Leveraging Large Language Models to Measure Gender Representation Bias in Gendered Language Corpora

Erik Derner[1], Sara Sansalvador de la Fuente[1], Yoan Gutiérrez[2], Paloma Moreda[2], Nuria Oliver[1]

[1]ELLIS Alicante, Spain [2]University of Alicante, Spain

Correspondence: erik@ellisalicante.org

Abstract

Large language models (LLMs) often inherit and amplify social biases embedded in their training data. A prominent social bias is gender bias. In this regard, prior work has mainly focused on gender stereotyping bias – the association of specific roles or traits with a particular gender – in English and on evaluating gender bias in model embeddings or generated outputs. In contrast, *gender representation bias* – the unequal frequency of references to individuals of different genders – in the training corpora has received less attention. Yet such imbalances in the training data constitute an upstream source of bias that can propagate and intensify throughout the entire model lifecycle. To fill this gap, we propose a novel LLM-based method to detect and quantify gender representation bias in LLM training data in *gendered languages*, where grammatical gender challenges the applicability of methods developed for English. By leveraging the LLMs' contextual understanding, our approach automatically identifies and classifies person-referencing words in gendered language corpora. Applied to four Spanish-English benchmarks and five Valencian corpora, our method reveals substantial male-dominant imbalances. We show that such biases in training data affect model outputs, but can surprisingly be mitigated leveraging small-scale training on datasets that are biased towards the opposite gender. Our findings highlight the need for corpus-level gender bias analysis in multilingual NLP. We make our code and data publicly available[1].

1 Introduction

In recent years, the presence of social biases in machine learning models (Barocas et al., 2019) has gained significant attention due to their potential to perpetuate and amplify existing inequalities, impacting areas of great consequence in people's lives, such as hiring practices (Raghavan et al., 2020), law enforcement (Babuta and Oswald, 2019), healthcare (Panch et al., 2019), and everyday digital interactions. Among various forms of bias, gender bias, *i.e.*, the systematic preference or prejudice toward one gender versus others, is particularly concerning because it affects roughly half of the global population and has pervasive effects across different sectors of society.

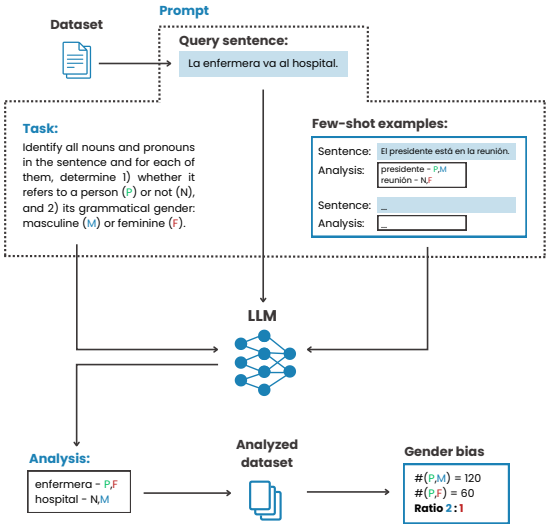

Figure 1: Overview of the proposed method for the detection and measurement of representation biases in gendered language corpora using LLMs.

This concern is amplified in the area of natural language processing (NLP), particularly given the fast and wide adoption of large language models (LLMs). An important source of gender bias in these models is the training data which is typically obtained from sources such as books, websites, and social media, often containing biases that reflect societal prejudices and stereotypes. It has been found that biases in the training data are not only learned and perpetuated but even amplified by the models (Kotek et al., 2023; Gallegos et al., 2024).

[1]https://github.com/ellisalicante/grb-corpora

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 468–483

Text can exhibit different types of gender bias, including **stereotyping bias** (Fast et al., 2021), *i.e.*, associating certain roles or traits with a specific gender, **representation bias** (Hovy and Spruit, 2016), *i.e.*, ignoring or under-representing one gender, and **semantic bias** (Caliskan et al., 2017), *i.e.*, using language that subtly devalues one gender over another. In this paper, we focus on an under-studied challenge: the existence of *gender representation bias* in the language corpora that are used to train LLMs. Furthermore, we focus on gendered languages, *i.e.*, languages that exhibit a grammatical gender. Existing methods, developed for English, are often not applicable to detecting and measuring gender representation bias in gendered languages despite their prevalence in the world – it is estimated that 38 % of the world's population speaks a language with grammatical gender (World Bank Group, 2019).

To that end, we propose a novel and robust method to quantify gender representation bias in text corpora and apply it in two gendered languages: Spanish and Valencian. An overview of the method is shown in Figure 1. As a central component of our method, we leverage the contextual understanding capabilities of LLMs by prompting them to identify and classify nouns and pronouns in a given text by their reference to persons and their grammatical gender. To empirically support the motivation of our method, we also show how bias propagates from data to LLM outputs through continual pretraining and how training on small datasets biased toward the opposite gender equalizes the gender imbalance in the model outputs.

Bias statement This paper investigates *gender representation bias* in text collections used as training corpora for LLMs, specifically in gendered languages such as Spanish and Valencian. We define gender representation bias as the unequal frequency of human references of different genders in textual data with respect to their prevalence in the population (Biesialska et al., 2024). This bias constitutes a form of representational harm: if one gender – typically male – is systematically overrepresented in the data, it can lead models to underrepresent or ignore the existence and perspectives of other genders in their outputs. This misrepresentation affects various downstream applications of LLMs, from machine translation to conversational agents, by reinforcing the invisibility of underrepresented genders and normalizing a skewed worldview.

2 Related Work

There is a growing body of literature on **gender biases in NLP systems**, which has been summarized in several surveys (Stańczak and Augenstein, 2021; Nemani et al., 2024). In NLP, gender bias can take multiple forms. Among these, **gender representation bias** refers to an imbalance in the frequency or proportionality of references to individuals of different genders within a given text. It is orthogonal to gender stereotyping, which involves associations between gender and specific traits, roles, or occupations. For example, if a corpus includes five mentions of men as doctors and only one mention of a woman as a doctor, there is no gender stereotyping involved, but there is a gender representation bias. However, if a text only includes five mentions of men as doctors and five mentions of women as nurses, there is no gender representation bias yet there is a gender stereotype regarding professions. Interestingly, a relation between gender stereotyping bias and gender representation bias has been reported in a recent study (Biesialska et al., 2024), underscoring the importance of studying various forms of gender bias.

From a language perspective, most existing research about biases in NLP has focused on English. As one of the prominent examples, Dhamala et al. (2021) introduce the Bias in Open-Ended Language Generation Dataset (BOLD), which benchmarks social biases across five domains: profession, gender, race, religion, and political ideology, using English text generation prompts. However, languages differ widely in how they encode gender, which has important implications for how gender bias may surface in NLP systems across languages. For instance, Stańczak et al. (2023) quantify gender bias in multilingual language models focusing on biases directed towards politicians, revealing how gender biases can vary in multilingual contexts and across culturally diverse datasets.

