuse any existing label if any matched the error type. Prompts and hyperparameters are available in Appendix E. Finally, we manually merged output labels with existing labels already applied for the RBS and fine-tuned models, and created high-level error label categories to have better visibility (e.g., “GENDER_AGREEMENT” and “NUMBER_AGREEMENT” error types were merged into a single “AGREEMENT” high-level label).

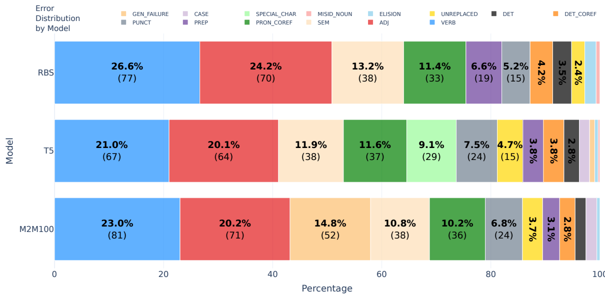
Across the RBS and the fine-tuned models, the most prominent error types are related to verbs and noun cases. Verbs account for 26.6% of errors for GeNRe-RBS, 21.1% for GeNRe-T5, and 22.9% for GeNRe-M2M-100. On the other hand, adjectives account for 24.2% of errors for GeNRe-RBS, and 20.1% for both GeNRe-T5 and GeNRe-M2M-100.

The M2M-100 model is highly prone to making token-specific generation errors (14.7%), this type of error being strictly specific to this model. Similarly, we find that the T5 model also makes specific errors related to the handling of special characters. We discuss these issues more in detail in Section 6.

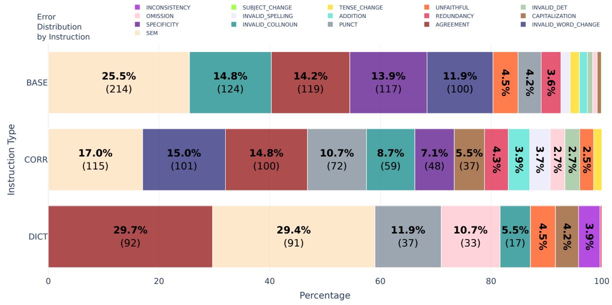
When it comes to instruction models, most errors are related to semantics (SEM) and agreement. SEM is the first error type for the BASE (25.5%; 214 occurrences) and CORR (17%; 115 occurrences) instruction types, and the second for the DICT instruction type (29.4%; 91 occurrences). AGREEMENT is the error type most found in the DICT instruction type (29.7%; 92 occurrences), and is the third most frequent error type for the BASE (14.2%; 119 occurrences) and CORR (14.8%; 100 occurrences) instruction types. Interestingly, the INVALID_COLLNOUN error type, which occurs when the corresponding MG member noun has not been replaced or the chosen/specified collective noun is not in accordance with the context, is the second most frequent error for the BASE instruction type. This may be due to the prompt being the least specific and leading to some nouns being incorrectly replaced by some collective noun equivalent, or left unreplaced.

A qualitative analysis of the generated sentences revealed that the RBS was making most of its errors when modifying adjectives and verbs. This is not surprising given that these two part-of-speech categories are the ones which require the most complex changes when transitioning from a member noun to a CN. Indeed, in French, adjectives undergo a certain number of changes when changing number or gender. Verbs can also have these same changes when used as past participles; otherwise, only number change will affect them. For instance, in Example F8, the verb “seront” (pl., *will be*) should have been changed to “sera” (sg.) to match with the new CN “citoyenneté” (*citizenry*).

[13] Human annotation for the Wikipedia corpus is ongoing.



(a) GeNRe-RBS, GeNRe-T5 and GeNRe-M2M-100



(b) Claude 3 Opus BASE, CORR, DICT

Figure 2: Error distribution for RBS, neutral and instruction models

Similarly, in Example F9, the adjective "chargés" (pl., *in charge of*) should match the new singular CN "parlement" (*parliament*) and be changed to "chargé".

