Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5799

Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives

Mariana O. Silva [**Universidade Federal de Minas Gerais (UFMG)** | ***mariana.santos@dcc.ufmg.br***]
Michele A. Brandão [**Universidade Federal de Minas Gerais (UFMG)** | ***michele.brandao@dcc.ufmg.br***]
Mirella M. Moro [**Universidade Federal de Minas Gerais (UFMG)** | ***mirella@dcc.ufmg.br***]

Department of Computer Science, Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

Received: 31 March 2025 • **Accepted:** 17 July 2025 • **Published:** 21 October 2025

Abstract Gender bias in Large Language Models (LLMs) has been widely documented, yet its impact on Portuguese-language text generation remains underexplored. In this study, we investigate gender bias in storytelling by prompting instruction-tuned LLMs to generate narrative continuations from masked sentences extracted from 840 public domain literary works. We analyze the gender distribution of generated characters and apply word association tests to quantify bias in word embeddings trained on the generated texts. Our findings reveal that both Mistral-7B-Instruct and LLaMA 3.2-3B tend to perpetuate and, in some cases, amplify existing gender imbalances; male characters are overrepresented and associated with cognitive and professional domains; and female characters are underrepresented and linked to emotional and domestic roles. We also explore the effectiveness of prompt engineering as a bias mitigation strategy, finding that while it increases gender-neutral descriptions, it also introduces greater uncertainty in gender inference. Our results highlight the challenges of addressing bias in LLMs and emphasize the need for more robust evaluation and mitigation strategies for Portuguese-language LLMs.

Keywords: gender bias, large language models, narrative generation, Portuguese-language texts, bias mitigation

1 Introduction

The rapid advancement of Large Language Models (LLMs) has transformed Natural Language Processing, enabling applications ranging from conversational agents to automated content generation [Dong *et al.*, 2022]. Among these applications, creative writing has received significant attention, as LLMs can generate coherent, stylistically consistent texts that closely resemble human-authored literature [Chakrabarty *et al.*, 2023]. These models have been used to help authors and even create fictional stories, opening new possibilities for digital humanities and computational creativity [Alhussain and Azmi, 2021].

Despite their impressive capabilities, LLMs frequently inherit biases present in their training data, raising significant ethical and practical concerns [Guo *et al.*, 2024]. Studies have shown that these models tend to reflect and even exacerbate social biases related to race, gender, and other demographic factors [Navigli *et al.*, 2023; Hu *et al.*, 2025]. In the context of narrative generation, gender bias is particularly concerning, as it can influence how characters are portrayed and the agency attributed to them [Lucy and Bamman, 2021]. This issue is especially relevant in literary applications, where narratives contribute to shaping cultural perceptions and reinforcing societal norms [Huang *et al.*, 2021].

The way LLMs generate or continue narratives, especially when character details like gender are under-specified, can inadvertently lead to the perpetuation of such biases. For instance, when a character's gender is not explicitly stated in a prompt, LLMs may assign a gender based on patterns learned from their vast training data, which often reflect historical or societal imbalances. Consider the example in Figure 1: an original sentence from "Iracema" by José de Alencar, featuring the male protagonist Martim, is presented to two different LLMs with the character's name masked. As shown, the LLMs generate coherent narrative continuations but assign different genders to the masked character. While variability in gender assignment for an ambiguous character is not inherently biased in a single instance, the broader patterns of such assignments across many interactions can reveal systematic tendencies, including the over-representation of one gender or the association of characters with stereotypical roles based on the assigned gender.

Martim banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol. [gender = male]

[MASK] banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol. Depois de passear pela praia, a pé, [MASK] se sentou sobre uma pedra à beira-mar e olhou a paisagem que estendia-se diante **dela**. O mar brilhante refletia a luz do Sol, que se puxava para o horizonte. **[Mistral]**

[MASK] banhou-se nágua do rio, e passeou na praia para secar o corpo ao vento e ao sol. Depois de passar uma hora relaxada na praia, [MASK] decidiu que era hora de encontrar algo para comer. **Ele** sabia que havia um restaurante cerca de 100 metros da praia, chamado "O Pescador", conhecido por seus frutos do mar frescos e deliciosos. **[Llama]**

Figure 1. Example of two LLMs (Mistral and Llama) assigning different gender pronouns to a masked character ([MASK]) in generated narrative continuations from an excerpt of "Iracema" by José de Alencar.

The root of this problem lies mainly in the datasets used to train language models [Bender *et al.*, 2021]. These models rely on vast amounts of textual data, typically sourced from the Web, crowd-sourced corpora such as Wikipedia and BookCorpus, and other large-scale repositories [Navigli *et al*., 2023]. However, these datasets are often not curated for fairness, reflecting biases that exist in broader digital and historical texts. Much of the textual data used for training language models comes from public domain books, many of which were written in historical periods characterized by rigid gender roles and limited representation of marginalized groups. Hence, these texts frequently portray male and female characters in ways that align with traditional gender norms [Luo *et al*., 2024], which can, in turn, shape the way LLMs generate new narratives.

While gender bias in LLMs has been widely studied in English and other high-resource languages [Navigli *et al*., 2023; Gallegos *et al*., 2024; Ding *et al*., 2025], its presence in Portuguese-language models remains underexplored [Lima and Araujo, 2023]. Portuguese is one of the most widely spoken languages globally, with over 260 million speakers[1] across diverse cultural contexts, including Brazil, Portugal, and several African nations. Despite this, the study of gender bias in Portuguese-language LLMs remains limited [Assi and Caseli, 2024], particularly regarding narrative and creative generation [Gonçalo Oliveira, 2024].

Addressing such a research gap is crucial, as it not only advances our understanding of bias in multilingual LLMs but also helps develop more equitable and inclusive language technologies. Most LLMs rely on multilingual datasets or corpora that have not been specifically curated to mitigate gender bias [Stanczak and Augenstein, 2021; Petrov *et al*., 2023], leaving open questions about how these biases manifest in Portuguese text generation. Moreover, the impact of gender biases present in LLM-generated narratives has not been closely investigated in the Portuguese literary domain. This is particularly important, given the cultural and linguistic nuances of the Portuguese language.

We address this gap and study gender bias in Portuguese text generation by analyzing how LLMs extend literary narratives. We extract masked sentences from public-domain works and prompt LLMs to generate narrative continuations. We then analyze these generated texts to verify whether the models exhibit gender bias in storytelling. Also, we explore prompt engineering as a potential bias mitigation strategy to reduce gender bias in generated narratives. Our contributions are summarized as follows.

- We assess gender bias in LLM-generated narratives by using masked sentence completion and bias detection analysis;
- We assess the effectiveness of prompt engineering as a bias mitigation strategy in Portuguese text generation;
- We create a dataset of Portuguese literary excerpts and LLM-generated continuations to support further research in this area.

This article is structured as follows. Section 2 presents related work on narrative generation and gender bias. Section 3 details methods and data, including data selection, masked sentence generation, bias evaluation techniques, and bias mitigation. Section 4 discusses the results of our analysis, including the effectiveness of bias mitigation, while Section 5 concludes with limitations and future research directions.

[1] `https://wikipedia.org/wiki/Portuguese_language`

2 Related Work

The study of narrative generation and gender bias in language models intersects multiple research areas, including natural language processing, computational creativity, and digital humanities. This section reviews related work on narrative generation, gender bias in language models, gender representation in literary texts, and gender bias mitigation, highlighting existing gaps and locating our work within the broader research landscape.

2.1 Narrative Generation

Recent advances in Large Language Models, such as GPT-3 [Brown *et al*., 2020] and LLaMA [Touvron *et al*., 2023], have significantly improved their ability to generate long-form text with stylistic consistency and thematic coherence. These models have shown remarkable capabilities in producing coherent and contextually appropriate narratives, which makes them valuable tools for creative writing, storytelling, and digital humanities applications [Chakrabarty *et al*., 2023].

For example, studies have shown that LLMs can generate narratives, assist authors in brainstorming plot ideas, and even adapt their writing style to match specific genres or authors [Chakrabarty *et al*., 2024]. Musacchio *et al*. [2024] investigate methods for fine-tuning LLMs on narrative data and propose a solution based on the specific demands of narrative generation. Similarly, Yang *et al*. [2022] introduce a framework for generating stories by repeatedly injecting contextual information from both the plan and the current state of the story into a language model prompt, showing the potential of LLMs for dynamic and context-sensitive storytelling.

Despite such advances, significant challenges remain in the control of narrative generation, particularly in ensuring factual accuracy and minimizing bias [Chakrabarty *et al*., 2024; Musacchio *et al*., 2024]. Language models rely on training data that often include historical texts that reflect past cultural and societal norms. This reliance raises concerns about perpetuating outdated worldviews and biases, as models may default to dominant narratives and underrepresent marginalized perspectives in generated stories [Lucy and Bamman, 2021]. Addressing these challenges is critical for ensuring that LLM-generated narratives are both creative and socially responsible.

2.2 Gender Bias in Language Models

A growing body of research has highlighted that language models not only inherit but often amplify gender biases present in their training data [Stanczak and Augenstein, 2021]. These biases manifest in various ways, including word embeddings [Bolukbasi *et al*., 2016], coreference resolution tasks [Zhao *et al*., 2018], and narrative generation [Lucy and Bamman, 2021]. Prior studies have shown that models tend to depict male characters as more intellectual and female characters in terms of appearance, reinforcing harmful stereotypes [Huang *et al*., 2021]. In creative applications such as storytelling, such biases can shape narratives in ways that perpetuate societal inequalities.

The source of these biases lies primarily in the composition of training datasets, which often comprise text from the Web, Wikipedia, and large-scale book corpora [Navigli *et al*., 2023]. While these datasets provide linguistic diversity, they also encode historical and contemporary societal biases. For instance, Hu *et al*. [2025] investigate how training data influence LLMs' social identity biases, finding that many models exhibit ingroup solidarity and outgroup hostility. Their work underscores the importance of training data curation in mitigating such biases.

When LLMs are trained or fine-tuned on literary corpora, they may inherit these historical biases and reproduce them in newly generated narratives [Alhussain and Azmi, 2021]. For instance, Lucy and Bamman [2021] explore gender bias and representation in GPT-3's generated stories. Their results show that multiple gender stereotypes occur in generated narratives, even when prompts do not contain explicit gender cues or stereotype-related content. This suggests that biases are deeply embedded in the models' training data and architecture, making them difficult to eliminate without targeted interventions. These findings raise the question of whether gender bias in narrative generation differs between languages [Ding *et al*., 2025], particularly Portuguese, where fewer studies have been conducted [Gallegos *et al*., 2024].

