# On the Relationship of Social Gender Equality and Grammatical Gender in Pre-trained Large Language Models

Magdalena Biesialska[1], David Solans[2], Jordi Luque[2] and Carlos Segura[2]

[1]TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Telefónica I+D Research, Barcelona, Spain

Abstract

Large Language Models pre-trained on vast amounts of text have demonstrated remarkable capabilities in modeling and generating human language, finding applications across a wide range of Natural Language Processing tasks. However, recent studies have unveiled the presence of biases in these models, inherited from social biases reflected in their training data. In this research article, we delve into the examination of grammatical gender's influence on four distinct languages exploring how the gender prejudices, exhibited by the LLMs, relate to their capacity to characterise social realities. We show that prevalence of gender biases differ not just in relation to the architecture and training data of the LLMs, as previously documented, but also vary with respect to the language and level of grammatical gender marking present in the language under study. Different LLM systems and languages are examined, ranging from a major grammatical gender language, such as Polish, up to English, which lacks most gender inflection, and throughout gendered languages, such as German and Spanish.

**Keywords**

gender bias, large language models, bias quantification

## 1. Introduction

Large Language Models (LLMs) are neural network systems that have been trained on massive amounts of text data by using deep learning techniques [1, 2]. These models seem capable of generating and comprehending human-like text, e.g., having reported remarkable performance across the majority of Natural Language Processing (NLP) benchmarks and tasks [3, 4]. Pre-trained LLMs are often adapted or fine-tuned to a specific NLP task (often referred to as *downstream tasks*) aiming at reducing the computationally expensive and time-consuming training stage. Downstream tasks can include a range of NLP tasks such as machine translation, question answering, semantic parsing, natural language inference or paraphrasing, among others [5] and often rely on extracted word embeddings [6] from pre-trained LLMs, e.g., in sentiment and gender bias towards politicians [7]. However, they are not immune to the biases that exist in the society, often reflected in their training corpora, as gender bias or other social clichés [8]. As LLMs are trained on not well-balanced data, in terms of gender or other attributes, they reflect societal stereotypes in many shapes, forms and times [9, 10]. The biases that are present in the massive amounts of linguistic data used to train LLMs are often incorporated by them, like the case of Virtual Assistants [11]. This could have a long-lasting effect on the behavior of society conditions. Leading to discriminatory responses and decisions about race, age, religion, geographical origins, or the specific case of gender [7, 12, 13, 14]. Thus, perpetuating mechanisms that create and maintain male dominance.

As a result of this, LLMs could not correlate female terms, e.g., with engineering professions, being prone to not promote female candidates for engineering positions even when being equally qualified [15, 9] as their male counterparts. A biased LLM may perpetuate harmful stereotypes and reinforce both bad preconceptions and prejudices which would limit chances and increase inequalities, then limiting opportunities for some groups [16].

Furthermore, online data is gathered from the specific group of population that uses online resources, which has particular characteristics, resulting in biased training samples that fail to effectively reflect the needs of marginalised social groups [17, 18, 2]. Detecting and characterizing biases becomes a crucial task, especially in the case that such models are used in high-risk domains [1], where NLP applications can easily limit human potential, e.g., by inducing biases against women in authority [19]; hamper economic growth, and, definitively, reinforce social inequity [20]. In the labour market domain, efforts to address gender biases include promoting diversity and inclusion in hiring and promotion processes, raising awareness of unconscious bias, and providing support to women and other underrepresented groups [21].

*SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.*
jordi.luque@telefonica.com (J. Luque)
0000-0001-7890-3523 (M. Biesialska); 0000-0001-6979-9330 (D. Solans); 0000-0002-4507-4930 (J. Luque); 0000-0001-5867-281X (C. Segura)


[1]European Commission, Regulatory framework proposal on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

But the problem does not only apply to the data. NLP systems are prone to amplify the gender bias exhibited in text corpora. Hence, the problem becomes multi-faceted and may be present at various stages of the development of NLP systems, including training data, resources, pre-trained models, and algorithms [22]. Further propagation of gender bias from NLP models to downstream applications is likely to reinforce harmful stereotypes and may result in, for example, discrimination of female candidates on the labour market.

Presence of LLMs' gender biases in the labour market domain has been previously investigated at the level of professions, assessing the correlation between labour census and LLMs' association scores for a subset of professions across genders [8, 23]. However, we argue that this form of bias evaluation does not consider relationships between professions, such as the economic sectors in which their activity is developed.

For this reason, we provide an alternative perspective that relies on the evaluation of biases in LLMs at the level of economic sectors. Using a higher level of granularity allows us to detect patterns that could not be observed before. The findings of this study have important implications for the development and use of cross-lingual language models. By quantifying gender bias, these models can be improved to provide more fair and unbiased representations of language. This research contributes to the broader goal of promoting gender equality and reducing bias in NLP applications.

## 2. Related Work

Gender bias is understood as the systematic preference or prejudice toward one gender over the other [17, 24, 25]. Previous work has studied the issue of quantifying social biases in language [26], NLP [27, 28], and specifically, gender biases elicited by LLMs or carried on by implicit associations in their word embeddings, for human work-related activities [15]. However, while the proposed methods work well for English-based LLMs, they fail to capture bias for languages with a rich morphology or gender-marking, such as German, Polish or Spanish [29]. Countries where gendered languages are spoken often evidence less gender equality compared to countries with other grammatical gender systems [30]. While previous work has centered on the English language, recent studies have explored bias in multilingual contexts and languages other than English [31, 32, 33, 34] .

There are weak evidences that *language shapes the way of thinking*. Previous argument is mentioned in the work of Whorf and Carroll [35] and such ideas have been the subject of debate and criticism. Whorf's work explores the idea that language shapes our cognition and perception proposing that language influences the way we perceive and think about the world, a concept known as the Sapir-Whorf hypothesis or linguistic relativity hypothesis. Whorf argues that different languages may lead to different ways of thinking and perceiving the world, suggesting that language not only reflects our thoughts but also shapes and constrains them, having the key argument that language affects our perception of time.

