Same Job, Different Gender: Uncovering LLM Stereotypes via Counterfactual Biographies

Chiara Manno[1], Martina Miliani[1,2,*]

[1]Dept. Human Sciences, University of Palermo, Italy
[2]CoLing Lab, Dept. of Philology, Literature, and Linguistics, University of Pisa, Italy

Abstract

Large language models absorb the occupational stereotypes present in their training data, but how this plays out across languages with different gender systems is still underexplored. We introduce Beliefs (Bias Evaluation in engLish and Italian gEnder counterFactual bioS), a bilingual English–Italian counterfactual dataset for occupation prediction built on top of the dataset Bias in Bios. In Beliefs each biography is paired with a minimal counterfactual twin that differs only in the referent's gender, so that any change in a model's prediction of the described occupation can be attributed to gender alone. We evaluate multilingual, Italian-specialised, and reasoning models on both languages, showing that multilingual models are strongest also on Italian but sometimes downgrade women into less prestigious occupations.

Keywords

gender bias, counterfactual fairness, occupation prediction, Italian, large language models

1. Introduction

In recent years, large language models (LLMs) have achieved remarkable performance across a wide range of NLP tasks [1]. However, the same learning process that enables such performances, when applied to large and largely unfiltered corpora, also leads models to absorb and reproduce social stereotypes, including gender stereotypes [2]. As a result, a growing body of recent work has investigated the extent to which LLMs exhibit gender bias, particularly in professional domains, and whether such biases vary across languages, model architectures, or model sizes [3, 4, 5].

The occupational domain is a particularly sensitive case, as it is historically marked by strong gender stereotypes: some professions have long been associated with men and others with women. Such stereotypes can be conceived as socially shared representations that may be automatically activated and influence subsequent judgments and behaviour [6, 7]. Bias, in turn, broadly refers to systematic tendencies in judgment, often associated with automatic cognitive processes [8]. Because such associations pervade human-written text, they are absorbed into the training data of language models, which can learn and reproduce them automatically, with concrete consequences when these systems are deployed in high-stakes settings such as recruitment [9]. This concern becomes sharper once we recognise that a model's behaviour with respect to gender can vary substantially with the structural properties of the input language. English and Italian, in particular, encode gender very differently: English marks it sparsely, mainly through the pronoun system, whereas Italian expresses it pervasively, through nominal, adjectival, and verbal inflection and through agreement [10, 11, 12]. It is therefore an open question whether, and how much, LLMs are swayed by the referent's gender when interpreting a text, and whether this tendency changes across languages.

Existing resources give only partial purchase on this question: most are English-only, target machine translation or coreference rather than occupation inference, or rely on synthetic templates and open-ended generation (Section 2). Our study adopts a *counterfactual fairness* perspective: a model is counterfactually fair with respect to a sensitive attribute if its output does not change when only

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy
*Corresponding author.
martina.miliani@fileli.unipi.it (M. Miliani)
0000-0003-1124-9955 (M. Miliani)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper20.pdf

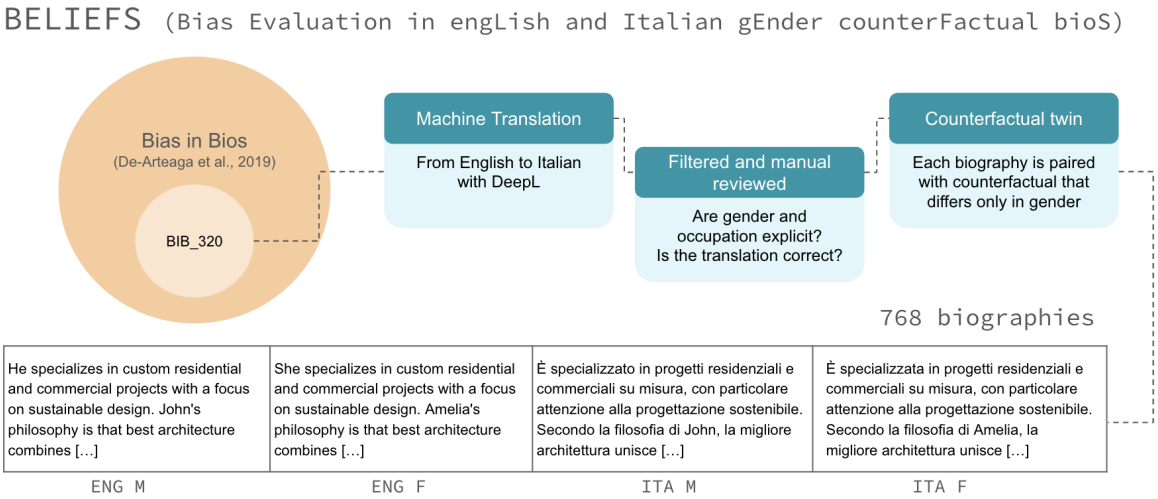


Figure 1: The process of creating BELIEFS, which is described in Sec. 3. We obtained BELIEFS by extracting a subsample from Bias in Bios, BIB_320, which we machine translated into Italian, manually revised and filtered, and paired with counterfactual twins that only differ in gender.

that attribute is altered, all else being equal, and any deviation from this invariance signals a bias in the model's behaviour [13]. To study gender bias in occupation prediction under controlled, cross-linguistic conditions, we introduce BELIEFS (see Fig. 1), a bilingual English–Italian counterfactual dataset built on a subset of manually selected instances of Bias in Bios [9], a dataset containing biography extracts annotated per profession and gender. More specifically, we frame occupation prediction as a classification task: given a biography, the model must return a single occupation label from the closed set of professions. In BELIEFS, each biography is paired with a minimal counterfactual twin that differs only in the referent's gender, so that any change in a model's occupation prediction can be attributed to gender alone.

Our results show that multilingual models are the strongest, including on Italian, and suggest that a masculine bias might appear more systematic in Italian than in English. The counterfactual pairs also point to gender-dependent flips, often towards less prestigious occupations for female referents, that barely move the aggregate F1: This suggests that a model with almost no gender gap may still not be gender-invariant at the level of the single biography.

Our main contributions are the following:[1]

- We release BELIEFS, a bilingual (English–Italian) counterfactual dataset for occupation prediction, derived from Bias in Bios through occupation filtering, automatic translation, and extensive *manual* revision of both the translations and the counterfactual transformation of the biographies, enabling controlled counterfactual probing of gender bias (Sec. 3).
- We conduct an in-depth, cross-lingual evaluation of a diverse set of models — multilingual, Italian-specialised, and reasoning ("thinking") models — combining aggregate performance, gender-gap analysis, and an error analysis that isolates gender-dependent mistakes (Sec. 4 and 5).

2. Related Work

Most existing studies on gender bias can be grouped into four broad paradigms. A central line of research examines gender bias through the analysis of texts translated by language models. Within the field of machine translation, one of the most comprehensive resources is MT-GenEval [14], a benchmark designed to evaluate gender translation accuracy from English into eight target languages, including Italian. Its main contribution lies in its hybrid counterfactual and contextual structure: it includes

[1]The dataset, prompts, and results are available at https://github.com/Unipisa/beliefs.

sentences extracted from Wikipedia, balanced across male and female referents, and rewritten into fully parallel counterfactual versions through the systematic replacement of gender markers. A related approach is adopted in WinoMT [15], which builds on the structure of the Winogender [16] and WinoBias [17] datasets to assess how machine translation systems handle gender in potentially ambiguous coreference contexts, providing sentences in eight different languages. The dataset introduced in the present work, Beliefs, adopts a cross-linguistic perspective as well, as it compares model performance on a bilingual dataset in English and Italian, however, its focus is on occupation classification.