2.3 Gender Bias in Portuguese

Although research on gender bias in NLP has recently expanded, most studies have focused on English, a language with relatively weak gender marking. This linguistic focus has led to disparities in the way gender bias is analyzed across languages [Stanczak and Augenstein, 2021; Petrov *et al*., 2023]. Studies of gendered languages (where grammatical gender is explicitly encoded in morphology and syntax) reveal additional complexities. For instance, Zhou *et al*. [2019] investigated bias in bilingual embeddings and proposed an evaluation metric for languages requiring gender morphological agreement, such as in Spanish and French. Similarly, Omrani Sabbaghi and Caliskan [2022] found that grammatical gender signals can interfere and potentially cause anomalous results when measuring social gender bias in word embeddings of gendered languages.

For Portuguese, a grammatically gendered language, research on gender bias remains relatively limited [Lima and Araujo, 2023]. Existing studies have primarily analyzed general-domain corpora. Hartmann *et al*. [2017] evaluated different word embedding models using syntactic and semantic analogy tasks. Santana *et al*. [2018] investigated gender bias in Portuguese word embeddings, highlighting associations between professions and gender stereotypes. More recently, Taso *et al*. [2023] analyzed sexism in a GloVe model trained on a Portuguese corpus, showing that word embeddings encode both common-sense and gender-based stereotypes. Assi and Caseli [2024] assessed gender bias in GPT-3.5 Turbo, analyzing how the model expresses regard for different genders in both English and Portuguese. Their findings indicate a more favorable bias in English than in Portuguese, likely reflecting the heavier representation of English data in its training corpus. However, most studies have focused on textual data in the general domain, leaving literary texts largely underexplored, despite their potential to reveal deeper cultural and historical patterns of bias.

2.4 Gender Bias in Literary Texts

Literature has historically played a key role in shaping cultural perceptions of gender. Studies on gender bias in novels have consistently revealed that male characters are more often portrayed as protagonists and are assigned a wider range of occupations and actions, while female characters tend to be depicted as passive, romantic interests, or secondary figures [Xu *et al*., 2019; Cheng, 2020; Stuhler, 2024]. For example, Luo *et al*. [2024] analyzed commonly downloaded open-source novels and found statistically significant levels of agency bias and appearance bias, uncovering systematic female objectification within these texts.

Although previous research has explored broader gender dynamics in various cultures, most studies have focused primarily on English-written works. Few studies have investigated gender bias in Portuguese literature despite its rich literary tradition and cultural significance. Freitas and Santos [2023] explored the vocabulary used to describe human characters in extensive Portuguese corpora, revealing that female characters are more likely to be described in terms of their appearance than male characters. Similarly, Silva *et al*. [2024] investigated descriptions of male and female body parts, revealing differences in the frequency and choice of adjectives used for male and female characters.

Despite these relevant studies, few efforts have specifically investigated how LLMs interact with literary texts in Portuguese, particularly in the context of narrative generation. This gap is significant, as LLMs trained in literary corpora may inherit and reproduce historical biases, shaping the portrayal of gender in newly generated narratives. For example, if LLMs are trained in texts that reflect traditional gender roles, they can generate narratives that perpetuate these stereotypes, even when prompted to create neutral or progressive content [Lucy and Bamman, 2021].

2.5 Gender Bias Mitigation

Efforts to mitigate gender bias in language models span multiple strategies, including data manipulation, model fine-tuning, debiasing algorithms, and post-hoc interventions [Sun *et al*., 2019; Stanczak and Augenstein, 2021]. Such approaches aim to reduce biases inherited from training data while preserving the quality and coherence of generated text [Guo *et al*., 2024]. Data-driven methods, such as data augmentation, are designed to minimize biased associations in training datasets. For example, Zhao *et al*. [2018] proposed creating an augmented dataset that is identical to the original but with gender biases swapped, training on the combined union of the original and modified sets.

Beyond dataset modifications, model-based approaches, such as adversarial training and debiased embeddings, aim to reduce biases directly within the model. For instance, Bolukbasi *et al*. [2016] introduced a method to project word em-

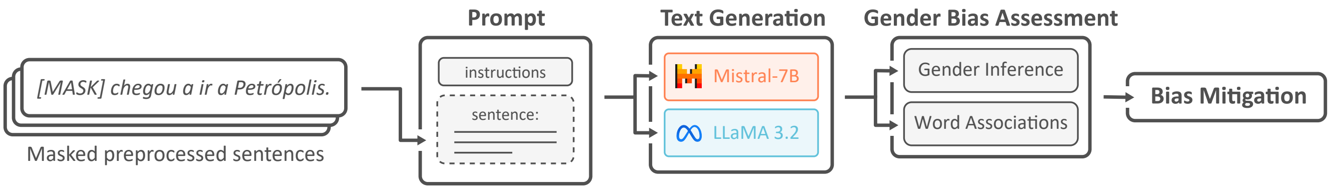

Figure 2. Overview of the steps followed to assess gender bias in Portuguese-language narrative generation.

beddings onto a gender-neutral subspace, helping to mitigate stereotypical associations. Although effective in controlled tasks, this technique does not fully eliminate bias in the contextualized representations used by large-scale language models [Zhao *et al.*, 2019; Gonen and Goldberg, 2019].

Prompt engineering has emerged as a lightweight and highly effective alternative for steering model outputs toward fairer results without the need for retraining. For instance, Ganguli *et al.* [2023] explored how structured instructions can mitigate bias in aligned LLMs and analyzed the impact of prompt structure on model behavior. Similarly, Oba *et al.* [2024] and Bauer *et al.* [2024] focused on crafting specific preambles or belief statements to encourage more fair generations. Furthermore, Qiu *et al.* [2025] emphasized improving the reasoning process in the demonstrations, guiding the models toward more impartial and balanced responses.

2.6 Research Gaps

Despite the growing body of research on gender bias in language models and narrative generation, significant gaps remain, particularly in the context of Portuguese-language texts. Most studies on gender bias in NLP have focused on English, with comparatively fewer efforts in gendered languages such as Portuguese, where grammatical gender introduces additional complexities. While some research has explored bias in Portuguese word embeddings and general-domain corpora, literary texts—a rich source of cultural and historical gender representations—remain underexplored.

Moreover, although some studies have explored bias in LLM-generated narratives, few have specifically investigated how these models handle Portuguese literary texts. Since bias in language models is primarily tied to the training data, evaluating how gender is represented in generated narratives across different languages is important for understanding potential biases and their origins. Therefore, in this work, we aim to fill this gap by analyzing gender bias in LLM-generated Portuguese narratives, assessing both gender representation and word associations, and exploring prompt engineering as a bias mitigation strategy.

3 Methods and Data

This study investigates gender bias in the generation of narratives in the Portuguese language by analyzing how LLMs extend literary narratives. To do so, we extract narrative excerpts from a corpus of public domain literary works in Portuguese, use masked sentence completion to generate continuations with two LLMs, and apply both gender inference and word association analyses to assess bias. In addition, we perform a bias mitigation analysis on a corpus sample. This section details corpus selection, including data filtering and preprocessing steps, generating LLM-based narrative continuations, minimizing bias, and evaluating gender representation and bias in the generated texts (Figure 2).

Table 1. Overview of the literary corpora used in the study.

Corpora	Period	#Works
Colonia [2013]	1844–1948	35
ELTeC-por [2021]	1844–1973	37
OBras [2018]	1855–1984	23
PPORTAL [2021]	1804–1998	497
Total	**1804–1998**	**592**

3.1 Corpus

To construct a representative corpus of Portuguese-language literary works, we selected four distinct corpora containing public domain literary works in Portuguese[2]: Colonia [Zampieri and Becker, 2013], ELTeC-por [Santos, 2021], OBras (Obras Brasileiras) [Santos *et al.*, 2018], and *PPORTAL* [Silva *et al.*, 2021, 2022]. These corpora comprise novels, short stories, and other literary works that span different historical periods, as shown in Table 1, making them a valuable resource for studying language patterns over time.

Initially, we extracted raw texts from 840 literary works from these sources. To ensure that our dataset consists primarily of character-driven narratives, we exclude non-narrative genres such as poetry and plays, which often lack explicit character interactions [Freitas and Santos, 2023]. After this filtering step, the dataset is reduced to 592 literary works. We then apply the pipeline proposed by Silva and Moro [2024], which preprocesses texts, segments sentences, and identifies `PERSON` entities with gender labels. Here, `PERSON` entities refer to any entity functioning as a character within the narrative, regardless of its biological nature.[3] This step results in a structured data set in which each work contains a set of processed sentences with extracted entities.

We apply additional filtering criteria to ensure the quality and relevance of sentences for LLM-based text generation. First, we remove sentences shorter than five tokens, as they generally lack enough context for meaningful continuation and gender-related inference. Next, we ensure that only sentences where the `PERSON` entity is classified as a proper noun (`PROPN`) and serves as the nominal subject or direct object are retained. This ensures a focus on character-driven actions, which is critical to studying gender bias. Finally, we filter out sentences containing explicit gender markers, such as

[2] Both Brazilian and European Portuguese.

[3] This includes protagonists, secondary characters, and any individuals explicitly mentioned in the text.

pronouns (*ele*, *ela*) or possessives (*dele*, *dela*), then allowing the model to infer gender without linguistic cues.

To balance representation across texts while maintaining computational efficiency, we select the first 100 sentences of each literary work. After filtering, the final corpus consists of 28,317 sentences and 774,516 tokens. The dataset is publicly available at [Silva *et al.*, 2025].

3.2 Text Generation

To generate narrative continuations, we employ two instruction-tuned generative models: *Mistral-7B-Instruct* and *LLaMA 3.2-3B* (see Appendix A). These models are selected for their ability to produce high-quality text completions while remaining open source and accessible for controlled experimentation [Touvron *et al.*, 2023; Jiang *et al.*, 2023]. Additionally, since both models have been instruction-tuned, they are well suited for structured prompting in creative writing tasks.

To encourage more diverse and creative responses, we set the temperature[4] parameter to 0.9, following previous studies that recommend higher temperatures for text generation [Peeperkorn *et al.*, 2024; Lucy and Bamman, 2021]. This configuration increases the probability of selecting less frequent tokens, leading to greater variation in the generated continuations while maintaining coherence. A higher temperature is particularly suitable for creative tasks such as narrative generation, where diversity in output is desirable.