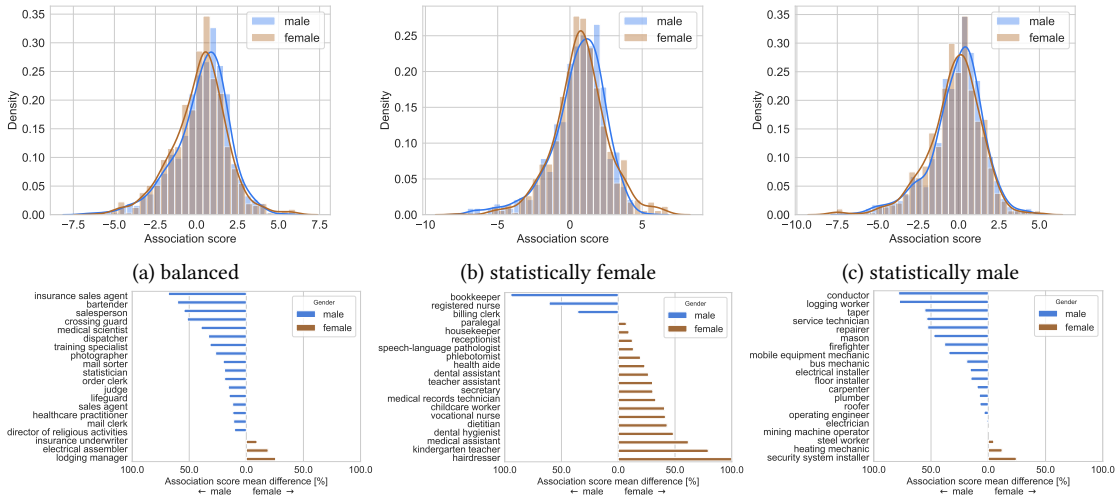
In contrary, some studies argued that the influence of language on thought is limited and that there are universal cognitive processes that are independent of language [36]. This pattern of perception, which is predicted by the asymmetry between space and time in linguistic metaphors, was reported also in [37] by tasks that do not involve any linguistic stimuli or responses, arguing that our mental representations and conceptualization of time are built upon our experiences with space and motion and not necessarily involving *the way we talk about time*, e.g., by using the spatial language from an idiom.

Nonetheless, recent research has provided evidence supporting the influence of the language we speak on our cognitive framework of the world we perceive. The work of Tan et al. [38] uses functional magnetic resonance imaging (fMRI) reporting that brain regions involved in language processing are also activated during perceptual decision-making tasks, which suggests that language and perception are closely intertwined. Finally, the study from Banaji and Hardin [39] supports the claim that gender information conveyed by word and sentences can automatically influence judgement, creating a form of automatic stereotyping in persons.

One of the primary objectives of this research is to investigate differences in gender equality among countries, across various economic sectors, and with regard to LLMs. This research will explore the correlation between gender-marked languages and gender equality and evaluate whether LLMs represent the world depending on the language they were trained on. Additionally, previous work in LLMs and labour sector, from computational linguistics [8, 15] has focused on a small fraction ($\sim$15.6%) of the complete list of professions available in the U.S. census to assess biases where differences in gender prevalence according to the census are maximised (e.g., female professions) or minimised (e.g., neutral professions). However, the previous analysis does not shed light on patterns that might be dependent on the economic sector where the full set of professions are located.

### 2.1. Contributions

This work makes the following contributions: (i) We extend previous definitions of gender biases in pre-trained LLMs to work with two different types: *stereotyping bias* and *representation bias* and characterise multiple items in the trade-off between them. (ii) We evaluate such




(a) balanced (b) statistically female (c) statistically male

(d) balanced (e) statistically female (f) statistically male

**Figure 1:** Density plots of female and male distributions of association scores for (a) balanced, (b) statistically female and (c) statistically male professions averaged for the Spanish RoBERTa-BSC large cased (upper) and bar plots of the corresponding association scores for each individual profession (bottom). Spanish professions are cherry-picked using the same criteria as the BEC-PRO dataset and using only gendered attributes. The association scores are estimated using the method in Section 3.3.2

biases in pre-trained LLMs across multiple languages, ranging from *languages without grammatical gender* (e.g., English) to rich morphological or gender-marking languages, which we name *gendered languages* (e.g., Spanish). (iii) For each language, we perform an evaluation on multiple pre-trained LLMs such as BERT [40] and RoBERTa [41]. (iv) With state-of-the-art focused in studying biases in labour market at the level of professions, we change the lenses and analyze them at the level of economic sectors, comparing results with gender statistics for the labour market. (v) We will release our code including templates.

## 3. Methodology

This section describes the methodology employed to measure gender bias in LLMs, with a focus on labour market stereotypes. In our work, we leverage pre-trained LLMs to quantify gender bias using a template-based approach to measure association scores between a token and a masked *target* or *attribute*.

### 3.1. Pre-trained Language Models

Pre-trained LLMs have been successfully employed to different tasks and numerous applications in NLP in recent years. Significant performance gains have led to the development of various architectures. One of the most prominent LLMs is BERT [40]. Later RoBERTa [41], a more robust version of BERT, was released. Both models rely on the Transformer architecture introduced in [42]. In a nutshell, BERT is trained to predict the original tokens in a sentence that have been randomly masked. Utilizing the Masked Language Model (MLM) objective, BERT evaluates the probability distribution of possible tokens that could fit the masked position in the sentence, attempting to correctly infer the original masked word. Additionally, the model predicts the next sentence. RoBERTa aimed to address some of the shortcomings of the BERT architecture, hence RoBERTa was trained with *dynamic masking* instead of the *static* variant when a sequence is input to the model.

In particular, we performed experiments with two types of LLMs: BERT [40] and RoBERTa [41]. All the BERT BASE models were trained using 110M parameters, while BERT LARGE with 340M parameters. RoBERTa BASE and LARGE models were trained with 125M and 355M parameters respectively. Importantly, we experimented with different LLM architectures (BERT and RoBERTa), model size (base and large), as well as uncased and cased variants for four languages. In the previous studies, such as [15] the evaluation of LLMs was limited only to two languages (English and German), only one model type (BERT BASE) and the authors did not analyze how casing influences the results. The diverse models and the corresponding languages and corpus they were trained on are outlined in table 4.

**Table 1**
Example of templates for the different languages. Depending on the grammatical gender system of the language, Polish, German and Spanish templates are changed accordingly in pronouns, verbs or articles, see section 3.2.

| English | Spanish |
|---|---|
| $< person >$ works in the $< economic\_sector >$ sector. | $< person >$ trabaja en el sector $< economic\_sector >$. |
| $< person >$ has a job in the $< economic\_sector >$ sector. | $< person >$ tiene un trabajo en el sector $< economic\_sector >$. |
| $< person >$ would like a job in the $< economic\_sector >$ sector. | $< person >$ quiere un trabajo en el sector $< economic\_sector >$. |
| $< person >$, who works in the $< economic\_sector >$ sector, had a good day at work. | $< person >$, que trabaja en el sector $< economic\_sector >$, tuvo un buen día en el trabajo. |
| $< person >$ applied to the position in the $< economic\_sector >$ sector. | $< person >$ solicitó una posición en el sector $< economic\_sector >$. |
| **German** | **Polish** |
| $< person >$ arbeitet im $< economic\_sector >$. | $< person >$ pracuje w sektorze $< economic\_sector >$. |
| $< person >$ hat einen Job im $< economic\_sector >$. | $< person >$ ma pracę w sektorze $< economic\_sector >$. |
| $< person >$ würde gerne im $< economic\_sector >$ arbeiten. | $< person >$ chciałby pracować w sektorze $< economic\_sector >$. |
| $< person >$, $< article >$ im $< economic\_sector >$ arbeitet, hatte einen guten Arbeitstag. | $< person >$, który pracuje w sektorze, $< economic\_sector >$ miał dobry dzień w pracy. |
| $< person >$ hat sich um die Stelle im $< economic\_sector >$ beworben. | $< person >$ ubiegał się o pracę w sektorze $< economic\_sector >$. |

