

Gender biases in GPT-4 short biographies.

A corpus study on Italian and French anthroponyms*

Anna-Maria De Cesare (Dresden)

Abstract

As has been shown in various studies considering different languages, in professional contexts women tend to be referred to differently than men. While men are typically referred to by their surname (e. g., *Fermi*), women are more often referenced with their full name (e. g., *Samantha Cristoforetti*) or first name alone (e. g., *Samantha*). The present study proposes an empirical case study investigating whether this gender-indexing bias is also present in texts generated by large language models (LLMs). Based on the analysis of a self-assembled data collection comprising 420 biographies produced by GPT-4 on 140 eminent Italian and French female and male personalities, our study reveals that the synthetic texts investigated not only reflect the gender biases found in human-authored texts but, in some cases, even amplify them.

1 Introduction

The literature investigating gender¹ biases in human-authored texts written in Italian (e. g., Sabatini 1993; Burr 1997) and English (cf. Atir/Ferguson 2018) shows a common trend: in professional contexts, women and men tend to be referenced with different forms of anthroponyms. While men are commonly referred to by their surname (e. g., *Fermi*, *Gassman*, *Obama*), women are most often mentioned either by their full name (e. g., *Marie Curie*, *Sophia Loren*, *Hillary Clinton*) or by their first name (e. g., *Marie*, *Sophia*, *Hillary*). In other words, men are indexed by a gender-neutral form of anthroponym, not revealing the sex of the referent, while women are indexed by gender-marking anthroponyms. In Italian and English, as well as in many other European languages, including French, most of the first names denote the sex of the person (as, e. g., in *Sophia* and *Vittorio*).

* I wish to thank two anonymous reviewers for their valuable comments and suggestions. A special thank you also to Tom Weidensdorfer for creating the text samples and to Luisa Burchardt and (again) Tom Weidensdorfer, respectively, for identifying the anthroponyms in the BioWiki and BioGPT-4 datasets (the relevant material used in this study is available in De Cesare/Weidensdorfer/Burchardt 2025).

¹ In this contribution, the term *gender* is used to refer to the biological sex of specific individuals and comprises exclusively two categories: females and males. This does not imply that the concept of *gender* is dichotomic and that there only are two biological sexes (for a detailed definition of *gender*, cf. e. g. Fusco 2024). Including other genders in the discussion would be important but given that other parameters would have to be considered, we leave the investigation of anthroponyms related to intersex and non-binary individuals for future research.

The goal of this contribution is to investigate whether this gender-indexing bias in the use of anthroponyms also occurs in texts generated by Large Language Models (henceforth: LLM). In order to shed light on this question, we propose a case study examining how males and females are referenced in short biographies generated by GPT-4 (an LLM developed by OpenAI) in two Romance languages: Italian and French. Considering these two Romance languages at once allows us to investigate the presence of cross-cultural differences and similarities in referring to eminent personalities and highlight common sexist cultural habits to mitigate or solve. From a linguistic point of view, both languages display the three anthroponomic forms illustrated above, *viz.* <surname>, <first name + surname> (i. e. full name) and <first name>. In addition to these three forms, Italian also refers to females with the gender-indexing pattern <fem. definite article + surname>, as e. g. in *la Loren* <the_{FEM.SING} Loren>. In written contemporary Italian, however, this pattern tends to be avoided in gender-sensitive communication (D'Achille 2016: 176) and previous research found that it occurs less often in GPT-3.5 short biographies than in Wikipedia entries (cf. De Cesare 2023 and 2026). In this study, this pattern will not be considered.

The present contribution is organized in five parts. Section 2 defines the category of anthroponyms from a formal point of view and presents the factors allowing to explain the choice of one form over the others, paying special attention to gender-related biases; importantly, besides outlining our research questions, it also provides new data from Wikipedia entries. Section 3 describes the BioGPT-4 data collection, a self-assembled corpus of short biographies generated in Italian and French by GPT-4 in January 2024. Section 4 presents the main quantitative and qualitative results obtained in relation to three parameters (data generation success rate, distribution of anthroponyms across the four components of the dataset and contextual preference in the use of anthroponyms corresponding to first names). Finally, Section 5 answers the research questions, highlighting how GPT-4 generated biographies mirror and even enhance gender biases present in human-authored texts. It also suggests some mitigation solutions and follow up case studies to deepen our understanding of gender-associated biases in LLM generated texts.

2 Anthroponyms: forms, (new) data and research questions

2.1 Form of anthroponyms and gender biases

An anthroponym is a person's name and belongs to the morphosyntactic class of proper nouns. It serves as a label to “identify one individual in relation to all the others” (Marcato 2010: 85; transl. mine). In semantic terms, anthroponyms have an inherently referential value (cf. Marcantonio 1988: 328). This definition applies to all forms of anthroponyms listed and illustrated in Table 1: Type I, corresponding to a full name (e. g., *Maria Monti/Mario Monti*); Type II, corresponding to a surname (e. g., *Monti*); and Type III, corresponding to a first name (e. g., *Maria/Mario*). All these forms of anthroponyms are commonly used across Western languages (and cultures), including Italian and French.

Types	FEMALE	MALE
I	Maria Monti	Mario Monti
II	Monti	Monti
III	Maria	Mario

Table 1: Types of anthroponyms referring to females and males

According to Marcantonio (1988), anthroponyms do not convey any descriptive content. Thus, in contrast to definite descriptions, such as the noun phrase *the author of Waverley*, the proper noun *Walter Scott* does not express any property characterizing the referent (Marcantonio 1988: 329). At the same time, one could argue that anthroponyms revealing the sex of the referent, as is commonly the case when the first names are present (e. g., *Maria* (female), *Mario* and *Walter* (males)), do in fact convey a referential feature.

In addition to referring to the sex of the referent, the form of an anthroponym can convey important social information, such as the degree of acquaintance (closeness vs. distance) between the speaker/writer and the person who is referred to by the anthroponym. Anthroponyms of Type III clearly denote social closeness, while anthroponyms of Type I and II denote social distance. The choice of one form of anthroponym over another also depends on other factors, such as the degree of accessibility of the referent (cf. Ariel 1990) and the degree of formality of the text (for details on the interplay between the form of an anthroponym and referential accessibility, see the case study presented in De Cesare 2025 on a small dataset of short biographies generated by GPT-4).

Another important factor involved in the choice of an anthroponym is the gender of the referent. As has been shown in several studies (focusing on different languages, empirical data and methodologies), females and males tend not to be referenced in the same way. Based on an extensive empirical study of Italian newspaper articles published in 1989, Burr (1997) shows that men tend to be referred to by their surname alone (55,76%), while women are referenced either with their full name or with their first name alone (85%). In other words, men are associated to the one “gender-neutral” form of anthroponym (Type II), women to the two gender-indexing forms of anthroponyms (Types I and III). Similar results were obtained by Sabatini (1993) based on a smaller corpus of Italian newspapers published several years earlier.