A second group of benchmarks relies on intrinsic evaluation measures, focusing on model representations rather than downstream predictions. Bias-STS-B [18], for instance, is used to analyse how models estimate semantic similarity between pairs of sentences drafted by inserting gender terms, such as *man* and *woman*, together with occupational terms from Rudinger et al. [16], into several templates. A similar paradigm is adopted by Ruzzetti et al. [3], who use templates to estimate the probability that a model associates a given profession with one gender or the other, using synthetic sentences such as "*engineer is a job for a man/woman*". More recently, An et al. [19] investigate the mutual influence of gender and occupation in LLM representations, showing that occupational context can systematically affect the gender representation associated with personal names and that such intrinsic representations are related, though not perfectly aligned, with biased behaviour in occupation prediction.

A more recent strand of research investigates gender bias in open-ended generation. This includes studies on cover letters [20], educational recommendations [21], sentence completions [22], and clinical summaries [5]. Rickman [5], for example, evaluates the potential impact of gender bias in the automatic generation of clinical summaries in long-term care. In summaries referring to male patients, the model tends to emphasise clinical complexity, foregrounding more severe or problematic conditions. By contrast, in summaries concerning female patients, Gemma tends to reduce the emphasis on relevant physical and psychological aspects, in some cases minimising the complexity of the clinical cases. A complementary contribution is offered by Ducel et al. [20], who analyse the generation of cover letters. The authors propose a system for automatically detecting morphosyntactic gender markers in model-generated letters and evaluate seven LLMs on more than 52,000 French texts and over 4,000 Italian texts. Their results reveal a marked preference for masculine forms: masculine markers occur twice as often as feminine markers in French and up to eight times more frequently in Italian.

Alongside translation-oriented datasets, a further line of work evaluates gender bias within downstream classification and decision-making tasks. Recent studies have documented gender-related differences in both automatic occupation coding [23] and occupation-related decision making [24]; notably, the direction and magnitude of these effects vary substantially across models, and in several cases favour female-marked inputs rather than penalising them. In a complementary setting, Yang et al. [25] show that classifiers still recover gender from authentic English CVs after data balancing and the removal of explicit markers such as pronouns and named entities, suggesting that self-presentation in professional texts encodes gender information beyond overt gender markers.

A key resource in the study of occupation-related gender bias is Bias in Bios [9]. Bias in Bios was built from English biographies extracted from Common Crawl and explicitly labelled with the individual's occupation. The dataset contains nearly 400,000 biographies covering 28 occupations, with gender distributions balanced within each occupation. It is primarily used for occupation classification, with the aim of assessing how models predict a person's occupation from biographical text.

Drawing inspiration from MT-GenEval, we constructed counterfactual versions of a subset of Bias in Bios. However, Beliefs differs from both MT-GenEval and Bias in Bios in an important respect. Whereas MT-GenEval relies on the automatic replacement of gender markers, a procedure that may introduce errors [26], and Bias in Bios is based on the automatic selection of biographies from Common Crawl, in Beliefs these operations underwent manual revision, allowing for greater control over the linguistic quality of the counterfactual pairs and over the selection of the biographical texts included in the dataset (Sec. 3).

3. Dataset

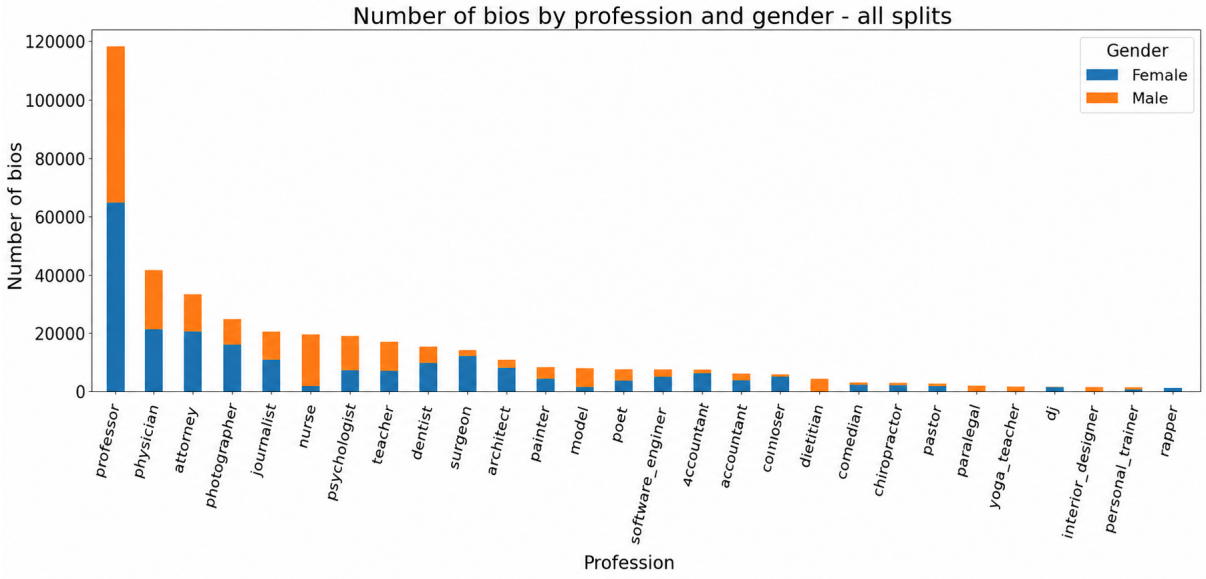


Figure 2: Distribution of biographies by profession and gender in Bias in Bios. Each stacked bar represents a profession, with segments showing the number of biographies labeled as Female and Male.

To investigate the effect of the referent's gender on occupation classification across English and Italian, we introduce Beliefs, a counterfactual bilingual dataset built upon Bias in Bios [9]. The dataset consists of occupation-labelled biographies organised into gender-balanced minimal pairs that differ only in the referent's gender. We first describe the original Bias in Bios dataset and the limitations relevant to our study (Section 3.1), then detail the selection of occupations and the construction of a balanced subset of biographies (Section 3.2). We next describe the translation of the selected English biographies into Italian (Section 3.3) and, finally, the manual revision and controlled gender-flipping procedure used to create the counterfactual pairs that constitute Beliefs (Section 3.4).

3.1. Bias in Bios

Bias in Bios (BiB) comprises 396,189 English biographies automatically extracted from Common Crawl.[2] The texts are written in the third person and therefore typically contain pronouns (*he*, *she*) and first names from which binary gender can be inferred. The dataset spans 28 occupations, selected by their frequency in Common Crawl and is very unevenly distributed across the occupations (see Fig. 2). Overall, 53.9% of the biographies refer to male and 46.1% to female subjects.