## 3.2. Grammatical and Natural Gender Languages

In the field of linguistics, a grammatical gender system represents a distinct form of a noun class system, wherein nouns are categorised based on gender attributes. In languages featuring a grammatical gender system, the majority of the nouns inherently bear one value of the grammatical category known as gender.

The Spanish language is considered a romantic language that falls within the grammatical gender language category as well as German and Polish languages. In Spanish, there are two genders: masculine and feminine, and both the noun and adjective systems exhibit these two genders [43]. In addition, articles and some pronouns and determiners have a neuter gender in their singular form. German is also an inflected language [44] with three genders: masculine, feminine and neuter. In Polish, the only non Indo-European language in this study, nouns belong to one of three genders: masculine, feminine and neuter. In this West-Slavic language, the masculine gender is also divided into subgenders: animate/inanimate in the singular, and human/nonhuman in the plural. Furthermore, adjectives agree with nouns in terms of gender and conjugated verb forms agree with their subject's gender in the case of past tense and subjunctive/conditional forms.

Nevertheless, English is considered a natural gender language and most of the nouns, with some exceptions, are considered genderless [44]. English has three gendered pronouns, but no longer has grammatical gender in the sense of noun class distinctions or inflections. Instead, gender is characterised through the language's pronouns [30], that is, the distinction between "he", "she", and other personal pronouns and "it".

## 3.3. Bias Quantification

To quantify biases in a particular context, it is important to first establish a clear definition of what a bias-free system would look like. This requires a thoughtful reflection on the desired behavior of the analysed model and the impact that potential biases might have. In our work, we approach this task from two different perspectives, what we name *Stereotyping Bias* ($S_b$) and *Representation Bias* ($R_b$). The former quantifies how a given LLM is far from gender neutrality given a context. The latter takes into account the LLM bias with respect to what is observed in society. For instance, in figure 1d where professions are supposed to be balanced among genders [15], we would expect that a BERT model with no bias will produce association scores around zero (see section 3.3.2 for more details on the association scores). Looking at figure 1a, any deviation from the observed perfect overlapping would account for stereotyping bias, see sections 3.3.3 and 3.3.4 for further details on the two perspectives on bias quantification. The applicability and preference for one notion over the other depends on the context of usage of the LLM at hand [45]. Existing studies quantify gender bias in pre-trained LLMs typically using tailored sets of synthetically generated sentences and implicit associations between word embbedings [46]. In the work of Kurita et al. [47], gender bias in BERT models is measured using a probability-based metric [25] and by using template sentences. Specifically, the LLM is directly queried for a particular token in a template sentence by sequentially masking of either *target* or *attribute* token, see table 1 in where $< person >$ and $< economic\_sector >$ stand for the *target* and the *attribute* words, respectively. In our analysis, the mask [TARGET] is replaced by gendered nouns and pronouns (e.g.: he/she/my sister) and the mask [ATTRIBUTE] is replaced with terms related to specific economic sectors (e.g., fishing/services/secondary). As contextualised embeddings of a given token are dependent on its context, a relative measure of bias for the *attribute* word can be evaluated by substituting target classes (e.g., male and female). In [47], the authors compare their evaluation method with the baseline cosine similarity measure among word embeddings.

However, applying Kurita's methodology confronts different challenges when applied for grammatically gendered languages such as Spanish or German. Previous work by Bartl et al. [15] demonstrated that the original association scores proposed in [47] were not effective for the German language due to its inherent gender suf-

fixes in attributes. In English a few gendered words exist (e.g., king/queen, waiter/waitress, actor/actress), and measuring the association score for sentences with those words, e.g., "[TARGET] is the waitress", with male or female options would yield misleading results when using word embedding projection methods [29], thus showing a gender bias against men instead of women. This phenomenon prevails into gendered languages, where different words are used for each gender. For instance, if we compare the distributions from the figures 1a to 1c, corresponding to the distributions of association scores for Spanish language, we notice that Kurita's method obtains overlapped distributions for the three groups of professions in Spanish language. This result is also confirmed by a drastic reduction of the p-values obtained by a Wilcoxon test statistic compared to English distributions. It is worth to mention that the same effect occurs for both German, as previously noticed by Bartl et al. [15], and for the Polish languages. The previous results motivate us to develop a new set of templates, aiming to avoid the effects of gendered attributes for the quantification of bias in this work.

#### 3.3.1. Templates

We adapt the idea of using templates to quantify and measure gender bias [47, 15]. Bartl et al. [15] used association scores to analyze gender biases across professions, releasing the BEC-Pro dataset for English and German languages. We follow a similar approach, but we extend the analysis to two additional languages: Spanish and Polish. More importantly, we shift the focus from individual professions to entire economic sectors.

Note that, as we discussed before in 3.3, the relation between the grammatical gender of the person word and the profession does influence the associations scores in gender-marking languages. In response to this, the novel approach of measuring biases across economic sectors instead of using a list of occupation words allows us to minimize potential complications stemming from grammatical gender inflections and pronouns. Additionally, by examining economic sectors, our investigation encompasses an aggregated view instead of limiting the analysis to a specific list of professions and, for instance, facilitating the relation of results to macroeconomic statistics.

Our templates are designed to assess gender bias in LLMs concerning economic sectors. To achieve that, we take into account changes in sentence structure (e.g., articles) depending on the female or male person word. These templates follow a standard structure, where a sentence contains an economic sector reference as the *attribute* with a specific gendered term as the *target*.

#### 3.3.2. Adapted Kurita's algorithm

The association scores methodology proposed by Kurita et al. [47] is employed to measure the likelihood of a masked word being associated with a specific gender. These scores quantify the gender bias present in the LLMs by evaluating the probability that the masked token is classified as male or female. Higher scores for a particular gender indicate a stronger bias towards that gender in the predictions of the evaluated LLM.

The aim of this method is to estimate the implicit association between specific *targets* and *attributes* using BERT's MLM objective. For example, using the template sentence "she works in the construction sector", the method can quantify the association between the target *female* (given by the pronoun "she") and the attribute *construction*. The distribution scores drawn by figure 1 are obtained in the same manner.