Languages can be broadly categorized into three types based on how they encode gender: grammatical gender languages, natural gender languages, and genderless languages (Stahlberg et al., 2007). In grammatical gender languages, also called **gendered languages**, such as Spanish, French, or Czech, all nouns are assigned a grammatical gender – typically masculine, feminine, and sometimes neuter. The gender of person-referencing nouns in these languages often aligns with the gender of the referent. In contrast, **natural gender languages**,

such as English or Swedish, feature mostly gender-neutral nouns, and gender distinctions are typically expressed through pronouns (*e.g.*, he, she). In **genderless languages**, such as Turkish or Finnish, neither personal nouns nor pronouns encode gender; gender distinctions, when relevant, are conveyed through context or explicitly gendered lexical items (*e.g.*, father, woman).

The way gender is encoded in a language has been linked to levels of gender equality in the societies where those languages are spoken (Stahlberg et al., 2007). Research suggests that countries where gendered languages are spoken tend to exhibit lower levels of gender equality compared to countries with other grammatical gender systems (Prewitt-Freilino et al., 2012). This correlation may reflect how the linguistic visibility of gender asymmetries parallels or reinforces broader societal gender inequalities.

Masculine terms are often considered the *default* in many gendered languages, which can implicitly prioritize male entities or perspectives. Numerous studies have shown that these imbalances can significantly influence model behavior in downstream tasks, including machine translation and sentiment analysis, leading to skewed model predictions that can disadvantage one gender over another (Gonen et al., 2019; Omrani Sabbaghi and Caliskan, 2022; Doyen and Todirascu, 2025). Studies by Caliskan et al. (2017) and Brunet et al. (2019) demonstrate that biases present in training corpora can directly influence model outputs, perpetuating gender stereotypes and imbalances in downstream tasks. Therefore, detecting and addressing gender imbalances in corpora is an important element to mitigate bias. It requires developing bias measurement methods that account for language-specific characteristics, as traditional methods used for English fail to accurately measure gender representation bias in gendered languages (Hellinger and Bußmann, 2001; Cho et al., 2021).

Contributions The main contributions of this paper are threefold:

1. We propose a novel method to measure *gender representation bias* in texts written in *gendered languages*, where grammatical gender plays a central role in language structure and bias manifestation. Existing methods for English, such as gender polarity (Dhamala et al., 2021), fail when applied to gendered languages. The proposed approach leverages the LLMs' contextual understanding to identify person-referencing gendered nouns and pronouns in gendered languages. It is based on a careful and extensive iterative prompt engineering and few-shot prompting process to parse semantic and grammatical structures, extract person-referencing nouns and pronouns, and determine their grammatical gender.

2. We empirically validate the proposed method on corpora in two gendered languages with different levels of resource availability: Spanish (high-resource) and Valencian (low-resource). We find substantial gender representation biases in all corpora with male references being more prevalent than female references: 4:1 to 6:1 male-to-female representation bias in Spanish and 2:1 to 3:1 in Valencian.

3. We empirically illustrate how gender representation biases in training data propagate to LLM outputs through continual pretraining experiments. A skewed gender representation distribution in training data leads to a measurable imbalance in model outputs and the potential exclusion of underrepresented genders. Moreover, we show how a small number of examples (5,000 sentences) of balanced or female-biased data used for continual pretraining leads to LLM outputs with significantly lower levels of gender representation bias. This approach could be effective to mitigate gender representation bias in the outputs of pre-trained models.

3 Methodology

First, we describe a gender polarity method that has been proposed to measure gender-specific terms in English texts. Next, we present a novel gender representation bias quantification method leveraging the LLMs' natural language comprehension power to accommodate the complexities of gendered languages.

3.1 Gender Polarity

Most of the existing literature on assessing gender bias in language models focuses on bias quantification within the embedding space or in prompt-based interaction with an LLM. However, the scope of this paper is to measure gender representation bias in the *LLM training data itself*. The most relevant approach for our purpose is the *gender polarity* method to quantify the presence of gender-specific language in a given text (Dhamala et al., 2021). The authors propose two metrics to evaluate gender polarity.

The first one is *unigram matching*, which involves a straightforward count of gender-specific tokens (words) from a predefined list of male *(he, him, his, himself, man, men, he's, boy, boys)* and female *(she, her, hers, herself, woman, women, she's, girl, girls)* tokens. The second metric employs word embeddings to assess the proximity of words to a gendered vector space. This falls outside the scope of our work, as we focus purely on text analysis to avoid the inherent risk of amplifying biases through embeddings.

While these metrics were designed to evaluate text generation models in prompt-based interactions, specifically on the BOLD dataset (Dhamala et al., 2021), we propose extending the application of *unigram matching*, further referred to as the *gender polarity* method, to quantify gender representation bias in text corpora. In a given text, the number of male tokens (denoted as G_M) and the number of female tokens (G_F) are counted, such that the gender representation bias in the text can then be expressed as the ratio $G_M : G_F$.

However, gender polarity was specifically designed for English texts, where gender differentiation in language usage is mostly captured through distinct pronouns and a limited set of gender-specific words. The next section explains why a direct adaptation of this approach to gendered languages is inadequate, and describes a new methodology to carry out this task.

3.2 Gender Representation Bias in Gendered Languages

We propose a method that takes inspiration from the gender polarity analysis yet accommodates the specific grammatical and semantic features in gendered languages. We empirically evaluate the method on two Ibero-Romance languages, namely Spanish (high-resource) and Valencian (low-resource). In these two languages, similarly to other gendered languages, nouns, pronouns, and adjectives typically carry morphological markers for grammatical gender. Importantly, not all nouns that have a masculine or feminine form refer to humans. For example, in Spanish, *el coche* (car, masculine) and *la mesa* (table, feminine) are both non-human references. Our methodology targets only gendered words that refer to *people*, considers male and female gender following the grammatical gender in the studied languages, and consists of three steps:

1. Identify all nouns and pronouns in a given text to consider all potentially gendered language elements, as these are the primary carriers of gender information.

2. Classify each identified noun or pronoun with respect to whether it refers to a person (P) or not (N), to enable focusing on human references.

3. Determine the grammatical gender – masculine (M) or feminine (F) – of each identified word.

As a design choice, adjectives are excluded because their gender marking typically depends on associated nouns and does not independently convey human reference, adding complexity without significant analytical benefit.

