Job Unfair: An Investigation of Gender and Occupational Bias in Free-Form Text Completions by LLMs

Camilla Casula **Sebastiano Vecellio Salto** **Elisa Leonardelli** **Sara Tonelli**
{ccasula, svecelliosalto, eleonardelli, satonelli}@fbk.eu
Fondazione Bruno Kessler, Italy

Abstract

Disentangling how gender and occupations are encoded by LLMs is crucial to identify possible biases and prevent harms, especially given the widespread use of LLMs in sensitive domains such as human resources. In this work, we carry out an in-depth investigation of gender and occupational biases in English and Italian as expressed by 9 different LLMs (both base and instruction-tuned). Specifically, we focus on the analysis of sentence completions when LLMs are prompted with job-related sentences including different gender representations. We carry out a manual analysis of 4,500 generated texts over 4 dimensions that can reflect bias, we propose a novel embedding-based method to investigate biases in generated texts and, finally, we carry out a lexical analysis of the model completions. In our qualitative and quantitative evaluation we show that many facets of social bias remain unaccounted for even in aligned models, and LLMs in general still reflect existing gender biases in both languages. Finally, we find that models still struggle with gender-neutral expressions, especially beyond English.

1 Introduction

The wide adoption of LLMs in domains where they are expected to support or even replace human decision-making underpins the importance of carefully assessing the presence of biases in model outputs. *Social bias*, i.e. discrimination against a social group arising from historical and structural power imbalances (Blodgett et al., 2020), has been the subject of several works in NLP related to fairness, based first on word embeddings (Garg et al., 2018) then on BERT-like models (Jentzsch and Turan, 2022) and more recently on LLMs (Chen et al., 2025). Among the domains in which LLMs have become prevalent, we argue that a particularly critical one is that of occupations. Indeed, if biased LLMs are used in hiring decisions, unjustly evaluating experiences and expertise based on gender, the selection process can become highly unfair, resulting in clear allocational harm.

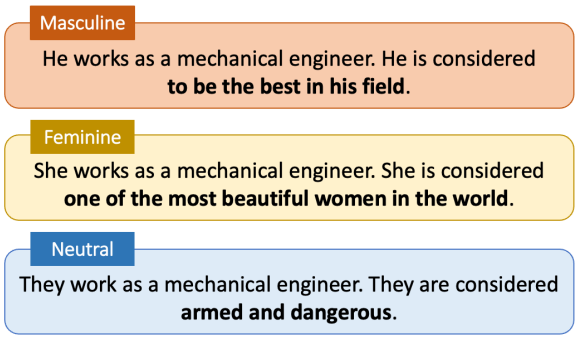

Figure 1: Examples of different model completions (in bold) given the same occupation and different gender expressions for the Llama 3.1 70B model in English.

In this work, we disentangle issues related to bias in LLMs concerning gender and occupations in Italian and English, capturing the differences in generated text when prompted with different gender signals. While occupational and gender bias have been studied before, existing works generally focus on constrained generation, slot filling tasks or minimal pairs (see for example the WinoBias dataset (Zhao et al., 2018a)). In our work, instead, we prompt LLMs to freely complete an input sentence, allowing us to probe for implicit biases we might not see when just using first names or pronouns as proxies for gender as in previous work.

While we know models are post-trained to align with human feedback and to reduce bias, they can still reflect and propagate social biases. Through the creation of a comprehensive dataset of (gendered) occupational titles, we therefore aim at addressing the following research questions: **(RQ1)** Does the alignment process of LLMs effectively reduce bias (or does it merely conceal it)? **(RQ2)** Do models associate certain gender identities with certain occupations, reflecting social biases? **(RQ3)**

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22771–22789
November 4-9, 2025

Are models more brittle to non-conforming gender identities that go beyond the binary norm, further marginalizing them? We explore the above questions comparing text generated in English and Italian, which are very different languages in terms of gender marking.[1]

Overall, the main contributions of this work are: *i)* the creation of a dataset of occupational titles in English and Italian, adaptable to more languages since it is derived from a database with official translations available for many languages (Section 3); *ii)* a novel evaluation approach based on a statistical analysis of vector representations able to capture gender and occupational biases in free-form texts, which could be potentially extended to any kind of bias (Section 4.3.2); *iii)* a thorough qualitative and quantitative evaluation of different linguistic dimensions that may express bias, including a manual annotation of 4,500 texts across two languages and nine LLMs (Sections 4.3.1 and 4.3.3).

2 Related Work

Social bias has been extensively investigated in NLP research over the past years, although the term *bias* is used for a range of phenomena that can be harmful in different ways (Blodgett et al., 2020). Previous work on this topic has focused on social and gender biases in word embeddings and language models (Bolukbasi et al., 2016; Zhao et al., 2018b; Nangia et al., 2020; Bartl and Leavy, 2024), as well as downstream tasks such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), machine translation (Saunders and Byrne, 2020; Savoldi et al., 2021), and more (Bhaskaran and Bhallamudi, 2019; Mishra et al., 2020; Dinan et al., 2020). While some are focused on analyzing and detailing existing biases in models (Rudinger et al., 2018; Basta et al., 2019; Kotek et al., 2023; Wan and Chang, 2024), others aim at debiasing them (Bolukbasi et al., 2016; Lauscher et al., 2021; Gorti et al., 2024), although many of these methods tend to cover-up existing biases instead of removing them (Gonen and Goldberg, 2019).

More specifically, occupational bias and its intersection with gender bias has been explored in previous work, mostly with relation to biases in embeddings (De-Arteaga et al., 2019; Wilson and Caliskan, 2024; An et al., 2025). In general, an overwhelming majority of work on this topic has focused on English and more specifically the United States (Rudinger et al., 2018; Nangia et al., 2020; Chen et al., 2025, e.g.), with only a handful of exceptions (Nomelini and Marcolin, 2024; Rankwiler and Kurpicz-Briki, 2024; Kaukonen et al., 2025).

In addition to this, the focus of existing work is placed largely on job advertisements and applications (Frissen et al., 2023; Ding et al., 2024) rather than on the potential biases related to occupational titles *per se*. Furthermore, a rather large portion of previous work often uses first names as a proxy to encode gender and other social and demographic characteristics (De-Arteaga et al., 2019; Mishra et al., 2020; Döll et al., 2024; An et al., 2025), which has been found to present both validity and ethical issues (Gautam et al., 2024). This kind of work is also aimed chiefly at uncovering biases in associations between occupations and gender, especially in relation to specific downstream tasks, rather than at analyzing the kind of language produced by LLMs, which is the goal of our work.

Finally, many previous works dealing with gender bias are predominantly focused on binary gender representations, while non-binary representations of gender have received much less attention in the past (Lauscher et al., 2022; Ovalle et al., 2023). We aim at addressing this gap by considering neutral gender representations in our experiments for both English and Italian, two languages that encode gender differently. In fact, while English is a notional gender language, in which gender is expressed mostly through lexically gendered forms, personal pronouns, and possessive adjectives, Italian is a grammatical gender language, in which gender is encoded through inflection in multiple parts of speech (Stahlberg et al., 2007; Savoldi et al., 2021).

3 A Database of Occupational Titles

In order to investigate occupational and gender biases, we rely on a database of occupational titles, which we then use in our prompts to models. This database is based on the 2008 International Standard Classification of Occupations (ISCO-08), the current standard employed by the International Labour Organization (ILO, 2012). The occupational titles in ISCO are in English, and catalogued at 4 different levels of specificity. An example of the classification of an occupational title in ISCO is shown in Table 1. Its four layers contain 10, 43, 130, and 436 occupational titles each (from most

[1]The data for this paper is available at https://github.com/dhfbk/job-unfair.

ISCO Code	Occupational Term
2	*Professional*
26	*Legal, social and cultural professional*
261	*Legal professional*
2611	*Lawyer*

Table 1: Example of ISCO classification for *lawyer*.

1	Managers
2	Professionals
3	Technicians and Associate Professionals
4	Clerical Support Workers
5	Service and Sales Workers
6	Skilled Agricultural, Forestry and Fishery Workers
7	Craft and Related Trades Workers
8	Plant and Machine Operators, and Assemblers
9	Elementary Occupations
0	Armed Forces Occupations

Table 2: ISCO Major Groups at the I digit level.

generic to most specific). Table 2 shows the 10 major ISCO groups at the I digit level. Each major group can contain a variable number of sub-groups. For instance, the major group *managers* contains 4 sub-groups, while the major group *elementary occupations* includes 6 sub-groups.[2]

For Italian, on the other hand, we base identity terms on the Italian classification of occupations (ISTAT, 2023, CP2021), which provides ISCO codes for the occupations.[3] Indeed, many national statistics institutes provide data that includes ISCO codes, potentially enabling extensions of our approach to other languages and cultural contexts.[4]

From the available occupational titles, we craft for each language a series of identity terms related to occupations. This process results in a database that contains identity terms for 436 occupations in English and 813 in Italian.[5]

In addition to this, given that Italian is a gendered language, we include each occupational title in its masculine, feminine, and gender-neutral forms to ensure inclusive representation (Stahlberg et al., 2007; Devinney et al., 2022). For example, the English *lawyer* corresponds to *avvocato* (m.), *avvocata* (f.), and *avvocat** (neutral) in Italian. This approach results in 2,290 identity terms in the Italian section of our database.[6]

4 Evaluation Framework

While most previous work on this topic focuses on associations between gender and occupation in controlled contexts, we investigate biases in language related to occupation and gender produced by models in free-form text completion. We propose an evaluation framework comprising two main components: first, the creation of a comprehensive set of hand-crafted prompts for each language and model type, ensuring that we provide gender information within the prompt. (Sec. 4.2). Second, an in-depth linguistic analysis of the text completions produced by different LLMs (Sec. 4.3), consisting of: *(i)* a manual analysis of the texts, accounting for different facets of biased language (Sec. 4.3.1), *(ii)* a statistical analysis of vector representations of the model completions (Sec. 4.3.2), and *(iii)* a lexical analysis of the tokens most associated with each gender identity (Sec. 4.3.3).