When it comes to the fine-tuned models (T5 and M2M-100), analysis shows that they were able to generalize linguistic rules and correct dependencies that were not properly modified by the RBS, especially verbs and adjectives, slightly reducing the number of errors for these POS. As a result, in spite of their lower results compared to the RBS, fine-tuned models may still prove useful in certain scenarios where the RBS struggles to apply linguistic rules correctly, such as in sentences with complex dependencies or nuanced contextual relationships. Example F10 shows a case where the verb "vouloir" (*want*) is correctly inflected by the fine-tuned model.

Additionally, the fine-tuned models were capable of utilizing different CN equivalents from the dictionary (some CNs being associated to the same member noun).

Errors observed in the fine-tuned models and different from the RBS included token generation failures (M2M-100, Example F11, where "Nebski" was generated instead of "Zemski"), and incorrect generation of special characters (T5, as in Example F12 where "main-d'uvre" was generated instead of "main-d'œuvre" [*labour*]). The first error might come from the multilingual aspect of the model, as it may generate words or mix tokens from other languages, while the second error is probably due to the model being mostly trained on English data. For both models, we also found cases where words were not uppercased correctly, as in Example F13.

As far as the instruction model is concerned, Claude 3 Opus-BASE and CORR were found to be highly prone to altering the formulation of sentences, as shown in Example F14. Claude 3 Opus-DICT was found to have a similar effect, but to a much lesser extent, likely due to the increased precision of the prompt.

Notably, the DICT prompt was observed to generate sentences with correct verbs and adjectives, indicating its ability to effectively leverage the CN dictionary to produce grammatically accurate sentences. We give such an example in Example F15.

Nonetheless, among the errors made by Claude 3 Opus-DICT, we identified instances of unreplaced nouns, where the model failed to substitute the MG with their corresponding CN equivalents, such as in Example F16.

7 Conclusion

Our work represents a step towards addressing gender-biased textual data in French. We make three key contributions to the task of gender rewriting in NLP: 1) a dictionary of French CNs and their corresponding member nouns, which serves as a resource for future research in this area; 2) a dataset of gender-neutralized and non-gender-neutralized sentences; and 3) a rule-based system that effectively gender-neutralizes French sentences using CNs, laying groundwork for further advancements for this task in that language. Our experiment combining our manually created dictionary with the Claude 3 Opus instruction model also shows promise for the use of such models for the task of gender rewriting. We strongly believe that future research further exploring the capabilities of these models for that task could lead to the development of effective solutions for mitigating gender bias in other languages with collective nouns (such as Spanish) or similar gender-neutralization techniques.

Limitations

French CNs adhere to specific semantic rules, which means that their usage may not be universally applicable to all sentences, sometimes resulting in constructions that appear asemantic. This limitation is further compounded by the fact that only a small subset of these nouns is actively employed in everyday language by native speakers, which restricts their versatility and adaptability in various linguistic contexts. We however believe that they are good candidates for gender neutralization, and the development of our system may help promote a broader use of such nouns. In addition, combining our system with a contextual or semantic analysis framework could help address these issues by ensuring that the CN equivalents are both contextually relevant and semantically appropriate.

Furthermore, even though collective nouns have not been tested specifically, recent research works from Spinelli et al. (2023) and Tibblin et al. (2023) showed that gender-neutralization appears to be less effective to counter gender biases induced by the use of MG. As previously stated, however, this writing technique is less contentious among the general population compared to others which explicitly highlight the feminine ending of words or separate it from the masculine ending.

Finally, this work is limited to the French language only, and the methodology we resorted to can only be used by languages with collective nouns acting as gender neutralizers (e.g., Spanish) and requires the creation of a language-specific human-member noun dictionary.

Ethics Statement

We did not filter the datasets that were used for the development of the RBS and for fine-tuning models for harmful, hateful, inappropriate or personal content. Considering the sources used to constitute these datasets (Wikipedia and Europarl), we believe it very unlikely for those to display such type of content. Similarly, when it comes to output sentences generated by the fine-tuned models, since those were trained on replacing specific words in sentences, the generation of such content seems unlikely. As discussed in the paper, instruction models are more prone to reformulating input sentences: while we did not find any inappropriate content in the Claude 3 Opus-generated sentences we evaluated, LLMs may be trained on such data, which might lead to the generation of harmful or hateful content.