The task associated with BiB is to predict a person's occupation from their biography. On this task, De-Arteaga et al. [9] show that simply removing explicit gender indicators (names, pronouns) does not eliminate bias — significant gaps between men and women persist and correlate with real-world occupational gender imbalances, suggesting that such classifiers may reinforce pre-existing inequalities.

Despite its strengths, BiB presents some limitations when reused for occupation classification. Since occupations are automatically extracted and biographies are not manually reviewed, some input passages provide insufficient or misleading evidence for the assigned label. Manual inspection revealed cases where biographies are truncated, uninformative, or inconsistent with their occupation label. As a result, plausible model predictions may be incorrectly counted as errors, introducing evaluation noise that can affect model comparisons and the interpretation of gender-related differences. This motivates the filtering procedure described next (Sec. 3.2).

[2] In our experiments we use the test split of the dataset version distributed through Hugging Face [27]: https://huggingface.co/datasets/LabHC/bias_in_bios

Occupation (It)	Occupation (En)	Frequency
giornalista	**journalist**	**1,103,257**
professore/professoressa	**professor**	**978,699**
avvocato/avvocata	**attorney**	**909,359**
insegnante	**teacher**	**834,033**
regista	**filmmaker**	**805,175**
poeta/poetessa	**poet**	**591,579**
architetto/architetta	**architect**	**484,731**
pittore/pittrice	**painter**	**410,419**
fotografo/fotografa	**photographer**	**336,444**
psicologo/psicologa	**psychologist**	**270,463**
infermiere/infermiera	**nurse**	**236,234**
compositore/compositrice	**composer**	**219,643**
dj	dj	173,967
chirurgo/chirurga	**surgeon**	**154,838**
commercialista	**accountant**	**94,713**
dentista	**dentist**	**94,651**
programmatore/programmatrice	**software engineer**	**81,253**
rapper	rapper	58,376
personal trainer	personal trainer	24,366
dietologo/dietologa	dietitian	16,177
attore comico/attrice comica	comedian	6,210
insegnante di yoga	yoga teacher	3,446
designer di interni	interior designer	603

Table 1
Absolute frequencies of candidate occupations in the *itTenTen20* corpus, together with their feminine Italian forms in parentheses and English equivalents. The 16 occupations retained for the study are shown in **bold**.

3.2. Filtering BiB

BiB underwent a filtering process involving two aspects: the selection of occupations to analyse, and the extraction of specific biography samples. These samples were used both for a preliminary round of English-language experiments (Section 5.1) and as the basis for building the Italian dataset.

Occupations were selected on empirical and linguistic grounds: we privileged professions that are commonly attested and easily recognisable also during translation and adaptation. To this end we used Sketch Engine,[3] and verified the candidate occupations against itTenTen20.[4] Both the masculine and the feminine form were counted whenever available (e.g. *professore* + *professoressa*), so as to estimate the overall frequency of each occupation regardless of grammatical gender.

Frequencies were computed for 23 of the 28 BiB occupations. The remaining five were left out either because they are uncommon in the Italian setting (e.g., *chiropractor*, *pastor*) or because their Italian rendering is polysemous (e.g., *model*, which overlaps with the sense of "template/scheme"). Of the 23, we selected the 16 most frequent (above 80,000 occurrences), listed in Table 1.[5]

Having selected the 16 occupations, we then removed the biographies in which the occupation was explicitly mentioned in the text, so that the classification task could not be solved by simple lexical matching, yielding a filtered set of 59,592 biographies. From it we drew a balanced two samples. A first sample of 200 biographies, BIB_200, was obtained for some preliminary tests (Section 5.1). BIB_320 is a second balanced sample of 320 biographies, 20 for each of the 16 occupations, evenly split by gender (10 masculine and 10 feminine); it was randomly extracted from BiB and served as the starting point for the Italian dataset, built through automatic translation and subsequent manual revision, as described in the following subsections.

[3] https://www.sketchengine.eu/
[4] https://www.sketchengine.eu/ittenten-italian-corpus/
[5] Although *dj* was among the 16 most frequent occupations in itTenTen20, it was excluded because it is not sufficiently represented in BiB.

3.3. Translating BiB

Translating English biographies into Italian is not a neutral operation: English marks gender mainly through pronouns, whereas Italian encodes it through agreement across articles, adjectives, and participles. As a result, gender that remains implicit in the source text often has to be made explicit in Italian, with the masculine typically selected as the unmarked default [28, 29].[6] Machine translation inherits and can amplify this masculine default [30, 31], which makes the choice of MT system consequential for the reliability of the Italian resource.

Recent work on commercial MT confirms a systematic masculine default, most pronounced for profession nouns lacking explicit gender markers — precisely the case at hand [32, 33, 34].

Albeit overall, no currently available MT system is gender-neutral, DeepL offers the best compromise between linguistic quality, context sensitivity, and limited gender bias, particularly for English-to-Italian [34, 32, 35]; we therefore selected it to build BELIEFS. Given the intrinsic limits of automatic translation, every translated biography then underwent systematic manual revision, ensuring that each Italian biography faithfully and grammatically encodes the gender of its English source.

3.4. BELIEFS

The new dataset we built, BELIEFS (*Bias Evaluation in engLish and Italian gEnder counterFactual bioS*), derived from the BIB_320 subset described above through a multi-step procedure that combines automatic translation, controlled selection, and manual revision. The biographies in BIB_320 were first machine-translated from English into Italian with DeepL (Section 3.3).

These biographies underwent manual revision aimed at checking the quality of the translation, the semantic coherence of the content, and the clarity of the information about the subject's occupation and gender, the latter being essential to assess whether models are swayed by the referent's grammatical gender when linking a biography to a profession. During this step we discarded biographies whose occupation was not clearly deducible from the text, was ambiguous, or was inconsistent with the content, as well as cases in which the referent's gender could not be disambiguated.

Where automatic translation preserved the informational content but left the subject's gender implicit, we intervened by introducing disambiguating cues, while avoiding edits that would sound unnatural in Italian (see Table 2). For example, we chose not to force explicit subject pronouns in contexts where Italian normally omits them, thereby preserving the syntactic naturalness of the text. The interventions mainly involved adjusting or inserting morphological, pronominal, and lexical choices that make gender and profession explicit: selecting gender-marked verbal or nominal forms (e.g. *ha conseguito la laurea* [*earned a degree*] → *si è laureata* [*she graduated*]), correcting the gender of professional-role nouns (*co-editore* [*co-editor*, m.] → *co-editrice* [*co-editor*, f.]), and restoring honorific or professional titles present in the source but dropped in the automatic rendering (*il signor* [*Mr.*], *la dottoressa* [*Dr.*, f.]), as well as omitted articles and determiners (*è un consulente* [*is a consultant*] with the masculine article). All edits were kept minimal and faithful to the source, avoiding reformulations that might alter style or add information. After revision, we retained, for each of the 16 occupations, the 12 biographies (6 masculine and 6 feminine) in which both gender and profession were clearly identifiable, yielding a gender-balanced base of 192 biographies.