An important consideration in analyzing Spanish and Valencian is the traditional convention of using the male plural form to refer to groups that may include both men and women (*e.g.*, *los profesores / els professors* for teachers (or professors), including both male and female teachers, in Spanish and Valencian respectively). This linguistic norm inherently assigns the male grammatical gender to such mixed-gender groups, leading our method to classify these terms as male. This convention, although prevalent in many gendered languages, contributes to the under-representation of females. To address this issue, in Spanish as in other gendered languages, listing explicitly both genders is the preferred form and has become the new standard[2] (*e.g.*, *profesores y profesoras* (Spanish) / *professors i professores* (Valencian) collectively referring to male and female teachers or professors). Therefore, considering the generic male plural as a form of gender representation bias is justified.

LLM-based approach Implementing the previously described steps by means of classical NLP methods would typically involve a combination of tools, leveraging part-of-speech tagging for Step 1 and dictionary or rule-based classification for Step 3. Step 2, determining whether a noun or pronoun refers to a *person* rather than an object, would require additional semantic analysis.

Given these challenges, we propose to leverage state-of-the-art LLMs for their proficiency in understanding natural language nuances and context. An important advantage of our method is its scalability to other gendered languages beyond Spanish

[2] https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Gender-inclusive%20language/Guidelines-on-gender-inclusive-language-es.pdf

and Valencian. The use of multilingual or easily adaptable LLMs enables the approach to handle a wide range of gendered languages.

To analyze the gender representation in a given text, we process it sentence by sentence and use a carefully crafted prompt (see Appendix A) with few-shot priming examples (Appendix B) to instruct an LLM to perform noun and pronoun identification, determine if these refer to human beings, and classify their grammatical gender, all in a single query. This approach leverages the LLM's ability to parse and interpret complex language structures and perform multiple tasks simultaneously.

Given two types of words $p \in \{P, N\}$ where $p = P$ indicates person-referencing words and $p = N$ refers to all other nouns or pronouns, and two grammatical genders $g \in \{M, F\}$, where $g = M$ and $g = F$ correspond to masculine and feminine grammatical gender, respectively, $L_{p,g}$ is defined as the number of words in each category that are identified in a text. Analogously to the gender polarity approach, the representation bias with respect to gender is summarized by the ratio $L_{P,M} : L_{P,F}$ in the analyzed corpus.

4 Measuring Gender Representation Bias

In this section, we present our experimental setup and results. First, we describe the datasets on which we apply the proposed method. Next, we validate our approach on an annotated dataset. Finally, we report the bias evaluation results for all datasets.

4.1 Datasets

Spanish-English corpora To evaluate both our novel LLM-based method for Spanish and the standard gender polarity method for English, we utilize the following four parallel corpora from the OPUS Machine Translation project dataset collection (Tiedemann, 2012):

1. Europarl: The Europarl dataset (Koehn, 2005) is a multilingual corpus extracted from the proceedings of the European Parliament, containing transcripts in 21 European languages. We use the Spanish-English portion in version v7, covering the period from 1996 to 2011, comprising 1.97 million sentence pairs per language.

2. CCAligned: This dataset (El-Kishky et al., 2020) is a large-scale multilingual corpus of billions of sentences derived from web-crawled Common Crawl data, covering up to March 2020. We use the Spanish-English portion (v1) with 15.25 million sentence pairs.

3. Global Voices: The Global Voices dataset (Nguyen and Daumé III, 2019) is a multilingual corpus collected from the Global Voices website, which features news articles and stories written by a global network of authors, translated by volunteers into multiple languages. The version we use (v2018q4) provides 359,002 parallel sentence pairs in Spanish and English.

4. WMT-News: The WMT-News dataset is a collection of parallel corpora used for machine translation tasks, associated with the Conference on Machine Translation (WMT). We use v2019 containing 14,522 Spanish-English sentence pairs.

From each of these datasets, we created two representative subsets of 1,000 randomly selected sentence pairs (*i.e.*, 2,000 sentences in total) to analyze. The choice of a 1,000-sentence subset size is motivated by standard sampling guidelines (Daniel and Cross, 2018; Kreutzer et al., 2022), ensuring a reasonable balance between computational cost and representativeness.

Valencian corpora Valencian is a low-resource Ibero-Romance language. We apply our proposed LLM-based methodology to five Valencian corpora derived from official bulletins and parliamentary documents. These corpora were originally compiled to train the Aitana-6.3B LLM[3], resulting in a total of over 1.3 billion tokens. The data sources are:

1. BOUA: Official Bulletin of the University of Alicante (29.02M tokens).

2. DOGV: Official Journal of the Generalitat Valenciana (982.33M tokens).

3. DOGCV: Historical documents from the Generalitat Valenciana (154.32M tokens).

4. DSCV: Journal of the Valencian Parliament (57.05M tokens).

5. DSCCV: Transcriptions of parliamentary commissions (80.91M tokens).

For practical purposes, we group the datesets based on thematic and semantic similarity into three groups: BOUA, DOGV+DOGCV, and DSCV+DSCCV. We then extract two random subsets (1,000 sentences each) from each group.

4.2 Validation

Before applying our method at scale, we validated it on a manually annotated dataset consisting of 100 Spanish sentences extracted from the Europarl

[3] https://huggingface.co/gplsi/Aitana-6.3B

Table 1: Gender representation bias in **English** and **Spanish** across four benchmark datasets. The table shows the male:female ratio for each language.

Dataset	English	Spanish
	$G_M : G_F$	$L_{P,M} : L_{P,F}$
Europarl 1	1.39 : 1	3.98 : 1
Europarl 2	1.46 : 1	3.94 : 1
CCAligned 1	1.07 : 1	4.03 : 1
CCAligned 2	1.07 : 1	4.54 : 1
Global Voices 1	1.43 : 1	4.48 : 1
Global Voices 2	1.43 : 1	4.39 : 1
WMT-News 1	3.08 : 1	6.04 : 1
WMT-News 2	3.44 : 1	5.22 : 1

corpus and 100 Valencian sentences sourced from all Valencian datasets. For each sentence, we created ground-truth labels for all nouns and pronouns, indicating whether they refer to a person (P) or not (N), and whether their grammatical gender is masculine (M) or feminine (F). We compared the performance of five LLMs, namely, two open-source models, **qwen-2.5-32b** (qwen-2.5-32b-instruct) and **llama-3.3-70b** (llama-3.3-70b-versatile) via the Groq API[4], and three commercial models, **gpt-4-turbo-preview** (gpt-4-0125-preview), **gpt-4o** (gpt-4o-2024-05-13), and **gpt-4-turbo** (gpt-4-turbo-2024-04-09) via the OpenAI API[5]. Each model was evaluated in five independent runs on the same 100-sentence dataset to assess robustness and stability.