Acknowledgements

We sincerely thank the reviewers for their constructive comments and suggestions, which contributed substantively in improving this paper.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022a. 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.

Bashar Alhafni, Nizar Habash, Houda Bouamor, Ossama Obeid, Sultan Alrowili, Daliyah Alzeer, Khawlah M. Alshanqiti, Ahmed ElBakry, Muhammad ElNokrashy, Mohamed Gabr, Abderrahmane Issam, Abdelrahim Qaddoumi, K. Vijay-Shanker, and Mahmoud Zyate. 2022b. The Shared Task on Gender Rewriting. http://arxiv.org/abs/2210.12410. *Preprint*, arXiv:2210.12410.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. http://arxiv.org/abs/2005.14165. *Preprint*, arXiv:2005.14165.

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

Adrien Chuttarsing. 2021. Inflecteur.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs. *TALN 2024*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. http://arxiv.org/abs/2010.11125. *Preprint*, arXiv:2010.11125.

Nelly Flaux. 1999. À propos des noms collectifs. *Revue de linguistique romane*, (63):471–502.

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

graelo. 2023. Graelo/wikipedia.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Pascal Gygax, Ute Gabriel, Arik Lévy, Eva Pool, Marjorie Grivel, and Elena Pedrazzini. 2012. The masculine form and its competing interpretations in French: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology*, 24(4):395–408.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, 23(3):464–485.

Pascal Mark 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.

Harris Interactive. 2017. L'écriture inclusive : La population française connaît-elle l'écriture inclusive ? Quelle opinion en a-t-elle ? Technical report.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. http://arxiv.org/abs/2109.11708. *Preprint*, arXiv:2109.11708.

Marsha B. Jacobson and William R. Insko. 1985. Use of Nonsexist Pronouns as a Function of One's Feminist Orientation. *Sex Roles*, 13(1-2):1–7.

Paul Kay and Chad K. McDaniel. 1978. The Linguistic Significance of the Meanings of Basic Color Terms. *Language*, 54(3):610–646.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Marie Lammert. 2010. *Sémantique et Cognition : Les Noms Collectifs*. Droz, Genève.

Marie Lammert and Michelle Lecolle. 2014. Les noms collectifs en français, une vue d'ensemble. *Cahiers de lexicologie*, (105):203–222.

Manuel Lardelli and Dagmar Gromann. 2023. Gender-Fair Post-Editing: A Case Study Beyond the Binary.

Michelle Lecolle. 2019. *Les Noms Collectifs Humains En Français. Enjeux Sémantiques, Lexicaux et Discursifs*. Lambert-Lucas, Université de Lorraine.

Paul Lerner and Cyril Grouin. 2024. INCLURE: A Dataset and Toolkit for Inclusive French Translation.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. http://arxiv.org/abs/2107.13586. *Preprint*, arXiv:2107.13586.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2024. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

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. http://arxiv.org/abs/2203.02155. *Preprint*, arXiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. From Inclusive Language to Gender-Neutral Machine Translation. http://arxiv.org/abs/2301.10075. *Preprint*, arXiv:2301.10075.

David Pomerenke. 2022. INCLUSIFY: A Benchmark and a Model for Gender-Inclusive German. http://arxiv.org/abs/2212.02564. *Preprint*, arXiv:2212.02564.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. http://arxiv.org/abs/1910.10683. *Preprint*, arXiv:1910.10683.

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

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

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.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *Preprint*, arXiv:2102.06788.

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

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. http://arxiv.org/abs/2109.06105. *Preprint*, arXiv:2109.06105.

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.

Jean-Philippe Watbled. 2012. Linguistique du genre. *L'Harmattan*, pages 167–179.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Preprint*, arXiv:1804.06876.

A Fine-Tuning Details

Models were trained on a single NVIDIA RTX 4090 GPU. Training time took approximately 3 hours for each model.