Starting from this base, we used Claude[7] to automatically build a counterfactual twin, which was then manually revised: an otherwise identical biography in which only the referent's gender is changed. The transformation involved converting personal pronouns (*he*/*she*, *him*/*her*; *lui*/*lei*), possessive pronouns (*his*/*her*), articles (*un*/*una* [*a*], *il*/*la* [*[*the]]), and adjectives; replacing first names with strongly gender-connoted names of the opposite gender (e.g. *Marc* → *Maria*); adapting titles (*Mr.* → *Ms.*; *signore* → *signora*; *Dott.* [*Dr.* m.] → *Dott.ssa* [*Dr.* f.]); and rewriting inherently gendered expressions (e.g. *she is a mother* → *he is a father*, *he is a man* → *she is a woman*). The flip was applied consistently in both

[6] This choice reflects and reinforces the dominance of the masculine, producing what has been described as a form of *distranslation*.

[7] The counterfactual variants were created in March 2026 with Sonnet 4.6.

Source (EN)	DeepL	Manual revision
…she received her B.A. in Creative Writing	…ha conseguito la laurea in scrittura creativa	…**si è laureata** in scrittura creativa
As counsel, Mr. Creadick assists clients …	In qualità di consulente, Creadick assiste i clienti …	In qualità di consulente, **il signor** Creadick assiste i clienti …
She is a co-editor of *Soundscape* …	È co-editore di *Soundscape* …	È **co-editrice** di *Soundscape* …

Table 2
Examples of manual revision for gender disambiguation: English source, initial DeepL translation, and revised version (edited spans in **bold**).

languages; in Italian it required propagating gender across every agreeing element (articles, adjectives, participles).

Each original biography is thus paired with a minimal counterfactual that differs only in gender, giving 24 biographies per occupation (12 original and 12 counterfactual) over the 16 occupations — 384 per language, and 768 in total across English and Italian. We refer to this resource as BELIEFS: Because the only systematically varying signal within a pair is the referent's gender, BELIEFS lets us measure in a controlled way how much a model's occupation prediction depends on gender.

4. Experimental settings

This section describes the models we evaluate, the prompting configuration adopted for the occupation-prediction task, and the metrics with which we score it.

All experiments were run on a compute node equipped with eight NVIDIA A100 GPUs, using a variable number of GPUs depending on the memory requirements of each model. Models were loaded in `bfloat16` precision, with no quantisation, and automatically distributed across the available devices using `device_map="auto"`. Inference was performed in evaluation mode using deterministic greedy decoding (`do_sample=False`); sampling parameters were disabled by setting both `temperature` and `top_p` to `None`. Input sequences were tokenised with truncation enabled.

For each input, the order of the candidate profession labels and of the two few-shot demonstrations was randomly shuffled, while generation itself was deterministic. We evaluated five models on the English version of the dataset and additionally included three Italian models, which were tested on the Italian portion of the dataset (Sec. 4.1). All models were evaluated under both zero-shot and few-shot settings (Sec. 4.2). For standard models, the maximum number of generated tokens was set to 10, as the task requires producing a single profession label. For reasoning models, we increased this limit to 1500 tokens to allow for intermediate reasoning before producing the final prediction.

We frame occupation prediction as single-label classification and evaluate it with precision, recall, and F1 score. For each occupation c, precision is the proportion of biographies predicted as c that are truly labelled c; recall the proportion of biographies labelled c that the model predicts as c, and F1 the harmonic mean of the two. We report macro-averages over the 16 occupations, which the balanced design of BELIEFS makes directly comparable across classes. Since BELIEFS pairs every biography with a gender-flipped counterfactual, we can break these scores down by gender, isolating the effect of gender on the model's predictions (Sec. 5).

4.1. Models

We evaluated a diverse set of open-weight large language models, selected to cover multilingual, Italian-specialized, and reasoning-oriented settings. As a multilingual baseline, we used `google/gemma-2-2b-it`, from the Gemma 2 family [36]. We further included three Italian-oriented models: `sapienzanlp/Minerva-7B-instruct-v1.0`, based on the Minerva family of LLMs trained from scratch on Italian data [37], and two Italian-adapted models, `galatolo/cerbero-7b` [38] and `LLaMAntino-3-ANITA-8B-Inst-DPO-ITA` [39]. In addition, we evaluated two larger mul-

tilingual models, namely `google/gemma-2-9b-it` [36] and `meta-llama/Llama-3.1-8B` [40]. Finally, we considered two reasoning-oriented models: `DeepSeek-R1-Distill-Llama-8B`, derived from the DeepSeek-R1 reasoning framework and distilled on top of the Llama architecture [41, 40], and `DeepSeek-R1-0528-Qwen3-8B`, which combines the DeepSeek-R1 reasoning paradigm with the Qwen3 model family [41, 42]. For English, we restricted the evaluation to the multilingual and reasoning-oriented models, namely `google/gemma-2-2b-it`, `google/gemma-2-9b-it`, `meta-llama/Llama-3.1-8B`, `DeepSeek-R1-Distill-Llama-8B`, and `DeepSeek-R1-0528-Qwen3-8B`. This setup allows us to compare Italian-specialized models against broader multilingual and reasoning-oriented architectures, while maintaining a consistent cross-lingual comparison on English.

We deliberately did not include encoder-based classifiers such as BERT-style models, although the task is framed as classification. Encoder-based approaches require task-specific fine-tuning on Bias in Bios, whose behaviour then reflects both the pre-trained representations and the gender distribution of the fine-tuning data, and bias in this setting has already been studied [9]. Our goal is instead to probe models without parameter updates, which also keeps the cross-lingual comparison controlled: the same model is evaluated on English and Italian, whereas encoder-based classifiers would have to be fine-tuned separately per language.

4.2. Prompts

Since both the wording of the instruction and the way the label set is presented can substantially affect a model's behaviour on our occupation classification task, we selected the prompt through a dedicated preliminary study on English data (`BIB_200`), comparing six instruction structures[8] crossed with four label-presentation variants. The best-performing configuration was a compact, instruction-repeated prompt that explicitly constrains the output to one of the listed professions and to a single word, with the label set presented in *randomised* order and enclosed in quotes. The two version of the zero-shot template, one in Italian and one in English, are the following:

> **EN.** `Given a biography, predict and return only ONE of these professions: “professor”, “attorney”, . . . Return ONLY ONE WORD.`
>
> **IT.** `Data una biografia, predici e restituisci solo UNA di queste professioni: “professore”, “avvocato”, . . . Restituisci SOLO UNA PAROLA.`

We adopted it as the reference configuration throughout: in English in its original form, and translated while keeping the instruction structure unchanged, so as to maximise comparability across the two languages, for the Italian and multilingual models.

We also conducted an additional preliminary experiment on the Italian data to assess whether explicitly prompting models to choose between the masculine and feminine forms of occupation labels could improve classification performance. As this strategy did not yield consistent improvements across models, it was not adopted in the main experiments.

5. Results and Discussion

We first evaluate the models on the English biographies, where gender is mainly conveyed through names and pronouns (Sec. 5.1), and then on the Italian biographies, where gender is also marked morphologically across the text (Sec. 5.2). For each language, we compare zero-shot and few-shot prompting and report macro-averaged F1 both overall and separately for female- and male-subject biographies. This setup allows us to distinguish gains due to better task compliance from systematic gender-related differences in model predictions. To this aim we also conducted an error analysis, examining the results by means of confusion matrixes (Sec. 5.3).

[8] The six prompt structures compared in the preliminary English-language study vary in the constraint imposed on the output, the role assigned to the model, and the way the class labels are presented.