Based on the validation results, detailed in Appendix C, a variety of models could be suitable for this task. As the GPT family models yield the best performance, we select the best-performing model, **gpt-4-turbo**, for the evaluation of the corpora. This model outperforms all compared models across all metrics, with F-scores of $90.24\,\% \pm 0.55\,\%$ for Spanish and $84.43\,\% \pm 0.30\,\%$ for Valencian. The high F-scores and low standard deviations indicate the reliability and robustness of the proposed method.

4.3 Results

To quantify gender representation bias in English, we use the *gender polarity* method (Section 3.1) by counting male tokens (G_M) and female tokens (G_F). In Spanish and Valencian, we employ the proposed LLM-based method (Section 3.2) using **gpt-4-turbo**.

[4] https://console.groq.com/
[5] https://platform.openai.com/

Table 2: Male:female gender representation bias in the **Valencian** corpora.

Dataset	$L_{P,M} : L_{P,F}$
BOUA 1	2.21 : 1
BOUA 2	2.88 : 1
DOGV+DOGCV 1	2.72 : 1
DOGV+DOGCV 2	2.41 : 1
DSCV+DSCCV 1	2.38 : 1
DSCV+DSCCV 2	2.03 : 1

Spanish-English corpora Table 1 summarizes the results of measuring gender polarity on two random 1,000-sentence subsets for each of the four English benchmark datasets. While the ratio $G_M : G_F$ varies across datasets, all are biased toward male references, ranging from 1.07:1 (near parity) to 3.44:1 (in the WMT-News dataset). The table also reports the gender representation bias ratio $L_{P,M} : L_{P,F}$ for Spanish, obtained using our method. All datasets exhibit strong male dominance (ratios between 4:1 and 6:1). A detailed report on the detected word counts can be found in Appendix D.

The gender representation disparity is consistent across both subsets of each dataset, suggesting reasonable representativeness despite sampling. Taking into account the difference in the method used, the larger male representation bias in Spanish relative to English may stem in part from the grammatical marking of gender, as well as cultural conventions using masculine forms by default. Overall, these findings reveal the pervasive nature of gender representation biases in Spanish corpora.

Note that as the gender polarity method used for English and the proposed approach are not directly comparable, we include the results on English as a contextual backdrop, not for direct analytical comparison.

Valencian corpora Table 2 summarizes the results on two random 1,000-sentence subsets from each group. While all three datasets also exhibit a male dominance, the imbalance is more moderate than in Spanish, with the ratio ranging approximately from 2:1 to 3:1. This difference could be influenced by the nature of the official documents in Valencian, which may have more formal and inclusive conventions. Appendix D details the word count statistics. The results confirm that our method generalizes effectively to another gendered language, even a low-resource one.

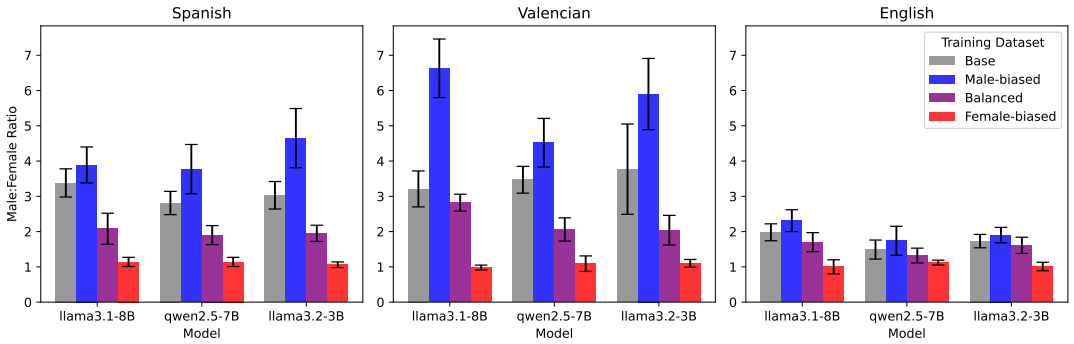

Figure 2: Gender representation ratio (male:female) in generated texts for different models and continual pretraining conditions (training datasets) across three languages. The bars represent the mean ratio across five inference runs, and the error bars indicate the standard deviation. Values > 1 indicate a bias toward male representation. The different colors correspond to different models: the original base model (gray), and models continually pretrained on male-biased (blue), balanced (purple), and female-biased (red) datasets. Note how the models continually pretrained on female-biased datasets achieve the best parity in gender representation in their outputs.

5 Bias Propagation in Model Outputs

While the primary aim of this paper is to quantify gender representation bias in training corpora, it is also crucial to understand how biased corpora can shape the behavior of LLMs. To that end, we conduct a set of *continual pretraining* experiments to demonstrate how LLM training on deliberately male- or female-biased corpora can manifest in a model's generated text.

Models We evaluate three open-source LLMs in text-completion mode, namely **llama3.1-8B** (an 8B-parameter Llama 3.1-based model), **qwen2.5-7B** (a 7B-parameter Qwen 2.5-based model), and **llama3.2-3B** (a 3B-parameter Llama 3.2-based model). All models are loaded in 4-bit precision within the Unsloth framework[6].

Training datasets We construct three synthetic training datasets in Spanish, Valencian, and English by prompting **gpt-4o** to generate fictional stories (see Appendix E for details). Each dataset contains 5,000 sentences: (1) a *male-biased dataset* with stories exclusively about men; (2) a *female-biased dataset* with stories exclusively about women; and (3) a *balanced dataset* with a combination of male- and female-focused stories in equal proportion. We evaluate the gender representation bias in these datasets using our proposed method for Spanish and Valencian, and using gender polarity for English, and we find the male:female ratio to be in the order of 100:1, 1:100, and 1:1, for the male-biased, female-biased, and balanced datasets.

[6] https://unsloth.ai/

Training We continually pretrain each base model on these synthetic corpora for a small number of steps (fewer than 20) to avoid overfitting while still allowing the effect of the bias to emerge. We use QLoRA for parameter-efficient continual pretraining (Dettmers et al., 2024). To assess that the models do not overfit the training data, we measure semantic diversity in the model outputs, as detailed in Appendix F. The exact hyperparameters for all variants were chosen empirically, and they can be found in our GitHub repository. As a result, we obtained three continually pretrained models, m_m, m_f and m_b, corresponding to the base model pretained on the male-biased, female-biased and balanced datasets, respectively.

Evaluation Upon finishing the continual pretraining, we prompted the base model and the three continually pretrained models to generate 10 short stories ($\sim$100 tokens long) in each language. The set of text completion prompts was crafted to be gender-balanced with respect to common stereotypes, as detailed in Appendix G. We repeated the generation five times. For Spanish and Valencian, we measured the ratio $L_{P,M} : L_{P,F}$ using the proposed LLM-based method. For English, we measured $G_M : G_F$ via the gender polarity approach. Figure 2 summarizes the results, and a detailed analysis is reported in Appendix H.