A.1 GeNRe-T5

```
BATCH_SIZE = 48
NUM_PROCS = 16
EPOCHS = 5
LEARNING_RATE = 0.0005
WEIGHT_DECAY = 0.02
```

A.2 GeNRe-M2M-100

```
BATCH_SIZE = 8
NUM_PROCS = 16
EPOCHS = 5
LEARNING_RATE = 0.0005
WEIGHT_DECAY = 0.02
```

B Instruction Details

B.1 Instruction Model Hyperparameters

```
model="claude-3-opus-20240229",
temperature=0,
messages=[
    {"role": "user",
    "content": f"{message}"},
    {"role": "assistant",
    "content": "Here is the
    output sentence:"}
]
```

B.2 Types of Instructions

Table 3 contains the different types of instructions given to Claude 3 Opus as well as their respective content.

"EXAMPLES" refers to the few-shot sentences given to the instruction model. See Tables 4 and 5 for more information.

"ORIGINAL SENTENCE" is replaced with the sentence containing one or several masculine generic nouns that we want to replace with their collective counterparts. It is part of the prompt in a similar way to the example sentences so that the instruction model is guided towards generating the final, gender-neutralized sentence.

C Few-shot sentences given to Claude 3 Opus

Tables 4 and 5 contain the few-shot sentences used respectively for the "BASE" and "DICT" instructions, and the "CORR" instruction. They were formatted as such in the prompt:
[Sentence with masculine generic] → [Gender-neutralized sentence].

Instruction Type	Content
BASE	Make this French sentence inclusive by replacing generic masculine nouns with their French collective noun equivalents. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
DICT-SG	Make this French sentence inclusive by replacing generic masculine noun {NM} with its respective French collective noun equivalent {NCOLL}. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
DICT-PL	Make this French sentence inclusive by replacing generic masculine nouns {NM1, NM2, . . . } with their respective French collective noun equivalents {NCOLL1, NCOLL2, . . . }. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
CORR	Correct grammar in this French sentence. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →

Table 3: Content of instructions per type given to Claude 3 Opus

Sentence with masculine generic	Gender-neutralized sentence
Le président de la FIFA Sepp Blatter rejette les accusations **des manifestants** en les accusant d'opportunisme. (FIFA President Sepp Blatter dismisses **the protesters**' accusatations as opportunism.)	Le président de la FIFA Sepp Blatter rejette les accusations de **la manifestation** en l'accusant d'opportunisme. (FIFA President Sepp Blatter dismisses **the protest**'s accusatations as opportunism.)
Les auteurs et **les spectateurs** ont été satisfaits des réponses des représentants. (**Authors** and **spectators** were pleased with **the representatives**' responses.)	**L'autorat** et **le public** ont été satisfaits des réponses de **la représentation**. (**The authorship** and **the audience** were pleased with the **representation**'s responses.)
Le vicaire général proposa de disperser **les religieux** dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments. (The vicar general suggested to disperse **religious people** to other houses of the order to repair the buildings.)	Le vicaire général proposa de disperser **le couvent** dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments. (The vicar general suggested to disperse **the convent** to other houses of the order to repair the buildings.)

Table 4: Few-shot sentences for "BASE" and "DICT" instructions. Bold indicates the differences between sentences with MG and gender-neutralized sentences.

RBS-generated sentence with errors	Manual sentence
Le président de la FIFA Sepp Blatter rejette les accusations de la manifestation en **les** accusant d'opportunisme.	Le président de la FIFA Sepp Blatter rejette les accusations de la manifestation en **l'**accusant d'opportunisme.
L'autorat et le public **a** été satisfaits des réponses des la représentation.	L'autorat et le public **ont** été satisfaits des réponses de la représentation.
Le vicaire **générale** proposa de disperser le couvent dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments.	Le vicaire **général** proposa de disperser le couvent dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments.

Table 5: Few-shot sentences for "CORR" instruction. Bold indicates the differences between the RBS-generated ssentences with error and the manual, correct sentences.