Similar results were also found across several experimental studies conducted on English data comparing female and male professionals (cf. Atir/Ferguson 2018). Overall, the research findings converge in showing that, when referring to professionals working in high-status fields (specifically STEM), literature and sports, male tend to be referenced with their surname (Type II), while women tend to be referred to either with their full name (Type I) or with their first name alone (Type III). Based on different linguistic features (in addition to anthroponyms, mention should be made to occupational names and the use of adjectives), many studies in Natural Language Processing (NLP) and social sciences show that LLMs do not only mirror the gender biases present in the text on which they have been trained, but sometimes even enhance and amplify them (cf. Bolukbasi et al. 2016; Nemani et al. 2024; Nghiem et al. 2024; Leong/Sung 2024; Wan et al. 2023).

Most of the studies available to date focus on English data. There are, however, some studies on other languages (cf. De Cesare 2023 on Italian; Garrido-Muñoz/Martínez-Santiago/Montejo-Ráez 2023 on Spanish; and Ding et al. 2025 on French, Spanish, Chinese, Japanese and Korean). To the best of my knowledge, there is currently no comparative study on Italian and French investigating the form and uses of anthroponyms in the output of LLM-generated texts, despite the fact that LLMs are known to reproduce the biases included in their training data or in the data added in the post-training and fine-tuning phases (for an overview of the typologies of biases associated to LLMs, in particular of gender-related asymmetries, cf. *inter alia* Ferrara 2023; Navigli/Conia/Ross 2023; and Nemani et al. 2024).

2.2 New data and research questions

In order to gain insights on the recent use of female and male anthroponyms in Italian and French, it is necessary to consider results from newer data, comparable to GPT-4 short biographies. I therefore conducted a first case study on 140 Wikipedia entries, including the biography of 140 eminent personalities (35 Italian women, 35 Italian men, 35 French women and 35 French men), active in different domains: arts, science, entertainment, sports, politics etc.² (the full list of personalities is provided in De Cesare/Weidensdorfer/Burchardt 2025). This case study is based on four text datasets (see Table 2) created with Trafilatura, a software that can automatically extract text from the web (cf. Barbaresi 2021). To ensure maximum comparability with the GPT-4 generated biographies, several contents typical of Wikipedia entries were discarded from the BioWiki files (e. g., endnotes, filmography or complete list of works of the personality described in the entry, references and external links).

A semi-manual³ search for Type I, II and III anthroponyms in the BioWiki data collection yielded the quantitative results shown in Table 2. Note that only the anthroponyms referring to the personality at the center of the biography were counted, excluding references to family members or societies named after the personality in question.

² There is no balance between gender and domain of activity (e. g., we did not select the same number of female and male sportspersons); in addition, in creating the list of personalities, we did not pay attention to the representation of each gender in each domain as this is not directly relevant to our research questions.

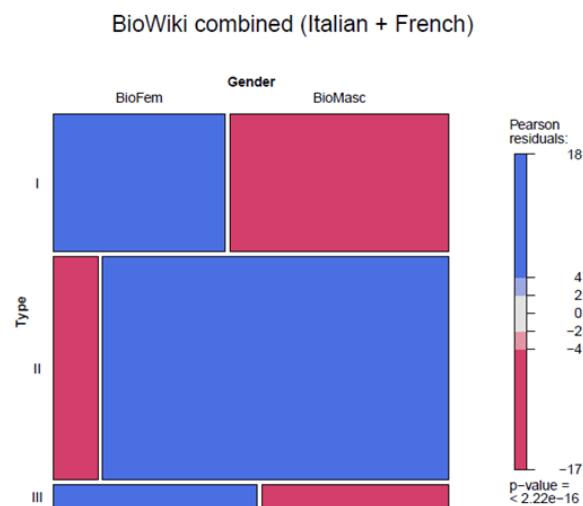
³ The anthroponyms were identified using different querying techniques: (i) searching each one of the 140 (last) names in the corpus uploaded on the Sketch Engine platform and (ii) searching each one of the 140 (last) names in a word document. These very time-consuming tasks have the advantage of yielding very solid results (i. e. of identifying all the forms correctly, without miscounting false positives and negatives).

	BioFemWiki		BioMascWiki	
Types	IT	FR	IT	FR
I	281	753	291	1034
II	126 ⁴	309	1472	1936
III	80	137	94	105
Tot.	591	1199	1857	3075
	1790		4932	

Table 2: Absolute frequency of anthroponyms

The most important quantitative result shown in Table 2 concerns the wide gap between the total number of occurrences across the BioFemWiki and BioMascWiki datasets: women are mentioned 1790 times, while men are referred to more than double the number of times (4932 occ.). Moreover, women (regardless of the language) are referred to less often than men with all Types of anthroponyms (I, II and III), although Italian female and male show a similar number of occurrences for Type I (281 vs. 291 occ.) and Type III (80 vs. 94 occ.). Finally, there is a similar ranking in the frequency of use of the three types of anthroponyms within each gender: Italian and French women are referred to most often with Type I, then Type II and lastly with Type III; Italian and French men, by contrast, are overwhelmingly referenced with Type II, then Type I and lastly with Type III.

To examine if the distribution of anthroponym types differs systematically between female and male in the BioWiki dataset, a chi-square test of independence was conducted. Effect sizes were interpreted based on the thresholds for Cramér's V proposed by Lee (2016), where values of 0.10, 0.30, and 0.50 represent weak, moderate, and strong associations, respectively. Results are presented Figure 1.

**Figure 1: Types of anthroponym and Gender in BioWiki**

⁴ This figure does not include the cases (104 occ.) in which the Surname is preceded by the definite article (e. g., *la Loren*).

The chi-square test shows a significant association between anthroponym type (I, II, III) and gender in the combined Italian and French BioWiki dataset (χ^2 (df = 2, N = 6651) = 980.38, $p < .001$, Cramér's $V = 0.39$). The distribution of anthroponym types differs significantly between female and male biographies. The effect size indicates a moderate association between type and gender.

The colored mosaic plot in Figure 1 shows that the association strength between anthroponymic type and gender is specular in the two datasets. In the BioMasc dataset, there is a very strong positive association with Type II (surname), while there is a very strong negative association of the same type in the BioFem dataset. Conversely, there is a very strong negative association of Type I (full name) and Type III (first name) in the BioMasc dataset, while there is a very strong positive association of both types of anthroponyms in the BioFem dataset. In the sample of Wikipedia entries, female thus tend to be referred to with gender-explicit forms of anthroponyms, while males tend to be referred to with the gender-neutral form. These findings perfectly mirror the results obtained in previous studies examining data from human-authored texts (see Section 2.1).