	Zero-shot			Few-shot		
Model	F1	$F1_F$	$F1_M$	F1	$F1_F$	$F1_M$
Multilingual						
Llama-3.1-8B	0.895	0.885	0.905	**0.952**	0.964	0.940
gemma-2-9b-it	**0.930**	0.934	0.925	0.931	0.933	0.929
Reasoning						
DeepSeek-R1-Distill-Llama-8B	0.236	0.235	0.235	0.882	0.877	0.886
DeepSeek-R1-0528-Qwen3-8B	0.692	0.687	0.694	0.854	0.850	0.857
Baseline						
gemma-2-2b-it	0.501	0.473	0.524	0.844	0.831	0.854

Table 3
Macro-averaged F1 on the English biographies, overall and split by the subject's gender ($F1_F$: female; $F1_M$: male), under zero-shot and few-shot prompting. Best overall F1 per setting in **bold**.

5.1. English Bios

Table 3 reports macro-F1 on the English biographies, overall and broken down by the subject's gender, under zero-shot and few-shot prompting. In the zero-shot setting, the two larger multilingual models are clearly the strongest: `gemma-2-2b` reaches an overall F1 of 0.930 and `Llama-3.1` reaches 0.895, both following the single-label format reliably. The other models struggle to comply with the output constraint without examples: the 2B baseline `gemma-2-2b` scores only 0.501, and the reasoning-oriented models are the weakest of all: `DeepSeek-Llama` collapses to 0.236 and `DeepSeek-Qwen3` reaches 0.692. Few-shot prompting closes most of this gap: It lifts `DeepSeek-Llama` by 0.65 to 0.882 and the 2B baseline `gemma-2-2b` by 0.34 to 0.844, brings every model above 0.84, and yields the best overall score (`Llama-3.1`, 0.952), while leaving the already-strong `gemma-2-9b` essentially unchanged ($0.930 \rightarrow 0.931$). The benefit of examples is thus largest precisely where zero-shot format compliance was poorest, and negligible for models that already follow the instruction.

Turning to gender, the F1 gap between female- and male-subject biographies is small for the competent configurations, within about 0.01 for `gemma-2-9b` and `DeepSeek-Qwen3`. Where it is non-negligible, it usually favours male subjects, consistent with a mild masculine bias. The clearest case is the 2B baseline `gemma-2-2b`, whose zero-shot gap is the largest in the table ($F1_F = 0.473$ vs. $F1_M = 0.524$, a 0.05 male advantage); few-shot roughly halves it (0.831 vs. 0.854). Two caveats temper this reading. First, the perfectly even zero-shot scores of `DeepSeek-Llama` (0.235 vs. 0.235) reflect uniformly near-random performance rather than fairness: both genders fail equally. Second, the direction of the gap is not stable: `gemma-2-9b` slightly favours female subjects in both settings, and `Llama-3.1` switches from a 0.02 male advantage in zero-shot to a 0.02 female advantage in few-shot. On balance, the gender gaps on English are modest, which is expected given that English signals gender only sparsely, through names and pronouns; the cross-lingual comparison with Italian, where gender is grammatically pervasive, is therefore the more revealing test (Section 5.2).

5.2. Italian Bios

Table 4 reports macro-F1 on the Italian biographies. As on English, `gemma-2-9b` is the strongest model (0.914 zero-shot), and few-shot leaves the results essentially unchanged (0.918). Two patterns from the English experiments recur. First, few-shot prompting rescues the models that cannot follow the task without examples: `gemma-2-2b`, which produces no valid label at all in zero-shot (0.000), reaches 0.739, and `LLaMAntino-3` climbs from 0.403 to 0.763. Second, it fails to help — and slightly harms — the reasoning models, with `DeepSeek-Qwen3` dropping from 0.871 to 0.849. More striking, the Italian-specialised models do not outperform the multilingual `gemma-2-9b` on their own language: `cerbero-7b` (0.785), `Minerva-7b` (0.740), and `LLaMAntino-3` (0.403) all trail it in zero-shot. `DeepSeek-Llama` is the clearest casualty of the switch to Italian, collapsing to 0.215 in zero-shot

	Zero-shot			Few-shot		
Model	F1	$F1_F$	$F1_M$	F1	$F1_F$	$F1_M$
Multilingual						
gemma-2-9b	**0.914**	0.902	0.926	**0.918**	0.924	0.912
Llama-3.1	0.689	0.679	0.698	0.827	0.822	0.831
Italian						
cerbero-7b	0.785	0.806	0.764	0.810	0.798	0.822
LLaMAntino-3	0.403	0.392	0.410	0.763	0.745	0.780
Minerva-7b	0.740	0.729	0.748	0.752	0.709	0.788
Reasoning						
DeepSeek-Qwen3	0.871	0.849	0.891	0.849	0.860	0.834
DeepSeek-Llama	0.215	0.205	0.216	0.485	0.489	0.467
Baseline						
gemma-2-2b	0.000	0.000	0.000	0.739	0.731	0.747

Table 4
Macro-averaged F1 on the Italian biographies of BELIEFS, overall and split by the subject's gender ($F1_F$: female; $F1_M$: male), under zero-shot and few-shot prompting. Best overall F1 per setting in **bold**.

Occupation	`gemma-2-9b`	`cerbero-7b`	`DeepSeek-Qwen3`
architetto (architect)	**1.000**	0.933	0.913
avvocato (attorney)	**0.923**	0.830	0.898
chirurgo (surgeon)	**0.941**	0.909	0.864
commercialista (accountant)	**0.960**	0.588	0.939
compositore (composer)	**0.957**	0.894	0.933
dentista (dentist)	0.980	**1.000**	0.960
fotografo (photographer)	**1.000**	0.898	0.889
giornalista (journalist)	**0.941**	0.884	0.800
infermiere (nurse)	**0.884**	0.780	0.818
ingegnere informatico (software eng.)	**0.979**	0.754	0.933
insegnante (teacher)	**0.737**	0.560	0.708
pittore (painter)	**1.000**	0.979	0.979
poeta (poet)	**0.979**	0.917	0.933
professore (professor)	0.500	0.286	**0.808**
psicologo (psychologist)	**0.958**	0.844	0.727
regista (filmmaker)	**0.960**	0.909	0.833

Table 5
Per-occupation F1 (zero-shot) on both the male and female Italian biographies for three representative models: a multilingual model (`gemma-2-9b`), an Italian-specialised model (`cerbero-7b`), and a reasoning model (`DeepSeek-Qwen3`). Best per row in **bold**. Worst results per model are in red.

(against 0.692 on English) and recovering only to 0.485 few-shot; together with the zero-scoring `gemma-2-2b`, its zero-shot gender figures are not interpretable.

The gender breakdown is more telling than on English. In zero-shot, almost every model assigns higher F1 to male-subject biographies: `gemma-2-9b` (0.902 vs 0.926), `Llama-3.1` (0.679 vs 0.698), `Minerva-7b` (0.729 vs 0.748), `LLaMAntino-3` (0.392 vs 0.410), and `DeepSeek-Qwen3` (the largest male-favouring gap, 0.849 vs 0.891), with only `cerbero-7b` favouring female subjects (0.806 vs 0.764).