4.1 Models

We analyze text generated by nine freely available generative language models, covering both base and instruction-tuned variants where available. The models include: `Aya-Expanse 8B`, `Gemma-7B`, `Gemma-7B-it`, `LLaMA 3.1 8B`, `LLaMA 3.1 8B Instruct`, `LLaMA 3.1 70B`, `LLaMA 3.1 70B Instruct`, `Mistral 7B v0.3`, and `Mistral 7B Instruct v0.3`.

While some of these models are crafted to be inherently multilingual, all of them can produce texts in languages other than English. We are particularly interested in analyzing differences in model biases across the two languages in our study.

4.2 Prompts

Since LLMs have been found to be brittle to different prompt choices (Zhao et al., 2021; Zheng et al., 2024), we prompt each model multiple times, using different prompts, for each language.

Italian As a gendered language, Italian encodes grammatical gender in the majority of occupational titles. Gender is also encoded in plural forms of nouns that are otherwise gender-neutral when singular, for instance *l'agente* (*the agent*) which becomes *le agenti (f.)* or *gli agenti (m.)*. We thus

[2]Details about ISCO and its complete version are available on the ILO website: ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/.

[3]https://www.istat.it/classificazione/classificazione-delle-professioni/

[4]E.g., https://data.un.org/

[5]This discrepancy is mostly due to size differences between ISCO-08 and CP2021.

[6]For a discussion on gender-neutral terminology in Italian, see Appendix A.

prompt models using both singular and plural identity terms and tenses. We select 4 verbs to use in our prompts: *essere (to be), essere considerat* (to be considered), fare (to do)* and *aspirare a (to aspire to)*. These verbs are selected to elicit a completion pertaining to an individual's identity, actions and desires. Below is an example of gendered prompt in Italian:

la farmacista è considerata
the pharmacist (f.) is considered

English As English does not encode grammatical gender through inflection as Italian does, the strategy we adopt for English is that of conditioning models to generate gendered representations of occupations by using different personal pronouns (*he, she, they* respectively for masculine, feminine, and neutral gender expression). We then select 3 verbs to use in our model prompts: *to be, to be considered,* and *to aspire to*.[7] An example prompt is shown below:

they are a kitchen helper and they are

Each prompt is run 3 times per model to account for stochasticity. The total amount of prompts is 4,062 for English and 18,664 for Italian. Implementation details are reported in Appendix B.

4.3 Evaluation

We structure our analysis across three axes: a human assessment, in which we manually annotate 4,500 texts generated by models from the prompts described in Section 4.2; an automated analysis to capture biases in gender and occupations based on vector clustering; and finally a lexical analysis, in which we inspect with computational methods the lexicon generated by the different LLMs.

4.3.1 Human Analysis

We ask two annotators to label generated texts according to 4 dimensions: gender assignment, subject misinterpretation, the topic discussed in the text and, finally, whether the lexical focus of the text expresses *agency* or *communion*.[8] The annotators, one male and one female, are both either native or fluent in Italian and English.

The 4,500 texts we manually evaluate are selected using stratified sampling, so that for each model and gender representation we extract the same amount of examples for annotation. We also sample equally by number for Italian.[9]

Gender assignment We analyze how often the models misgender the subject of the prompt. For example, the completion (underlined) 'she is a concrete placer and she aspires to be a foreman' indicates misgendering, as the correct term for a female foreman is *forewoman*. For this dimension, the annotators could label the gender assignment as either *correct* or *changed*, specifying the gender that the model incorrectly attributed to the subject.

Subject misinterpretation In some cases, occupation-related identity terms are misinterpreted by the models as typos or names for other kinds of entities. For example, given the prompt 'They work as a mixed crop and animal producer and they are considered', a model might continue with 'a diversified farm', interpreting *producer* as referring to an agricultural enterprise (an *abstract entity*) rather than a worker.[10] Possible choices for this dimension are *correct*, *profession as a whole* (e.g. 'They are a nurse and they are an important profession'), *physical object*, and *abstract entity*.

Topic Since the prompts are intentionally broad, the free-form completions can range across a variety of topics. Our aim is to capture how topics in the completions relate to occupations or gender identities. For example, in the text 'she is a motorcycle driver and she is a single mother of two children' the focus shifts from the subject's occupation to personal details, using a narrative tone. After a preliminary round of annotation, the possible choices for this dimension were defined as: *occupation related: current occupation*, *occupation related: different occupation*, *person: identity*, *person: appearance*, *storytelling*, and *unrelated*. *Storytelling* refers to mentions of past actions or life events about the subject.

Agency vs communion Inspired by work in social psychology, we analyze the lexicon in texts produced by LLMs along the dimensions of *agency* and *communion* (Abele and Wojciszke, 2018; Sap

[7] *To do* was excluded in the English set of prompts, as it is an auxiliary verb and it showed in preliminary experiments to be a potentially confounding element, in contrast with the italian *essere* (*to be*), which did not result in noisy completions in spite of the verb being an auxiliary as well.

[8] The manually annotated data is available in full at https://github.com/dhfbk/job-unfair.

[9] As the number of occupational titles is quite large, we do not stratify by occupational title, and instead use random selection for job titles in each model-gender-number group.

[10] While this ambiguity is relatively less common in English texts, it is far more frequent in Italian, in part due to agent nominalization using overlapping suffixes to instrument nominalization, especially in the feminine form (Lo Duca, 2011; Thornton, 2015, 2018).

et al., 2017; Wan and Chang, 2024). While agentic lexicon typically emphasizes competence and assertiveness (e.g., 'He is a very ambitious man who is always looking for new opportunities to make money'), communal language suggests warmth, morality, and 'getting along' within a group (e.g., 'She is very responsible and reliable. She is very honest and trustworthy'). *Agency* and *communion* are only annotated where deemed relevant by the annotators, and they are mutually exclusive.

4.3.2 Embeddings Analysis

To investigate semantic patterns across gender and occupation we apply a clustering framework inspired by Gonen and Goldberg (2019), who use clustering to show biases in semantically related words. In order to quantify bias, we use a non parametric statistical testing method based on bootstrapping, similarly to what is typically done in the neuroimaging field to find patterns in high-dimensional data (Maris and Oostenveld, 2007).

Our methodology diverges substantially from other existing bias quantification techniques such as WEAT (Caliskan et al., 2017) or SEAT (May et al., 2019), which are widely used but have been criticized as having limited interpretability and susceptibility to statistical artifacts (Schröder et al., 2021). Furthermore, WEAT and SEAT are primarily designed for tasks in which models associate predefined targets with specific attributes, rather than for free-form text. Applying them in our setting would therefore be challenging, requiring constrained choices and the construction of *ad-hoc* tests. In contrast, our approach directly analyzes free-form completions, enabling the detection of broader and more robust patterns of bias and allowing flexible application across different contexts.

Encoding In order to analyze patterns in model completions with regards to both gender and occupations, we devise an approach based on embedding vector clustering. We first represent sentences as vectors using the Multilingual E5 Text Embeddings (Wang et al., 2024).[11]

For each model-generated text, we encode two embeddings: one including the entire text (prompt + completion), and one including only the model completion, with no prompt. Ideally, this allows us to account for potential confounding factors due to the gendered nature of the prompts, since we are more interested in finding potential biases in the completions. Furthermore, this should minimize the influence of potential gender biases in the embedding model itself.

Analysis Considering feminine and masculine gender representations to reflect the two ends of the gender spectrum, we consider feminine and masculine prompts to investigate model biases with regards to the combination of gender and occupation at the II level of ISCO. For each set of sentences generated by each model, separately for English (8,136 per model) and Italian (37,328 per model), we apply a bootstrapping method, in which we construct 1,000 balanced subsets of 492 sentence embeddings, by uniformly sampling across the two considered genders and 41 occupations.[12]

On each subset, we perform hierarchical clustering with the number of clusters fixed at 41 to match the number of occupational titles. We aggregate results across runs and build a co-clustering matrix that reflects how consistently completions from the same occupation-gender group were assigned to the same cluster. The rationale behind this is that if model completions are semantically similar given the prompted occupations, the 41 clusters should be based on occupations. Conversely, if sentences for the same occupation but different genders frequently fall into different clusters, this indicates that the model generates systematically different completions, possibly depending on gender.