Based on the results obtained in the BioWiki data reported above, the research questions to which this empirical case study seeks to provide some answers are the following:

1. With what forms of anthroponyms (Type I, Type II and/or Type II) are women and men referred to in GPT-4 short biographies? In high status professional contexts, are women referred to differently than men?
2. Are there any meaningful cross-linguistic differences in the use of female and male anthroponyms? How do we explain them?
3. Do GPT-4 short biographies mirror or even enhance the gender biases present in human-authored texts, in particular in Wikipedia entries? If so, what can be done to mitigate the observed biases?

3 The BioGPT-4 datasets

To answer the research questions listed above, we used a self-assembled collection of GPT-4 generated texts representing the genre 'short biography', which belongs to the macro-category of 'informative texts' (according to Werlich's 1975 text types). We chose to investigate this specific text genre for three main reasons: a) biographies typically include many tokens of anthroponyms, in particular referring to the individual at the center of the biography;⁵ b) this procedure allows us to easily select and analyze the profile of eminent personalities; and c) the data on which GPT-4 is trained includes a large amount of Wikipedia entries (cf. Navigli/Conia/Ross 2023: 104; Baack 2024), one fifth of which correspond to biographies (cf. Flekova/Ferschke/Gurevych 2014: 855).⁶

⁵ From a semantic point of view, this study only considers specific individuals (i. e. known referents); for a semantic categorization of discourse referents, cf. De Cesare/Giusti (2024: 7f.).

⁶ We do not have information on the training data of GPT-4; however, we know that GPT-3 was trained on a dataset comprising 3% of Wikipedia pages (cf. Brown et al. 2020: 9; for details, see Figure 6 in Zhao et al. 2023: 17).

In this case study, we used the same list of 140 personalities described above (see Section 2.2). All profiles respond to two objective criteria: (i) they concern well-known and eminent personalities (mostly) born in the 20th century; and (ii) there is a matching biography written in the respective language (Italian and French) on Wikipedia; this guarantees – at least to a certain extent – sufficient publicly available material to generate a successful biography of the personality in question. As we will point out below, in some cases an English Wikipedia entry is sometimes also available.

Our self-assembled collection of short biographies was generated by GPT-4, a closed commercial LLM released in March 2023.⁷ Specifically, we used the model gpt-4-0613, trained on texts collected on the internet up to September 2021. As mentioned in footnote 6, we do not know the exact composition of the training data, nor how old the texts are, but we can assume that, besides Wikipedia entries, the datasets include digitalized books, newspapers, magazines, as well as contents from social media (see Brown et al. 2020: 9 in relation to GPT-3). We also know that multilingual LLMs, such as GPT-4, are trained on datasets in which English is over-represented. There is thus a clear “English-language bias” (Ferrara 2023: 5). GPT-3, for instance, was trained on a dataset including 93% of English words, 1,8% of French words and 0,6% of Italian words. Noteworthy is also that both Romance languages are among the top five languages used to train GPT-3 (for details on these figures, see Johnson et al. 2022: 3).

The biographies were generated in January 2024 accessing GPT-4 with an API key⁸ and using a python script.⁹ More specifically, we used a two-level, zero-shot prompting technique, a type of instruction that does not provide any examples of the desired output (for details, cf. e. g. Chen et al. 2023). As can be seen in Table 3, the second prompt is the most important one: it identifies the text genre (biography) and includes a gap corresponding to the pattern <first name + surname>. In each prompt, the gap is filled with one Italian or French female or male an-throponym (e. g., *Scrivi una biografia su Dacia Maraini* ‘Write a biography on Dacia Maraini’).

Prompting level	Prompt
System prompt	Il tuo compito è generare testi informativi. ‘Your task is to generate informative texts.’
User prompt	Scrivi una biografia su {nome + cognome}. ‘Write a biography on {first name + surname}.’

Table 3: Prompting technique used to generate GPT-4 short biographies

Each prompt was used three times for the same personality. This multiple prompting technique allows evaluating the stability of the output. Also note that the temperature hyperparameter is

⁷ At the time this study was conducted, OpenAI’s GPT-4 was one of the latest and best LLMs available. Today, many other LLMs have been released (cf. Chiang et al. 2024), including GPT4o, GPT-4.5 (OpenAI), Mistral Le Chat (Mistral AI), DeepSeek R1 (DeepSeek), Teuken7B (OpenGPT-X).

⁸ API, which stands for “Application Programming Interface Key”, is a unique code informing the server who you are and that you want to communicate with the LLM.

⁹ I wish to thank Valerie Hekkel for writing the python script.

set to the default value of 0.7. This provides a balance between randomness and determinism in the generated text. Table 4 presents an overview of the internal composition of our data collection, which comprises again four datasets.

BioFem_IT	BioFem_FR	BioMasc_IT	BioMasc_FR
35 x 3 Italian females	35 x 3 French females	35 x 3 Italian males	35 x 3 French males
70 female personalities/210 texts		70 male personalities/210 texts	
140 personalities/420 outputs			

Table 4: Biographies generated with GPT-4: data collection

With the help of the Sketch Engine corpus enquiry platform (cf. Kilgarriff et al. 2014), we calculated the size of each dataset in terms of the total number of words. As can be observed in Table 5, the GPT-4 datasets show an unexpected asymmetry in the average text length across genders: the BioMasc dataset includes ca. 7000 more words than the BioFem dataset. As we will explain in Section 4, this asymmetry is due to “failed outputs” in the Italian and French BioFem datasets.

BioFem_IT	BioFem_FR	BioMasc_IT	BioMasc_FR
31'809	34'079	35'646	37'606
65'888		73'252	
139'140			

Table 5: Size (n. of words) of the GPT-4 data collection

4 Results and discussion

4.1 Generation success rate

A notable finding concerns the asymmetry in the success rate of biographies generated across the BioFem and BioMasc datasets, i. e., in the actual number of biographies generated on the female and male personalities. As shown in Table 6, the success rate for male biographies is consistently 100% both in Italian and French. In contrast, female biographies show lower success rates, with 94% for French women and 89% for Italian women. This discrepancy highlights a distinct form of gender bias associated to the two datasets.

GPT-4 datasets	Success rate	Success rate in %
BioFem_IT	93/105	89%
BioFem_FR	99/105	94%
BioMasc_IT	105/105	100%
BioMasc_FR	105/105	100%

Table 6: Generation success rate (105 = 35 personalities x 3 texts)

Biographies were never generated on four Italian women (Clara Lollini, Anna Marabini, Emma Strada, and Elisabetta Vignotto) and two French women (Anne Chopinet and Jeanne Lévy).