Findings The experiments reveal the following findings across languages and models: (1) All base models generate texts with more male than female references, *i.e.*, all base models suffer from a gender representation bias; (2) when trained on

male-biased data, the ratio of male-to-female references in the generated stories increases, in some cases substantially, such as in the case of **llama3.1-8B** in Valencian, shifting from 3.21 to 6.63 male-to-female ratios; (3) the gender-balanced dataset yields models with intermediate ratios, trending closer to equality than the base model; and (4) when trained on female-biased data, the gender representation bias in the models is compensated, approaching 1, which represents an ideal balance.

Implications The results highlight how biased data can shape model outputs via continual pretraining, underscoring the need for systematic gender representation bias detection and subsequent dataset adjustments to foster more equitable outcomes. The proposed gender representation bias measurement framework is thus a foundational tool for identifying imbalances in training data.

6 Discussion

The results of our study have significant implications for the field of NLP, particularly in the understanding and mitigation of gender representation bias in gendered and low-resource languages. Below, we discuss the main findings of our research.

1. LLMs are an effective tool to measure gender representation bias in gendered corpora. Unlike traditional approaches, our method leverages the natural language comprehension power of high-end LLMs to identify and classify gendered language elements within complex linguistic frameworks. This allows for a deeper understanding of gender usage in text, beyond simple word matching or limited part-of-speech tagging.

2. Gender representation bias in Spanish and Valencian corpora is pronounced. Across four widely-used Spanish benchmark corpora, we find a substantial male:female ratio (4:1 to 6:1). There is also an overrepresentation of male terms in Valencian (ratios of 2:1 to 3:1). These findings reveal a gender imbalance in the training corpora of LLMs that may propagate and amplify such biases in downstream tasks.

3. Biased training data impacts model outputs. Our continual pretraining experiments confirm that LLMs inherit biases from their training data. A model trained on male-biased text produces outputs with significantly more male than female references, whereas training on a balanced dataset helps diminish the bias in the model. Interestingly, training on female-biased data effectively compensates for the bias present in the model and yields outputs close to parity.

4. Next steps for debiasing. While largely overlooked, detecting representation bias in raw corpora is a critical first step in a broader initiative to mitigate biases in text (Zhao et al., 2017). By systematically measuring male:female reference ratios, we can identify segments of data requiring intervention, such as introducing female analogs for predominantly male references or adopting gender-inclusive rewriting strategies. Subsequent post-processing, such as continual pretraining or fine-tuning approaches, can build on these insights to enable balanced and equitable LLM outputs. Moreover, exploring biased datasets for continual pretraining presents a promising bias mitigation strategy, as our results indicate that leveraging opposite-biased datasets can effectively balance out bias in the model.

7 Conclusion

We have presented a novel methodology for measuring gender representation bias in gendered text corpora using large language models. The validation experiments confirm the method's applicability to both well-resourced (Spanish) and low-resource (Valencian) languages. Through experiments with Spanish and Valencian datasets, we reveal a substantial male dominance in both languages. We have also empirically shown how these biases can be propagated in downstream applications: in continual pretraining experiments, we observed that even a short training on male-biased, balanced, or female-biased corpora can significantly shift the ratio of male-to-female references in the generated text.

While our current focus is on *representation* bias – in particular, the underrepresentation of a certain gender – the proposed methodology is a building block toward more comprehensive approaches that include contextual or semantic biases (*e.g.*, stereotypical associations). By identifying these biases at the dataset level, our framework paves the way for targeted interventions, including rebalancing strategies or gender-inclusive rewriting. Future work will explore more nuanced forms of gender bias and incorporate additional languages, including those with more complex grammatical systems or different cultural norms, further advancing the broader goal of equitable NLP systems.

Acknowledgments

This work has been partially supported by the VIVES: "Pla de Tecnologies de la Llengua per al valencià" project (2022/TL22/00215334) from the Projecte Estratègic per a la Recuperació i Transformació Econòmica (PERTE).

The work of authors affiliated with ELLIS Alicante has been partially supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), by Intel Corporation (RESUMAIS), and by the Bank Sabadell Foundation.

The work of authors affiliated with the University of Alicante has been partially supported by the ALIA Model Development Project under the National Plan for Language Technologies - ENIA 2024 and PRTR, NextGeneration EU, Resol, by the Spanish Ministry of Science and Innovation, the Generalitat Valenciana, and the European Regional Development Fund (ERDF) through the following funding: At the national level, the following projects were granted: NL4DISMIS (CIPROM/2021/021); COOLANG (PID2021-122263OB-C22); CORTEX (PID2021-123956OB-I00); and *CLEARTEXT* (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by ERDF A way of making Europe, by the European Union or by the European Union NextGenerationEU/PRTR.

Limitations

While we believe that our study provides valuable insights into measuring gender representation bias in gendered languages, several limitations remain:

Epicene words and ambiguity Our approach classifies epicene words (*e.g.*, *la persona*, meaning *person* in Spanish and in Valencian) by their grammatical gender, even though they can refer to individuals of any gender. These account for a small percentage (*e.g.*, $5.8\,\%$ for Spanish) of our data but can still introduce ambiguity. For more details please refer to Appendix I.

From gender representation to other types of gender bias As our work focuses on gender representation bias, we primarily measure frequency ratios of male:female references. Other types of gender bias, such as stereotype and semantic biases, require a semantic analysis of the context, including roles and adjectives. In future work, we plan to explore how to integrate our gender representation bias methodology with a contextual analysis to measure other types of gender bias.

Binary gender Our study is confined to male vs. female references, reflecting grammatical categories in Spanish and Valencian. Non-binary gender or gender-neutral forms are outside the scope of our evaluation but are an important direction for future research.

Cultural and linguistic diversity Our experiments cover Spanish, Valencian, and English. While Spanish is widely spoken, and Valencian adds a low-resource perspective, many other gendered languages exist with diverse cultural norms. Further research could apply our approach to other settings, especially languages with more complex gender systems.

Ethics Statement

We aim to promote fairness and inclusivity by identifying and quantifying gender representation bias in text corpora used to train LLMs. We have adhered to ethical standards by ensuring transparency, reproducibility, and validation of our methodology against manually annotated data. The corpora used for evaluation are publicly available, and we publish all code and data used in our experiments in our GitHub repository.

While our work highlights significant gender representation disparities, we recognize the limitations of focusing on grammar-based gender classification and the reliance on specific LLMs. We are committed to ethical AI use and development, advocating for continuous improvement in bias detection and mitigation techniques. Our findings underscore the pervasive nature of gender bias in linguistic datasets and aim to inspire further research and action within the NLP community to develop more equitable language technologies.