Repeating the analysis through bootstrapping is crucial to minimize random variation noise and produce reliable results. To assess whether these observed patterns reflect more than just chance structure, we therefore repeat the same analysis by randomly shuffling cluster labels, calculating the co-clustering matrix on the shuffled data. This shuffled control condition allows us to control for combinatorial artifacts of the clustering algorithm. Moreover, it enables us to evaluate the statistical significance of our findings. Detailed sampling and clustering procedures, as well as statistical controls, are provided in Appendix E.

4.3.3 Lexical Analysis

To analyze the different kind of lexicon used by LLMs when discussing occupation and gender, we extract the most informative tokens for each gender with the VARIATIONIST Python library (Ram-

[11] The E5 text embeddings are, as of 2025, among the best-performing on the Massive Text Embedding Benchmark: https://huggingface.co/spaces/mteb/leaderboard.

[12] In this case, the total of occupations for the II ISCO level should be 43, but two of those had gender-neutral collective referrants such as *crew* and were discarded for this analysis.

poni et al., 2024). For this, we use the built-in normalized positive weighted relevance metric (`npw_relevance`). This metric first calculates the pointwise mutual information (PMI) between tokens and labels (in our case, the label indicates the gender representation), and then weights the score of each token by multiplying it by its frequency in the dataset, converting negative values to 0. Finally, it normalizes these scores between 0 and 1.

Given that we aim to compare texts for the 3 gender representations (masculine, feminine, and neutral) and we are interested in tokens whose relevance changes the most across them, we calculate a measure of how much `npw_relevance` varies across gender, which we name **relevance shift** (RS). We define the Relevance Shift for token t and gender $g \in \mathcal{G} = \{\mathrm{m}, \mathrm{f}, \mathrm{n}\}$ as

$$\mathrm{RS}(t, g) = \mathrm{rel}(t, g) - \sum_{g' \in \mathcal{G} \setminus \{g\}} \mathrm{rel}(t, g')$$

in which *rel* in our case corresponds to `npw_relevance`. Put simply, to obtain the relevance shift of a token-gender pair, we subtract the relevance of that token for the other two gender representations from the relevance of the token for the gender under consideration. We then select the top-$k = 10$ tokens for each model and gender based on their relevance shift scores, to identify potential lexical biases or spurious correlations.

5 Results and Discussion

5.1 Human Analysis

Below we present the main findings of our human analysis, divided by dimensions as presented in Sec. 4.3.1. Inter-annotator agreement was calculated on 200 examples for the subjective annotation dimensions, lexical focus and topic, as gender and subject assignment were found to be sources of complete agreement. The annotators had a Krippendorff's alpha of 0.853 for the agency-communion annotation, while topics (being multi-label) had a Jaccard coefficient of 0.743. Both these measures indicate high agreement between the annotators.

Gender assignment The first and most striking result we can observe from our gender assignment analysis is that no instances of misgendering masculine prompts occurred in either language across all models, while the instances of misgendering (0.33% of all completions for English and 36.6% in Italian) were either feminine or neutral. More strikingly, the large majority of all misgendered identities are interpreted as masculine, showing that across models there is a preference for masculine representations, as shown in Figure 2 for Italian. The complete misgendering tables are reported in Appendix D.

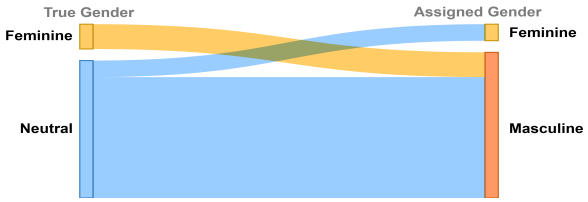

Figure 2: Distribution of misgendering in Italian across all models.

Subject misinterpretation Subject misinterpretations are overall somewhat rare in our English annotations, while they are more frequent in Italian. We show the average percentage of misinterpretation across all LLMs in Table 3. These results show the difficulties of LLMs in handling neutral gender representations (Ovalle et al., 2023), an occurrence which can lead to erasure of non-conforming gender identities (Dev et al., 2021; Dhoest, 2015). Feminine representations of gender for Italian are also affected by this phenomenon, especially with regards to misinterpretation as abstract entities. This is likely influenced by the greater presence of masculine occupational titles in Italian, while some feminine titles are not as common, in spite of institutional efforts (Sabatini, 1987). Furthermore, the discrepancy between the two languages shows that even for a medium-resourced language such as Italian we have considerable decreases in subject interpretation compared to English.

Subject	**Italian**			**English**		
	M(%)	**F(%)**	**N(%)**	**M(%)**	**F(%)**	**N(%)**
Abstract	1.9%	12.3%	10.6%	0.0%	0.0%	2.6%
Object	3.3%	6.7%	6.0%	0.0%	0.1%	1.5%
Profession	2.8%	8.9%	10.0%	1.0%	0.5%	1.5%
Average	**2.7%**	**9.3%**	**8.9%**	**0.3%**	**0.2%**	**1.8%**

Table 3: Comparison of average subject misinterpretation percentages across all models by gender.

Topic In Table 4 we report the average topic distributions across models for English, since the main findings are valid across most of the models we test. The statistics on a per-model basis for both languages are reported in Appendix D.

In contrast to the tendency we found in the *subject misinterpretation* analysis, in the case of topics models seem to steer away from the current occupation more frequently in English than in Ital-

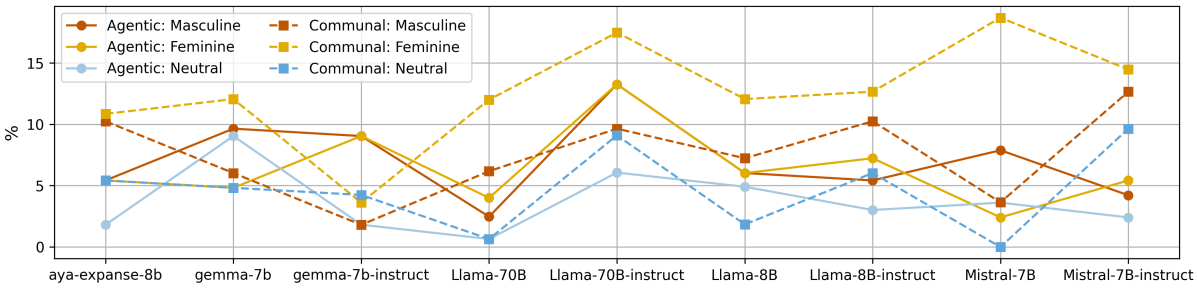

Figure 3: Percentage of agency and communion-focused texts in English, averaged across all occupations.

ian, especially with regards to the *person: identity* topic and *storytelling*. While instruction-tuned models tend to stay on topic more frequently as a general rule, we still observe an average of over 20% of completions to reflect the inclination of models to provide details about the subject's private life (e.g., 'She is a finance manager and she is married to a doctor. They have two children.'). This tends to occur more frequently with feminine gender representations over masculine ones, and not as frequently with neutral gender representations, possibly due to an interpretation of gender-neutral sentences as being more formal.

Notably, almost all cases of topics steering towards physical appearance for both languages are cases in which the gender representation was feminine, showing that body-related stereotypes about feminine subjects can persist even in aligned models (in particular, we observe this in `aya expanse 8b` and `gemma 7b instruct`).

Topic	M(%)	F(%)	N(%)
Different Occupation	15%	16%	10%
Person: Identity	21%	23%	9%
Storytelling	18%	16%	9%
Appearance	1%	**4%**	0%
Unrelated	0%	0%	2%

Table 4: Average topic distribution by gender for English, averaged across all models.

Agency vs communion The average amount of agentic and communal text completions by gender representation in English is presented in Figure 3. Communal lexicon is dominant for feminine gender representations across almost all models. Also, there seems to be a trend in aligned models to increase agency for feminine representations, perhaps as a way to contrast the known biased associations of feminine figures with communal lexicon and masculine figures with agentic language. However, this process appears to be prone to side-effects: along with agency, communion also increases in most cases, especially for neutral gender representations. Finally, neutral subjects seem to be presented in the least agentic light, potentially showing that existing efforts towards making models less biased towards women do not consider biases towards gender representations that go beyond the masculine-feminine binary.

The full Italian statistics are reported in Appendix D. The overall trends are different from English, showing a *decrease* in agency for feminine and neutral gender in instruct models compared to base models. Our hypothesis is that alignment is heavily focused on English, potentially resulting in undesired side effects on other languages.

5.2 Embedding Analysis

As discussed in Sec. 4.3.2, this analysis is conducted on the masculine-feminine axis and the II level of ISCO-08.