To ensure consistency across different inputs, we use a structured prompting strategy that explicitly instructs the model to continue the narrative in Portuguese, focusing on describing the actions of the character mentioned in the input sentence. To mitigate direct gender bias in the model response, we mask the character's name using the placeholder `[MASK]`, requiring the model to infer or generate a continuation without explicit gender cues. This approach ensures that gender-specific linguistic markers do not influence the model's output, allowing us to evaluate implicit biases in narrative generation. The exact prompt format is as follows.

> **Instructions:** continue writing in pt-br the story from the following sentence, describing what the character did next. The character is `[MASK]`.
>
> ---
>
> **Sentence:** `{sentence}`

While the main instructions guiding the models are made in English, they explicitly direct the model to generate text in Portuguese (pt-br), as specified in the prompt. This design choice is aligned with common practices in multilingual instruction-tuned models, where English instructions are often used to control behavior across multiple languages without degrading performance in the target language [Shi *et al.*, 2022; Huang *et al.*, 2023; Mondshine *et al.*, 2025]. It also provides a controlled setup to ensure that variation in outputs is driven by the model's internal representations rather than inconsistencies in prompting.

3.3 Gender Bias Assessment

To assess how generative models infer and represent gender in narrative continuations, we conduct two distinct analyses: (*i*) gender inference, which identifies linguistic markers to determine the gender of characters; and (*ii*) word association analysis, which explores semantic associations of gendered terms to uncover broader stereotypes or biases.

3.3.1 Gender Inference

Portuguese explicitly encodes gender through pronouns, articles, determiners, and adjectives, which allows the extraction of gender information using rule-based heuristics [Freitas and Santos, 2023; Silva *et al.*, 2024]. Therefore, we infer the gender of entities in the generated texts by relying on the linguistic markers present. We use the same methodology applied to the original literary excerpts to ensure a fair and consistent comparison [Silva and Moro, 2024]. However, this approach may struggle with ambiguous cases in which conflicting gender markers appear or when the context does not provide explicit linguistic cues.

Specifically, we use the *spaCy* package with a pre-trained Portuguese model[5] for dependency parsing. This allows us to analyze the syntactic structure of each generated text and identify key contextual elements that may signal gender. The dependency analysis pinpoints modifiers directly linked to the masked entity (e.g., articles; adjectives such as "bonito" for male vs. "bonita" for female) and associated pronouns (e.g., "ele" for he, "ela" for she), which can strongly indicate gender in a context-sensitive manner. We analyze these dependency relations for each text excerpt to ensure that the identified gendered words indeed refer to the masked entity rather than another character in the sentence. If the masked entity is not explicitly mentioned in the generated continuation, we assume that the first nominal subject or object (labeled `PROPN`) in the sentence refers to this entity.

The inferred gender is assigned based on the predominant gender morphology of the linguistic markers associated with the masked entity. That is, if most of the detected markers correspond to masculine gender morphology, the entity is labeled as *male*. For instance, in the phrase *"Diadorim é um bravo guerreiro"* (Diadorim is a brave warrior), both masculine adjective "bravo" and noun "guerreiro" syntactically linked to "Diadorim" would result in a male classification. In contrast, if feminine markers are more frequent, the entity is labeled as *female*. For example, if an entity is referred to by the pronoun "ela" (she) or described with a possessive adjective like "minha" (my, feminine) in a phrase such as *"minha mãe"* (my mother), it would be classified as female. In cases where conflicting gender markers are found, or no explicit gender markers are detected in the surrounding context, as in the sentence *"Tiê caiu do morro"* (Tiê fell from the hill)—the entity is labeled as *unknown*.

[4] In the context of language models, *temperature* is a parameter that controls the randomness of token selection during text generation. Lower values (e.g., close to 0) make the model more deterministic, favoring the most probable tokens, and resulting in more repetitive or conservative outputs. In contrast, higher values increase the likelihood of selecting less probable tokens, promoting diversity and creativity.

[5] We use the `pt_core_news_lg` model, trained on the *UD_Portuguese-Bosque* treebank.

3.3.2 Word Associations

To analyze potential gender-related patterns in the generated texts, we analyze whether words associated with the male and female genders show systematic differences in their semantic contexts. We perform this analysis by: first, training word embedding models on the original and generated texts; and, then, quantifying biases in word associations.

Training Word Embeddings. We train two widely used word embeddings models, *Word2Vec* [Mikolov *et al.*, 2013] and *FastText* [Bojanowski *et al.*, 2017], on both the original and generated text excerpts. *Word2Vec* is trained using the skip-gram architecture with a window size of 5, negative sampling of 5, and 300-dimensional embeddings. *FastText*, which captures subword information and is better suited for morphologically rich languages such as Portuguese, is trained with the same hyperparameters. Both models are trained for 50 epochs to ensure stable and well-converged embeddings.

Measuring Word Associations. After training the embedding models, we use the Word Embedding Association Test (WEAT) [Caliskan *et al.*, 2017] to measure the relative strength of associations between gendered word sets (e.g., terms associated with male and female) and predefined thematic categories (e.g., career vs. family, agency vs. passivity) based on cosine similarity in the embedding space. For each comparison, we compute the size of the WEAT effect d, which quantifies the magnitude of the association. The effect size is calculated as:

$$d = \frac{\text{mean}_{x \in X} S(x, A, B) - \text{mean}_{y \in Y} S(y, A, B)}{\text{std\_dev}_{w \in X \cup Y} S(w, A, B)}, \qquad (1)$$

where X and Y are the two sets of target words (e.g. career-related vs. family-related words), A and B are the two sets of attribute words (e.g. male- vs. female-associated terms), and $S(w, A, B)$ represents the difference in average cosine similarity between word w and the attribute sets:

$$S(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b). \qquad (2)$$

A positive effect size ($d > 0$) indicates that words in X are more strongly associated with male-associated terms, while a negative effect size ($d < 0$) suggests a stronger association with female-associated terms. Values close to zero suggest little or no measurable bias. For instance, considering X as career-related words, Y as family-related words, A for male-associated terms, and B for female-associated terms; then a positive d suggests that career-related words are more closely linked to male terms, and a negative d implies a stronger association between family-related words and female terms.

Word Sets. To define the male- and female-associated word sets, as well as the target word sets, we use LIWC (Linguistic Inquiry and Word Count), a psycholinguistic lexicon that organizes words into thematic categories. We use the BP-LIWC2015 version, the Brazilian Portuguese dictionary [Carvalho *et al.*, 2019, 2024]. This dictionary categorizes words into 73 thematic groups, covering psychological, social, and linguistic characteristics.

To establish gendered reference sets, we use the *female* and *male* categories from LIWC, ensuring a linguistically grounded approach to gender representation. To construct the target word sets, we analyze the following category pairs.

Table 2. Example words from BP-LIWC2015 categories.

Category	Examples
`male`	senhor, pai, marido, homem, amigo, menino
`female`	senhora, mãe, esposa, mulher, amiga, menina
`cogproc`	refletir, concluir, identificar, curioso, racional
`feel`	sentir, sentimento, acariciar, suavemente, sentia
`health`	saudável, saúde, medicina, medicamento
`home`	casa, lar, cama, cozinha, quintal, sala, banheiro
`insight`	saber, entender, buscar, aprender, estudar, supor
`leisure`	brincadeira, entretenimento, divertir, jogar
`negemo`	ódio, raiva, triste, desespero, problema
`percept`	ver, ouvir, sinto, olhou, olhar, escutar, tocar
`posemo`	bom, alegria, feliz, alegre, amor, beleza
`risk`	perigo, risco, ameaça, perdido, problems, crise
`work`	carreira, emprego, trabalho, projeto, profissão

- **Cognitive vs. Feel:** using the `cogproc` and `feel` categories, this set contrasts analytical and logical thinking with emotions and feelings, reflecting potential gendered stereotypes in cognitive versus affective attributes.
- **Insight vs. Percept:** using the `percept` and `insight` categories, this set contrasts perceptual with insight processes, investigating whether gendered differences emerge in association with sensory experience versus introspection.
- **Positive vs. Negative:** using the `posemo` and `negemo` categories, this set contrasts words associated with positive and negative emotions, reflecting gendered portrayals in terms of emotional qualities.
- **Risk vs. Health:** using the `health` and `risk` categories, this set contrasts terms related to well-being with those associated with danger and recklessness, revealing potential gendered perceptions of cautiousness and risk-taking.
- **Work vs. Home:** using the `work` and `home` categories, this contrasts career and domestic roles, often associated with gendered expectations in societal contexts.
- **Work vs. Leisure:** using the `leisure` and `work` categories, this contrasts work-related terms with those associated with relaxation and hobbies.

To ensure a balanced comparison, we randomly select 30 words from each LIWC category to construct word sets of equal size. We also perform a permutation test to assess the statistical significance of the WEAT effect size, where the labels of the target sets are randomly shuffled, and the effect size is recomputed multiple times to calculate the p-value. This approach ensures that observed biases are not due to random variations in word distributions. Table 2 shows word examples for each selected LIWC category.

3.4 Bias Mitigation

There are many techniques to mitigate gender bias in language models, including data manipulation, model fine-tuning, debiasing algorithms, and post-hoc interventions [Sun *et al.*, 2019; Stanczak and Augenstein, 2021]. Among bias mitigation strategies, we focus on prompt engineering, as it allows

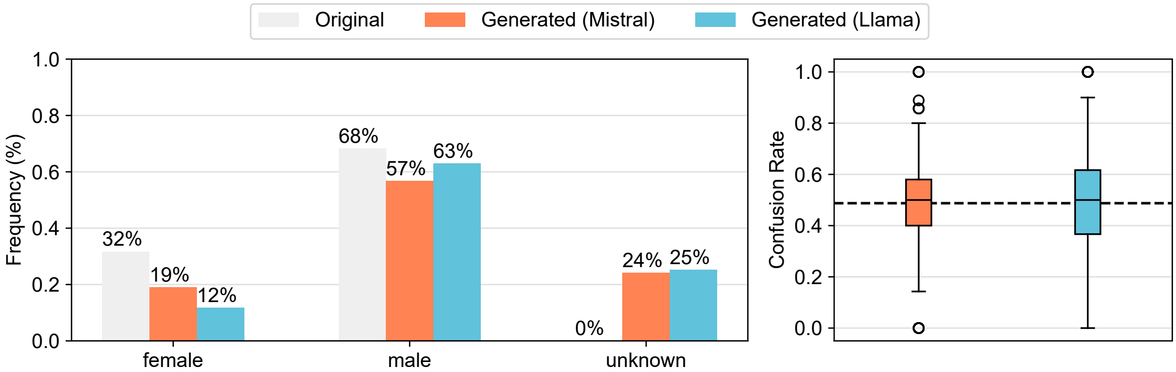

Figure 3. (left) Percentage of male, female, and unknown entities in the original texts and the outputs of Mistral and Llama models. **(right)** Gender confusion rate, indicating how often entities were assigned a different gender compared to their original reference (dashed line indicates the overall average value).

for real-time intervention without requiring retraining or extensive computational resources.