This male advantage is far more consistent than on English, where the direction of the gap changed from one model to the next. `gemma-2-9b` is a clear example: it slightly favours female subjects on English but switches to favouring male subjects on Italian. This fits our main hypothesis — since Italian marks gender on many words, gender information is always present in the text, and the models rely on it more. Few-shot prompting changes the size of these gaps but does not remove them: `Minerva-7b`, for instance, ends up with a clear male advantage (0.709 vs 0.788). Examples therefore help the models follow the output format, but they do not reduce gender bias.

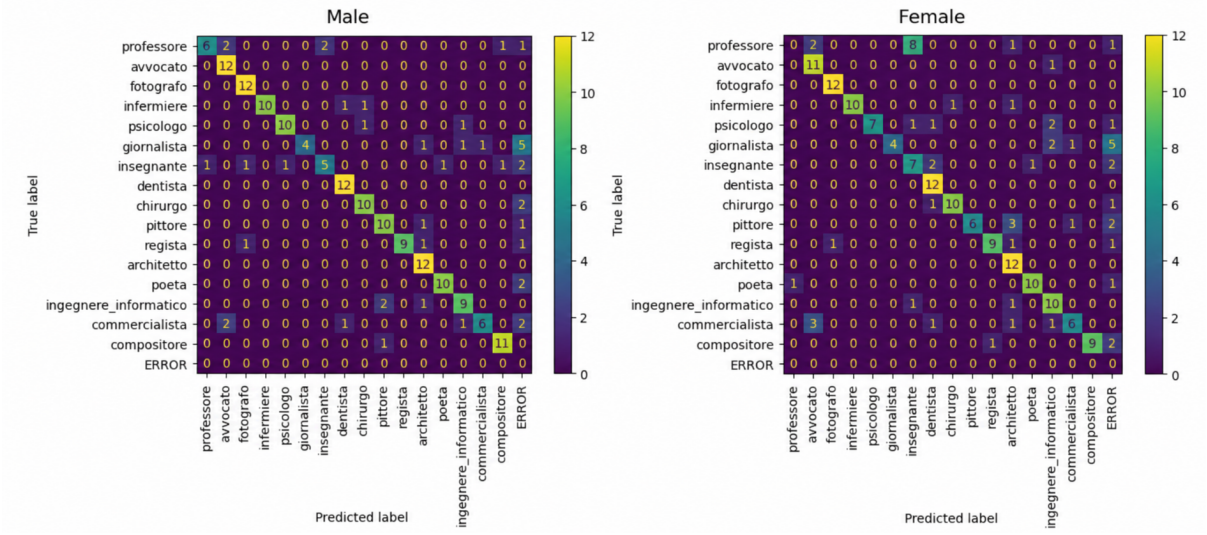


Figure 3: `Minerva-7b` confusion matrices on the Italian biographies, split by the subject's gender (left: male; right: female). Occupation labels are reported in the canonical masculine form, since experiments with feminine labels lead to lower results (see Sec. 4.2).

Table 5 breaks zero-shot F1 down by occupation for three representative models. `gemma-2-9b` is at or near ceiling on most classes (e.g. 1 on *architect*, *photographer*, and *painter*), with the teaching-related labels standing out as the hardest: *teacher* (0.737) and especially *professor* (0.500), which the two non-reasoning models handle poorly (down to 0.286 for `cerbero-7b`). These two roles are mutually confusable, and *professor* further subsumes a broad range of academic profiles, so a biography rarely states the exact label. The one model that copes well with *professor* is `DeepSeek-Qwen3` (0.808): plausibly, the very ambiguity that defeats the other models is where its step-by-step reasoning pays off, disambiguating among the candidate academic roles before committing to a label.

To investigate whether the very low zero-shot performance of the reasoning models could be partly due to output-processing failures rather than classification errors, we manually inspected the structure of generations that were returned as `error` (Table 8). The effect is particularly pronounced for `DeepSeek-Llama` on Italian: 292 of its 335 zero-shot errors show a structurally problematic output, and in 284 cases the final `<answer>` tag is not closed; the same phenomenon, although less severe, is also observed in the few-shot setting (158 of 242 errors, with 149 unclosed answers). In contrast, these failures are rare on English and substantially less frequent for `DeepSeek-Qwen`. This indicates that part of the exceptionally low Italian performance of `DeepSeek-Llama`, especially in the zero-shot setting, is attributable to generation or output-format failures: the model often does not reach a well-formed final answer from which the target label can be extracted. The observed pattern is consistent with generations being interrupted before the reasoning process is completed, potentially also as a consequence of the 1500-token output limit. However, since the available logs do not allow us to verify token exhaustion directly, we conservatively classify these cases as structurally incomplete generations rather than attributing them specifically to the output-length limit.

5.3. Error analysis

Looking at *where* the models go wrong, and whether the errors depend on gender, tells us more than the aggregate scores. For most models the errors are content-driven and similar for both genders, but we also observed some patterns that might be a clue of the presence of bias. Splitting `Minerva-7b` confusion matrix by gender (Figure 3) shows that its errors are concentrated on the *professor*/*teacher* pair and depend on gender: female *professor* biographies are never kept as *professor* (0 of 12) and are relabelled *teacher* 8 times, while the same biographies in their male version stay *professor* 6 times. The minimal pairs in Table 6 show this directly: changing only the subject's gender flips the prediction

Italian biography excerpt	English translation	Gold	Pred (M)	Pred (F)
*…special***izzato/izzata** *in letteratura inglese medievale… insegna corsi su razza, genere e sessualità…*	*…special***ized** *in medieval English literature… teaches courses on race, gender, and sexuality…*	professor	professor	**teacher**
…il **dottor***/la* **dottoressa** *Ghosh… dottorato in Economia… si occupa di insegnare…*	*…***Dr.** *Ghosh… PhD in Economics… is involved in teaching…*	professor	professor	**teacher**
*…esperienza di insegnamento decennale… è stat***o/a** *premiat***o/a***…*	*…ten years of teaching experience… has been awarded…*	teacher	**professor**	teacher

Table 6
`Minerva-7B` on Italian minimal pairs: changing only the subject's gender flips the prediction between *professor* and *teacher*. English translations of the biography excerpts are provided for reference. Boldface in the Italian excerpts marks the gendered cues manipulated across the minimal pairs, while wrong predictions are shown in **bold**.

Biography (gendered cue)	Gold	Pred (M)	Pred (F)
Cynthia/Ryan*… offering breast exams and teaching a health class… a women's health curriculum…*	nurse	**dentist**	nurse
Mr./Mrs. *Schwartz… NPI Number… practice location…*	nurse	**dentist**	nurse
Dr. Levy teaches in the social work programs… **he/she** *recently completed a course…*	professor	professor	**teacher**
Reed… professional curriculum writing for education led **him/her***… Common Core study guide…*	teacher	**poet**	teacher

Table 7
`Llama-3.1` on English minimal pairs: changing only the name and pronouns flips predictions along stereotypical lines (a male *nurse* becomes a *dentist*; a female *professor* becomes a *teacher*). Wrong predictions in **bold**.