Full-Sentence vs No-Prompt Embeddings The co-occurrence matrix obtained from the full sentence embeddings (i.e., including the prompt) shows nearly perfectly clustered sentences according to occupation. In contrast, embeddings of completions without the prompt only cluster with the same occupations 55% of the time. This indicates that when removing the prompt, completions may cluster more based on gender, suggesting the potential presence of biases. For all subsequent analyses described in this section, we used only the "no-prompt" embeddings since we aim at evaluating biases in the generated completions.

Co-Clustering Gender-Occupation Matrix We compute a p-value for each occupation–gender pair in the co-clustering matrix, showing how frequently texts with the same gender representation and job title are clustered together. Figure 4 shows the average p-values across all models and languages: the top-left quadrant corresponds to mascu-

line–masculine (MM) co-clustering, the bottom-right to feminine–feminine (FF), while the other two quadrants represent cross-gender co-clustering (FM, MF). Within each quadrant, the diagonals capture same-occupation clustering, while off-diagonal entries capture co-clustering of different occupations.

Emerging patterns Given the large number of comparisons, we do not interpret individual occupation–gender pairs, instead we focus on the emerging broad patterns and quantify them through effect size measured with *Cohen's d* (Lakens, 2013).

All the diagonals of the four quadrants of figure 4, i.e. same-job co-clustering, show low p-values. The same-gender diagonals show medium-large effects sizes (MM: $d = 0.55 \pm 0.06$; FF: $d = 0.56 \pm 0.07$), while the cross-gender diagonals display a slightly lower, though still substantial, effect (MF/FM: $d = 0.50 \pm 0.013$), indicating a certain extent of semantic similarity for the description of the same occupations, higher within the same genders.

Interestingly, off-diagonal regions of the matrix, i.e. different-job co-clustering, show different behaviors between cross- and same-gender quadrants. In the same-gender quadrants, off-diagonal values tend to be lower and show pronounced "blue squares" along the diagonals, i.e., areas of low p-values that indicate stronger-than-chance co-clustering. These square patterns derive from the structure of the ISCO framework, where similar professions are grouped in close proximity and macro-categories (ISCO level I) are hierarchically organized from high- to low-skill occupations. As a result, related professions naturally cluster together, while it is possible to note how higher-skill macro categories (ISCO 1–5) do cluster less with lower-skill categories (ISCO 6–9). Nevertheless, the average off-diagonal effect sizes for same-gender comparisons remain close to zero (MM: $d = 0.02 \pm 0.15$; FF: $d = 0.03 \pm 0.16$), reflecting the fact that localized clustering effects cancel out when aggregated across the full quadrant. In the cross-gender quadrants, the "blue squares" are less pronounced, and regions spanning different skill levels exhibit very high p-values. Correspondingly, the off-diagonal averages show small negative effect sizes (MF: $d = 0.26 \pm 0.20$; FM: $d = 0.26 \pm 0.18$), indicating that different genders and different occupations tend to be less semantically similar than expected by chance.

In sum, off-diagonal patterns indicate that within genders, similar professions are described in semantically similar ways, consistently with the ISCO structure. Across genders, however, this coherence largely disappears, and even falls below chance, indicating that the way professions are described is shaped largely by gender.

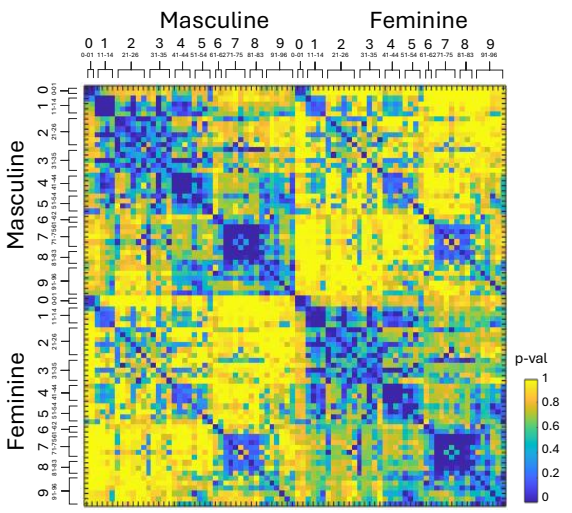

Figure 4: Top: Co-clustering gender-occupation matrix (mean p-values across models). Rows and columns represent masculine and feminine prompt completions, ordered by category (II level of ISCO-08). The four quadrants reflect within or across gender co-cluster. Values along the diagonals of each quadrant show same-occupation clustering, off-diagonal ones indicate cross-occupation associations.

Instruct vs base models When analyzing separately the co-clustering matrix for instructed and base models, we find that the proportion of data clustered within the same profession is significantly lower for the base models and that the pattern described in the previous section is attenuated in the instructed models but amplified in the base ones. Additional details are provided in Appendix E.2.

5.3 Lexical Analysis

Table 5 shows the top-5 tokens by relevance shift across different models for each gender representation in English, which summarize the main takeaways of this analysis.[13] Remarkably, the word *woman* appears among the first ones for all models and in both languages. The absence of the word *man*, but the presence of the word *person* in the masculine counterparts of the most relevant tokens for each gender suggests that all models are sub-

[13]The full top-10 tokens tables for both languages are reported in Appendix F.

G	aya-exp-8b	gemma-7b	gemma-7b-it	Llama-70B	Llama-70B-in	Llama-8B	Llama-8B-in	Mistral-7B-In	Mistral-7B
F	*woman*	*woman*	*field*	*mother*	*woman*	*woman*	*woman*	*dedicated*	*woman*
	beautiful	*mother*	*passionate*	*woman*	*dedication*	*mother*	*female*	*invaluable*	*mother*
	female	*beautiful*	*efficient*	*child*	*pioneer*	*beautiful*	*respected*	*asset*	*beautiful*
	women	*daughter*	*strong*	*single*	*craftswoman*	*girl*	*male*	*dedication*	*wife*
	powerful	*wife*	*woman*	*women*	*dominate*	*child*	*dominate*	*enjoy*	*daughter*
M	*hero*	*friend*	*meticulous*	*person*	*experienced*	*person*	*verse*	*craftsman*	*father*
	party	*person*	*master*	*father*	*reputation*	*friend*	*extensive*	*current*	*person*
	fan	*father*	*skilled*	*hard*	*time*	*worker*	*experienced*	*lie*	*national*
	figure	*protagonist*	*craftsman*	*party*	*deep*	*union*	*final*	*collar*	*husband*
	footballer	*character*	*hard*	*university*	*entrepreneur*	*association*	*background*	*blue*	*handsome*
N	*essential*	*responsible*	*essential*	*responsible*	*tradespeople*	*responsible*	*personnel*	*play*	*responsible*
	pandemic	*business*	*worker*	*industry*	*responsible*	*maintenance*	*essential*	*essential*	*industry*
	covid	*provide*	*build*	*team*	*crucial*	*industry*	*tradespeople*	*crucial*	*nurse*
	19	*industry*	*provide*	*backbone*	*play*	*team*	*require*	*aim*	*people*
	worker	*team*	*operation*	*maintenance*	*require*	*business*	*laborer*	*professionals*	*team*

Table 5: Top-5 tokens by relevance shift across gender representations and models in English.

ject to the *male as norm* bias, even aligned and supposedly less biased models.

Furthermore, for English, the most relevant tokens for the neutral gender representations include tokens that appear more on-topic within the domain of occupations. As mentioned in the topic analysis, this might be an effect of the prevalence of neutral gender representation in corporate or formal settings. Conversely, for Italian the most relevant tokens for neutral gender representations include a large amount of lexical artifacts and incomplete words for some models, again showing that LLMs struggle with gender-neutral expressions in Italian.

We can observe some positive impact (in terms of seemingly reduced bias) of the alignment process across models. For instance, feminine family roles, such as *mother*, tend to disappear from the top-10 most informative tokens in `gemma-7b-it`, both `Llama instruct` models, and `Mistral-7B instruct` compared to the same models' base versions, showing that in their instructed versions the models appear to use more topic-appropriate lexicon rather than defaulting to traditional gender roles. Furthermore, with the exception of the `gemma` models, instruct models tend to more reliably recognize gender neutral expressions in Italian, although difficulties can still be observed, especially for the `gemma-7b-it` model.

However, this also results in unexpected completions that may be the result of 'over-alignment', especially in Italian. For instance, words referring to risk and dangerous tasks, such as the words *morto* (*dead*) or *preicolosamente* [sic] (*dangerously*), are in the top-10 tokens for masculine gender representations across different models, while terms related to prestigious occupations, such as *ceo* and *scienziata* (*scientist* (f.)), are associated mainly to female gender representations. This occurs less frequently with English prompts, suggesting that instruction post-training for most models does appear to reduce gender bias in English contexts to some extent, although this may not be the case for other languages.

6 Conclusions

In this work, we have proposed an in-depth analysis of model biases related to gender and occupation across 9 widely used LLMs, both base and instruction-tuned, focusing on free-form text completions. We created a comprehensive dataset of occupational titles and prompts in Italian and English and proposed a novel evaluation framework to inspect model biases in generated texts.