Interestingly, as shown in the outputs reproduced in (1) and (2), in these “failed biographies” various factors are mentioned to explain why the task could not be performed.

- (1) Mi dispiace, non posso fornire una biografia su Anna Marabini perché non ci sono informazioni pubblicamente disponibili su questa persona. Potrebbe essere un nome comune o una persona non pubblica. Per tutelare la privacy di tutti, solo le informazioni su individui altamente pubblici, come celebrità, politici o persone rilevanti in termini storici, sono generalmente disponibili e possono essere discusse. Se stai cercando informazioni su un individuo noto con un nome diverso, fornitemelo e sarò lieto di aiutarvi.
 ‘I am sorry, I cannot provide a biography on Anna Marabini because there is no publicly available information about this person. It could be a common name or a non-public person. To protect everyone's privacy, only information about highly public individuals, such as celebrities, politicians, or historically relevant people, is generally available and can be discussed. If you are looking for information on a known individual with a different name, please provide it and I will be glad to help you.’¹⁰
- (2) Je suis désolé, mais je ne trouve aucune information sur une personne nommée Anne Chopinet. Il est possible que vous ayez mal orthographié le nom ou que cette personne ne soit pas suffisamment connue ou publique pour avoir des informations disponibles. Si vous avez plus de détails qui pourraient aider à identifier la personne, j'apprécierais vraiment.
 ‘I’m sorry, but I can't find any information about a person named Anne Chopinet. It is possible that you have misspelled the name or that this person is not well known or public enough to have any information available. If you have more details that might help identify the person, I would really appreciate it.’

The two main reasons provided in the output – lack of publicly available information on the person and misspelled name (the second reason is only provided for Anne Chopinet) – are both incorrect. For each one of the 140 personalities included in our list, there is publicly available information, at least in the form of a Wikipedia entry in Italian and French, respectively (see the selection criterion in Section 3). In two cases (Emma Strada and Elisabetta Vignotto), there is even a Wikipedia entry in English dating back to 2020 and 2021, respectively (remember that GPT-4 was trained on texts collected on the internet up to September 2021). Noteworthy is also the claim in the third sentence of (1): “only information about highly public individuals, such as celebrities, politicians, or historically relevant people, is generally available and can be discussed”. This is of course also true for Anna Marabini, an influential and internationally renowned Italian chemist, active in the second half of the 20th c.

Examples in (1) and (2) also show that “failed biographies” consist of very short texts. While successful biographies comprise between 300 and 500 words (see ex. 3 below), failed biographies only include around 50–80 words.

- (3) Alda Merini è stata una delle voci poetiche più riconosciute e apprezzate del panorama letterario italiano.
 Nata a Milano il 21 marzo 1931, Alda crebbe in una famiglia modesta. Fin da piccola mostrò un grande interesse per la scrittura e la poesia, ogni volta che poteva, si dedicava con passione a leggere testi di poeti come Rilke, Montale, Ungaretti e altri autori del ‘900. Tuttavia, non poté

¹⁰ When not indicated otherwise, the example is translated in English with the help of DeepL.com (free version).

frequentare il liceo a causa delle difficoltà economiche della sua famiglia e fu costretta ad interrompere gli studi dopo la scuola media.

La sua carriera poetica iniziò molto presto, attirando l'attenzione di scrittori noti come Giorgio Manganelli e Pier Paolo Pasolini, che predissero il suo straordinario talento. Merini pubblicò la sua prima raccolta di poesie, "La presenza di Orfeo", nel 1953.

Merini fu affetta da problemi di salute mentale che l'avrebbero seguita per tutta la vita, con periodi di ricovero in manicomio. Le sue battaglie con la malattia mentale influenzarono in modo significativo la sua poesia, arricchendola di introspezione, sensibilità e una visione unica della realtà. Uno dei suoi libri più noti, "La Terra Santa", racconta proprio la sua esperienza in ospedale psichiatrico.

Alda Merini tornò sotto i riflettori della critica e del pubblico alla fine degli anni '80, quando l'editore Scheiwiller pubblicò una raccolta di sue poesie. Da allora, la sua fama crebbe sempre di più, tanto da essere proposta come candidata per il Premio Nobel per la Letteratura nel 1996. Le sue opere, contraddistinte da un linguaggio essenziale, diretto ed evocativo, parlano d'amore, di sofferenza, di follia e di fede, temi che resteranno centrali in tutte le sue poesie. Alda Merini è stata anche autrice di aforismi, passati alla storia per il loro significato profondo e la loro bellezza.

Alda Merini morì a Milano il 1 novembre 2009. È sepolta nel cimitero monumentale della città, luogo che amava profondamente e che ha spesso celebrato nei suoi versi. La sua morte rappresenta una grande perdita per il panorama culturale italiano, ma il suo importante contributo alla poesia e alla letteratura rimane indelebile. [341 words]

'Alda Merini was one of the most recognized and appreciated poetic voices in Italian literature. Born in Milan on March 21, 1931, Alda grew up in a modest family. From an early age, she showed a keen interest in writing and poetry, and whenever she could, she devoted herself passionately to reading the works of poets such as Rilke, Montale, Ungaretti, and other 20th-century authors. However, she was unable to attend high school due to her family's financial difficulties and was forced to interrupt her studies after middle school.

Her poetic career began very early, attracting the attention of well-known writers such as Giorgio Manganelli and Pier Paolo Pasolini, who predicted her extraordinary talent. Merini published her first collection of poems, "La presenza di Orfeo", in 1953.

Merini suffered from mental health problems that would follow her throughout her life, with periods of hospitalization in mental institutions. Her struggles with mental illness significantly influenced her poetry, enriching it with introspection, sensitivity, and a unique vision of reality. One of her best-known books, "La Terra Santa", recounts her experience in a psychiatric hospital.

Alda Merini returned to the spotlight of critics and the public in the late 1980s, when the publisher Scheiwiller published a collection of her poems. From then on, her fame grew steadily, so much so that she was nominated for the Nobel Prize for Literature in 1996.

Her works, characterized by an essential, direct, and evocative language, speak of love, suffering, madness, and faith, themes that remain central to all her poems. Alda Merini was also the author of aphorisms, which have gone down in history for their profound meaning and beauty. Alda Merini died in Milan on November 1, 2009. She is buried in the city's monumental cemetery, a place she loved deeply and often celebrated in her verses. Her death represents a great loss for the Italian cultural scene, but her important contribution to poetry and literature remains indelible.'