	English						Italian					
Model	**Corr.**	**Wrong**	**Poss. exh.**	**Think**	**Ans.**	**Post**	**Corr.**	**Wrong**	**Poss. exh.**	**Think**	**Ans.**	**Post**
DeepSeek-Llama few	326	58	5	0	5	0	142	242	158	0	149	9
DeepSeek-Llama zero	353	31	3	0	3	0	49	335	292	0	284	8
DeepSeek-Qwen few	354	30	8	8	8	0	315	69	18	18	6	12
DeepSeek-Qwen zero	360	24	4	4	4	0	325	59	13	13	8	5

Table 8
Structural error analysis of reasoning model generations. *Poss. exh.* denotes possible exhaustion without token-count information; *Think* and *Ans.* denote unclosed `<think>` and `<answer>` tags, respectively; *Post* denotes continued generation after `</answer>`.

between the two labels, with the higher-status academic title (*professor*) going to the male version and the lower-status teaching role (*teacher*) to the female one. The reverse also happens, with a male *teacher* promoted to *professor*.

The same instance-level view matters on English too, where the F1 gender gaps were small (Section 3). On its minimal pairs, the best English model, `Llama-3.1`, still flips predictions along stereotypical lines (Table 7): a *nurse* biography is labelled correctly when the subject is female but becomes *dentist* when only the name and pronouns are made male; a *professor* bio stays *professor* when male but is downgraded to *teacher* when female; a *teacher* bio is correct when female but becomes *poet* when male. Each flip is rare on its own and they partly cancel out, so they barely change the group F1 — but they show that a model with almost no F1 gender gap might still be far from gender-invariant on the single biography. Notably, the female *professor*→*teacher* downgrade appears in both the Italian `Minerva-7b` and the English-based model `Llama-3.1`, suggesting a stereotype that is not tied to one language.

Finally, we investigated the reason behind the very low zero-shot performance of the reasoning models, by inspecting the structure of generations that were returned as `error` (Table 8). The effect

is particularly pronounced for `DeepSeek-Llama` on Italian: 292 of its 335 zero-shot errors show a structurally problematic output, and in 284 cases the final `<answer>` tag is not closed; the same phenomenon, although less severe, is also observed in the few-shot setting (158 of 242 errors, with 149 unclosed answers). In contrast, these failures are rare on English and substantially less frequent for `DeepSeek-Qwen`. Thus, part of the exceptionally low Italian performance of `DeepSeek-Llama`, especially in the zero-shot setting, is attributable to generation or output-format failures. The observed pattern is consistent with generations being interrupted before the reasoning process is completed, potentially also as a consequence of the 1500-token output limit, which might be low and extended in future work.

6. Conclusion

We introduced Beliefs, a bilingual English–Italian counterfactual dataset for occupation prediction, built on top of Bias in Bios through occupation filtering, automatic translation, and manual revision of both the translations and the gender-flipped counterfactual pairs. Using Beliefs, we evaluated a diverse set of models — multilingual, Italian-specialised, and reasoning ("thinking") models — on both languages, looking at overall performance, the gender gap, and the errors the models make.

Several findings stand out. Multilingual models are the strongest, and they remain the strongest on Italian too: training a model from scratch on Italian, or specialising it for the language, does not give it an edge on its own language. The masculine bias is also more systematic in Italian than in English, in line with the fact that Italian marks gender pervasively while English marks it only sparsely. Even the best models are not gender-invariant: on the minimal pairs they sometimes move a woman into a less prestigious occupation, a status-linked stereotype that the aggregate scores tend to hide. Neither specialised Italian training nor reasoning helps with this: reasoning models bring an advantage only on the most ambiguous occupations, such as *professor*, where step-by-step disambiguation pays off, but not on gender bias as such. Finally, few-shot prompting mainly improves weaker models' ability to follow the output format. Overall, a counterfactual, cross-lingual evaluation surfaces gender-dependent behaviour that single-language, aggregate metrics miss, and this behaviour is more pronounced in a grammatically gendered language such as Italian.

For future work, we aim to extend the analysis beyond binary gender to neutral and inclusive forms, which remain particularly challenging for current language technologies [43, 44]. This could also involve developing a gender-neutral version of Beliefs and pairing the evaluation with concrete mitigation strategies. A further direction concerns the metrics. Counterfactual fairness comes with dedicated pair-level measures that we do not exploit here and that could be explored [45]. Moreover, per-gender F1 is a group-level measure and can look almost identical across genders while the underlying errors systematically move female referents down and male referents up the prestige scale. Computing on the Beliefs pairs how much the model's prediction shifts when only gender changes and in which direction with respect to prestige would quantify what we can currently only illustrate with examples. Finally, enlarging Beliefs and including larger open and commercial models would test how far the patterns we report generalise.

7. Limitations

A first limitation concerns the lack of systematic repetition of the experiments. Although generation itself was deterministic, the order of the candidate labels and, in the few-shot setting, of the demonstrations was randomised, while each configuration was evaluated only once. Repeating the experiments across multiple random seeds and reporting averaged results would provide more robust estimates of model performance and reduce the possible influence of prompt ordering. For computational reasons, this was not feasible within the present study.

A second limitation is the relatively small size of Beliefs, and of its Italian portion in particular. Building the dataset required a time-consuming process of translation, selection, and manual revision

aimed at ensuring that both gender and occupation were always clearly identifiable. This choice favoured quality and control over the relevant variables rather than quantity, but resulted in a limited number of biographies per occupation (12 in the final balanced setting). Consequently, some of the observed differences between gender conditions may partly reflect sampling variability, and caution is therefore required when interpreting small performance gaps. Moreover, while we carried out a preliminary analysis, we did not systematically test the significance of all observed differences. Future work should therefore validate these patterns on larger samples and complement descriptive results with systematic statistical testing.

A further limitation concerns the translation process. Although the Italian biographies were manually revised to preserve their original content and to control gender-related information as closely as possible, translation cannot guarantee full equivalence between the English and Italian versions. Linguistic and cultural differences between the two languages may affect how occupations, personal attributes, or biographical information are expressed and interpreted.

The present study is also restricted to binary gender. This design choice was motivated by the structure of the source dataset and by the need to construct tightly controlled counterfactual pairs, but it excludes non-binary and gender-diverse identities, which may be subject to distinct and potentially stronger forms of discrimination in real-world applications. Similarly, our analysis isolates gender from other social dimensions and therefore does not address intersectional biases involving, for instance, ethnicity, age, or socioeconomic background. Extending the counterfactual framework to a broader range of gender identities and to intersectional settings represents an important direction for future work.

Another limitation concerns the set of occupations included in the dataset. The selected professions are skewed towards relatively high-status and highly skilled occupations and therefore do not fully represent the broader labour market. The patterns observed here may not generalise to lower-status, manual, or service-sector occupations, where gender stereotypes and their interaction with other social categories may take different forms. A broader and more diverse occupation inventory would make it possible to assess the extent to which the effects identified in this study hold across different sectors and levels of occupational prestige.

A further limitation concerns the evaluation of reasoning models. In particular, it remains unclear whether the poor performance observed for these models, especially in Italian, is primarily due to token-limit constraints. Experiments allowing larger token budgets could help determine whether the observed failures reflect genuine reasoning limitations or insufficient generation capacity, thereby enabling a more reliable assessment of their performance.