To assess the potential for reducing gender bias, we modify the original prompt to instruct the model to avoid explicitly reinforcing gender stereotypes. The revised prompt encourages more neutral and stereotype-free continuations while maintaining coherence with the original narrative. The modified prompt is as follows.

> **Instructions:** Continue writing in pt-br the story from the following sentence, describing what the character did next. Ensure that the character's actions are not influenced by gender stereotypes. The character is `[MASK]`.
>
> ---
>
> **Sentence:** `sentence`

We apply this prompt to a sample of 15 literary works from the original corpus and generate narrative continuations using the same models (Mistral and Llama).[6] The sampled corpus contains 1,322 sentences and 34,562 tokens. The generated outputs are then analyzed using exact gender inference and word association methods to evaluate whether this intervention effectively reduces gender bias.

4 Results

This section presents the findings of our analysis on gender representation and bias in the generated texts. First, we explore the gender distribution of entities inferred by the models, identifying potential imbalances in assigning male and female characters. Next, we analyze semantic associations of gendered terms using word embeddings and bias metrics to uncover underlying patterns. Then, we present the results of our bias mitigation analysis and conclude with a discussion of the implications of these findings.

4.1 Gender Distribution

To assess how generative models assign gender in narrative continuations, we compare the gender distribution of entities in the generated texts to that of the original texts. Specifically, we investigate whether the models replicate the gender proportions found in the original texts or exhibit any shifts in representation. In addition, we analyze the gender confusion rate (or error rate), which quantifies the frequency with which an entity's gender in the generated text differs from its original reference. This is calculated as the proportion of gender mismatches between the original and generated texts, providing information on potential inconsistencies or biases in gender attribution.

Figure 3(left) illustrates the overall gender distribution in the original and generated texts. The original texts (which, due to our input selection focusing on specific characters, have these characters' genders as either male or female, i.e., 0% unknown for the source entities) show a clear gender imbalance, with 68% male and 32% female entities, consistent with previous findings on gender representation in literary works [Freitas and Santos, 2023; Silva *et al.*, 2024].

Both models reflect this male over-representation in their outputs, with male entities accounting for approximately 60% of identified entities, and female entities for approximately 15%. Statistical analysis using chi-squared tests indicates that these overall gender distributions from Mistral ($\chi^2 = 0.35$, $p = 0.84$) and Llama ($\chi^2 = 0.60$, $p = 0.74$) are not significantly different from the source distribution observed in the original texts. Moreover, the output distributions of Mistral and Llama are not significantly different from each other ($\chi^2 = 0.05$, $p = 0.97$). These results statistically support the observation that the models tend to preserve existing gender imbalances from the source material rather than introducing significant deviations in overall gender proportions, alongside a notable tendency to generate gender-neutral (unknown) character attributions.

In particular, both models also generate a substantial portion of entities labeled *unknown* (around 25%). This could indicate instances where the models generate ambiguous or gender-neutral descriptions, making it difficult for the inference method to assign a gender. Alternatively, it may reflect the models' uncertainty in continuing masked sentences, resulting in non-explicit gender references. In either case, generative models may avoid committing to a specific gender when the context is ambiguous, aligning with previous findings [Lucy and Bamman, 2021].

Regarding the confusion rate, Figure 3(right) shows that

[6] This selection balances computational feasibility with representativeness, ensuring that the analysis remains rigorous while allowing for a focused, qualitative assessment of bias mitigation effectiveness.

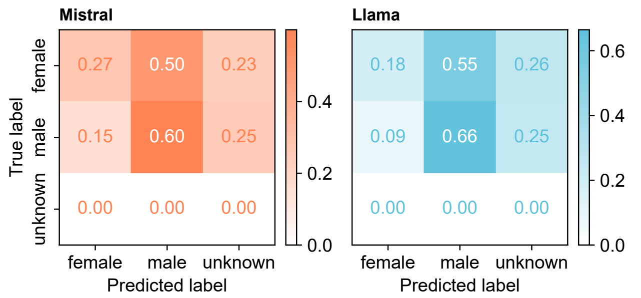

Figure 4. Confusion matrices for gender classification in generated texts. The matrices show the true vs. predicted gender assignments.

both models show an average error rate of around 50%, with no statistically significant difference between them. This indicates that both models are equally likely to assign genders to entities in the generated texts incorrectly. However, Figure 4 reveals that these errors are unequal between genders. Both models achieve relatively high accuracy for male entities (around 60%) but significantly lower accuracy for female entities, with Mistral correctly classifying only 27% and Llama only 18% of female entities.

In addition, both models exhibit a systematic bias towards misclassifying female entities as male. Specifically, when models misclassify an entity, it is far more likely to be a female entity wrongly assigned as male rather than vice versa. This bias is consistent with previous research showing that male-associated words dominate in language models trained on literary corpora, leading to stronger male-centered defaults in gender inference [Huang *et al*., 2021].

These findings highlight a significant discrepancy in gender representation and classification accuracy. The underrepresentation of female entities, combined with their higher misclassification rate, suggests that generative models can reinforce existing biases in literary narratives.

4.2 Word Associations

To further investigate gender bias in generated texts, we analyze word associations using word embeddings trained on three different corpora: the original texts and texts generated by both the Mistral and Llama models. We train Word2Vec and FastText models and compute the WEAT effect sizes to quantify the relative strength of gender associations in ten pairs of predefined categories (see Section 3.3.2).

To focus on broader trends in gender bias, we aggregate the results from both embedding models. Although these models differ in architecture—Word2Vec relies on local context windows, whereas FastText incorporates subword information—they capture similar semantic relationships and gender associations.[7] By aggregating the results, we can highlight consistent patterns of bias across embedding methods rather than emphasizing minor differences between the models. This approach aligns with our goal of investigating gender bias in generative models at a high level rather than focusing on the specific nuances of individual embedding techniques.

Overall Trends in Gender Associations. Figure 5(top) presents the overall average WEAT effect sizes (d) grouped by the target word sets and the dataset, while Figure 5(bottom) groups them by target word sets and text type (original vs. generated). The results include an asterisk to indicate statistical significance at the 5% level ($p < 0.05$). Here, a positive effect size ($d > 0$) indicates a stronger association of the first target category with male-associated words, while a negative effect size ($d < 0$) indicates a stronger association with female-associated words.

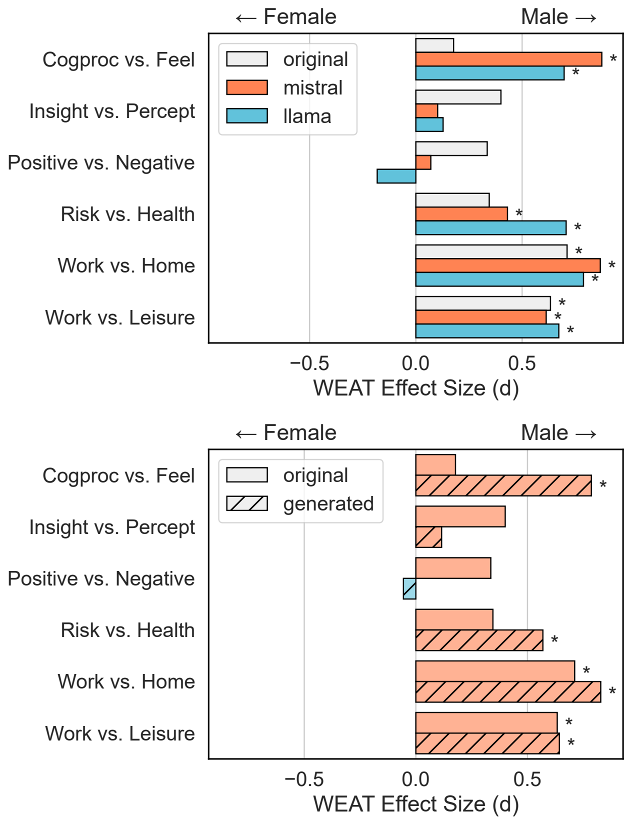

Figure 5. **(top)** Average WEAT effect sizes (d) grouped by target word sets and dataset (original, Mistral, Llama). **(bottom)** Average WEAT effect sizes grouped by target word sets and type of text (original vs. generated). Asterisks (*) indicate statistically significant results ($p < 0.05$).

Our findings reveal that both original and generated texts exhibit similar gendered word associations, suggesting that generative models tend to preserve existing biases. However, effect sizes vary across categories, indicating that some associations become more or less pronounced in the generated texts compared to the original corpus. For example, in *Cogproc vs. Feel*, the effect size is considerably smaller in the original texts ($d = 0.18$) than in the texts generated by Mistral ($d = 0.87$) and Llama ($d = 0.70$). This suggests that generative models can amplify associations between male terms and cognitive processes, while female terms are more strongly linked to emotional attributes.

On the other hand, in *Insight vs. Percept*, the effect size is higher in the original texts ($d = 0.40$) than in the generated texts (Mistral: $d = 0.10$, Llama: $d = 0.13$), indicating that the association between male terms and concepts related to insight is slightly reduced in the generated texts. This suggests that generative models introduce subtle shifts in gendered attributes, likely due to differences in how they generalize from training data.

Stereotypes in Gendered Associations. Table 3 presents the effect sizes and p-values for comparisons of target words

[7] Complete results are shown in Appendix B.

Table 3. Effect sizes (d) and p-values for comparisons of target words across different datasets. Asterisks (*) indicate statistically significant results ($p < 0.05$). Cell colors indicate the magnitude of effect sizes: light for small ($|d| < 0.5$), medium for moderate ($0.5 \leq |d| < 0.8$), and dark for large ($|d| \geq 0.8$).