D Error Types Labels

We give additional information about some of the error types below.

Error Type	Description
ADJ	Errors related to adjective agreement with the modified noun. Past participles used as adjectives are included in this category.
CASE (N)	Errors related to an incorrect use of lowercase/uppercase characters.
DET	Errors related to determiner agreement with the modified noun.
DET_COREF	Errors related to coreferent possessive determiner agreement with the modified noun.
ELISION	Errors related to elision.
GEN_FAILURE (N)	Errors related to incorrect text-to-text model generation, most particularly with proper nouns or words that are not part of the model's vocabulary.
MISID_NOUN	Errors occurring when a member noun's form in the collective-member noun dictionary was wrongly detected as a noun in the original sentence, and was thus incorrectly changed into a CN.
PREP	Errors related to preposition usage.
PRON_COREF	Errors related to coreferent pronoun agreement with the modified noun.
PUNCT	Errors related to punctuation (e.g. missing or double spaces).
SEM	Errors occurring when changing the member noun into its CN counterpart leads to an asemantic sentence.
SPECIAL_CHAR (N)	Errors related to special characters (e.g. accents).
UNREPLACED	Errors occurring when the member noun was not replaced with its CN counterpart.
VERB	Errors related to verb or auxiliary agreement.

Table 6: Error types and descriptions for the RBS and fine-tuned models

The ELISION error is related to how elision works in French: in the sentences that we are modifying, the masculine determiner "le" and the feminine determiner "la" (*the*) should be elided and written as "l'" when the word that follows begins with a vowel or a mute "h".

The MISID_NOUN error may occur when the form of a member noun shares several different grammatical categories. For example, "jeunes" (*young*), the member noun's form of the CN "jeunesse" (*youth*), can be both a noun and an adjective. When the adjective form was wrongly detected as a noun, it was included in our dataset and produced an ungrammatical result sentence.

Finally, when it comes to the SEM error type, as discussed by Lecolle (2019), CNs in French, and more specifically human CNs, feature specific semantic characteristics due to how they are used to group human beings under a common denomination, based for example on their profession ("le professorat" [*professorate*]), their social status ("l'aristocratie" [*the aristocracy*]), or their political leaning ("la gauche" [*the left*]). Combining human CNs with specific verbs or contexts may thus not be considered semantically correct, and may occur when transforming a sentence. We labeled such transformed sentences with this error.

E GPT 4o-mini Automatic Error Type Labelling

F Generation Examples

(F8) a. Cette démarche fera progresser les droits **des citoyens**, car, par l'intermédiaire du Parlement, **les citoyens seront** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [original sent.]
(This approach will increase **citizens' [masc.]** rights, because, through the Parliament, **citizens will [pl.]** have a direct line to the Commission thereby generating considerable legitimacy.)

b. Cette démarche fera progresser les droits **de la citoyenneté**, car, par l'intermédiaire du Parlement, **la citoyenneté seront** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [GeNRe-RBS]
(This approach will increase the rights of **the citizenry**, because, through the Parliament, **the citizenry will [pl.]** have a direct line to the Commission thereby generating considerable legitimacy.)

c. Cette démarche fera progresser les droits **de la citoyenneté**, car, par l'intermédiaire du Parlement, **la citoyenneté sera** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [manual sent.]
(This approach will increase the rights of **the citizenry**, because, through the Parliament, **the citizenry will [sg.]** have a direct line to the Commission thereby generating considerable legitimacy.)

(F9) a. Je vous invite à informer **les députés européens chargés** des dossiers agricoles de l'avancement des négociations. [original sent.]
(I urge you to inform **the Members of European Parliament [masc] in charge of [pl.]** the agricul-

tural issues about the progress of negotiations.)

b. Je vous invite à informer **le parlement européen chargés** des dossiers agricoles de l'avancement des négociations. [GeNRe-RBS]

(I urge you to inform **the European parliament in charge of [pl.]** the agricultural issues about the progress of negotiations.)

c. Je vous invite à informer **le parlement européen chargé** des dossiers agricoles de l'avancement des négociations. [manual sent.]