Finally, the range of models considered in this study is necessarily limited. Among open models, we focused primarily on medium-sized architectures because of computational constraints, which prevented us from systematically including larger models. We also did not evaluate the latest commercial systems, such as ChatGPT, Claude, or commercial DeepSeek models, because of their querying costs. Extending the analysis to larger open models and commercial systems would broaden the comparison and help determine whether the dynamics observed here also hold for the models most widely deployed in real-world applications.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and ChatGPT (OpenAI) in order to: improve writing style (English text revision and proofreading); formatting assistance (LaTeX tables and bibliography); and assistance with code writing for running the experiments. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).

[2] S. G. Haez, M. Dragoni, Neutral is not unbiased: evaluating implicit and intersectional identity bias in llms through structured narrative scenarios, Findings of the Association for Computational Linguistics: EMNLP 2025 (2025) 15060–15088.

[3] E. S. Ruzzetti, D. Onorati, L. Ranaldi, D. Venditti, F. M. Zanzotto, Investigating gender bias in large language models for the italian language, in: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), 2023, pp. 562–569.

[4] G. Giachino, M. Rondina, A. Vetrò, R. Coppola, J. C. De Martin, An empirical investigation of gender stereotype representation in large language models: The italian case, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2025, pp. 228–242.

[5] S. Rickman, Evaluating gender bias in large language models in long-term care, BMC Medical Informatics and Decision Making 25 (2025) 274.

[6] P. G. Devine, Stereotypes and prejudice: Their automatic and controlled components, Journal of Personality and Social Psychology 56 (1989) 5–18. doi:10.1037/0022-3514.56.1.5.

[7] A. G. Greenwald, M. R. Banaji, Implicit social cognition: Attitudes, self-esteem, and stereotypes, Psychological Review 102 (1995) 4–27. doi:10.1037/0033-295X.102.1.4.

[8] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (1974) 1124–1131. doi:10.1126/science.185.4157.1124.

[9] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, 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, 2019, pp. 120–128.

[10] G. G. Corbett, Gender, Cambridge University Press, 1991.

[11] A. Curzan, Gender shifts in the history of English, Cambridge University Press, 2003.

[12] S. Luraghi, O. Anna, et al., Linguaggio e genere., Carocci, 2006.

[13] A. Parziale, G. Voria, V. Pontillo, G. Catolino, A. De Lucia, F. Palomba, Toward systematic counterfactual fairness evaluation of large language models: The caffe framework, arXiv preprint arXiv:2512.16816 (2025).

[14] A. Currey, M. Nadejde, R. R. Pappagari, M. Mayer, S. Lauly, X. Niu, B. Hsu, G. Dinu, Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 4287–4299.

[15] G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1679–1684.

[16] R. Rudinger, J. Naradowsky, B. Leonard, B. Van Durme, 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), 2018, pp. 8–14.

[17] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, 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), 2018, pp. 15–20.

[18] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, S. Petrov, Measuring and reducing gendered correlations in pre-trained models, arXiv preprint arXiv:2010.06032 (2020).

[19] H. An, C. Baumler, A. Sancheti, R. Rudinger, On the mutual influence of gender and occupation in llm representations, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 1663–1680.

[20] F. Ducel, A. Névéol, K. Fort, " you'll be a nurse, my son!" automatically assessing gender biases in autoregressive language models in french and italian, Language Resources and Evaluation (2024) 1495–1523.
[21] L. Benedetto, A. Donvito, A. Lucchetti, A. Cappelli, P. Buttery, Beyond names: How grammatical gender markers bias llm-based educational recommendations, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 5648–5668.
[22] C. Casula, S. V. Salto, E. Leonardelli, S. Tonelli, Job unfair: An investigation of gender and occupational bias in free-form text completions by llms, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 22770–22788.
[23] O. Kononykhina, A.-C. Haensch, F. Kreuter, Mind the gap: Gender-based differences in occupational embeddings, in: Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 2025, pp. 83–91.
[24] D. Zhang, Y. Zhang, G. Bihani, J. Rayz, Hire me or not? examining language model's behavior with occupation attributes, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7891–7911. URL: https://aclanthology.org/2025.coling-main.529/.
[25] J. Yang, S. Njoto, M. Cheong, L. Ruppanner, L. Frermann, Professional presentation and projected power: A case study of implicit gender information in English CVs, in: D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, S. Volkova (Eds.), Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), Association for Computational Linguistics, Abu Dhabi, UAE, 2022, pp. 140–150. URL: https://aclanthology.org/2022.nlpcss-1.15/. doi:10.18653/v1/2022.nlpcss-1.15.
[26] M. Gupta, C. M. Parra, D. Dennehy, Questioning racial and gender bias in ai-based recommendations: do espoused national cultural values matter?, Information systems frontiers 24 (2022) 1465–1481.
[27] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg, Null it out: Guarding protected attributes by iterative nullspace projection, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7237–7256.
[28] D. Bressan, Gender-related ambiguity and stereotyping in english-italian, italian-english translation, Italian Studies in Southern Africa/Studi d'Italianistica nell'Africa Australe 12 (1999) 81–96.
[29] U. K. Nissen, Aspects of translating gender, Linguistik online 11 (2002).
[30] E. Vanmassenhove, J. Monti, gender-it: An annotated english-italian parallel challenge set for cross-linguistic natural gender phenomena, in: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, 2021, pp. 1–7.
[31] J. E. Font, M. R. Costa-Jussa, Equalizing gender bias in neural machine translation with word embeddings techniques, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 2019, pp. 147–154.
[32] A. Rescigno, J. Monti, et al., Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german, in: INTERNATIONAL CONFERENCE Human-informed Translation and Interpreting Technology (HiT-IT 2023) Proceedings, Incoma Ltd., 2023, pp. 1–11.
[33] S. A. Piazzolla, B. Savoldi, L. Bentivogli, Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems, HERMES-Journal of Language and Communication in Business (????) 209–225.
[34] F. Bianchi, T. Fornaciari, D. Hovy, D. Nozza, Gender and age bias in commercial machine translation, in: Towards responsible machine translation: Ethical and legal considerations in machine translation, Springer, 2023, pp. 159–184.
[35] S. A. Piazzolla, B. Savoldi, L. Bentivogli, Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems, HERMES - Journal of Language and Communication in Business (2023) 209–225. URL: https://tidsskrift.dk/her/article/view/137553.

doi:10.7146/hjlcb.vi63.137553.

[36] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).

[37] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[38] F. A. Galatolo, M. G. Cimino, Cerbero-7b: A leap forward in language-specific llms through enhanced chat corpus generation and evaluation, arXiv preprint arXiv:2311.15698 (2023).

[39] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[40] A. G. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[41] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al., Deepseek-r1 incentivizes reasoning in llms through reinforcement learning, Nature 645 (2025) 633–638.

[42] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[43] A. Piergentili, D. Fucci, B. Savoldi, L. Bentivogli, M. Negri, Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges, in: Proceedings of the First Workshop on Gender-Inclusive Translation Technologies, 2023, pp. 71–83.

[44] A. Piergentili, B. Savoldi, M. Negri, L. Bentivogli, Gender-neutral rewriting in italian: Models, approaches, and trade-offs, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025, pp. 899–910.

[45] V. Steinborn, P. Dufter, H. Jabbar, H. Schuetze, An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 921–932.