Considering that the BioFem datasets include 18 failed biographies (6 personalities x 3 generated texts), we can estimate that this sample is made of ca. 1260 words (18 texts x 70 words),

instead of the 9000 words corresponding to the successful biographies (18 texts x 500 words). The difference between the two samples (failed vs successful biography) explains the gap of ca. 7000 words between the BioFem and the BioMasc datasets observed based on Table 5.

4.2 Form of anthroponyms: differences across genders and languages

A semi-manual search (similar to the one described in Section 2.2) for Type I, II and III anthroponyms in the BioGPT-4 data collection yielded the quantitative data shown in Table 7. Here, too, we only considered anthroponyms referring to the personality at the center of the biography, excluding references to family members or societies named after the personality in question (as, e. g., in *Fosco Maraini, la famiglia Maraini* ‘the Maraini family’, *Metodo Montessori* ‘Montessori method’).

Types	BioFemGPT-4		BioMascGPT-4	
	IT	FR	IT	FR
I	282	328	243	274
II	298 ¹¹	203	573	508
III	52	35	23	8
Tot.	650	566	839	790
	1216		1629	

Table 7: Absolute frequency of anthroponyms

The last line of Table 7 shows a gap between the BioFemGPT-4 and BioMascGPT-4 datasets: there are 1216 female anthroponyms and 1629 male anthroponyms. This can be explained by the fact that the BioFem datasets include 18 failed biographies, in which there is only one or two mentions of the personality in question (as shown in the supplementary data provided in De Cesare/Weidensdorfer/Burchardt 2025, we did not count these mentions), instead of the roughly 20 mentions occurring in a successful biography (which would amount to roughly 360 anthroponyms).

The most important quantitative result shown in Table 7 concerns the different distribution of the three forms of anthroponyms across the BioFem and BioMasc datasets: women (regardless of the language) are referred to more often than men with Types I and III, while men are referred to more often than women with Type II. Moreover, there is a similar ranking in the frequency of use of the three types of anthroponyms within each gender: women are referred to most often with Type I (note that in BioFemIT the data is very similar for Types I and II), then Type II and lastly with Type III; men with Type II, then Type I and lastly with Type III.

Running the chi-square test on the data in Table 7 shows a significant association between anthroponym type and gender in the combined Italian and French BioGPT-4 dataset (χ^2 (df = 2, N = 1830) = 185.49, $p < .001$, Cramér’s V = 0.26). The distribution of anthroponym types differs significantly between female and male biographies. The effect size indicates a moderate association between type and gender. More details are provided in Figure 2.

¹¹ For Italian Type II anthroponyms, we did not include the 18 cases in which the surname is preceded by a definite article, e. g. *la Meloni* (for details, cf. De Cesare 2023).

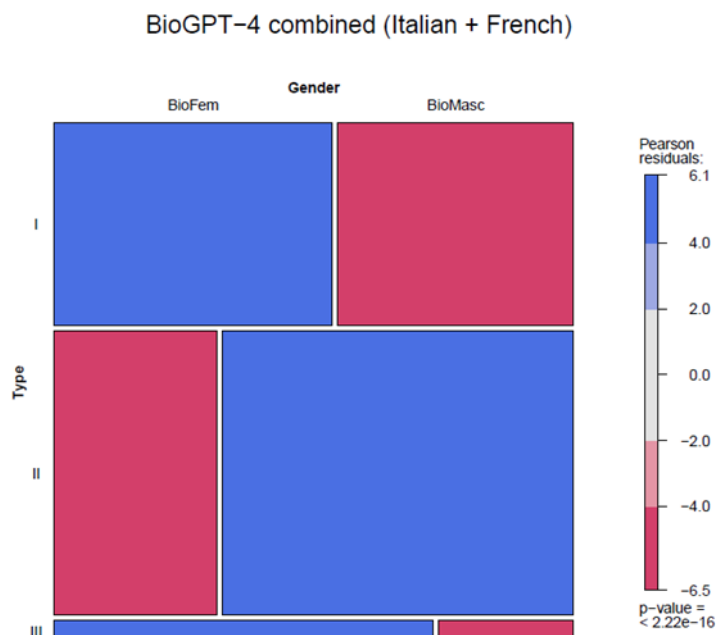


Figure 2: Types of anthroponym and gender in BioGPT-4

Interestingly, the colored mosaic plot in Figure 2 shows that the results perfectly align with the ones found in the Wikipedia entries (see Figure 1 in Section 2.2). In the BioMasc dataset, there is a very strong positive association with Type II (surname), while there is a very strong negative association of the same form in the BioFem dataset. Conversely, there is a very strong negative association of Type I (full name) and Type III (first name) in the BioMasc dataset, while there is a very strong positive association of both types of anthroponyms in the BioFem dataset. Thus, exactly as in the case of the BioWiki datasets, in the sample of BioGPT-4 short biographies, female tend to be referred to with gender-explicit forms of anthroponyms, while males tend to be referred to with gender-neutral forms.

If we consider the results across Italian and French (in Figures 3 and 4, respectively), no significant difference can be identified, except for the fact that the strengths of association between types and gender are weaker in the Italian than in the French dataset.

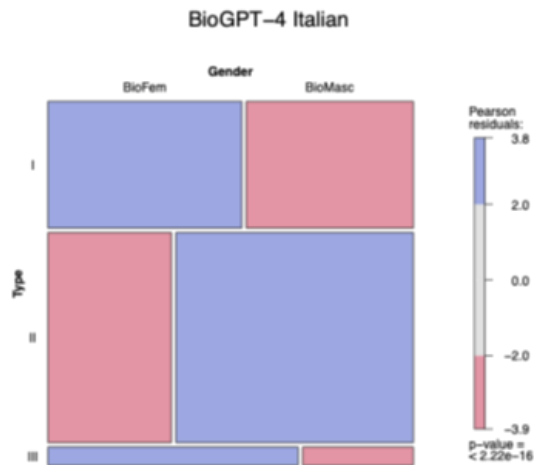


Figure 3: Types of anthroponym and gender
(Italian Dataset)

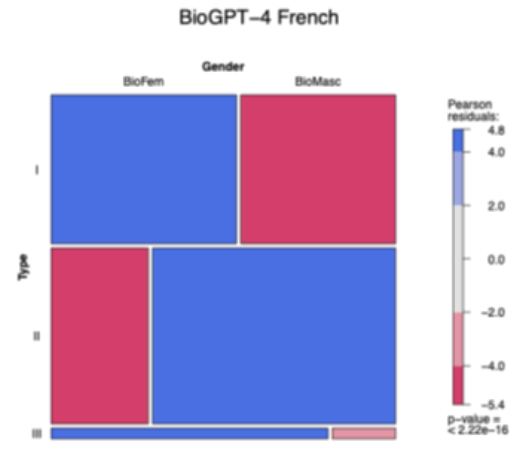


Figure 4: Types of anthroponym and gender
(French Dataset)

4.3 A closer look at Type III anthroponyms

Beyond their frequency (87 occ. in the BioFem datasets and 31 occ. in the BioMasc datasets), a closer look at Type III anthroponyms in the GPT-4 short biographies reveals an additional form of asymmetry, related to (i) their dispersion and (ii) their internal text distribution.