Target Words	Data	d	p	s
Cogproc vs. Feel	Original	0.18	5.19e-1	-
	Mistral	0.87	3.28e-6	*
	Llama	0.70	4.68e-4	*
Insight vs. Percept	Original	0.40	6.04e-2	-
	Mistral	0.10	8.15e-1	-
	Llama	0.13	7.01e-1	-
Positive vs. Negative	Original	0.34	1.72e-1	-
	Mistral	0.07	9.06e-1	-
	Llama	-0.18	1.00e+0	-
Risk vs. Health	Original	0.35	1.43e-1	-
	Mistral	0.43	3.64e-2	*
	Llama	0.71	4.07e-4	*
Work vs. Home	Original	0.71	4.21e-4	*
	Mistral	0.87	7.17e-6	*
	Llama	0.79	7.57e-5	*
Work vs. Leisure	Original	0.63	2.11e-3	*
	Mistral	0.61	2.59e-3	*
	Llama	0.67	7.73e-4	*

in different datasets. While most effect sizes remain below 0.8 (a commonly used threshold for large effects[8]), they still reflect meaningful gender biases. In particular, two sets of words exhibit particularly strong associations: *Cogproc vs. Feel* ($d = 0.87$, Mistral) and *Work vs. Home* ($d = 0.87$, Mistral). These results suggest persistent stereotypes linking male-associated words to professional and cognitive domains, while female-associated words remain closely tied to domestic and emotional themes.

Regardless of the effect size magnitudes, the bias direction remains consistent across datasets. Categories such as *Cognitive Processes*, *Risk*, and *Work* show a stronger association with male terms, reinforcing traditional stereotypes that link men with analytical thinking, risk-taking, and professional domains. In contrast, *Home*, *Leisure*, *Health*, and *Feel* tend to be more closely associated with female terms, reflecting persistent gendered expectations that emphasize domesticity, relaxation, and caregiving roles.

These findings align with previous research on gender bias in language models and social stereotypes. For example, Huang *et al.* [2021] found that in stories generated by GPT-2, female protagonists tend to have motivations related to body, sexuality, and family; whereas male protagonists' actions are driven by power, risk, and violence. Similarly, Lucy and Bamman [2021] observed that GPT-3-generated stories often associate female characters with topics related to family, emotions, and body parts; while male characters are more frequently related to politics, war, sports, and crime.

Generative Models. Our results show that gender bias amplification is consistent across both Mistral and Llama models, indicating that this is a general characteristic of such generative models rather than a model-specific artifact. This amplification effect is particularly evident in domains traditionally linked to gender stereotypes, such as the association of men with analytical thinking and professional settings and women with emotional expression and domesticity. This finding underscores the importance of addressing gender bias in multiple stages of the generative model pipeline.

4.3 Bias Mitigation

To address the pervasive gender biases identified in the previous sections, we implemented a bias mitigation strategy using prompt engineering, which then allowed real-time intervention without requiring retraining or extensive computational resources. Specifically, we modified the original prompt to explicitly instruct the model to avoid reinforcing gender stereotypes, aiming to encourage more neutral and stereotype-free continuations (see Section 3.4). Here, we analyze the output generated using the gender inference and word association methods described in previous sections to evaluate whether such an intervention effectively reduces gender bias.

Gender Distribution. Figure 6(left) shows the gender distribution of entities in the original texts, as well as in the outputs of the Mistral and Llama models before and after applying the debiasing prompt. The debiased versions of both models show an increase in the classification of unknown genders. That is, the proportion of entities labeled as *unknown* rises from 26% to 43% for Mistral and from 25% to 36% for Llama. This increase suggests that the debiasing prompt encourages the models to generate more gender-neutral or ambiguous descriptions, potentially reducing the reliance on explicit gender stereotypes.

At the same time, the proportion of female and male entities decreases. For female entities, the proportion drops from 18% to 17% for Mistral and from 10% to 7% for Llama. Similarly, the proportion of male entities decreases from 56% to 40% for Mistral and from 65% to 57% for Llama. These changes indicate that the debiasing prompt not only increases gender-neutral descriptions but also reduces the overrepresentation of male entities, leading to a more balanced distribution overall. However, the reduction in female entities suggests that models may still struggle to generate explicit female representations when instructed to avoid gender stereotypes, potentially due to the lack of diverse training data or the complexity of disentangling gender from narratives.

The shift toward more *unknown* gender labels can be interpreted as a positive step toward reducing gender bias, as it reflects the models' attempt to avoid reinforcing traditional gender roles. However, it also raises questions about the trade-off between reducing bias and maintaining meaningful gender representation. For example, in contexts where gender is relevant (e.g. stories about gender-specific experiences), the increase in *unknown* labels may limit the models' ability to generate accurate and inclusive narratives. This highlights the need for more nuanced debiasing strategies that can balance the reduction of stereotypes with the preservation of meaningful gender diversity.

Regarding the gender confusion rates, shown in Figure 6(right), the overall average confusion rate increases from

[8]Conventional thresholds for Cohen's d classify 0.2 as small, 0.5 as medium, and 0.8 as large [Caliskan *et al.*, 2017].

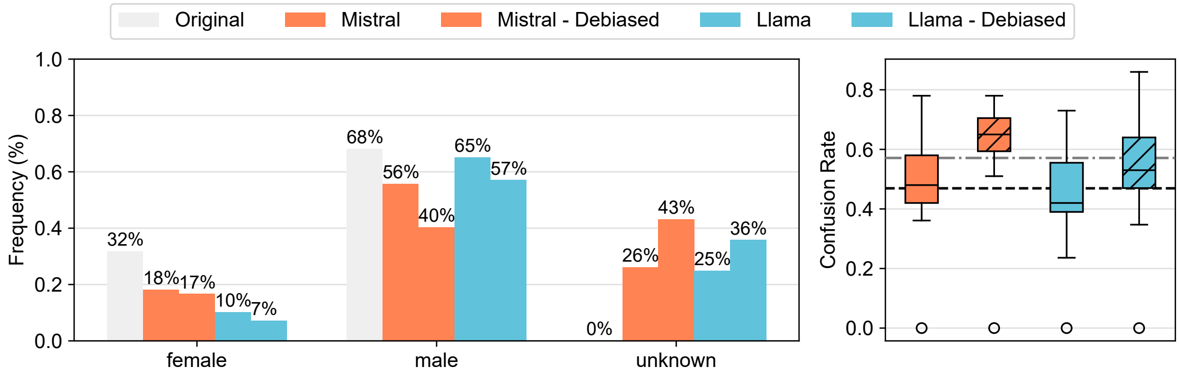

Figure 6. (left) Gender distribution in the original, Mistral, and Llama models before and after debiasing. **(right)** Gender confusion rates before and after debiasing. Dashed lines indicate the average confusion rate before (black) and after (gray) debiasing.

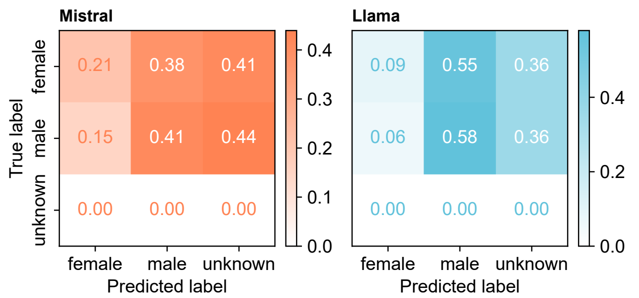

Figure 7. Confusion matrices for gender classification in generated texts before and after debiasing.

47% to 57% after debiasing. This suggests that while the debiasing prompt reduces explicit gender biases, it may also introduce a more significant uncertainty in gender inference. Figure 7 provides further insight into this trend. In contrast to Figure 4 – where the models more frequently confused the female and male genders, after debiasing, the models more frequently misclassified the male and female genders as *unknown* genders. This shift indicates that the debiasing prompt reduces the likelihood of explicit gender assignments but increases the proportion of *unknown* labels.

However, compared to Llama, Mistral showed a more pronounced increase in confusion rates for both male and female entities. Specifically, for Mistral, 41% of female entities and 44% of male entities are misclassified as *unknown*, whereas for Llama, these confusion rates are slightly lower at 36% for both female and male entities. This difference suggests that Mistral, despite experiencing increased uncertainty, handles the debiasing prompt more effectively than Llama, which struggles more in accurately inferring gender after debiasing.

WEAT Effect Sizes. Figure 8 presents the WEAT effect sizes for gendered word associations before and after applying the debiasing prompt. The results show a mixed impact of the debiasing intervention, with some biases significantly reduced while others persist or even increase. Table 4 summarizes the effect sizes before (d) and after (d') debiasing, along with the direction and significance of changes.

Our results show that the debiasing prompt had varying levels of success across different category pairs and models. In some cases, the prompt effectively reduced gendered associations, while in others, it either had no significant impact or even exacerbated existing biases. For example, in the *Cogproc vs. Feel* contrast, Mistral's effect size decreased

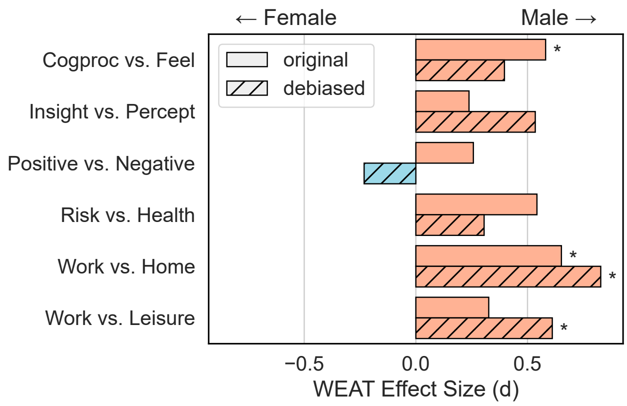

Figure 8. WEAT effect sizes (d) for gendered word associations before and after debiasing. Positive values indicate stronger associations with male terms, while negative values indicate stronger associations with female terms. Asterisks (*) indicate statistically significant results ($p < 0.05$).

significantly from $d = 0.92$ to $d' = 0.04$, indicating that the debiasing prompt successfully weakened the association between male terms and cognitive processes. However, Llama's effect size for the same contrast increased from $d = 0.24$ to $d' = 0.76$, suggesting that the prompt was less effective for this model and may have inadvertently reinforced the bias.

Similarly, in the *Work vs. Home* contrast, Llama's effect size increased significantly from $d = 1.18$ to $d' = 1.35$, indicating a stronger association between male terms and work-related domains after debiasing. In comparison, Mistral's effect size for the same contrast showed a modest increase from $d = 0.12$ to $d' = 0.31$, suggesting that the debiasing prompt had a limited impact on this model.