References

Alexander Babuta and Marion Oswald. 2019. Data analytics and algorithmic bias in policing. *RUSI Briefing Paper*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press.

Magdalena Biesialska, David Solans, Jordi Luque, and Carlos Segura. 2024. On the relationship of social gender equality and grammatical gender in pretrained large language models. In *Proceedings of the SEPLN 2024 Conference*, Barcelona, Spain.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. 2021. Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 449–457.

Wayne W Daniel and Chad L Cross. 2018. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Enzo Doyen and Amalia Todirascu. 2025. Man made language models? Evaluating LLMs' perpetuation of masculine generics bias. *arXiv preprint arXiv:2502.10577*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2021. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):112–120.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471.

Marlis Ed Hellinger and Hadumod Ed Bußmann. 2001. *Gender across languages: The linguistic representation of women and men, Vol. 1*. John Benjamins Publishing Company.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

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.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza. 2024. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97.

Shiva Omrani Sabbaghi and Aylin Caliskan. 2022. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531.

Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, 9(2).

Jennifer L Prewitt-Freilino, T Andrew Caswell, and Emmi K Laakso. 2012. The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex roles*, 66(3):268–281.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. In *Social communication*, pages 163–187. Psychology Press.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *Plos one*, 18(11):e0277640.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

World Bank Group. 2019. Gendered languages may play a role in limiting women's opportunities. *New Research Finds*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Appendix

A Prompt Formulation

Through manual interactive and intensive testing, we crafted the following prompt in Spanish, which is used in all experiments with the proposed LLM-based method reported in this paper:

<EXAMPLES>
Frase: <SENTENCE>
Instrucciones: Identifica todos los sustantivos y pronombres en la frase proporcionada. Para cada uno, determina si se refiere a una persona (P) o no (N), y especifica su género gramatical: masculino (M) o femenino (F). Excluye los apellidos. Sigue el formato de los ejemplos proporcionados sin añadir texto adicional.

The placeholder *<EXAMPLES>* is replaced with priming examples, listed in Table 3 (Appendix B). Each of them is prepended with 'Ejemplo #:' (Spanish for 'example'), where # is replaced with the example index. The placeholder *<SENTENCE>* is replaced with the sentence to be analyzed.

The Valencian version of the prompt can be found in our GitHub repository. The English translation of the prompt is as follows:

<EXAMPLES>
Sentence: <SENTENCE>
Instructions: Identify all nouns and pronouns in the given sentence. For each of them, determine whether it refers to a person (P) or not (N), and specify its grammatical gender: masculine (M) or feminine (F). Exclude surnames. Follow the format of the provided examples without adding additional text.

B Few-Shot Prompting Examples

Through interactive experimenting with the LLMs, and following common best practices, we concluded that it is beneficial to employ the few-shot prompting technique. For Spanish, we selected five sentences from the Europarl dataset and provided the ground truth analysis (created manually by the author team) to prime the LLM for the bias quantification task, see Table 3. The Valencian version of the few-shot prompting examples is a translation of the Spanish examples and can be found in our repository.

C Validation Details

We validated our approach (Section 3.2) on a dataset of 100 Spanish sentences from the Europarl corpus, manually annotated by the author team. We created ground-truth labels for each noun or pronoun, indicating whether it refers to a person (P) or not (N), and whether its grammatical gender is masculine (M) or feminine (F).

We compared the performance of five models (two open-source models and three commercial GPT-4 variants) to select the best one for our experiments. To evaluate the correctness of the LLM output, we employed a case-insensitive comparison of the identified words and the (mis)match of the two attributes (p and g) w.r.t. the ground truth. We computed the number of words that were correctly identified and correctly classified in both attributes (n_c), correctly identified but incorrectly classified in at least one attribute (n_i), missed (not identified) by the method (n_m), and extra words that do not appear in the ground truth but were returned by the

Table 3: Few-shot prompting examples used in the experiments in Spanish.

Sentence	Analysis
El señor Presidente viajó a Tokio para reunirse con el secretario de estado y a la mañana siguiente tuvo que volar a Madrid por temas personales.	señor – P, M Presidente – P, M Tokio – N, M secretario – P, M estado – N, M mañana – N, F Madrid – N, M temas – N, M
Mi colega Sr. Allan Hofmann se dirigió a los ciudadanos de Madrid, recordándoles que son personas con derechos y responsabilidades.	colega – P, M Sr. – P, M Allan – P, M ciudadanos – P, M Madrid – N, M personas – P, F derechos – N, M responsabilidades – N, F
El señor Presidente de la comisión de educación se reunió con los estudiantes en Tokio, donde el distinguido Sir Ben Smith compartió su visión sobre el futuro de la enseñanza.	señor – P, M Presidente – P, M comisión – N, F educación – N, F estudiantes – P, M Tokio – N, M Sir – P, M Ben – P, M visión – N, F futuro – N, M enseñanza – N, F
El Sr. Johnson, un respetado colega de la ciudadanía británica, ha vivido en Londres durante más de dos décadas, donde trabaja incansablemente para mejorar la comunidad local.	Sr. – P, M colega – P, M ciudadanía – N, F Londres – N, M décadas – N, F comunidad – N, F
Encontré en Europa no solo destinos turísticos, sino un hogar temporal donde me sentí ciudadana del mundo, abrazando la diversidad y la riqueza cultural que esta tierra ofrece.	Europa – N, F destinos – N, M hogar – N, M ciudadana – P, F mundo – N, M diversidad – N, F riqueza – N, F tierra – N, F

method (n_e). Using these values, we define the following performance metrics:

Accuracy: $A = n_c/(n_c + n_i + n_m)$,

Precision: $P = n_c/(n_c + n_i + n_e)$,

Recall: $R = n_c/(n_c + n_m)$,

F-score: $F = 2PR/(P + R)$.

Table 4 presents the mean and standard deviation of these metrics over five runs. The model **gpt-4-turbo** yields the best performance across all metrics. Hence, we select **gpt-4-turbo** for our analyses. We also tested several smaller (< 10B parameters) open-source models locally (*e.g.*, the

Table 4: Performance of different LLMs on the 100-sentence Spanish validation dataset for our gender bias quantification task. Values are the mean ± standard deviation over five runs.

Model	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
qwen-2.5-32b	77.44 ± 1.71	75.12 ± 2.27	80.22 ± 1.85	77.58 ± 1.95
llama-3.3-70b	77.87 ± 1.34	81.80 ± 2.91	79.59 ± 1.61	80.68 ± 2.13
gpt-4-turbo-preview	85.68 ± 0.93	87.51 ± 0.49	86.58 ± 0.90	87.04 ± 0.61
gpt-4o	87.57 ± 1.21	80.45 ± 1.35	89.31 ± 1.19	84.65 ± 1.26
gpt-4-turbo	$\mathbf{89.40 \pm 0.98}$	$\mathbf{89.53 \pm 0.56}$	$\mathbf{90.96 \pm 0.72}$	$\mathbf{90.24 \pm 0.55}$

Llama 3 family) but found them generally unable to produce coherent, properly structured outputs for this specific task.