As far as their dispersion is concerned, Table 8 allows to highlight an important discrepancy between the BioFem and BioMasc datasets. The use of anthroponyms of Type III is much more spread in the BioFem datasets: we find them in 31 female biographies, as opposed to 17 male biographies (the figure in parentheses besides each name refers to the number of occurrences of Type III anthroponym). Interestingly, some profiles are referenced more often than others with the first name.

	BioFem: 87 occ./31 biographies
BioFemIT (52 occ./17 biographies)	Lina Bo Bardi (7*); Sophia Loren (6); Natalia Ginzburg (5); Dacia Maraini (5); Rita Levi-Montalcini (5); Samantha Cristoforetti (4); Paola Cortellesi (3); Giuliana (Cavaglieri) Tesoro (3); Elsa Morante (3); Liliana Segre (3); Rosy Bindi (2); Margherita Hack (1); Anna Magnani (1); Alda Merini (1); Maria Luisa Spaziani (1); Oriana Fallaci (1); Antonietta Brandeis (1)
BioFemFR (35 occ./14 biographies)	Irène Joliot-Curie (9); Jeanne Hébuterne (6); Denise Grey (3); Juliette Gréco (3); Simone Veil (3); Édith Piaf (2); Simone Signoret (2); Françoise Dolto (1); Claudie Haigneré (1); Françoise Sagan (1); Marie Marvingt (1); Sarah Bernhardt (1); Simone de Beauvoir (1); Suzanne Lenglen (1)
	BioMasc: 31 occ./17 biographies
BioMascIT (23 occ./11 biographies)	Ettore Bugatti (9); Alberto Ascari (3); Terence Hill (2); Adriano Celentano (2); Guglielmo Marconi (1); Luciano Pavarotti (1); Reinhold Messner (1); Luis Trenker (1); Alberto Tomba (1); Bruno Pontecorvo (1); Renato Guttuso (1)
BioMascFR (8 occ./6 biographies)	Marcel Marceau (3); André Citroën (1); Jean Dausset (1); Gérard Depardieu (1); Jean-Paul Belmondo (1); Louis de Broglie (1)

Table 8: Dispersion of Type III anthroponyms (*n. of occurrences)

As for the second issue, in the BioMasc datasets, first names tend to appear in textual spaces associated with the personal sphere of the individual at the center of the biography. As illustrated in (4), the singer Adriano Celentano is referred to by his first name in a sentence of the biography describing his marital status and family members. By contrast, when the biography describes the education (in particular at the university level) and professional career of a male, as in (5), the individual's surname is used:

- (4) **Adriano** è sposato con Claudia Mori dal 1964, famosa cantante e attrice anche lei, con la quale ha tre figli: Rosita, Rosalinda e Giacomo. La coppia ha sempre mantenuto un profilo basso e mantenuto un forte legame familiare.
'**Adriano** is married to Claudia Mori since 1964, also a famous singer and actress, with whom he has three children: Rosita, Rosalinda and Giacomo. The couple has always kept a low profile and maintained strong family ties.'
- (5) Enrico Fermi¹², noto fisico, professore e uno dei principali contributori alla scoperta della fissione nucleare e al progetto Manhattan, è nato a Roma, in Italia, il 29 settembre del 1901. Mostrando una notevole abilità per le materie scientifiche fin dalla sua giovinezza, **Fermi** ha poi proseguito la sua istruzione presso l'Università degli Studi di Pisa dove ha ottenuto un dottorato in fisica nel 1922.
'Enrico Fermi, a noted physicist, professor and a major contributor to the discovery of nuclear fission and the Manhattan Project, was born in Rome, Italy, on September 29, 1901. Showing a remarkable ability for scientific subjects from his youth, **Fermi** went on to further his education at the University of Pisa, where he earned a doctorate in physics in 1922.'

In the BioFem datasets, by contrast, women are referred to by their first name not only in sections describing their childhood and family, as illustrated in (6), but also in professional contexts, particularly in passages detailing their careers and achievements, as shown in (7):

- (6) Samantha Cristoforetti è una figura di notevole rilevanza nel campo dell'esplorazione spaziale. Il suo operato come astronauta europea e il suo ruolo in diverse missioni spaziali la rendono uno dei principali volti dell'agenzia spaziale europea (ESA).
Nata a Milano il 26 aprile 1977, **Samantha** ha trascorso la sua infanzia tra Malè, in Trentino, e Monaco di Baviera. Dopo il diploma al Liceo Scientifico, ha conseguito una laurea in ingegneria meccanica presso l'Università Tecnica di Monaco in Germania nel 2001. [...]
'Samantha Cristoforetti is a notable figure in the field of space exploration. Her work as a European astronaut and her role in several space missions make her one of the main faces of the European Space Agency (ESA).
Born in Milan, Italy, on April 26, 1977, **Samantha** spent her childhood between Malè, Trentino, Italy, and Munich, Bavaria. After graduating from high school, she earned a degree in mechanical engineering from the Technical University of Munich in Germany in 2001.'

¹² All biographies, regardless of gender, start with the full name of the person (e. g. *Enrico Fermi*). This is a fixed pattern. In the datasets analyzed, there is an asymmetry in the use of the full name in other textual spaces: female full names also appear in paragraph initial or even in paragraph internal positions. In the latter case, the female referent is generally highly accessible (cf. Ariel 1990) and the full name would not be necessary. This leads to a phenomenon of over-codification of the referent (for details related to empirical data drawn from GPT-4 Italian short biographies, cf. De Cesare 2025).

- (7) Contrariamente alle aspettative del padre, che vedeva nel matrimonio e nella maternità il destino delle sue figlie, **Rita** [Levi Montalcini] decise di iscriversi alla facoltà di medicina dell'Università di Torino, dove si laureò nel 1936. A seguito delle leggi razziali del 1938 che impedivano agli ebrei di insegnare presso università e istituti di ricerca, **Rita** diede vita a un laboratorio domestico, dove continuò a svolgere ricerche sul sistema nervoso.
- ‘Contrary to her father's expectations, who saw marriage and motherhood as her daughters' destiny, **Rita** [Levi Montalcini] decided to enroll in the medical school at the University of Turin, where she graduated in 1936. Following the racial laws of 1938, which prevented Jews from teaching at universities and research institutes, **Rita** started a home laboratory, where she continued to conduct research on the nervous system.’