The main steps of the method are as follows:

1. Prepare a template sentence
   e.g. "[TARGET] works in the [ATTRIBUTE] sector".
   For example this may be "she works in the construction sector".
2. Mask the [TARGET] word and compute the target probability $p_{tgt}$ which corresponds to the likelihood of the target word given an unmasked attribute.
   For the updated example, the sentence becomes "[MASK] works in the construction sector" and $p_{tgt}$ measures how likely the LLM is to predict "she" as the missing word.
3. Compute the prior probability $p_{prior}$, which is the likelihood of the target word when the attribute is also masked.
   The example sentence would be "[MASK] works in the [MASK] sector", and $p_{prior}$ is the probability of predicting "she" without the influence of the attribute.
4. Compute the association between target and attribute as $a_s = \log \frac{p_{tgt}}{p_{prior}}$.

This logarithmic ratio is the association score, $a_s$. To measure gender bias, we compute the gender bias by comparing these scores for different targets, such as "he" and "she", averaging for all templates and taking the difference between female and male association score averages. This method, as evidenced in the original paper, outperforms traditional cosine-based measures like WEAT [8] in detecting gender biases.

#### 3.3.3. Stereotyping Bias

*Stereotyping bias* ($\mathcal{S}_b$) quantifies the extent to which a given LLM is far away from gender neutrality in a given

context given by a specific language ($\mathcal{L}$) and a LLM model ($\mathcal{M}$). To do so, it quantifies the disparities in average association score across genders for each of the economic sectors: primary, secondary and tertiary.

To calculate the overall $S_b$ across sectors we first calculate the stereotyping bias for a specific sector $s$ as the inner disparity $\mathcal{ID}_s(\mathcal{L}, \mathcal{M})$ by first computing the model's average difference between association scores $a_s$ between females and males. This difference is calculated for each i-*th* sentence generated for females $a_s(f_i)$ and males $a_s(m_i)$ between the total number $n$ of male and female oriented sentences generated for $\mathcal{L}$ and $s$.

$$\mathcal{ID}_s(\mathcal{L}, \mathcal{M}) = \frac{1}{n} \sum_{i=0}^{n} \big(a_s(f_i) - a_s(m_i)\big) \quad (1)$$

The overall stereotyping bias $\mathcal{S}_b$ across sectors for $\mathcal{M}$ and $\mathcal{L}$ is computed as the average inner disparity across all three economic sectors:

$$\mathcal{S}_b(\mathcal{L}, \mathcal{M}) = \frac{1}{3} \sum_{s=1}^{3} \mathcal{ID}_s(\mathcal{M}, \mathcal{L}) \quad (2)$$

A model $\mathcal{M}$ trained for language $\mathcal{L}$ without stereotyping bias $S_b(\mathcal{L}, \mathcal{M}) = 0$, would produce equal average association scores for male and female targets in economic sectors. Negative values indicate bias favoring males, while positive values indicate bias favoring females. Stereotyping bias is specific to each model and the language in which it was trained.

#### 3.3.4. Representation Bias

With a the broader view, *representation bias* in a given domain generally refers to the underrepresentation or overrepresentation of certain groups (such as genders or ethnicities) as compared to their prevalence in the overall target population. However, in the context of our research, we adopt a definition of *representation bias* ($\mathcal{R}_b$), particularly tailored to the context of our analysis. Here, $\mathcal{R}_b$ is understood as the divergence of a model's internal representation of genders from the actual societal gender distributions in the workforce.

We define the overall representation bias across economic sectors for a given LLM $\mathcal{M}$ trained for language $\mathcal{L}$ as:

$$\mathcal{R}_b(\mathcal{L}, \mathcal{M}) = \frac{1}{3} \sum_{s=1}^{3} \big(\mathcal{ID}_s(\mathcal{L}, \mathcal{M}) - \mathcal{OD}_s(\mathcal{L})\big) \quad (3)$$

Here, $\mathcal{ID}_s(\mathcal{M}, \mathcal{L})$ is the model's inner disparity score as defined above, and $\mathcal{OD}_s(\mathcal{L})$ represents the observed gender ratio in economic sector $s$ in the country associated with the language $\mathcal{L}$. $\mathcal{OD}_s(\mathcal{L})$ is calculated by comparing the percentages of females and males in sector $s$ using data from the Global Gender Gap Index across different countries.

A balanced model ($\mathcal{R}_b = 0$) perfectly reproduces societal gender distributions. Negative values indicate a model preference for males compared to the real prevalence, while positive values indicate a preference for females. This metric helps language modelers ensure accurate societal representations in their models.

### 3.4. Labour Market Data

We evaluate gender bias by investigating the relation between gender-denoting target words and sectors names in English, Spanish, Polish and German. As mentioned before, our proposed method avoids using professions names, gendered or not, to keep the attribute unchanged among languages, thus making the results comparable among different languages under study. The observed bias is compared to real-world through the gender statistics across activity sectors provided by the World Bank (see table 2), which describe the prevalence of males and females across economic sectors. Comparisons are done based on specific countries and their relative languages (e.g., Spanish models are compared with statistics from Spain) so that we are able to specifically compare each model's outputs with its social reality, regarding the level of gender equality represented at the workforce statistics and per each economic sector. We also report the Global Gender Gap Index[2] by the World Economic Forum (WEF). As seen in table 2, the bigger gender gap among sectors is found for the secondary sector, a common trend in the four countries, where is mostly occupied by male workers. The contrary case is found in the tertiary sector, in where the relative difference favours the female gender. For the primary sector, a similar statistic is found except for Poland where gender is almost balanced.