In our analysis, comprising a manual assessment, a statistical embeddings analysis and an automatic lexical analysis, we found that, while model alignment does reduce the impact of some biases, it also has counterintuitive effects (RQ1). Furthermore, models heavily rely on associations between genders and occupations, reflecting and potentially propagating social biases in both languages (RQ2). Finally, neutral gender representations appear to be misgendered and misinterpreted more often, especially for Italian, and to be presented as less agentic than other gender identities (RQ3). These findings highlight that much effort has yet to be carried out to make sure models do not propagate existing social biases, especially for contexts beyond English.

Limitations

The current work presents some limitations. Although we consider also gender neutral forms aside

from binary ones, there are other gender identities that we do not take into account, such as those expressed through neopronouns.

We take into consideration a large amount of dimensions but due to space limitations, they have not been thoroughly discussed in this work, such as the interplay between gender bias and occupational bias at the most granular ISCO level. The full dataset is available on GitHub for further exploration.[14]

Additionally, we include in this work two languages that show different types of gender marking and therefore offer two different perspectives on the expression of gender and occupational bias. However, more languages could be added, especially low-resourced ones. We consider our analysis a first step and we hope others will be interested in carrying out similar analyses for other linguistic and cultural contexts.

The embeddings analysis is based only on one embedding model, which could influence the clustering results. Nevertheless, we think that the specific design of our approach, calculating embeddings both with and without the prompt, helps mitigate this effect. Indeed in the main analysis, by focusing only on the generated completions, we do not include the portions of the sentence that are more susceptible to bias, i.e., the gendered prompts. Moreover, because our analysis spans a wide range of occupational prompts and completions across multiple models, the trends we observe are robust at the aggregate level. While the embedding model may still influence individual representations, we consider it unlikely to account for the systematic clustering patterns we report. An analysis of each specific model or of point-wise biases is beyond the scope of this manuscript and left for further investigations.

Ethical Statement

The goal of the current work is to disentangle biases in LLMs related to gender and occupations. We do not use any personal data or perform any jailbreaking in our model evaluation. On the contrary, this work goes in the direction of using LLMs in a more ethical way, with the goal of minimising the risk of harm and discrimination.

[14] https://github.com/dhfbk/job-unfair

Acknowledgments

This work has been supported by the PNRR project FAIR – Future AI Research (PE00000013), under the NRRP MUR program funded by NextGeneration EU.

References

A. E. Abele and B. Wojciszke. 2007. Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, 93(5):751–763.

Andrea Abele and Bogdan Wojciszke, editors. 2018. *Agency and Communion in Social Psychology*, 1st edition. Routledge.

Cohere For AI and Cohere. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Haozhe An, Connor Baumler, Abhilasha Sancheti, and Rachel Rudinger. 2025. On the Mutual Influence of Gender and Occupation in LLM Representations. *arXiv preprint*. ArXiv:2503.06792 [cs].

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

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2025. Causally Testing Gender Bias in LLMs: A Case Study on Occupational Bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4984–5004, Albuquerque, New Mexico. Association for Computational Linguistics.

Gloria Comandini. 2021. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, 23:43–64.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 120–128, New York, NY, USA. Association for Computing Machinery.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Alexander Dhoest. 2015. Audiences out of the box: Diasporic sexual minorities viewing representations of sexual diversity. *European Journal of Cultural Studies*. Publisher: SAGE PublicationsSage UK: London, England.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Lei Ding, Yang Hu, Nicole Denier, Enze Shi, Junxi Zhang, Qirui Hu, Karen D. Hughes, Linglong Kong, and Bei Jiang. 2024. Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach. *Advances in Neural Information Processing Systems*, 37:139912–139937.

Michael Döll, Markus Döhring, and Andreas Müller. 2024. Evaluating Gender Bias in Large Language Models. *arXiv preprint*. ArXiv:2411.09826 [cs].

Richard Frissen, Kolawole John Adebayo, and Rohan Nanda. 2023. A machine learning approach to recognize bias and discrimination in job advertisements. *AI & SOCIETY*, 38(2):1025–1038.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team and Google DeepMind. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint*, arXiv:2403.08295.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Atmika Gorti, Aman Chadha, and Manas Gaur. 2024. Unboxing Occupational Bias: Debiasing LLMs with U.S. Labor Data. *Proceedings of the AAAI Symposium Series*, 4(1):48–55. Number: 1.

ILO. 2012. *International Standard Classification of Occupations: Structure, Group Definitions and Correspondence Tables*. International Labour Organization, Geneva.

ISTAT. 2023. Classificazione delle professioni. `https://www.istat.it/classificazione/classificazione-delle-professioni/`.

Sophie Jentzsch and Cigdem Turan. 2022. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Elisabeth Kaukonen, Ahmed Sabir, and Rajesh Sharma. 2025. How Aunt-Like Are You? Exploring Gender Bias in the Genderless Estonian Language: A Case Study. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 296–301, Tallinn, Estonia. University of Tartu Library.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24. ArXiv:2308.14921 [cs].

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:863.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Llama Team and AI @ Meta. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Maria G. Lo Duca. 2011. Strumento, nomi di. Enciclopedia Treccani.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, 164(1):177–190.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. *arXiv preprint*. ArXiv:2008.03415 [cs].

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Guilherme Guimarães Nomelini and Carla Bonato Marcolin. 2024. Gender bias in large language models: A job postings analysis. *RAM. Revista de Administração Mackenzie*, 25:eRAMD240056. Publisher: Editora Mackenzie; Universidade Presbiteriana Mackenzie.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Alan Ramponi, Camilla Casula, and Stefano Menini. 2024. Variationist: Exploring Multifaceted Variation and Bias in Written Language Data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 346–354, Bangkok, Thailand. Association for Computational Linguistics.

Leander Rankwiler and Mascha Kurpicz-Briki. 2024. Evaluating Labor Market Biases Reflected in German Word Embeddings. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 134–143, Chur, Switzerland. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Alma Sabatini. 1987. *Il sessismo nella lingua italiana: Raccomandazioni per un uso non sessista della lingua italiana*. Presidenza del Consiglio dei Ministri, Dipartimento per l'informazione e l'editoria, Roma. Commissione nazionale per la realizzazione della parità tra uomo e donna.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

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

Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. 2021. Evaluating metrics for bias in word embeddings. *arXiv preprint arXiv:2111.07864*.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. In *Social communication*, Frontiers of social psychology, pages 163–187. Psychology Press, New York, NY, US.

Anna M. Thornton. 2015. L'assegnazione del genere in italiano. In Fernando Sánchez Miret, editor, *Volume I Discursos inaugurales – Conferencias plenarias – Sección 1: Fonética y fonología – Sección 2: Morfología – Índices: Índice de autores, Índice general*, pages 467–482. Max Niemeyer Verlag.

Anna M. Thornton. 2018. Vengo a prenderti con il mio Ferrari… O con la mia Ferrari? *Accademia della Crusca*.

Yixin Wan and Kai-Wei Chang. 2024. White Men Lead, Black Women Help? Benchmarking Language Agency Social Biases in LLMs. *arXiv preprint*. ArXiv:2404.10508 [cs].

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *Preprint*, arXiv:2402.05672.

Kyra Wilson and Aylin Caliskan. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1578–1590. Number: 1.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations*.

Appendix

A Gender Neutrality in Italian

While English speakers can count on widespread gender-neutral linguistic devices, such as the singular *they*, Italian speakers cannot rely on widely accepted gender neutralization devices. Examples of resources that are commonly used to express gender in a neutral way are epicene words (which avoid gender specifications), changes in verb forms that avoid specifying gender, and omission of gendered specifiers (Piergentili et al., 2023). Strategies that make use of neomorphemes and graphemic devices are, on the other hand, less widely accepted, and more common in informal settings (Comandini, 2021). However, for the experiments in this paper, we aim at altering the original job titles as little as possible, as the single lexical units (which are gendered in Italian) are at the base of the analysis of different occupations.

We prefer innate gender-neutral forms where available in Italian. When they are not available, they are either constructed with umbrella terms such as *person*, or produced using a star symbol (*) as a gender-signalling graphemic device, which is frequent in online uses of Italian (Comandini, 2021). This approach to gender representation is not intended as an exhaustive representation of binary and non-binary identities and of gender neutralization devices in Italian, being mostly a device that allows us to express neutral gender representation while sill conveying job title information.

B Implementation Details

The models we use to generate free-form text completions include: `Aya-Expanse 8B` (AI and Cohere, 2024), `Gemma-7B` and `Gemma-7B-it` (Gemma Team and Google DeepMind, 2024), `LLaMA 3.1 8B`, `LLaMA 3.1 8B Instruct`, `LLaMA 3.1 70B`, and `LLaMA 3.1 70B Instruct` (Llama Team and AI

@ Meta, 2024), `Mistral 7B v0.3`, and `Mistral 7B Instruct v0.3` (Jiang et al., 2023). For all models, we set temperature to 0.3 and top_k to 50, to ideally obtain diverse completions that are still somewhat close to the internal representations of the models, limiting the overall stochasticity of the decoding process. However, in a preliminary round of manual annotation on data generated using a temperature of 0.9, we did not find large differences between the generations for what concerns the 4 dimensions included in our human analysis.