(I urge you to inform **the European parliament in charge of [sg.]** the agricultural issues about the progress of negotiations.)

(F10) a. Un deuxième élément concerne le soutien apporté à la Commission à **l'actorat local** qui **veulent** participer à ces programmes afin d'avoir accès aux sources de financement correspondantes. [GeNRe-RBS]

(A second factor is the Commission's support for **local actors [coll. sg.]** who **want [pl.]** to take part in these programmes, so that they can access the corresponding funding mechanisms.)

b. Un deuxième élément concerne le soutien apporté à la Commission à **l'actorat local** qui **veut** participer à ces programmes afin d'avoir accès aux sources de financement correspondantes. [GeNRe-FT-M2M-100]

(A second factor is the Commission's support for **local actors [coll. sg.]** who **want [sg.]** to take part in these programmes, so that they can access the corresponding funding mechanisms.)

(F11) a. Juin, Russie : le **Nebski** sobor prend des décisions importantes. [GeNRe-FT-M2M-100]

(June, Russia: the **Nebski** Sobor makes important decisions.)

(F12) a. Il est allé à Cologne, où il est devenu président de l'association de la **main-d'uvre** et a aidé à propager les idées marxistes parmi ses membres. [GeNRe-FT-T5]

(He went to Cologne, where he became president of the **labour** organization and helped spread Marxist ideas among its members.)

(F13) a. **l'**armée arriva avec une lance à eau pour disperser les détenus. [GeNRe-FT-T5]

(**the** army arrived with a water hose to disperse the prisoners.)

(F14) a. Dans une lettre à la **famille** datée du 13 juin 1861, Zeng Guofan a ordonné à **ses propres navires** de surveiller les navires commerciaux britanniques après avoir remarqué que des marchands étrangers déchargeaient du riz **à la rébellion** à Anqing. [GeNRe-RBS]

(In a letter addressed to the **family** and dated June 13, 1861, Zeng Guofan ordered **his own vessels** to monitor British commercial vessels after noticing that foreign sellers were giving rice to **the rebellion** in Anqing.)

b. Dans une lettre à la **parenté** datée du 13 juin 1861, Zeng Guofan a ordonné à **sa propre flotte** de surveiller les navires commerciaux britanniques après avoir remarqué que des marchands étrangers déchargeaient du riz **aux rebelles** à Anqing. [Claude 3 Opus-BASE]

(In a letter addressed to the **kinfolk** and dated June 13, 1861, Zeng Guofan ordered **his own fleet** to monitor British commercial vessels after noticing that foreign sellers were giving rice to **rebels** in Anqing.)

(F15) a. Mais l'armée protestante, toujours agressive, **restaient** à la charge des habitants et **constituaient** une lourde charge. [GeNRe-RBS]

(But the Protestant army, still aggressive, **remained [pl.]** in the care of the local people and **constituted [pl.]** a heavy burden.)

b. Mais l'armée protestante, toujours agressive, **restait** à la charge des habitants et **constituait** une lourde charge. [Claude 3 Opus-DICT]

(But the Protestant army, still aggressive, **remained [sg.]** in the care of the local people and **constituted [sg.]** a heavy burden.)

(F16) a. Paradoxalement, cette progression en voix s'accompagne d'un recul en nombre d'élus, du fait de la poussée des candidats indépendants (pour la plupart de la **représentation** de la communauté kurde) et du CHP. [GeNRe-RBS]

(Paradoxically, this increase in votes paralleled a decrease in the number of elected representatives due to better results for the independent candidates (most of them **coming from the representation** of the Kurdish community) and CHP.

b. Paradoxalement, cette progression en voix s'accompagne d'un recul en nombre d'élus, du fait de la poussée des candidats indépendants (pour la plupart des **représentants** de la communauté kurde) et du CHP. [Claude 3 Opus-DICT]

(Paradoxically, this increase in votes paralleled a decrease in the number of elected representatives due to better results for the independent candidates (most of them **being representatives** of the Kurdish community) and CHP.