## 4. Results

The results obtained for LLMs trained in English, German, Spanish and Polish languages reveal intricate patterns of the two notions ($S_b$, $R_b$)of gender bias that emerge and fluctuate across languages, economic sectors, and model types. The analysis based on results depicted in figure 2, representing gender bias, separated by primary, secondary, and tertiary sectors, exhibits diverse trends across languages and sectors. As can be observed, the uncased BERT models generally exhibit less stereotyping

[2]The GGG index "assesses countries on how well they are dividing their resources and opportunities among their male and female populations, regardless of the overall levels of these resources and opportunities", https://www3.weforum.org/docs/WEF_GGGR_2022.pdf

[3]https://genderdata.worldbank.org/indicators/sl-empl-zs/

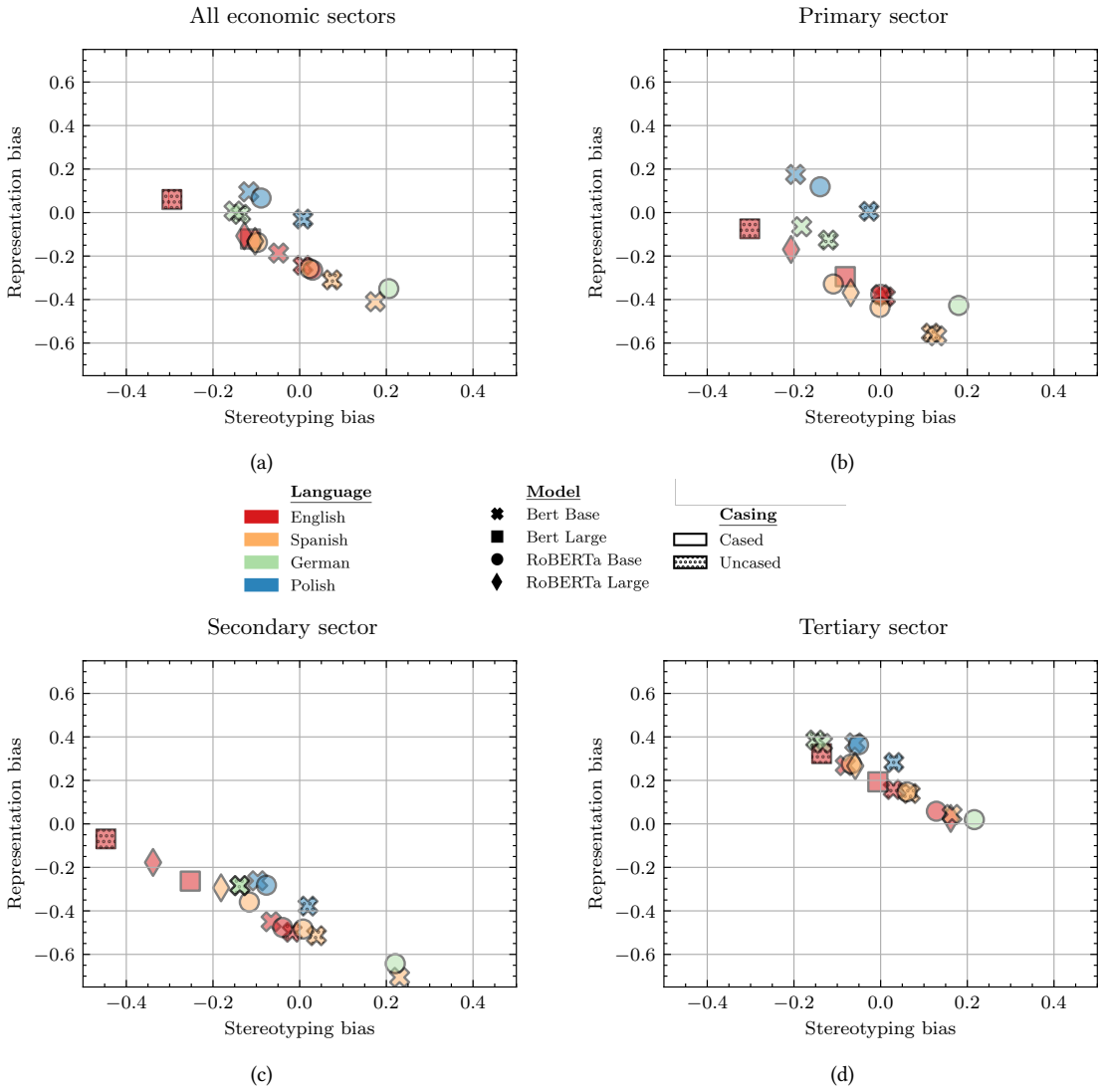



**Figure 2:** Stereotyping vs representation bias trade-offs on the analysed models. Negative values indicate bias towards male gender.

bias compared to their cased counterparts in all languages. Within the range of languages analysed, Polish demonstrates the lower stereotyping bias, followed by Spanish, with German and English showing similar average values, but with English models demonstrating a broader range of stereotyping bias.

When analyzing results both aggregated and across sectors, an interesting pattern emerges. In the 2D scatterplot all models, for a given language, align along a specific trajectory, revealing a clear trade-off between how accurate a language model represents the social reality (representation bias) and the stereotyping bias that the language model exhibits.

Specifically, we are interested in the specific domains for $R_b$ in which grammatical gendering of the LLM's language might be a proxy for predicting gender inequality in the country of the spoken language. As reported by Prewitt-Freilino et al. [30] , countries predominated by a natural gender language, like English, evidence greater gender equality than countries with other grammatical gender systems. Albeit, as seen in table 2, that is not the case for the English GGG index (only accounting for England), lower than GGG indexes reported for Germany and Spain, both gendered languages.

**Table 2**
The male/female columns refer to the % of workforce in the sector. The Rel. diff. column stands for the relative difference between both genders, normalised in the range [-1,1], where negative values indicate real-world bias towards male gender. The Global Gender Gap index (GGG) reported by the World Economic Forum 2022[2] and employment by sector with respect to gender statistics from 2019 by the World Bank[3]. The values of the index range from 0 to 1, with higher values indicating greater gender equality.

| Language | GGG | Sector | Female | Male | Rel. diff. |
|---|---|---|---|---|---|
| German | 0.801 | Primary | 37.64 | 62.36 | -0.25 |
| | | Secondary | 28.90 | 71.10 | -0.42 |
| | | Tertiary | 61.80 | 38.21 | 0.24 |
| Spanish | 0.788 | Primary | 28.13 | 71.87 | -0.44 |
| | | Secondary | 26.23 | 73.77 | -0.48 |
| | | Tertiary | 60.40 | 39.60 | 0.21 |
| Polish | 0.709 | Primary | 49.00 | 51.05 | -0.02 |
| | | Secondary | 32.10 | 67.94 | -0.36 |
| | | Tertiary | 65.65 | 34.35 | 0.31 |
| English | 0.780 | Primary | 31.19 | 68.81 | -0.38 |
| | | Secondary | 24.22 | 75.78 | -0.52 |
| | | Tertiary | 59.34 | 40.66 | 0.19 |

For the case of the English LLMs, see figure 2a, the BERT base uncased and RoBERTa base cased models report the lowest $|\mathcal{R}_b|$ and $|\mathcal{S}_b|$, thus being both a good proxy for real-world data and low stereotyping of LLMs. Note that values of $S_b$ can be understood as the LLM's *perception* of the world once the LLM is trained in a specific language, whereas $R_b$ describes its capacity for predicting the real-world data or gender gap. Similar results are observed for Spanish, where the Base models outperform Large models in both bias metrics, $\mathcal{R}_b$ and $\mathcal{S}_b$. Note that the main difference between Large and Base models resides in the number of parameters employed for the architecture, being the number of tokens in the training corpus the same for both LLM systems. Table 4 in annex, summarizes the training data used and number of parameters for each LLM.