5 Conclusion

Our empirical case study based on the BioWiki and BioGPT-4 text collections, comprising Wikipedia entries and GPT-4 generated biographies in Italian and French, reveals several important asymmetries in relation to anthroponyms referring to well-known and eminent personalities. In what follows, we provide some answers to the following research questions:

1. With what forms of anthroponyms (Type I, Type II and/or Type II) are women and men referred to in GPT-4 short biographies? In high status professional contexts, are women referred to differently than men?
2. Are there any meaningful cross-linguistic differences in the use of female and male anthroponyms? How do we explain them?
3. Do GPT-4 short biographies mirror or even enhance the gender biases present in human-authored texts, in particular in Wikipedia entries? If so, what can be done to mitigate the observed biases?

As far as the first question is concerned, the answer is clear: in GPT-4 short biographies, women and men are referred to very differently. In high status professional contexts, women are referred to more frequently than men by their full name (*Samantha Cristoforetti*) or first name (*Samantha*), whereas men are referred to more frequently than women by their surname (*Fermi*). Statistical tests indicate a strong association between women and gender-marking anthroponyms (particularly Type I, i. e. <first name + surname>), which explicitly convey information on the gender of the referent. In contrast, men are more strongly associated with gender-neutral anthroponyms (Type II, i. e. <surname>), which do not reveal the gender of the referent.

This study also highlighted an important asymmetry beyond the mere frequency of anthroponyms. A closer look at the dispersion and distribution of first names across different textual spaces reveals that, in male biographies, these names tend to appear in text passages describing the personal sphere (childhood and family life) of the main character. In female biographies, by contrast, first names are also used in textual spaces detailing the academic and professional life of the main character. This is an important finding, as first names denote social closeness and are less appropriate in stretches of texts related to the professional life of individuals. These findings reveal that GPT-4 generated biographies might reinforce the notion that, in high-status positions and professions, men are better known and more respect-deserving as they are predominantly referred to by the most formal type of anthroponym, expressing social distance – their surname. GPT-4 generated biographies might also perpetuate the habit of referring to

women in high status professional contexts by what Sabatini (1993) described as a “familiar and paternalistic form” of anthroponym.

As far as the second question is concerned, we did not find any meaningful cross-linguistic difference in the distribution of anthroponymic types in the BioFem and BioMasc datasets. In both Italian and French, women are more frequently associated with Type I anthroponyms (i. e., full name) than men, who are more often referred to by Type II anthroponyms (i. e., surname). This finding points to a common gender-bias, independent of the source language.

The results obtained in this study for GPT-4 short biographies perfectly align with the findings on the use of female and male anthroponyms in human-authored texts observed in the literature and in our case study on Wikipedia entries. Specifically, they reveal the same form of biased distribution between types of anthroponyms and gender of the referent. We can thus claim that our findings mirror the biases present in the training data, at least in the Wikipedia entries. Parts of our results, however, even suggest that GPT-4 short biographies amplify existing biases and asymmetries present in human-authored texts. Specifically, we found that GPT-4 consistently fails to generate biographies on six important women, mostly active in the STEM fields: Anne Chopinet and Emma Strada are engineers; Clara Lollini, Anna Marabini and Jeanne Lévy are chemists. Failed biographies were never encountered for males. In GPT-4 short biographies, several eminent women are thus more invisible than men, a fact that does not reflect their actual presence on the internet, in particular in Wikipedia entries. This result should be kept in mind considering that LLM-generated texts are themselves used in the training data of new LLMs, particularly of the ones for which there is not sufficient material (cf. Basile et al. 2024 in relation to Italian).

Our findings on GPT-4 short biographies can have significant societal implications. This is why we need to continue analyzing LLM-generated outputs in different fields (linguistics, computational linguistics, social sciences etc.). We need to monitor the extent to which asymmetries are present in the output and investigate the reasons explaining their presence. We also ought to reflect on efficient ways to mitigate the asymmetries and biases revealed in the present study as well as in other case studies. As far as mitigation is concerned, it is to the NLP community to intervene by training open LLMs on more gender-inclusive datasets, in particular by fine-tuning these tools with datasets balanced in terms of language, gender representation and gender-fair linguistic forms. In the words of Nemani et al. (2024: 12), “To tackle these issues, the NLP community should prioritize inclusivity, diversify language focuses, embed early-stage bias evaluations, and adopt standardized benchmarks to address gender bias holistically”.

In order to further deepen our understanding on the presence of biases and gender asymmetries in the output of LLM-generated texts, it would be essential to reproduce (using the same data) and/or replicate (with new datasets, e. g. using GPT-4 short biographies generated with different temperatures¹³) the research design used in this case study on GPT-4 short biographies produced in Italian and French. Future research should also expand the scope of the present case study by comparing the results obtained based on larger empirical datasets generated by other LLMs in two or more languages. In order to better understand the relation between the training

¹³ As mentioned in Section 3, in this study we generated text with the default temperature of 0.7.

data and internal architecture of LLMs, on the one hand, and their outputs, on the other hand, one of the most promising milestones in the field of generative artificial intelligence is the increasing availability of truly open LLMs, such as Teuken7B (from OpenGPT-X) and Minerva7B (by Minerva). All the research lines mentioned above allow gaining insights into the ethical profile of LLMs and could further improve the development of benchmarks assessing successful gender-inclusive communication.¹⁴

References

- Ariel, Mira (1990): *Accessing Noun Phrase Antecedents*. London/New York: Routledge.
- Atir, Stav/Ferguson, Melissa Jay (2018): “How gender determines the way we speak about professionals”. *PNAS* 115/28: 7278–7283. doi.org/10.1073/pnas.1805284115.
- Baack, Stefan (2024): “A critical analysis of the largest source for generative AI training data: Common Crawl”. *FACCT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery: 2199–2208.
- Barbaresi, Adrien (2021): “Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction”. In: Ji, Heng/Park, Jong C./Xia, Rui (eds.), *Proceedings of ACL/IJCNLP 2021: System Demonstrations*: 122–131.
- Basile, Pierpaolo et al. (2024): “LLaMAntino: Large Language Models per la lingua italiana”. doi.org/10.48550/arXiv.2312.09993 [18.01.2026].
- BioWiki (as well as: BioFemWiki, BioMascWiki, BioGPT-4, BioFemGPT-4, BioMascGPT-4): Datasets published in De Cesare, Anna-Maria/Weidensdorfer, Tom/Burchardt, Luisa (2025).
- Bolukbasi, Tolga et al. (2016): “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: Lee, Daniel/von Luxburg, Ulrike (eds.): *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*. Red Hook, NY, Curran Associates Inc.: 4356–4364.
- Brown, Tom et al. (2020): “Language Models are Few-Shot Learners”. *Advances in Neural Information Processing Systems* 33. doi.org/10.48550/arxiv.2005.14165.
- Burr, Elisabeth (1997): “Neutral oder Stereotyp. Referenz auf Frauen und Männer in der italienischen Tagespresse”. In: Dahmen, Wolfgang et al. (eds): *Sprache und Geschlecht in der Romania*. Tübingen, Narr: 133–179.
- Chen, Banghao et al. (2023): “Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review”. arxiv.org/pdf/2310.14735 [16.09.2024].
- Chiang, Wei-Lin et al. (2024): “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”. doi.org/10.48550/arXiv.2403.04132
- D'Achille, Paolo (2016): “Architettura dell'italiano di oggi e linee di tendenza”. In: Lubello, Sergio (ed.): *Manuale di linguistica italiana*. Berlin/Boston, De Gruyter: 165–189.
- De Cesare, Anna-Maria (2023): “Giorgia Meloni, Meloni o la Meloni? La codifica degli antropimi femminili in biografie generate da ChatGPT e pubblicate su Wikipedia”. *Lingue e Culture dei Media* 7: 1–20. doi.org/10.54103/2532-1803/22388.