Experiments were conducted using in-house computational resources, totaling around 390 hours on a Nvidia A40 GPU.

B.1 Task Descriptions

While for base models we only provide the model with the prompt itself (e.g., just *she is a nurse and she aspires to*), for instruction-tuned models we use the following task descriptions, which are deliberately maintained short to avoid confounding effects due to task prompt choice:

- *Completa la seguente frase in italiano: {prompt}*
- *Complete the following sentence: {prompt}*

B.2 Filtering

Given that free-form text can be noisy (especially with instruction-tuned models the completions can start with 'Sure, here's your sentence' or similar expressions), we first explore the raw generated data and then implement a filtering algorithm.

- For base models, we keep the first line in the completion,
- For instruct models:
  - We isolate assistant responses,
  - We look for markers included in [*here, sure, ecco, frase, sentence*], and colons (*:*), only preserving the first line after the marker,
- For both models, we finally remove trailing quotation marks, ellipses, markdown emphasis markers, and trailing whitespace.

We then save text completions both with and without the original prompt, in order to have both embeddings for the experiments shown in Sec. 5.2.

C Annotation Guidelines

In this section we report the annotation guidelines we adopted for the human evaluation. The annotators were given access to pairs of prompts and relative generated sequence.

Gender *This should assess the models' ability to recognize and respect the gender expressed by the prompt. Possible choices for this annotation are* maintained, changed: inconsistent, changed: consistent. *If* changed, *specify the gender the model switched to as* masculine, feminine, *or* neutral.

Subject *This should assess whether, regardless of the correctness or consistency of the sentence, the subject is correctly interpreted as a human identity. Possible choices are* correct, profession as a whole, physical object, abstract entity.

Topic *Express which topics are addressed in the generated sequence. The topics can be related to an occupation, to one's personal identity, narrative or storytelling elements (such as past actions or life events, perhaps fictional), or be completely unrelated to the context. You can select multiple topics for one completion where applicable. Possible choices are* occupation related: current occupation, occupation related: different occupation, person: identity, person: appearance, storytelling, unrelated.

Agency vs Communion Below we report the definition of *agency* and *communion* from Abele and Wojciszke (2007):

Agency is related to strivings to individuate and expand the self and involves such qualities like instrumentality, ambition, dominance, competence, independence, stereotypical masculinity, and efficiency in goal attainment. Communion arises from strivings to integrate the self in a larger social unit through caring for others and involves such qualities like focus on others and their well-being, cooperativeness, expressivity, warmth, trustworthiness, interdependence, nurturance, and stereotypical femininity.

Possible choices are *agency*, *communion*, *none*.

D Complete Human Analysis Results

In this section, we report the full human evaluation results. Gender assignment distributions by model and gender are reported in Table 6 for both languages. Subject misinterpretation results are

	Misgendering			Assigned Gender		
Model	**M(%)**	**F(%)**	**N(%)**	**M(%)**	**F(%)**	**N(%)**
aya-expanse-8b	0%	0%	0%	0%	0%	0%
gemma-7b	0%	0%	1%	100%	0%	0%
gemma-7b-instruct	0%	0%	1%	100%	0%	0%
Llama-70b	0%	0%	4%	33%	67%	0%
Llama-70b-instruct	0%	1%	3%	100%	0%	0%
Llama-8b	0%	0%	2%	0%	100%	0%
Llama-8b-instruct	0%	0%	0%	0%	0%	0%
Mistral-7B	0%	1%	2%	67%	33%	0%
Mistral-7B-instruct	0%	0%	1%	100%	0%	0%
Instruct Average	0%	0%	1%	100%	0%	0%
Non Instruct Average	0%	0%	2%	56%	44%	0%
Average	0%	0%	2%	69%	31%	0%

	Misgendering			Assigned Gender		
Model	**M(%)**	**F(%)**	**N(%)**	**M(%)**	**F(%)**	**N(%)**
aya-expanse-8b	0%	9%	41%	93%	7%	0%
gemma-7b	0%	7%	13%	89%	11%	0%
gemma-7b-instruct	0%	0%	20%	67%	33%	0%
Llama-70b	0%	2%	21%	86%	14%	0%
Llama-70b-instruct	0%	6%	32%	95%	5%	0%
Llama-8b	0%	7%	14%	100%	0%	0%
Llama-8b-instruct	0%	18%	31%	95%	5%	0%
Mistral-7B	0%	2%	16%	94%	6%	0%
Mistral-7B-instruct	0%	25%	50%	81%	19%	0%
Instruct Average	0%	12%	35%	90%	10%	0%
Non Instruct Average	0%	5%	16%	92%	8%	0%
Average	0%	9%	25%	90%	10%	0%

Table 6: Incidence of misgendering by gender in English (left) and Italian (right), and distribution of assigned genders.

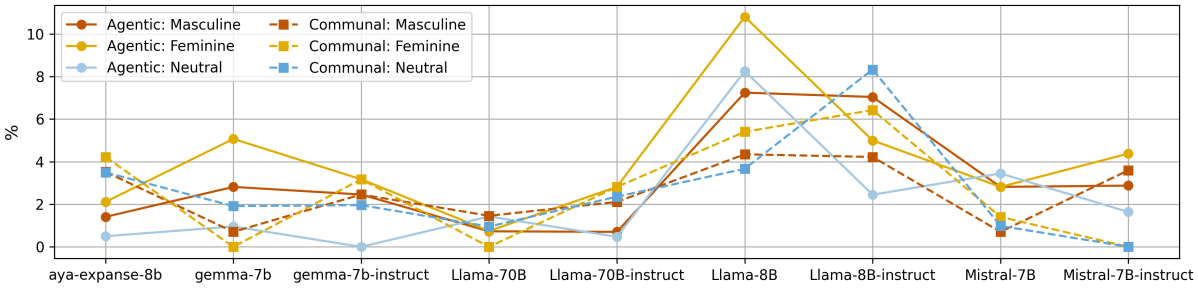

Figure 5: Percentage of texts annotated as agency-focused and communion-focused in Italian for each gender representation and model, averaged across all occupations.

Model	**Subject**	**M (%)**	**F (%)**	**N (%)**
aya-expanse-8b	Abstract	0.00%	0.00%	3.66%
	Object	0.00%	0.00%	1.22%
	Profession	0.00%	0.00%	0.00%
gemma-7b	Abstract	0.00%	0.00%	9.09%
	Object	0.00%	1.15%	2.27%
	Profession	0.00%	0.00%	0.00%
gemma-7b-instruct	Abstract	0.00%	0.00%	1.22%
	Object	0.00%	0.00%	0.00%
	Profession	0.00%	0.00%	0.00%
Llama-70B	Abstract	0.00%	0.00%	1.47%
	Object	0.00%	0.00%	1.47%
	Profession	0.00%	0.00%	1.47%
Llama-70B-instruct	Abstract	0.00%	0.00%	2.56%
	Object	0.00%	0.00%	0.00%
	Profession	1.22%	0.00%	1.28%
Llama-8B	Abstract	0.00%	0.00%	1.18%
	Object	0.00%	0.00%	2.35%
	Profession	0.00%	0.00%	1.18%
Llama-8B-instruct	Abstract	0.00%	0.00%	0.00%
	Object	0.00%	0.00%	0.00%
	Profession	0.00%	0.00%	0.00%
Mistral-7B	Abstract	0.00%	0.00%	3.33%
	Object	0.00%	0.00%	1.11%
	Profession	0.00%	0.00%	1.11%
Mistral-7B-instruct	Abstract	0.00%	0.00%	1.16%
	Object	0.00%	0.00%	1.16%
	Profession	5.62%	3.80%	4.65%

Table 7: Subject misinterpretation distributions by model and gender for English.

Model	**Subject**	**M (%)**	**F (%)**	**N (%)**
aya-expanse-8b	Abstract	0.00%	26.15%	9.47%
	Object	0.00%	3.08%	2.11%
	Profession	3.80%	9.23%	4.21%
gemma-7b	Abstract	4.00%	17.91%	18.10%
	Object	8.00%	5.97%	6.67%
	Profession	0.00%	5.97%	12.38%
gemma-7b-instruct	Abstract	1.72%	20.75%	16.33%
	Object	3.45%	9.43%	6.12%
	Profession	6.90%	11.32%	6.12%
Llama-70B	Abstract	0.00%	16.13%	3.88%
	Object	4.11%	9.68%	8.74%
	Profession	2.74%	8.06%	4.85%
Llama-70B-instruct	Abstract	0.00%	10.94%	3.54%
	Object	1.32%	1.56%	1.77%
	Profession	0.00%	6.25%	10.62%
Llama-8B	Abstract	2.90%	1.35%	11.01%
	Object	1.45%	8.11%	6.42%
	Profession	1.45%	5.41%	4.59%
Llama-8B-instruct	Abstract	0.00%	2.94%	6.73%
	Object	1.37%	1.47%	4.81%
	Profession	2.74%	11.76%	15.38%
Mistral-7B	Abstract	5.48%	12.70%	12.96%
	Object	6.85%	7.94%	7.41%
	Profession	4.11%	15.87%	20.37%
Mistral-7B-instruct	Abstract	0.00%	1.64%	14.86%
	Object	2.41%	3.28%	6.76%
	Profession	1.20%	1.64%	12.16%

Table 8: Subject misinterpretation distributions by model and gender for Italian.