On the other hand, some contrasts showed reductions in bias after debiasing. For instance, in the *Risk vs. Health* contrast, both Mistral and Llama exhibited decreases in effect sizes (Mistral: $d = 0.58$ to $d' = 0.19$, Llama: $d = 0.51$ to $d' = 0.42$). This suggests that the debiasing prompt was somewhat effective in reducing the association between male terms and risk, as well as female terms and health. Such behavior is also observed in the *Positive vs. Negative* contrast.

The mixed results of the debiasing prompt highlight the complexity of mitigating gender biases in language models. While the prompt successfully reduced biases in some cases (e.g., *Cogproc vs. Feel* for Mistral), it exacerbated biases in others (e.g., *Work vs. Home* for Llama). This variability suggests that the effectiveness of debiasing techniques may

Table 4. Effect sizes before (d) and after (d') debiasing, with a summary of changes. Arrows indicate the direction of change (↑: increase; ↓: decrease). Statistically significant changes ($p < 0.05$) are marked with an asterisk (*). Cell colors indicate the magnitude of effect sizes: light for small ($|d| < 0.5$), medium for moderate ($0.5 \leq |d| < 0.8$), and dark for large ($|d| \geq 0.8$).

Query	Model	d	d'	s
Cogproc vs.	Mistral	0.92	0.04	↓*
Feel	Llama	0.24	0.76	↑*
Insight vs.	Mistral	0.22	0.30	↑
Percept	Llama	0.26	0.77	↑*
Positive vs.	Mistral	-0.11	0.05	↓
Negative	Llama	0.62	-0.51	↓*
Risk vs.	Mistral	0.58	0.19	↓
Health	Llama	0.51	0.42	↓
Work vs.	Mistral	0.12	0.31	↑
Home	Llama	1.18	1.35	↑*
Work vs.	Mistral	-0.05	0.01	↓
Leisure	Llama	0.70	1.21	↑*

depend on the specific model architecture, training data, and the nature of the biases being addressed.

Qualitative Analysis. To complement the quantitative results and look for deeper insights into the effects of the debiasing prompt, we also conducted a manual qualitative analysis. We analyzed a sample of narrative continuations generated by both Mistral and Llama, comparing outputs produced with the original prompt against those generated using the debiasing prompt for the same input sentences. This involved manually inspecting the texts for changes in gendered language, and choosing examples where the debiasing worked and where the debiasing did not work, as illustrated in the selected interactions as follows.[9]

[MASK] (Martim) partia para a caça com Poti.

Mistral: *Quando partiu para a caça com Poti, Anita, uma caçadora resoluta e experiente[...] Ela colocou o chapéu-de-cowboy em cima de seu cabelo castanho, quebrando a tensão da noite.*

Mistral-debiased: *Continuando a história, [MASK] arrolou a faca no guião de madeira e montou o arco ao lado de Poti, mostrando a mesma atenção e calma que tinha aprendido nos anos passados.*

Llama: Quando chegou à casa de Poti, [MASK] estava ansioso para começar a aventura. Ele se apresentou ao seu amigo e disse:[...]

Llama-debiased: Eles partiram cedo da manhã, quando o sol ainda estava nascente, e a brisa matinal carregava um cheiro de verde fresco e terra úmida.

The initial examples, featuring the male character *Martim*, show instances where the debiasing prompt encouraged a shift towards gender neutrality. Mistral, which originally personified [MASK] as a new female character (*"Anita"*, explicitly using "Ela" - her), produced a gender-neutral continuation focusing on Martim's actions when the debiasing prompt was applied. Similarly, Llama, which initially used a male pronoun ("Ele" - he) for Martim, shifted to a plural pronoun ("Eles" - them) in its debiased version. This plural could refer to Martim and Poti collectively, thereby obscuring Martim's individual gender in the immediate continuation and effectively neutralizing the character's explicit gendering. These instances qualitatively support our quantitative findings of an increased proportion of *unknown* gender classifications after debiasing. The next example considers a female character.

[MASK] (Iracema) parou em face do jovem guerreiro:

Mistral: *[MASK], sem pausa nem dúvida, acionou uma habilidade escondida em seu peito, a poderosa energia da luz que estava dentro dele começou a brilhar.*

Mistral-debiased: *[MASK] ergueu as mãos e deu um forte abraço ao guerreiro jovem, mostrando a ele uma cara resoluta. "Vou te ajudar", disse ela.*

Llama: Enquanto o jovem guerreiro olhava para ele com uma mistura de curiosidade e desconfiança, [MASK] tentou manter a calma e não mostrar seu trueiro nervosismo. Ele sabia que [...]

Llama-debiased: O rosto de [MASK] ficou sério, mas não surpreso, pois sabia que estava prestes a enfrentar um adversário formidável. Ele olhou para o jovem guerreiro com atenção, notando as marcas de luta em sua pele e a confiança brilhante nos olhos.

The outcomes for the female character *Iracema* present a more nuanced scenario. With Mistral, the original prompt led to [MASK] being gendered as male ("dele" - his). The debiased prompt, interestingly, resulted in a female gendering ("ela" - she). In contrast, Llama's output for [MASK] remained male-gendered ("Ele" - he) even after applying the debiasing prompt. Both outputs show the debiasing prompt being less effective in altering the model's established gender assignment or in promoting a neutral or alternative gender portrayal for this specific character.

4.4 Discussion

Our results highlight the presence of gender bias in generative language models. The analysis of gender distribution in generated texts indicates that both Mistral and Llama reproduce the gender imbalances observed in the original texts, reinforcing the predominance of male characters and the underrepresentation of female entities. This suggests that, rather than mitigating historical disparities, both considered models tend to perpetuate and potentially amplify existing biases learned from their training data. Indeed, while the specific literary texts used for our narrative continuations may not have been part of the models' exact training sets, it is reasonable to assume these LLMs were trained on similar datasets that invariably include texts reflecting historical and societal gender imbalances; such training data can subsequently influence their outputs when generating new narratives.

Moreover, the high rate of gender misclassification, particularly the tendency to misattribute female entities as male, illustrates how gender is asymmetrically inferred in generated narratives. The word association analysis further supports these observations, showing that both considered models maintain and, in some cases, intensify gendered word associations. For instance, associations linking male terms with cognitive processing, professional success, and risk-taking are more pronounced in the generated texts, while connections between female terms and emotional expression, domesticity,

[9] Both selected examples are from the work Iracema, by José de Alencar.

and caregiving remain strong. These patterns align with existing societal stereotypes and echo prior findings on gender bias in language models [Caliskan *et al.*, 2017; Lucy and Bamman, 2021; Huang *et al.*, 2021]. If left unchecked, such biases can influence the narratives produced by AI-driven storytelling and content generation, reinforcing traditional gender roles in ways that may be subtle yet impactful.

Another notable aspect of our findings is the presence of a substantial proportion of entities classified as *unknown* in the generated texts. This could indicate instances where the models generate more ambiguous descriptions, potentially reflecting a shift toward gender-neutral language in uncertain contexts. While this may suggest some level of neutrality, further investigation is needed to explore whether these cases stem from an unbiased generative process or simply reflect model uncertainty and avoidance. The increase in *unknown* labels after debiasing, as observed in our bias mitigation analysis, further complicates this issue, as it raises questions about the trade-off between reducing explicit biases and maintaining meaningful gender representation.

The consistency of these patterns across both models suggests that these biases are not unique to a specific architecture but are somewhat inherent to many generative language models trained on historical and contemporary corpora. This aligns with prior research, such as Stanczak and Augenstein [2021], which found that gender biases in language models are often deeply rooted in the training data and the societal norms these datasets reflect. Addressing these multifaceted issues, from inherent data biases to challenges in model transparency, requires a combination of approaches. These include fostering more transparent data curation practices, developing bias-aware training strategies and controlled fine-tuning methods, and establishing robust evaluation frameworks that systematically assess gender representation across different linguistic and cultural dimensions.

5 Conclusion

In this study, we investigate gender bias in Portuguese-language text generation by analyzing how LLMs extend literary narratives. In particular, we extract masked sentences from public domain works and prompt Mistral and Llama models to generate narrative continuations. We then analyze these generated texts to verify whether the models exhibit gender bias in storytelling. Additionally, we explore prompt engineering as a potential bias mitigation strategy to reduce gender bias in generated narratives.

Overall, our findings reveal that both considered models tend to perpetuate and, in some cases, amplify existing gender imbalances and stereotypes present in their training data. Specifically, male characters are overrepresented and more frequently associated with cognitive and professional domains, while female characters are underrepresented and linked to emotional and domestic roles. These patterns align with societal stereotypes and highlight the challenges of addressing gender bias in generative language models.

Also, our results indicate potential for bias mitigation, particularly through prompt-based interventions, but also reveal inherent trade-offs. The observed increase in gender-neutral (*unknown*) classifications following debiasing raises critical questions about whether these interventions truly promote neutrality or merely introduce uncertainty into the models' outputs. This highlights the complexity of addressing bias in generative AI: reducing explicit biases must be carefully balanced with preserving meaningful gender representation.

Limitations. While our study provides valuable insights into gender bias in Portuguese-language text generation, it is not free of limitations. First, our analysis focuses on a specific set of public domain works, which may not fully capture the diversity of narratives present in contemporary literature. Additionally, our findings are limited to the selected multilingual models (Mistral-7B-Instruct and LLaMA 3.2-3B) and the specific prompt engineering techniques applied, meaning that results may vary for other models and mitigation strategies. A direct comparative analysis with a monolingual Portuguese model was not included in our scope, primarily due to the relative scarcity of open-source, instruction-tuned models developed exclusively for Portuguese at the time of this research. Still, such a comparison is open as future step: to verify whether bias manifestations differ between multilingual and dedicated monolingual architectures.

Also, a key challenge in this study (and in the broader field of LLM evaluation) is the often-limited transparency surrounding training datasets of many models, including the versions of Mistral-7B-Instruct and LLaMA 3.2-3B employed. The proliferation of models, often released with proprietary or inadequately documented training data, harms tracing the origins of observed biases or to ascertain the extent to which evaluation materials might overlap with training sets. This opacity hinders a deeper understanding of bias and the development of more targeted mitigation strategies.