Regarding the Spanish model BERTIN, it corresponds to a RoBERTa base model trained with 100B tokens more than the RoBERTa-BSC model. In the figures, both models are denoted with an orange circle. BERTIN portrays the lowest error in term of bias, skewed toward negative values of $R_b$ for the sector aggregation graph 2a. Note that the graph for all sectors is computed as a weighted average using WEF data proportions from the other three sectors results, see table 2.

Overall, the gendered languages exhibit a smaller variance around the $(\mathcal{S}_b, \mathcal{R}_b) = (0, 0)$ compared to English LLMs. Nonetheless, by removing the Large models from the analysis, we can realize that English LLMs, trained with a natural gender language, are closer in average to the non-bias point $(0, 0)$ compared to the rest of languages, except for Polish. Previous result is diluted depending on the specific economic sector we look at, in where the English $R_b$ in primary and secondary sector is compensated by the tertiary sector, the latter biased towards the female direction.

We also observe interesting results in the case of the Polish language, which belongs to a family of Slavic languages. Polish, as a West Slavic language, is gendered; however, there is a very limited number of studies investigating bias even in the broader Slavic language family [48] . Hence, the lack of such analysis is addressed in our work. The results for Polish LLMs, as shown in figure 2, in general demonstrate lower representation bias scores as compared to other languages.

We hypothesize that one of the reasons might be attributed to the gender-sensitive grammar structures in Polish. Unlike many other languages, Polish modifies not only pronouns but also verb forms to correspond with gender. For example, in certain sentence templates, the conditional and past tenses of verbs, and relative pronouns, alter according to the gender of the person being referred to. This linguistic feature may potentially impact how LLMs learn and represent gender-related concepts in Polish, thus influencing the extent of bias observed in these models. Given the gendered nature of Polish and how the provided patterns reflect that, we conclude that the representativeness (as indicated by the representation bias) of Polish LLMs is slightly better than its English, Spanish and German counterparts due to the necessary agreement of verbs and pronouns with the gendered subject.

## 5. Conclusion

In this work, we have used the idea of association scores [47] to quantify gender biases in LLMs in the labour market at the level of economic sectors. We distinguish between two different notions of biases: (i) Representativity bias and (ii) Stereotyping bias. The first quantifies the extent to which the model is able to learn patterns that can be observed in society, whereas the latter studies how far from gender-neutral its internal representation is.

By conducting this cross-linguistic analysis, we contribute to the understanding of biases in LLMs, highlighting the nuanced interplay between language structure, training data, and the biases exhibited by these models. Our study underscores the importance of comprehending how biases are captured or amplified within LLMs, paving the way for future efforts to mitigate and address these biases.

We use these definitions of bias to characterise multiple state-of-the-art pre-trained LLMs, comparing results

among different languages, from languages with no grammatical gender, or natural gender, to gendered languages. Among other results, the conducted analysis reveals interesting and consistent trends where biases vary across languages and economic sectors, being Polish the language whose models systematically showcase less biases and the tertiary sector, the unique case for which the models exhibit a biased preference towards the female gender.

Additionally, we observed a quasi-linear relation between both types of biases, with most of the models exhibiting representation, stereotyping biases or a combination of both and Large models reporting higher biases. We expect these results to contribute to building a better understanding on the presence of systematic gender biases in LLMs.

### 5.1. Limitations and Future Work

This study uses a multi-language dataset, synthetically created with equivalent examples across the studied languages. However, as other datasets used in the related work, it is still limited in the sense that few templates are used to generate it. Additionally, it is important to note that cultural biases might affect the understanding of the translated templates to each language, leading to differences that could be reflected in the obtained results. This raises the possibility that unintended biases may be present in the results derived from the data.

Furthermore, the dataset is composed by a tailored collection of terms that are descriptors of economic sectors for which the results are then aggregated. Although aggregation is a powerful tool to observe patterns, at the same time it has the drawback of restricting the visibility of interesting patterns that occur at a lower granularity level, for instance, using professions related to each economic sector. This means that using solely the results reported in this work might not be sufficient to understand all the possible types of biases in LLMs in the domain of the labour market, but correspond to another set of information to be accounted.

Moreover, we are comparing individual census with results of different languages that are spoken in multiple countries. As a future work, it could be interesting to compare results across multiple countries that use those languages. Additionally, more research could be done on the effects of other demographic factors or covariates.

## 6. Ethics Statement

This research provides a deepened insight into the influence of grammatical gender on gender biases within LLMs across multiple languages. The broader societal impact of understanding and quantifying these biases is significant for several reasons:

1. **Enhancing Awareness**: By bringing attention to the variances in gender biases across languages, we can enhance the broader community's awareness of potential pitfalls when deploying LLMs in diverse linguistic settings. This awareness is crucial for developers, policymakers, and users to make informed decisions about the application and potential limitations of LLMs in different linguistic contexts.
2. **Informed Deployment**: Knowledge about the biases inherent to these models can guide decision-making processes for institutions and industries that utilize LLMs. By being aware of the biases, stakeholders can make better decisions regarding where and how to deploy these models, especially in applications that may have real-world implications for individuals or groups.
3. **Influence on Future Research**: Our study can pave the way for future research into the mitigation of gender biases in LLMs. By understanding the nuanced interplay between language structure, training data, and model bias, the community can work towards developing techniques and best practices to address and reduce such biases.

### 6.1. Ethical Considerations

1. **Dataset Limitations**: While our study utilizes a multi-language dataset, it is synthetically created with equivalent examples across languages. As with any synthetic dataset, there's a risk of unintended biases, potentially affecting the results. We recognize and caution that translating templates across languages can introduce cultural biases, which might inadvertently influence the outcomes.
2. **Scope of Findings**: It is important to understand that our findings, while indicative of trends, may not extrapolate seamlessly to other new LLMs or to every application scenario. For example, we are only reporting results for a narrow set of languages and modeled by non-causal LLMs, that is, no autoregressive models, as GPT-like, have been evaluated. Biases are intricately linked to specific training data, model architecture, and application context. Our study should be viewed as a piece in the larger puzzle of understanding and addressing biases in LLMs, rather than a conclusive assessment of all possible instances of gender bias in every LLM.
3. **Aggregation of Results**: Our use of aggregation, while powerful in discerning patterns, might also mask more granular biases present in LLMs, particularly within specific economic sectors or professions. Users and developers should be aware

of this and consider more detailed analyses when appropriate.