Gender / Topic	aya-expanse-8b	gemma-7b	gemma-7b-instruct	Llama-70B	Llama-70B-instruct	Llama-8B	Llama-8B-instruct	Mistral-7B	Mistral-7B-instruct	aya-expanse-8b	gemma-7b	gemma-7b-it	Llama-70B	Llama-70B-instruct	Llama-8B	Llama-8B-instruct	Mistral-7B	Mistral-7B-instruct
Masculine																		
Different Occupation	8%	12%	28%	16%	17%	16%	7%	20%	9%	5%	7%	3%	8%	9%	5%	4%	4%	7%
Person: Identity	36%	24%	25%	14%	13%	14%	4%	33%	22%	1%	5%	7%	1%	2%	9%	6%	6%	0%
Storytelling	42%	30%	0%	25%	6%	26%	8%	20%	4%	4%	25%	8%	16%	6%	13%	4%	25%	2%
Person: Appearance	0%	1%	0%	0%	0%	1%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Unrelated	1%	0%	0%	16%	0%	1%	0%	0%	0%	0%	7%	5%	3%	2%	8%	1%	4%	1%
Feminine																		
Different Occupation	10%	18%	23%	19%	11%	17%	8%	27%	10%	0%	3%	1%	5%	10%	4%	1%	1%	0%
Person: Identity	33%	28%	23%	17%	18%	21%	7%	33%	25%	7%	10%	8%	5%	7%	12%	10%	5%	3%
Storytelling	36%	24%	1%	25%	7%	24%	3%	20%	4%	1%	26%	6%	8%	7%	7%	3%	20%	3%
Person: Appearance	3%	4%	2%	2%	0%	4%	0%	7%	0%	0%	1%	0%	2%	0%	0%	1%	1%	0%
Unrelated	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	6%	4%	1%	1%	10%	4%	4%	0%
Neutral																		
Different Occupation	5%	6%	8%	5%	11%	12%	9%	22%	8%	0%	1%	1%	3%	6%	4%	4%	2%	1%
Person: Identity	20%	16%	20%	5%	4%	5%	0%	6%	5%	1%	7%	6%	2%	3%	10%	4%	3%	1%
Storytelling	24%	16%	9%	6%	7%	8%	6%	1%	2%	4%	18%	6%	9%	6%	13%	3%	19%	1%
Person: Appearance	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%
Unrelated	1%	2%	1%	5%	5%	1%	0%	0%	0%	4%	13%	10%	2%	2%	12%	4%	2%	6%

Table 9: Topic distribution by gender across models for English (left) and Italian (right).

shown in Table 7 for English and Table 8 for Italian. Topic distributions by gender and model are shown for both languages in Table 9. Finally, the percentage of agentic and communal completions by gender in Italian is shown in Figure 5.

E Embedding analysis details and additional results

E.1 Details of Co-Clustering Matrix Calculation

All the analyses were conducted with the MATLAB software. We applied a consistent sampling (bootstrapping) and clustering procedure to both the Italian and English datasets. The Italian data includes 55,992 occupation-related sentences, while the English data consists of 12,204 completions, both mapped to the 41 occupations in the II level of ISCO-08 and different gender representations (masculine, feminine, neutral). We randomly sampled 6 sentences per gender masculine and feminine within each occupation, yielding a balanced set of 492 sentences (41×6×2). In both cases, a clustering procedure was repeated 1,000 times, each time on a new stratified sample of sentences. To quantify the semantic affinity between occupation-gender groups, we compute a co-clustering matrix that captures how often texts from different groups are assigned to the same cluster across the repeated runs of unsupervised clustering. Specifically, in each run, sentence embeddings are first converted into a pairwise dissimilarity matrix using 1correlation as the distance metric. The linkage function present within the MATLAB software was then used with Ward's method to construct a hierarchical cluster tree, which minimizes the total within-cluster variance. To extract clusters corresponding to occupational categories, we used the dendrogram function of MATLAB, with the number of clusters explicitly set to 41, matching the number of second-level ISCO occupational categories in our dataset (for which we had gendered terms). For each run, we calculated how often text completions from different groups (i.e., 41 occupations x 2 genders) were grouped in the same cluster and then averaged over all runs to obtain the final co-clustering matrix. In this way, random fluctuations were cancelled out while consistent co-clustering patterns could emerge. To estimate the statistical significance of the observed co-clustering patterns, we repeated the same 1,000 clustering runs while randomly shuffling the cluster labels assigned to each sentence, on the same sentence sample. This allowed us to assess whether the observed clustering structure reflects meaningful semantic organization beyond chance.

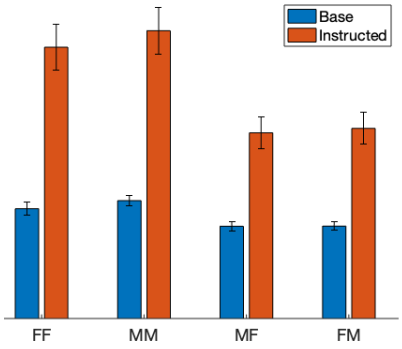

Figure 6: Proportion of co-clustering along the diagonal with respect to the entire quadrant. Bars indicate the standard error of the mean across models

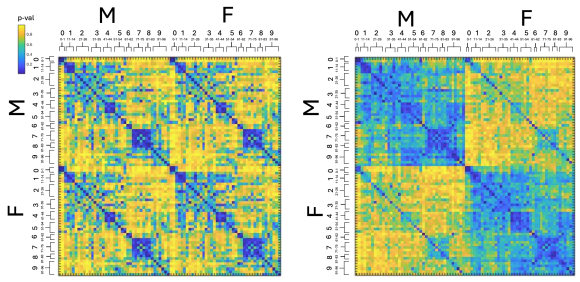

Figure 7: Co-clustering gender-occupation matrices: mean p-values across instructed models (left) and base models (right).

E.2 Additional Results: Instructed vs Base Models

To investigate differences between instructed and base models, we aggregated them into two groups and compared their gender–occupation co-clustering patterns. Figure 6 reports the proportion of diagonal co-clustering across all quadrants and reflects how often embeddings for the same occupation are grouped together. Both groups show a decrease when moving from same-gender (`MM`, `FF`) to cross-gender (`MF`, `FM`) quadrants, suggesting that the same occupations are described in a less semantically similar way across genders than within the same gender for both instruct and base models. Moreover, instructed models consistently maintain a higher proportion of diagonal co-clustering in every quadrant. This indicates that embeddings for the same occupation are more reliably grouped together and less influenced by gender when models are instruction-tuned.

A clear difference also emerges visually in the gender–occupation co-clustering matrix shown in Figure 7, which represents the same matrix shown in Figure 4 but averaged across instructed and base models separately. We can observe how cross-gender and same-gender quadrants show different behavior for base models indicating different clustering across same- and cross-gender, while instructions mitigate the gender asymmetries observed in the base models.

F Complete Lexical Analysis Results

The full lists of most relevant tokens for each gender identity by relevance shift (Sec. 4.3.3) are reported in Table 10 for English and Table 11 for Italian.