¹⁴ Current existing benchmarks, such as WinoBias and Winogender (described in Zhao et al. 2018 and Rudinger et al. 2018, respectively), are currently designed for specific tasks (such a coreference resolution).

- De Cesare, Anna-Maria (2025): “LLM referential chain generation. A qualitative case study based on Italian biographies produced by GPT-4”. *Linguistik Online* 136, 4/25: 25–52. doi.org/10.13092/8tppdv47.
- De Cesare, Anna-Maria (2026): *L’italiano sintetico dell’intelligenza artificiale generativa*. Firenze: Cesati.
- De Cesare, Anna-Maria/Giusti, Giuliana (2024): “Lingua inclusiva e variazione linguistica”. In: De Cesare, Anna-Maria/Giusti, Giuliana (eds): *Lingua inclusiva: forme, funzioni, atteggiamenti e percezioni. LiVVal Linguaggio e Variazione/Variation in Language* 6. Venezia, Ca’ Foscari Digital Editions: 3–17. doi.org/10.30687/978-88-6969-866-8/001.
- De Cesare, Anna-Maria/Weidensdorfer, Tom/Burchardt, Luisa (2025): *Supplementary material for: Gender biases in GPT-4 generated biographies. A corpus study on Italian and French anthroponyms*. doi.org/10.5281/zenodo.15705482.
- Ding, Yitian et al. (2025): “Gender Bias in Large Language Models across Multiple Languages. A case study of ChatGPT”. In: Cao, Trista et al. (eds.), *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*: 552–579.
- Ferrara, Emilio (2023): “Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models”. *First Monday* 28/11: 1–39. doi.org/10.5210/fm.v28i11.13346.
- Flekova, Lucie/Ferschke, Oliver/Gurevych, Iryna (2014): “What makes a good biography? Multidimensional quality analysis based on Wikipedia article feedback data”. In: Chung, Chin-Wan (ed.): *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*. New York, Association for Computing Machinery: 855–865.
- Fusco, Fabiana (2024): *Lingua e genere*. Roma: Carocci.
- Garrido-Muñoz, Ismael/Martínez-Santiago, Fernando/Montejo-Ráez, Arturo (2023): “MarIA and BETO are sexist: evaluating gender bias in large language models for Spanish”. *Language Resources and Evaluation* 58: 1387–1417. doi.org/10.1007/s10579-023-09670-3.
- Johnson, Rebecca L. et al. (2022): “The Ghost in the Machine has an American accent: value conflict in GPT-3”. doi.org/10.48550/arXiv.2203.07785.
- Kilgarriff, Adam et al. (2014). “The Sketch Engine: ten years on”. *Lexicography ASIALEX* 1: 7–36. doi.org/10.1007/s40607-014-0009-9
- Lee, Dong Kyu (2016): “Alternatives to P value: confidence interval and effect size.” *Korean journal of anesthesiology* 69/6: 555–562. doi.org/10.4097/kjae.2016.69.6.555.
- Leong, Kelvin/Sung, Anna (2024): “Gender stereotypes in artificial intelligence within the accounting profession using large language models”. *Humanities and Social Science Communication* 11: article 1141. doi.org/10.1057/s41599-024-03660-8.
- Marcantonio, Angela (1988): “Nomi propri”. In: Renzi, Lorenzo (ed.): *Grande grammatica italiana di consultazione*. Vol. 1. Bologna, Il Mulino: 328–332.
- Marcato, Carla (2010): “Antroponimi”. *Enciclopedia dell’italiano*. Roma, Treccani: 84–87.
- Navigli, Roberto/Conia, Simone/Ross, Björn (2023): “Biases in Large Language Models: Origins, Inventory, and Discussion”. *ACM J. Data Inform. Quality* 15/2: 101–121. doi.org/10.1145/3597307.
- Nemani, Praneeth et al. (2024): “Gender bias in transformers: A comprehensive review of detection and mitigation strategies”. *Natural Language Processing Journal* 6. Article 100047. doi.org/10.1016/j.nlp.2023.100047

- Nghiem, Huy et al. (2024): “*You Gotta be a Doctor, Lin: An Investigation of Name-Based Bias of Large Language Models in Employment Recommendation*”. In: Al-Onaizan, Yaser/Bansal, Mohit/Chen, Yun-Nung (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Association for Computational Linguistics: 7268–7287.
- Rudinger, Rachel et al. (2018): “Gender Bias in Coreference Resolution”. In: Walker, Marilyn/Ji, Heng/Stent, Amanda (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 2. New Orleans, Louisiana, Association for Computational Linguistics: 8–14.
- Sabatini, Alma (1993): *Il sessismo nella lingua italiana*. Roma: Presidenza del Consiglio dei Ministri, Dipartimento per l’informazione e l’editoria.
- Wan, Yixin et al. (2023): “*Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters*”. In: Bouamor, Houda/Pino, Juan/Bali, Kalika (eds.). *Findings of the Association for Computational Linguistics: EMNLP 2023*: 3730–3748. doi.org/10.18653/v1/2023.findings-emnlp.243.
- Werlich, Egon (1975): *Typologie der Texte. Entwurf eines Textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Heidelberg: Quelle & Meyer.
- Zhao, Jieyu et al. (2018): “Gender bias in coreference resolution: Evaluation and debiasing methods”. In: Walker, Marilyn/Ji, Heng/Stent, Amanda (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 2. New Orleans, Louisiana, Association for Computational Linguistics: 15–20.
- Zhao, Wayne X. et al. (2023): “A Survey of Large Language Models”. arXiv:2303.18223v13 [11.01.2024]