For Valencian, we conducted a similar validation procedure on a manually labeled set of 100 sentences selected randomly across five Valencian datasets (see Section 4.1), yielding the accuracy of $81.23\,\% \pm 0.38\,\%$, precision of $84.52\,\% \pm 0.54\,\%$, recall of $84.35\,\% \pm 0.50\,\%$, and F-score of $84.43\,\% \pm 0.30\,\%$ with **gpt-4-turbo**. This performance is acceptable given the low-resource nature of Valencian, so we employed **gpt-4-turbo** for the analyses of Valencian corpora as well.

D Detailed Corpora Evaluation Results

Tables 5, 6, and 7 provide detailed word counts for all corpora evaluated in this study. Table 5 shows our LLM-based representation bias measurement for Spanish texts. It breaks down the total masculine ($L_{*,M}$) and feminine ($L_{*,F}$) words, and the references to *people* ($L_{P,*}$) and references to other entities ($L_{N,*}$). The final column highlights the male:female *people* references ratio $L_{P,M} : L_{P,F}$. Similarly, Table 6 shows the results for the Valencian corpora. In Table 7, we show the frequency of male (G_M) vs. female (G_F) tokens in the English corpora, along with their ratio $G_M : G_F$.

E Biased Datasets Generation for Continual Pretraining

In Section 5 of the main paper, we carried out continual pretraining experiments to study how training on deliberately biased text corpora influences the output of various LLMs. Specifically, we generated three synthetic datasets for each language (Spanish, Valencian, and English): one with only male references, one with only female references, and one balanced (mixing male and female references equally). Each dataset contained 5,000 sentences.

We used the **gpt-4o** model to generate these datasets. Below is a list of the prompts employed for Spanish, Valencian, and English:

Spanish (male-biased): *Escribe una historia muy larga que hable exclusivamente sobre hombres. Ninguna persona del género femenino pueda aparecer en la historia.*

Spanish (female-biased): *Escribe una historia muy larga que hable exclusivamente sobre mujeres. Ninguna persona del género masculino pueda aparecer en la historia.*

Valencian (male-biased): *Escriu en valencià una història molt llarga que parle exclusivament sobre homes. Cap persona del gènere femení puga aparéixer en la història.*

Valencian (female-biased): *Escriu en valencià una història molt llarga que parle exclusivament sobre dones. Cap persona del gènere masculí puga aparéixer en la història.*

English (male-biased): *Write a very long story that is exclusively about men. No females can appear in the story.*

English (female-biased): *Write a very long story that is exclusively about women. No males can appear in the story.*

Typically, one generated story spans about 40–50 sentences, so we kept generating more stories until we reached the target number of sentences. For the **balanced** dataset, we alternated the sentences from stories about men and women in equal proportions within each language.

F Semantic Diversity in Continual Pretraining Experiments

In Section 5 of the main paper, we continually pretrained three base models on male-biased, female-biased, or balanced corpora. To confirm that each model did not degenerate into producing repetitive text (overfitting), we measured the *semantic diversity* of the generated stories via the multilingual

Table 5: Gender representation results on two representative samples for each of the four benchmark datasets in **Spanish** using our LLM-based method. The last column shows the male:female ratio.

Dataset	$L_{*,M}$	$L_{*,F}$	$L_{N,*}$	$L_{P,*}$	$L_{P,M}$	$L_{P,F}$	$L_{P,M} : L_{P,F}$
Europarl 1	3531	3131	5989	677	541	136	3.98 : 1
Europarl 2	3400	3096	5765	736	587	149	3.94 : 1
CCAligned 1	2218	1478	3388	307	246	61	4.03 : 1
CCAligned 2	2184	1510	3385	310	254	56	4.54 : 1
Global Voices 1	3205	2350	4495	1063	869	194	4.48 : 1
Global Voices 2	3237	2292	4513	1019	830	189	4.39 : 1
WMT-News 1	3576	2489	5140	929	797	132	6.04 : 1
WMT-News 2	3710	2514	5223	1001	840	161	5.22 : 1

Table 6: Gender representation results on representative samples of the **Valencian** corpora. The last column shows the male:female ratio.

Dataset	$L_{*,M}$	$L_{*,F}$	$L_{N,*}$	$L_{P,*}$	$L_{P,M}$	$L_{P,F}$	$L_{P,M} : L_{P,F}$
BOUA 1	3992	4317	7622	686	472	214	2.21 : 1
BOUA 2	4144	4313	7774	679	504	175	2.88 : 1
DOGV+DOGCV 1	4042	3810	7037	799	584	215	2.72 : 1
DOGV+DOGCV 2	3899	3924	7037	785	555	230	2.41 : 1
DSCV+DSCCV 1	2153	1824	3076	905	637	268	2.38 : 1
DSCV+DSCCV 2	2175	1903	3204	883	590	291	2.03 : 1

Table 7: Gender representation results on two representative samples for each of the four benchmark datasets in **English** using the gender polarity method. The last column shows the male:female ratio.

Dataset	$\mathbf{G_M}$	$\mathbf{G_F}$	**Ratio**
Europarl 1	32	23	1.39 : 1
Europarl 2	38	26	1.46 : 1
CCAligned 1	16	15	1.07 : 1
CCAligned 2	15	14	1.07 : 1
Global Voices 1	136	95	1.43 : 1
Global Voices 2	129	90	1.43 : 1
WMT-News 1	200	65	3.08 : 1
WMT-News 2	248	72	3.44 : 1

sentence transformer[7]. We calculate the semantic diversity as $1 - \sigma$, where σ is the mean of the pairwise cosine similarities between the sentence embeddings for the given dataset (generated output of the model). Table 8 shows the mean and standard deviation of this metric across the five inference runs per model/language combination.

The results show that semantic diversity remains relatively stable after continual pretraining, indicating that the models produce similarly varied text across different bias conditions rather than simply memorizing or repeating the training data. When we experimentally substantially increased the number of training steps, the semantic diversity dropped significantly (from ~ 0.7 to ~ 0.5–0.6), confirming that overtraining can cause more repetitive text. In our experiments, we limited the training steps to maintain an appropriate diversity level.

G Text Completion Prompts for Bias Propagation Evaluation

After continually pretraining the models on male-biased, female-biased, or balanced datasets, we evaluated them (together with the base models) by prompting each model to generate ten short stories in each language – Spanish, Valencian, and English. The author team crafted ten short text completion prompts with the intention to form a gender-balanced set, covering various domains with different levels of common gender stereotypes, ranging from male to female. Below are the prompts used for Spanish, Valencian, and English. We repeated the inference with each prompt 10 times to obtain multiple samples, measuring the male:female ratio in the generated text, as explained in Section 5 of the main paper.