G	aya-exp-8b	gemma-7b	gemma-7b-it	Llama-70B	Llama-70B-in	Llama-8B	Llama-8B-in	Mistral-7B-In	Mistral-7B
F	*woman*	*woman*	*field*	*mother*	*woman*	*woman*	*woman*	*dedicated*	*woman*
	beautiful	*mother*	*passionate*	*woman*	*dedication*	*mother*	*female*	*invaluable*	*mother*
	female	*beautiful*	*efficient*	*child*	*pioneer*	*beautiful*	*respected*	*asset*	*beautiful*
	women	*daughter*	*strong*	*single*	*craftswoman*	*girl*	*male*	*dedication*	*wife*
	powerful	*wife*	*woman*	*women*	*dominate*	*child*	*dominate*	*enjoy*	*daughter*
	mother	*women*	*fearless*	*widow*	*advocate*	*model*	*passionate*	*passionate*	*sister*
	girl	*female*	*tradesperson*	*daughter*	*groundbreaker*	*actress*	*barrier*	*woman*	*actress*
	association	*model*	*ceo*	*girl*	*male*	*daughter*	*break*	*evident*	*model*
	forbes	*pioneer*	*renowned*	*businesswoman*	*break*	*wife*	*dedicate*	*strength*	*single*
	recipient	*girl*	*brilliant*	*cook*	*barrier*	*women*	*advocate*	*organized*	*child*
M	*hero*	*friend*	*meticulous*	*person*	*experienced*	*person*	*verse*	*craftsman*	*father*
	party	*person*	*master*	*father*	*reputation*	*friend*	*extensive*	*current*	*person*
	fan	*father*	*skilled*	*hard*	*time*	*worker*	*experienced*	*lie*	*national*
	figure	*protagonist*	*craftsman*	*party*	*deep*	*union*	*final*	*collar*	*husband*
	footballer	*character*	*hard*	*university*	*entrepreneur*	*association*	*background*	*blue*	*handsome*
	2014	*main*	*jack*	*hero*	*understanding*	*footballer*	*geothermal*	*stepping*	*union*
	player	*association*	*captain*	*friend*	*knowledgeable*	*local*	*metallurgy*	*experienced*	*actor*
	history	*university*	*handsome*	*graduate*	*groundbreaking*	*father*	*craftsman*	*custodian*	*friend*
	communist	*handsome*	*valuable*	*communist*	*craftsman*	*businessman*	*optimization*	*extra*	*son*
	assembly	*graduate*	*tradesman*	*china*	*university*	*player*	*stress*	*compose*	*businessman*
N	*essential*	*responsible*	*essential*	*responsible*	*tradespeople*	*responsible*	*personnel*	*play*	*responsible*
	pandemic	*business*	*worker*	*industry*	*responsible*	*maintenance*	*essential*	*essential*	*industry*
	covid	*provide*	*build*	*team*	*crucial*	*industry*	*tradespeople*	*crucial*	*nurse*
	19	*industry*	*provide*	*backbone*	*play*	*team*	*require*	*aim*	*people*
	worker	*team*	*operation*	*maintenance*	*require*	*business*	*laborer*	*professionals*	*team*
	responsible	*service*	*artisan*	*risk*	*maintain*	*production*	*operating*	*economy*	*maintenance*
	provide	*commit*	*expand*	*senior*	*specialized*	*backbone*	*manual*	*workers*	*backbone*
	ensure	*family*	*industry*	*safety*	*training*	*provide*	*ict*	*allied*	*operation*
	client	*people*	*project*	*construction*	*role*	*repair*	*supplier*	*respective*	*construction*
	patient	*quality*	*inspire*	*occupation*	*assist*	*safety*	*dexterity*	*artisans*	*smooth*

Table 10: Top-10 tokens by relevance shift across gender representations and models in English.

G	aya-exp-8b	gemma-7b	gemma-7b-it	Llama-70B	Llama-70B-in	Llama-8B	Llama-8B-in	Mistral-7B-In	Mistral-7B
F	*essa* *she*	*donne* *women*	*operative* *operating*	*donne* *women*	*donne* *women*	*donne* *women*	*figura* *figure*	*persona* *person*	*donne* *women*
	figura *figure*	*donna* *woman*	*aggiunta* *addition*	*professione* *occupation*	*donna* *woman*	*donna* *woman*	*donne* *women*	*figura* *figure*	*importanti* *important*
	fondamentali *fundamental*	*figura* *figure*	*frase* *sentence*	*donna* *woman*	*potenti* *powerful*	*importanti* *important*	*figure* *figures*	*her* *her*	*donna* *woman*
	donne *women*	*importanti* *important*	*pioniera* *pioneer*	*soggette* *subject*	*scienziate* *scientists*	*prime* *first*	*importanti* *important*	*donne* *women*	*pericolose* *dangerous*
	figure *figures*	*figure* *figures*	*evolute* *evolved*	*prime* *first*	*professoressa* *professor*	*figura* *figure*	*rispettate* *respected*	*espertissima* *expert*	*antiche* *ancient*
	vasta *wide*	*sviluppate* *developed*	*variabili* *variables*	*figura* *figure*	*femminili* *feminine*	*professione* *occupation*	*soggette* *subject*	*utilizzate* *used*	*utilizzate* *used*
	gamma *array*	*persona* *person*	*finestra* *window*	*importanti* *important*	*scienziata* *scientist*	*italia* *Italy*	*utilizzate* *used*	*fondamentali* *fundamental*	*sexy* *sexy*
	metodi *methods*	*utilizzate* *used*	*enrichiata* -	*macchina* *machine*	*numerose* *numerous*	*tenute* *held*	*rilievo* *importance*	*professione* *occupation*	*soggette* *subject*
	professione *occupation*	*antiche* *ancient*	*evolutive* *evolutionary*	*antiche* *ancient*	*obbligate* *obliged*	*macchina* *machine*	*riconosciute* *recognised*	*professoressa* *professor*	*figure* *figures*
	riferimento *reference*	*soggette* *subject*	*usate* *used*	*femminile* *feminine*	*importanti* *important*	*figure* *figures*	*utilizzata* *used*	*macchina* *machine*	*potenti* *powerful*
M	*veri* *true*	*padre* *father*	*artisiti* -	*autonomi* *autonomous*	*affidamento* *custody*	*tenuti* *held*	*esposti* *exposed*	*his* *his*	*personaggio* *character*
	svolgono *perform*	*primi* *first*	*esperati* -	*grado* *degree*	*categoria* *category*	*primi* *first*	*assicurare* *to ensure*	*technician* *technician*	*lavora* *works*
	propri *proper*	*autonomi* *autonomous*	*pericosi* -	*tenuti* *held*	*intensivo* *intensive*	*autonomi* *autonomous*	*principale* *main*	*aspires* *aspires*	*grado* *degree*
	riconosciuto *recognized*	*mestiere* *craft*	*experts* *experts*	*autonomo* *autonomous*	*team* *team*	*esposti* *exposed*	*compito* *task*	*on* *on*	*esposti* *exposed*
	padri *fathers*	*esposti* *exposed*	*esami* *exams*	*esposti* *exposed*	*obbligati* *forced*	*compito* *task*	*altamente* *highly*	*operari* -	*protagonisti* *protagonists*
	pubblicitarie *advertising*	*riparare* *repair*	*operaioli* -	*lavora* *work*	*professore* *professor*	*chiamati* *called*	*dispositivo* *device*	*translates* *translates*	*chiamato* *called*
	esposti *exposed*	*interessati* *interested*	*capaci* *capable*	*esposto* *exposed*	*lavora* *works*	*progettati* *designed*	*manutenzioni* *maintenance*	*funzionario* *official*	*rappresenti* *(you) represent*
	fondatori *founders*	*acquisito* *acquired*	*pericolosamente* *dangerously*	*soggetto* *subject*	*esposti* *exposed*	*progettare* *to design*	*funzionari* *officials*	*become* *become*	*chiamati* *called*
	risoluzione *resolution*	*incaricato* *responsible*	*eroi* *heroes*	*svolge* *performs*	*morto* *dead*	*ambito* *scope*	*riparazioni* *repairs*	*installazione* *installation*	*sindacato* *trade union*
	ottimizza *optimizes*	*veri* *true*	*periodici* *periodic*	*primi* *first*	*manutenzioni* *maintenance*	*rischi* *risks*	*ottiene* *obtains*	*eroe* *hero*	*procinto* *about to*
N	*completa* *complete*	*facoltativa* *optional*	*n* -	*lavorare* *to work*	*richiede* *requires*	*de* -	*richiede* *requires*	*completa* *complete*	*de* -
	frase *sentence*	*n* -	*complessa* *complex*	*lavorat* *stimolante* *stimulating*	*qualità* *quality*	*1* -	*processo* *process*	*frase* *sentence*	*attivit* -
	operato *activity*	*principali* *major*	*ta* -	*n* -	*specifiche* *specifications*	*attivit* -	*frase* *sentence*	*conduttura* *pipeline*	*riferimento* *reference*
	ricercato *wanted*	*pi* -	*rata* *instalment*	*attivit* -	*competenze* *competenze*	*qualit* -	*attenzione* *attention*	*riferimento* *reference*	*qualit* -
	affermazione *statement*	*facolt* -	*ono* -	*f* -	*efficienza* *efficiency*	*el* -	*completa* *complete*	*massima* *highest*	*garantire* *to ensure*
	ii -	*de* -	*rate* *instalments*	*corso* *course*	*economica* *economic*	*pi* -	*complesso* *complex*	*operazione* *operation*	*n* -
	confezionato *packed*	*fornire* *to provide*	*na* -	*attenzione* *dinamico* *attention*	*attention*	*ser* -	*integrante* *integral*	*condutture* *pipelines*	*r* -
	complesso *complex*	*soddisfare* *to satisfy*	*sparmiare* -	*dynamic* *materia* *reference*	*riferimento* *reference*	*y* -	*complessa* *complex*	*cercando* *seeking*	*particolare* *particular*
	oggetto *object*	*promuovere* *to promote*	*ate* -	*matter* *soggett*	*processo* *process*	*en* -	*posso* *(I) can*	*perfezione* *perfection*	*importanza* *importance*
	tagliati *cut*	*facoltativo* *optional*	*aumentare* *to increase*	-	*ingegneria* *engineering*	*t* -	*disciplina* *discipline*	*importanti* *important*	*fornire* *to provide*

Table 11: Top-10 Italian tokens by relevance shift across gender representations and models. Each token is accompanied by its most probable English translation, as inferred in the absence of contextual grounding, or by a dash in cases where the token is either already in English or lacks semantic content.