4. **Comparative Analyses**: Our study compares census data with results from languages spoken across different countries. The cultural, economic, and social dynamics of each country can vary widely, even if the same language is spoken. Future work may benefit from a more localized approach, considering the multifaceted nature of biases in each country.
5. **Potential Misuse**: Recognizing that biased systems can perpetuate stereotypes or reinforce societal prejudices, it is ethically imperative for developers and users to ensure that LLMs are not misused, especially in critical domains where biases can lead to tangible harms or injustices.

## Acknowledgments

Funded by the European Union's Horizon 2020. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

## References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ArXiv abs/2304.13712 (2023).

[3] S. Tedeschi, J. Bos, T. Declerck, J. Hajič, D. Hershcovich, E. Hovy, A. Koller, S. Krek, S. Schockaert, R. Sennrich, E. Shutova, R. Navigli, What's the meaning of superhuman performance in today's NLU?, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12471–12491. URL: https://aclanthology.org/2023.acl-long.697. doi:10.18653/v1/2023.acl-long.697.

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023. arXiv:2307.03109.

[5] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446. doi:10.18653/v1/W18-5446.

[6] R. Bommasani, K. Davis, C. Cardie, Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4758–4781. URL: https://aclanthology.org/2020.acl-main.431. doi:10.18653/v1/2020.acl-main.431.

[7] K. Stańczak, S. R. Choudhury, T. Pimentel, R. Cotterell, I. Augenstein, Quantifying gender bias towards politicians in cross-lingual language models (2023). URL: https://doi.org/10.1371/journal.pone.0277640.

[8] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186. URL: https://www.science.org/doi/abs/10.1126/science.aal4230. doi:10.1126/science.aal4230.

[9] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 4356–4364.

[10] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences 115 (2018) E3635–E3644. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1720347115. doi:10.1073/pnas.1720347115.

[11] K. Seaborn, S. Chandra, T. Fabre, Transcending the "male code": Implicit masculine biases in nlp contexts, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581017. doi:10.1145/3544548.

3581017.

[12] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016. URL: https://books.google.es/books?id=NgEwCwAAQBAJ.

[13] M. R. Costa-jussà, An analysis of gender bias studies in natural language processing, Nature Machine Intelligence 1 (2019) 495–496. URL: https://doi.org/10.1038/s42256-019-0105-5. doi:10.1038/s42256-019-0105-5.

[14] D. Nozza, F. Bianchi, D. Hovy, HONEST: Measuring hurtful sentence completion in language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2398–2406. URL: https://aclanthology.org/2021.naacl-main.191. doi:10.18653/v1/2021.naacl-main.191.

[15] M. Bartl, M. Nissim, A. Gatt, Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias, in: M. R. Costa-jussà, C. Hardmeier, W. Radford, K. Webster (Eds.), Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1–16. URL: https://aclanthology.org/2020.gebnlp-1.1.

[16] P. Kahn, Rising tide: Gender equality and cultural change around the world, Perspectives on Politics 2 (2004) 407–409. doi:10.1017/S1537592704770978.

[17] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, Science faculty's subtle gender biases favor male students, Proceedings of the National Academy of Sciences 109 (2012) 16474–16479. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1211286109. doi:10.1073/pnas.1211286109.

[18] M. McKinnon, C. O'Connell, Perceptions of stereotypes applied to women who publicly communicate their stem work, Humanities and Social Sciences Communications (2020). doi:10.1057/s41599-020-00654-0.

[19] S. Marjanovic, K. Stańczak, I. Augenstein, Quantifying gender biases towards politicians on reddit, PLOS ONE 17 (2022) 1–36. URL: https://doi.org/10.1371/journal.pone.0274317. doi:10.1371/journal.pone.0274317.

[20] A. H. Bailey, A. Williams, A. Cimpian, Based on billions of words on the internet, people=men, Science Advances 8 (2022) 2463. URL: https://www.science.org/doi/abs/10.1126/sciadv.abm2463. doi:10.1126/sciadv.abm2463.

[21] J. Zou, L. Schiebinger, Ai can be sexist and racist — it's time to make it fair, Nature (2018). doi:10.1038/d41586-018-05707-8.

[22] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1630–1640. URL: https://aclanthology.org/P19-1159. doi:10.18653/v1/P19-1159.

[23] A. Konnikov, N. Denier, Y. Hu, K. D. Hughes, J. Alshehabi Al-Ani, L. Ding, I. Rets, M. Tarafdar, et al., Bias word inventory for work and employment diversity,(in) equality and inclusivity (version 1.0), SocArXiv (2022).

[24] M. E. Heilman, Gender stereotypes and workplace bias, Research in Organizational Behavior 32 (2012) 113–135. URL: https://www.sciencedirect.com/science/article/pii/S0191308512000093. doi:https://doi.org/10.1016/j.riob.2012.11.003.

[25] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, 2024. arXiv:2309.00770.

[26] K. M. White Smolinski, Gender Bias in Natural Gender Language and Grammatical Gender Language within Children's Literature, PhD dissertation, Liberty University, 2024. URL: https://digitalcommons.liberty.edu/doctoral/5294.

[27] D. Cirillo, H. Gonen, E. Santus, A. Valencia, M. R. Costa-jussà, M. Villegas, Sex and gender bias in natural language processing, in: D. Cirillo, S. Catuara-Solarz, E. Guney (Eds.), Sex and Gender Bias in Technology and Artificial Intelligence, Academic Press, 2022, pp. 113–132. URL: https://www.sciencedirect.com/science/article/pii/B9780128213926000091. doi:https://doi.org/10.1016/B978-0-12-821392-6.00009-1.

[28] P. Czarnowska, Y. Vyas, K. Shah, Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics, Transactions of the Association for Computational Linguistics 9 (2021) 1249–1267. URL: https://doi.org/10.1162/tacl_a_00425. doi:10.1162/tacl_a_00425.

[29] P. Zhou, W. Shi, J. Zhao, K.-H. Huang, M. Chen, R. Cotterell, K.-W. Chang, Examining Gender Bias in Languages with Grammatical Gender, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5276–5284. URL: https://aclanthology.org/D19-1531. doi:10.18653/v1/D19-1531.

[30] J. L. Prewitt-Freilino, T. A. Caswell, E. K. Laakso, The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages, Sex Roles (2011). doi:10.1007/s11199-011-0083-5.

[31] M. Kaneko, A. Imankulova, D. Bollegala, N. Okazaki, Gender bias in masked language models for multiple languages, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 2740–2750.

[32] S. Liang, P. Dufter, H. Schütze, Monolingual and multilingual reduction of gender bias in contextualized representations, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5082–5093.

[33] J. Zhao, S. Mukherjee, S. Hosseini, K.-W. Chang, A. H. Awadallah, Gender bias in multilingual embeddings and cross-lingual transfer, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2896–2907.