Another limitation relates to the scope and nature of our evaluation metrics. We explore gender distribution, misclassification rates, and word associations; still, bias in storytelling is a multifaceted phenomenon that could be further analyzed using alternative methods, such as discourse analysis or more nuanced contextual assessments. Specifically, our word association analysis relied on LIWC (BP-LIWC2015) to define thematic word sets. Although LIWC offers a structured psycholinguistic lexicon, we acknowledge that dictionary-based approaches inherently operate with a degree of context-insensitivity. The words within these predefined categories carry meanings and connotations that can shift significantly based on their specific usage within the generated narratives.

Consequently, the associative biases identified via WEAT using these static, decontextualized word sets might not fully capture the subtleties or the precise contextual manifestation of these themes. The WEAT metric itself, while a standard method for quantifying biases in word embeddings, also has known limitations [Ethayarajh *et al.*, 2019], such as potentially conflating different types of associations into a single measure and sensitivity to the choice of word sets. Alternative or complementary bias quantification methods have been proposed and could offer additional perspectives. Finally, our bias mitigation experiments focus on prompt-based interventions, leaving open the question of how more advanced techniques—such as fine-tuning or adversarial training—might impact bias reduction.

Future Work. These limitations inspire interesting directions for future work. Expanding the analysis to a broader and more diverse corpus, including contemporary texts and different genres, would provide a more comprehensive understanding of gender bias in generative models. Further research could also involve evaluating a wider range of LLMs. For instance, comparing multilingual models with monolingual models specifically trained or adapted for Portuguese could reveal whether similar bias patterns persist across different model types. Additionally, exploring models with advanced reasoning capabilities, such as DeepSeek, might offer insights into how enhanced reasoning influences the manifestation of bias and the effectiveness of debiasing techniques.

Methodologically, future studies could enhance bias assessment. This includes exploring more dynamic or context-aware methods for defining thematic attribute sets for association tests, moving beyond static lexicons and thereby addressing some of the contextual limitations of tools like LIWC. Employing alternative or complementary bias quantification metrics to WEAT would also be beneficial for gaining a more nuanced understanding of learned biases. Finally, exploring alternative bias mitigation strategies, such as adversarial training, dataset augmentation, or debiasing algorithms, could help identify more effective and scalable solutions for reducing gender bias in LLM-generated narratives.

Declarations

Acknowledgements

We acknowledge the computational resources that enabled this study and recognize the environmental impact associated with LLM inference. The total generation process required 47 hours of computation on an NVIDIA GeForce RTX 4050 GPU, resulting in an estimated carbon footprint of 1.51 kg CO_2, considering the GPU's power consumption of approximately 115W and the average carbon intensity of electricity (0.4 kg CO_2/kWh).

Authors' Contributions

Mariana O. Silva is the main contributor and writer of this manuscript, contributing to the conception of this study, the data curation, and the experiments. Michele A. Brandão and Mirella M. Moro contributed to the supervision, review, and editing of the final manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by *Fundação de Amparo à Pesquisa do Estado de Minas Gerais*, (FAPEMIG), and supported by *Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação* (INCT-TILD-IAR).

Availability of data and materials

The datasets generated and analyzed during the current study are available at [Silva *et al.*, 2025]. The code developed during the current study will be made available upon request.

References

Alhussain, A. I. and Azmi, A. M. (2021). Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.*, 54(5):103:1–103:38. DOI: 10.1145/3453156.

Assi, F. M. and Caseli, H. d. M. (2024). Biases in GPT-3.5 Turbo model: a case study regarding gender and language. In *Simp. Bras. de Techologia da Informação e da Linguagem Humana*, STIL, pages 294–305. SBC. DOI: 10.5753/stil.2024.245358.

Bauer, L., Mehrabi, N., Goyal, P., Chang, K.-W., Galstyan, A., and Gupta, R. (2024). BELIEVE: Belief-enhanced instruction generation and augmentation for zero-shot bias mitigation. In *Procs of Workshop on Trustworthy Natural Language Processing*, TrustNLPf, pages 239–251. ACL. DOI: 10.18653/v1/2024.trustnlp-1.20.

Bender, E. M., Gebru, T., McMillan-Major, A., *et al.* (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Procs ACM Conference on Fairness, Accountability, and Transparency*, FAccT, pages 610–623. ACM. DOI: 10.1145/3442188.3445922.

Bojanowski, P., Grave, E., Joulin, A., *et al.* (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. DOI: 10.1162/tacl_a_00051.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., *et al.* (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29 of *NIPS*. Available at:`https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html`.

Brown, T. *et al.* (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33 of *NIPS*, pages 1877–1901. Available at:`https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_source=transaction&utm_medium=email&utm_campaign=linkedin_newsletter`.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. DOI: 10.1126/science.aal4230.

Carvalho, F., Junior, F. P., Ogasawara, E., *et al.* (2024). Evaluation of the brazilian portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources and Evaluation*, 58(1):203–222. DOI: 10.1007/s10579-023-09647-2.

Carvalho, F., Rodrigues, R., Santos, G., *et al.* (2019). Avaliação da versão em português do LIWC lexicon 2015 com análise de sentimentos em redes sociais. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Min-*

ing, BraSNAM, pages 24–34. SBC. DOI: 10.5753/brasnam.2019.6545.

Chakrabarty, T., Padmakumar, V., Brahman, F., *et al*. (2024). Creativity Support in the Age of Large Language Models: An Empirical Study Involving Professional Writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 132–155. ACM. DOI: 10.1145/3635636.3656201.

Chakrabarty, T., Padmakumar, V., He, H., *et al*. (2023). Creative Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, EMNLP, pages 34–40. ACL. DOI: 10.18653/v1/2023.emnlp-tutorial.6.

Cheng, J. (2020). Fleshing Out Models of Gender in English-Language Novels (1850 – 2000). *Journal of Cultural Analytics*, 5(1). DOI: 10.22148/001c.11652.

Ding, Y., Zhao, J., Jia, C., Wang, Y., *et al*. (2025). Gender Bias in Large Language Models across Multiple Languages: A Case Study of ChatGPT. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579, Albuquerque, New Mexico. Association for Computational Linguistics. Available at:`https://aclanthology.org/2025.trustnlp-main.36/`.

Dong, C., Li, Y., Gong, H., *et al*. (2022). A Survey of Natural Language Generation. *ACM Comput. Surv.*, 55(8):173:1–173:38. DOI: 10.1145/3554727.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705. ACL. DOI: 10.18653/v1/P19-1166.

Freitas, C. and Santos, D. (2023). Gender depiction in portuguese. In *Conference Reader: 2nd Annual Conference of Computational Literary Studies*, CCLS, pages 4–30. DOI: 10.48694/jcls.3576.

Gallegos, I. O., Rossi, R. A., Barrow, J., *et al*. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179. DOI: 10.1162/coli_a_00524.

Ganguli, D., Askell, A., Schiefer, N., *et al*. (2023). The Capacity for Moral Self-Correction in Large Language Models. *arXiv preprint*. DOI: 10.48550/arXiv.2302.07459.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. ACL. DOI: 10.48550/arXiv.1903.03862.

Gonçalo Oliveira, H. (2024). Automatic generation of creative text in Portuguese: an overview. *Language Resources and Evaluation*, 58(1):7–41. DOI: 10.1007/s10579-023-09646-3.

Guo, Y., Guo, M., Su, J., *et al*. (2024). Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv preprint*. DOI: 10.48550/arXiv.2411.10915.

Hartmann, N. S. *et al*. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Simp. Bras. de Techologia da Informação e da Linguagem Humana*, STIL, pages 122–131. SBC. DOI: 10.48550/arXiv.1708.06025.

Hu, T., Kyrychenko, Y., Rathje, S., *et al*. (2025). Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75. DOI: 10.1038/s43588-024-00741-1.

Huang, H., Tang, T., Zhang, D., *et al*. (2023). Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. *arXiv preprint*. DOI: 10.48550/arXiv.2305.07004.

Huang, T., Brahman, F., Shwartz, V., *et al*. (2021). Uncovering Implicit Gender Bias in Narratives through Commonsense Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873. ACL. DOI: 10.18653/v1/2021.findings-emnlp.326.

Jiang, A. Q., Sablayrolles, A., Mensch, A., *et al*. (2023). Mistral 7b. *arXiv preprint*. DOI: 10.48550/arXiv.2310.06825.

Lima, L. F. F. P. d. and Araujo, R. M. d. (2023). A call for a research agenda on fair NLP for Portuguese. In *Simp. Bras. de Techologia da Informação e da Linguagem Humana*, STIL, pages 187–192. SBC. DOI: 10.5753/stil.2023.233763.

Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55. ACL. DOI: 10.18653/v1/2021.nuse-1.5.

Luo, K., Mao, Y., Zhang, B., *et al*. (2024). Reflecting the Male Gaze: Quantifying Female Objectification in 19th and 20th Century Novels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, LREC-COLING 2024, pages 13803–13812. ELRA and ICCL. DOI: 10.48550/arXiv.2403.17158.

Mikolov, T., Sutskever, I., Chen, K., *et al*. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. Available at:`https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html`.

Mondshine, I., Paz-Argaman, T., and Tsarfaty, R. (2025). Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico. Association for Computational Linguistics. DOI: 10.48550/arXiv.2502.09331.

Musacchio, E., Siciliani, L., Basile, P., *et al*. (2024). Adapting Large Language Models to Narrative Content. In *Procs. Workshop on Artificial Intelligence and Creativity co-located with 27th European Conference on Artificial Intelligence*, ECAI, pages 17–29. CEUR-WS. Available at:`https://ceur-ws.org/Vol-3810/paper2.pdf`.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality*. DOI: 10.1145/3597307.

Oba, D., Kaneko, M., and Bollegala, D. (2024). In-Contextual Gender Bias Suppression for Large Language Models. In *Findings of the Association for Computational Linguistics*, EACL, pages 1722–1742. ACL. Available at:`https://livrepository.liverpool.ac.uk/3178835/1/EACL24_PromptBias_CameraReady.pdf`.

Omrani Sabbaghi, S. and Caliskan, A. (2022). Measuring

Gender Bias in Word Embeddings of Gendered Languages Requires Disentangling Grammatical Gender Signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES, pages 518–531. ACM. DOI: 10.1145/3514094.3534176.

Peeperkorn, M., Kouwenhoven, T., Brown, D., *et al.* (2024). Is temperature the creativity parameter of large language models? *arXiv preprint*. DOI: 10.48550/arXiv.2405.00492.

Petrov, A., Malfa, E. L., Torr, P. H., *et al.* (2023). Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 36963–36990. Curran Associates Inc.. DOI: 10.48550/arxiv.2305.15425.