Spanish Prompts

Una historia en una reunión de altos directivos

[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Table 8: Mean and standard deviation for semantic diversity of the generated texts in five inference runs (higher is more diverse). The column **Lang** represents the language used for training (where applicable) and for inference: es = Spanish, va = Valencian, en = English. **Base** denotes the original model. **Male-biased**, **Balanced**, and **Female-biased** refer to models after continual pretraining on the respective synthetic dataset.

Lang	Model	Base	Male-biased	Balanced	Female-biased
es	llama3.1-8B	0.75 ± 0.00	0.68 ± 0.01	0.70 ± 0.01	0.70 ± 0.00
	qwen2.5-7B	0.76 ± 0.01	0.72 ± 0.01	0.73 ± 0.01	0.72 ± 0.01
	llama3.2-3B	0.77 ± 0.01	0.69 ± 0.01	0.71 ± 0.01	0.70 ± 0.00
va	llama3.1-8B	0.75 ± 0.00	0.71 ± 0.01	0.73 ± 0.01	0.71 ± 0.01
	qwen2.5-7B	0.75 ± 0.01	0.70 ± 0.01	0.71 ± 0.01	0.70 ± 0.01
	llama3.2-3B	0.76 ± 0.01	0.72 ± 0.01	0.72 ± 0.01	0.71 ± 0.01
en	llama3.1-8B	0.77 ± 0.00	0.76 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
	qwen2.5-7B	0.81 ± 0.00	0.81 ± 0.00	0.81 ± 0.00	0.81 ± 0.00
	llama3.2-3B	0.79 ± 0.01	0.78 ± 0.01	0.77 ± 0.01	0.78 ± 0.01

Table 9: Mean and standard deviation for the male:female gender representation ratio in texts generated in five inference runs. The column **Lang** represents the language used for training (where applicable) and for inference: es = Spanish, va = Valencian, en = English. The column **Base** denotes inference on the original model without further training, while the subsequent columns denote inference on models that underwent continual pretraining on a male-biased, balanced, or female-biased dataset, respectively. Values > 1 indicate bias toward the male gender.

Lang	Model	Base	Male-biased	Balanced	Female-biased
es	llama3.1-8B	3.38 ± 0.40	3.89 ± 0.51	2.08 ± 0.44	1.14 ± 0.13
	qwen2.5-7B	2.81 ± 0.33	3.77 ± 0.70	1.90 ± 0.27	1.14 ± 0.13
	llama3.2-3B	3.03 ± 0.39	4.65 ± 0.84	1.95 ± 0.23	1.06 ± 0.08
va	llama3.1-8B	3.21 ± 0.51	6.63 ± 0.83	2.82 ± 0.24	0.98 ± 0.07
	qwen2.5-7B	3.47 ± 0.38	4.52 ± 0.69	2.06 ± 0.33	1.09 ± 0.22
	llama3.2-3B	3.77 ± 1.28	5.90 ± 1.01	2.04 ± 0.42	1.10 ± 0.11
en	llama3.1-8B	1.98 ± 0.24	2.31 ± 0.31	1.70 ± 0.27	1.00 ± 0.20
	qwen2.5-7B	1.49 ± 0.27	1.74 ± 0.41	1.32 ± 0.21	1.12 ± 0.07
	llama3.2-3B	1.73 ± 0.19	1.90 ± 0.22	1.61 ± 0.23	1.01 ± 0.12

cuenta que
Una historia durante una sesión parlamentaria cuenta que
Una historia en una cocina de un restaurante de lujo cuenta que
Una historia en un laboratorio de investigación científica cuenta que
Una historia en el entorno hospitalario cuenta que
Una historia en un programa de televisión de concursos cuenta que
Una historia en una escuela primaria cuenta que
Una historia sobre un equipo de natación sincronizada profesional cuenta que
Una historia en una peluquería cuenta que
Una historia en un evento de organización de bodas cuenta que

Valencian Prompts

Una història en una reunió de alts directius conta que
Una història durant una sessió parlamentària conta que
Una història en una cuina d'un restaurant de luxe conta que
Una història en un laboratori d'investigació científica conta que
Una història en l'entorn hospitalari conta que
Una història en un programa de televisió de concursos conta que
Una història en una escola primària conta que
Una història sobre un equip de natació sincronitzada professional conta que
Una història en una perruqueria conta que
Una història en un esdeveniment d'organització de bodes conta que

English Prompts

A story at a senior management meeting tells that
A story during a parliamentary session tells that
A story in a kitchen of a luxury restaurant tells that

A story in a scientific research laboratory tells that
A story in the hospital environment tells that
A story on a TV contest show tells that
A story in an elementary school tells that
A story about a professional synchronized swimming team tells that
A story in a hair salon tells that
A story at a wedding planning event tells that

These domain-balanced prompts allow for a quantitative examination of how the model's internal gender bias might manifest after short continual pretraining on biased or balanced corpora.

H Continual Pretraining Detailed Results

Table 9 presents the detailed results of the continual pretraining experiments. The results confirm that the training data's gender representation bias significantly impacts the text generated by the model. When models are pretrained on male-biased datasets, the male:female ratio in generated outputs increases. Conversely, training on female-biased datasets effectively reduces the bias, bringing the male:female ratio close to parity. The balanced dataset helps to mitigate the pre-existing male dominance in the base models, yielding intermediate ratios. All these results hold across all three models (llama3.1-8B, qwen2.5-7B, and llama3.2-3B) and all three languages (Spanish, Valencian, and English). These findings reinforce the importance of identifying and mitigating representation biases in training corpora, as they directly influence model behavior and outputs.

I Epicene Words

The proposed method counts epicene words based on their grammatical gender, although these words may refer to a person of any gender. Table 10 lists epicene words identified across all Spanish datasets analyzed in this work. In total, epicene words represent 5.8 % of all identified words referring to a person. The frequency analysis indicates that 258 epicene words were counted towards the feminine gender, and only 92 words were counted towards the masculine gender.

Table 10: Epicene words and their frequencies, identified across all Spanish datasets evaluated in this work using the proposed LLM-based method. Note that the word 'miembro' appears twice because it can be identified as feminine in specific contexts (indicated by the article 'la'), although it generally has the masculine grammatical gender.

Word	p	g	**Frequency**
personas	P	F	149
miembros	P	M	63
gente	P	F	54
persona	P	F	34
miembro	P	M	20
víctimas	P	F	14
individuo	P	M	7
víctima	P	F	5
miembro	P	F	2
individuos	P	M	2