[34] X. Huang, Easy adaptation to mitigate gender bias in multilingual text classification, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 717–723.

[35] B. L. Whorf, J. M. Carroll, Language, thought, and reality: Selected writings of benjamin lee whorf, The American Journal of Psychology (1957). doi:10.2307/1419256.

[36] S. C. Levinson, Language and cognition: The cognitive consequences of spatial description in guugu yimithirr, Journal of Linguistic Anthropology 7 (1997) 98–131. URL: https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/jlin.1997.7.1.98. doi:https://doi.org/10.1525/jlin.1997.7.1.98.

[37] D. Casasanto, L. Boroditsky, Time in the mind: Using space to think about time, Cognition 106 (2008) 579–593. URL: https://www.sciencedirect.com/science/article/pii/S001002770700087X. doi:https://doi.org/10.1016/j.cognition.2007.03.004.

[38] L. H. Tan, A. H. D. Chan, P. Kay, P.-L. Khong, L. K. C. Yip, K.-K. Luke, Language affects patterns of brain activation associated with perceptual decision, Proceedings of the National Academy of Sciences 105 (2008) 4004–4009. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0800055105. doi:10.1073/pnas.0800055105.

[39] M. R. Banaji, C. D. Hardin, Automatic stereotyping, Psychological Science 7 (1996) 136–141. doi:10.1111/j.1467-9280.1996.tb00346.x.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[43] R. Kramer, The location of gender features in the syntax, Language and linguistics Compass 10 (2016) 661–677.

[44] P. M. Gygax, D. Elmiger, S. Zufferey, A. Garnham, S. Sczesny, L. von Stockhausen, F. Braun, J. Oakhill, A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men, Frontiers in Psychology 10 (2019). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.01604. doi:10.3389/fpsyg.2019.01604.

[45] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485. doi:10.18653/v1/2020.acl-main.485.

[46] D. Lei, Y. Dengdeng, X. Jinhan, G. Wenxing, H. Shenggang, B. Yanchun, J. Bei, Word embeddings via causal inference: Gender bias reducing and semantic information preserving, Proceedings of the Aaai Conference on Artificial Intelligence (2022). doi:10.1609/aaai.v36i11.21443.

[47] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 166–172. URL: https://aclanthology.org/W19-3823. doi:10.18653/v1/W19-3823.

[48] S. Martinková, K. Stanczak, I. Augenstein, Measuring gender bias in West Slavic language models, in: Proceedings of the 9th Workshop on

Slavic Natural Language Processing 2023 (SlavicNLP 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 146–154. URL: https://aclanthology.org/2023.bsnlp-1.17.

[49] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015 IEEE International Conference on Computer Vision (ICCV) (2015) 19–27.

[50] B. Staatsbibliothek, German bert, 2023. URL: https://github.com/dbmdz/berts.

[51] B. Minixhofer, F. Paischer, N. Rekabsaz, WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3992–4006. URL: https://aclanthology.org/2022.naacl-main.293. doi:10.18653/v1/2022.naacl-main.293.

[52] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[53] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[54] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[55] D. Kłeczek, Polbert: Attacking polish nlp tasks with transformers, in: M. Ogrodniczuk, Łukasz Kobyliński (Eds.), Proceedings of the PolEval 2020 Workshop, Institute of Computer Science, Polish Academy of Sciences, 2020.

[56] S. Dadas, M. Perełkiewicz, R. Poświata, Pre-training polish transformer-based language models at scale, in: L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J. M. Zurada (Eds.), Artificial Intelligence and Soft Computing, Springer International Publishing, Cham, 2020, pp. 301–314.

# A. Employment by Sector with Respect to Gender

**Table 3**
Employment by sector with respect to the gender statistics reported for the year 2019 by the World Bank.

| Country | Sector | Gender | Total (%) |
|---|---|---|---|
| Germany | Agriculture | Female | 0.8 |
| | | Male | 1.5 |
| | Industry | Female | 13.9 |
| | | Male | 38.7 |
| | Services | Female | 85.3 |
| | | Male | 59.7 |
| Spain | Agriculture | Female | 2.0 |
| | | Male | 5.7 |
| | Industry | Female | 9.4 |
| | | Male | 29.5 |
| | Services | Female | 88.6 |
| | | Male | 64.8 |
| Poland | Agriculture | Female | 8.1 |
| | | Male | 10.0 |
| | Industry | Female | 17.5 |
| | | Male | 43.9 |
| | Services | Female | 74.4 |
| | | Male | 46.1 |
| United Kingdom | Agriculture | Female | 0.6 |
| | | Male | 1.5 |
| | Industry | Female | 7.7 |
| | | Male | 27.3 |
| | Services | Female | 91.7 |
| | | Male | 71.2 |

# B. Evaluated Pre-trained Large Language Models

**Table 4**
Evaluated pre-trained large language models for four different languages.

| Language | Model class | Model | Training data |
|---|---|---|---|
| English | BERT [40] | Base uncased<br>Base cased<br>Large uncased<br>Large cased | BooksCorpus (0.8B words) [49] and English Wikipedia (2.5B words; excluding lists, tables and headers). Size: 16GB |
| | RoBERTa [41] | Base cased<br>Large cased | BooksCorpus (0.8B words) [49], English Wikipedia (2.5B words; excluding lists, tables and headers), CC-News (September 2016-February 2019), OpenWebText, Stories. Size: 161GB |
| German | BERT [50] | Base uncased<br>Base cased | Wikipedia, EU Bookshop, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. Size: 16GB, Tokens: 2.35B |
| | RoBERTa [51] | Base cased | BooksCorpus (0.8B words) [49], English Wikipedia (2.5B words; excluding lists, tables and headers), CC-News (September 2016-February 2019), OpenWebText, Stories. Size: 161GB |
| Spanish | BERT [52] | Base uncased<br>Base cased | Wikipedia, EU Bookshop, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. Size: 4GB, Tokens: 3B |
| | RoBERTa [53, 54] | Base cased (BERTIN)<br>Base cased (BSC)<br>Large cased (BSC) | Spanish mC4 corpus. Size: 1TB, Tokens: 235B<br>National Library of Spain (Biblioteca Nacional de España) crawls. Size: 570GB, Tokens: 135B |
| Polish | BERT [55] | Base uncased<br>Base cased | Wikipedia, Open Subtitles, ParaCrawl and Polish Parliamentary Corpus. Tokens: 1.9B<br>Wikipedia, Open Subtitles (deduplicated), ParaCrawl and Polish Parliamentary Corpus. Tokens: 0.7B |
| | RoBERTa [56] | Base cased | CommonCrawl, Wikipedia and Polish Parliamentary Corpus. Size: 135GB |