Qiu, H., Xu, Y., Qiu, M., *et al.* (2025). DR.GAP: Mitigating Bias in Large Language Models using Gender-Aware Prompting with Demonstration and Reasoning. *arXiv preprint*. DOI: 10.48550/arXiv.2502.11603.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there Gender bias and stereotype in Portuguese Word Embeddings? *arXiv preprint*. DOI: 10.48550/arXiv.1810.04528.

Santos, D. (2021). Portuguese Novel Corpus (ELTeC-por): April 2021 release. *Zenodo*. DOI: 10.5281/zenodo.4288235.

Santos, D., Freitas, C., and Bick, E. (2018). Obras: a fully annotated and partially human-revised corpus of brazilian literary works in public domain. *CorLex 24 de setembro de 2018*. Available at:`https://www.linguateca.pt/Diana/download/CorLex.pdf`.

Shi, F., Suzgun, M., Markus, F., *et al.* (2022). Language Models are Multilingual Chain-of-Thought Reasoners. *arXiv preprint*. DOI: 10.48550/arXiv.2210.03057.

Silva, M., Brandão, M., and M. Moro, M. (2025). Rewriting Stories with LLMs: Gender Bias in Generated Portuguese-language Narratives. Zenodo. DOI: 10.5281/zenodo.15756454.

Silva, M. and Moro, M. (2024). NLP Pipeline for Gender Bias Detection in Portuguese Literature. In *Anais do LI Seminário Integrado de Software e Hardware*, SEMISH, pages 169–180. SBC. DOI: 10.5753/semish.2024.2914.

Silva, M. O., de Melo-Gomes, L., and Moro, M. M. (2024). From words to gender: Quantitative analysis of body part descriptions within literature in portuguese. *Information Processing & Management*, 61(3):103647. DOI: 10.1016/j.ipm.2024.103647.

Silva, M. O., Scofield, C., de Melo-Gomes, L., *et al.* (2022). Cross-collection dataset of public domain portuguese-language works. *Journal of Information and Data Management*, 13(1). DOI: 10.5753/jidm.2022.2349.

Silva, M. O., Scofield, C., and Moro, M. M. (2021). PPORTAL: Public domain Portuguese-language literature Dataset. In *Anais do III Dataset Showcase Workshop*, pages 77–88. SBC. DOI: 10.5753/dsw.2021.17416.

Stanczak, K. and Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing. *arXiv preprint*. DOI: 10.48550/arXiv.2112.14168.

Stuhler, O. (2024). The gender agency gap in fiction writing (1850 to 2010). *Proceedings of the National Academy of Sciences*, 121(29):e2319514121. DOI: 10.1073/pnas.2319514121.

Sun, T., Gaut, A., Tang, S., *et al.* (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv preprint*. DOI: 10.48550/arXiv.1906.08976.

Taso, F. T. d. S., Reis, V. Q., and Martinez, F. V. (2023). Sexismo no Brasil: análise de um Word Embedding por meio de testes baseados em associação implícita. In *Simp. Bras. de Techologia da Informação e da Linguagem Humana*, STIL, pages 53–62. SBC. DOI: 10.5753/stil.2023.233845.

Touvron, H. *et al.* (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. DOI: 10.48550/arXiv.2302.13971.

Xu, H., Zhang, Z., Wu, L., *et al.* (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PLOS ONE*, 14(11):e0225385. DOI: 10.1371/journal.pone.0225385.

Yang, K., Tian, Y., Peng, N., *et al.* (2022). Re3: Generating Longer Stories With Recursive Reprompting and Revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 4393–4479. ACL. DOI: 10.18653/v1/2022.emnlp-main.296.

Zampieri, M. and Becker, M. (2013). Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5. Available at:`https://romanistik.phil-fak.uni-koeln.de/sites/romanistik/Mitarbeiter/Becker/PDFs_zum_Download/colonia.pdf`.

Zhao, J., Wang, T., Yatskar, M., *et al.* (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. ACL. DOI: 10.18653/v1/N18-2003.

Zhao, J., Wang, T., Yatskar, M., *et al.* (2019). Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634. ACL. DOI: 10.18653/v1/N19-1064.

Zhou, P. *et al.* (2019). Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP, pages 5276–5284. ACL. DOI: 10.18653/v1/D19-1531.

A Text Generation Details

This appendix provides additional details on the text generation process, including model execution, inference settings, hardware specifications, and an illustrative example of input-output pairs.

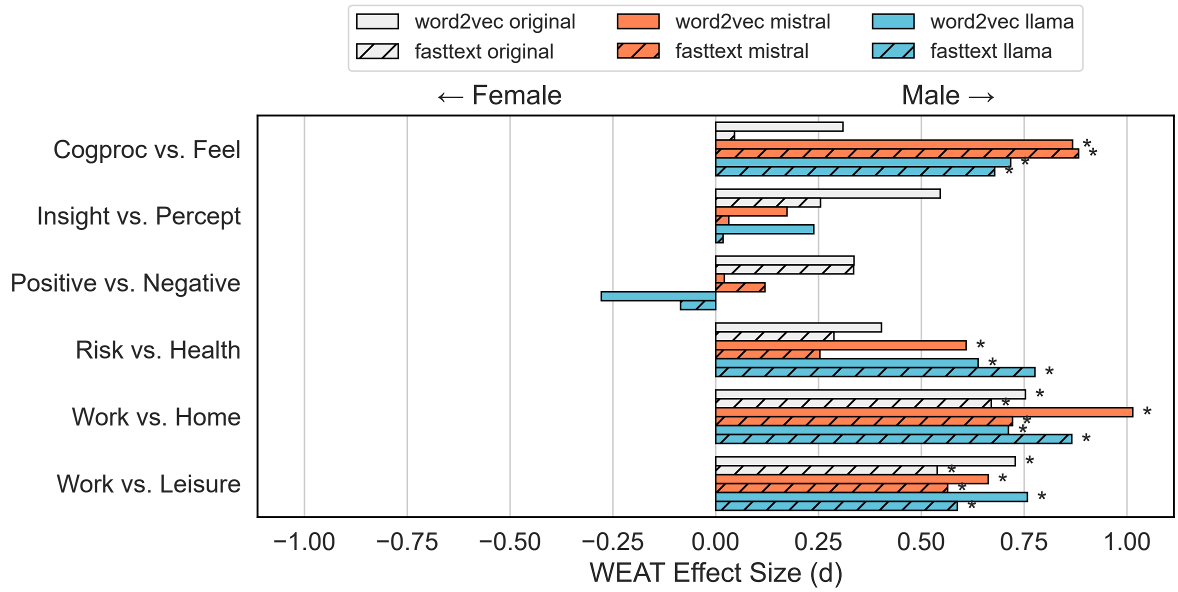

Figure 9. WEAT effect sizes (d) for gendered word associations across all category pairs and datasets. Positive values indicate stronger associations with male terms, while negative values indicate stronger associations with female terms. Asterisks (*) denote statistically significant results ($p < 0.05$).

A.1 Model Execution and Configuration

We employed two instruction-tuned LLMs for text generation: *Mistral-7B-Instruct* and *LLaMA 3.2-3B*. Both models were executed using *Ollama*, a framework optimized for running large language models locally with efficient resource management. Ollama provides a streamlined inference pipeline, allowing rapid text generation without requiring cloud-based APIs. The models were downloaded as pre-trained checkpoints and loaded into the Ollama framework.

The inference process was conducted with the following configuration parameters:

- **Model:** Mistral-7B-Instruct / LLaMA 3.2-3B
- **Temperature:** 0.9 (for creative variation in responses)
- **Top-p:** 0.9 (to balance coherence and diversity)
- **Stop:** `\n` (to ensure single-sentence continuations)
- **Number of generations per input:** 1

A.2 Hardware Specifications

The text generation was conducted on a local machine with the following hardware specifications:

- **Processor:** 13th Gen Intel(R) Core(TM) i7-13650HX
- **RAM:** 32GB
- **GPU:** NVIDIA GeForce RTX 4050

A.3 Example of Text Generation

To illustrate the model's behavior, Table 5 presents an example of an input-output pair for one of the processed sentences.

This example highlights differences in narrative style between the two models, showcasing their ability to produce coherent and contextually relevant continuations.

Table 5. Example of input-output pairs for text generation.

Original sentence: "Capitu foi ao muro, e, com o prego, disfarçadamente, apagou os nossos nomes escritos."
Input Prompt: **Instructions:** continue writing in pt-br the story from the following sentence, describing what the character did next. The character is `[MASK]`. **Sentence:** "`[MASK]` foi ao muro, e, com o prego, disfarçadamente, apagou os nossos nomes escritos."
Model Output (Mistral-7B-Instruct): "Depois de apagar sutilmente os nossos nomes do muro com o prego, `[MASK]` segurou a respiração para não desmascará-lo e olhou em volta da sala. Com um sorriso leve, ele voltou ao seu lugar no banco e voltou a ficar absorvido no livro. Aparentemente, ele estava tão envolvido com o romance de sua leitura que ninguém notou a mudança na sala ou percebeu que os nomes dos presentes haviam desaparecido."
Model Output (LLaMA 3.2-3B): "Enquanto fazia isso, seus olhos se moviam rapidamente para garantir que ninguém estivesse assistindo, e sua mente estava em constante alerta para qualquer sinal de perigo. Com um lastimeiro gesto, apagou os nomes escritos no muro, deixando apenas uma marca escurecida onde havia uma vez as palavras: *Eles não estão mais aqui*. Sua mão esfregou o prego contra a camisa, para garantir que nenhum indício de sua atividade tivesse sido deixado atrás."

A.4 Processing Time

The average inference time per sentence was approximately 4 seconds for *Mistral-7B-Instruct* and 2 seconds for *LLaMA 3.2-3B*. The total generation process for the entire dataset was completed within 47 hours.

B Additional Results

This appendix provides supplementary results from the lexicon-based analysis, including detailed WEAT effect sizes for all category pairs and datasets. These results offer a more granular view of the gendered word associations in the original and generated texts.

Figure 9 presents the WEAT effect sizes for all ten category pairs, comparing the original texts with the outputs of Mistral and Llama. The results reveal consistent patterns of gendered associations, with male terms more strongly linked to cognitive processes, work, and risk, while female terms are associated with emotions, home, and health. These findings align with the broader trends discussed in Section 3.3.2 and highlight the persistence of gender stereotypes in LLM-generated narratives.