Clustering Discourses: Racial Biases in Short Stories about Women Generated by Large Language Models

Gustavo Bonil[1]*, João Gondim[2]*, Marina dos Santos[1], Simone Hashiguti[1], Helena Maia[2], Nadia Silva[3], Helio Pedrini[2], Sandra Avila[2]

[1]Instituto de Estudos da Linguagem [2]Instituto de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas – SP – Brasil

[3]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

g237221@dac.unicamp.br, {joao.gondim,helio,sandra}@ic.unicamp.br

Abstract. *This study investigates how large language models, in particular LLaMA 3.2-3B, construct narratives about Black and white women in short stories generated in Portuguese. From 2100 texts, we applied computational methods to group semantically similar stories, allowing a selection for qualitative analysis. Three main discursive representations emerge: social overcoming, ancestral mythification and subjective self-realization. The analysis uncovers how grammatically coherent, seemingly neutral texts materialize a crystallized, colonially structured framing of the female body, reinforcing historical inequalities. The study proposes an integrated approach, that combines machine learning techniques with qualitative, manual discourse analysis.*

Resumo. *Este estudo investiga como grandes modelos de linguagem, em particular o LLaMA 3.2-3B, constroem narrativas sobre mulheres negras e brancas em contos gerados em português. A partir de 2100 textos, aplicamos métodos computacionais para agrupar contos semanticamente semelhantes, permitindo uma seleção para análise qualitativa. Três principais representações discursivas/tipos de narrativas emergem: superação social, mitificação ancestral e autorrealização subjetiva. A análise revela como textos gramaticalmente coerentes e aparentemente neutros materializam um enquadramento cristalizado e colonialmente estruturado do corpo feminino, reforçando desigualdades históricas. O estudo propõe uma abordagem integrada, que combina técnicas de aprendizado de máquina com análise qualitativa e manual do discurso.*

1. Introduction

The rise of Large Language Models (LLMs) as chatbots has led to their broad use across domains — from therapy and professional settings to programming, content creation, and education [Zao-Sanders 2025]. Critical readings on the uncritical use of these technologies in sensitive social contexts (such as education and therapy, especially for

[1]Equal contribution.

school-age children, highly susceptible to influences) highlight the importance of discussing biases in these models, which can be detrimental to the functioning of society, as in the case of so-called *algorithmic racism* [Silva 2022] that exposes the tendency of algorithmic systems to amplify already existing social inequalities. In view of this, this article builds on our previous work, which, from the perspective of language studies, proposed a qualitative and interpretive manual analysis of texts generated by LLMs [Bonil et al. 2025]. The objective then was to investigate whether there were noticeable differences in the representation of social minorities, contributing to the growing literature on bias [Blodgett et al. 2020]. In the previous work, conducted by a transdisciplinary team composed of researchers from computational studies and applied linguistics, a systematic differentiation in the representation of women with different skin tones was identified in a smaller and controlled dataset using seven different LLMs.

In this paper, we build on the previous qualitative analysis by combining computational and discursive methods to explore a larger dataset (25×) generated with LLMs. This integrated approach enables us to identify and interpret discourses at scale, revealing how they can be visualized as clusters. This is achieved by applying clustering algorithms to text embeddings, assuming that semantically similar texts will be assigned to the same group. Consequently, only representative samples are selected for manual verification during the qualitative analysis. Our central research question is: *how do LLMs construct, differentiate, and hierarchize white and Black*[1] *female characters in the narratives they generate?* We also examine what discourses are activated in these texts and how they shape possible social boundaries. In a responsible scenario, we understand that LLMs should produce equivalent textual outputs regardless of the protagonist's racial identity, preserving stylistic diversity while promoting a plurality of representations and plots.

Our main contributions are: (i) the development of a mixed-methodology that combines quantitative computational techniques with qualitative, manual analysis of data from large datasets, enabling a partially automated discourse analysis; (ii) an investigation into how LLMs handle the Portuguese language, highlighting how biases are linguistically manifested in this context. By doing so, we contribute to addressing the scarcity of studies examining this phenomenon, particularly within the Brazilian context, and (iii) a pipeline showing that the use of encoders' outputs for discursive analysis can be a valuable tool for computer scientists and discourse analysts when assessing biases in LLMs.

2. Related Work

The existence of biases in language models has been pointed out by various authors. For instance, [Abid et al. 2021] and [Venkit et al. 2023] contend that certain models generate negative outputs for underrepresented nationalities , while [Bordia and Bowman 2019] argue gender bias can be documented as operating in a lexical dimension, and [Lucy and Bamman 2021]demonstrate that GPT-3 associates women with less prestigious roles, regardless of prompts. [May et al. 2019] show that ELMo and BERT associate African-American names with negative terms. [Sheng et al. 2019] demonstrate that GPT-2 links women and Black individuals with stigmatized occupations, underscoring the need to consider intersectional bias.

Recent audits reinforce these concerns. [Salinas et al. 2024] analyze GPT-4 out-

[1]We capitalize "Black" to affirm identity and resist objectification [Grant and Grant 1975, Tharps 2014].

puts across 168,000 responses and show consistent favoritism toward white men. They adopt a legal framework to interpret model behavior, while [Assi and Caseli 2024] evaluate GPT-3.5 Turbo and find gender and language-based disparities, with English prompts favored over Portuguese. Human sciences complement this view by highlighting discursive harm. [Araújo 2024] critiques an AI-generated image reinforcing racial stereotypes. [Corazza et al. 2024] show how ChatGPT narratives reproduce cis-heteronormative and Western norms, excluding marginalized identities. These studies help address the gap pointed out by [Blodgett et al. 2020], identifying who is harmed and why. They reframe bias as not merely a technical flaw, but a reproduction of systemic social inequality.

Unlike works focused on quantitative metrics [Salinas et al. 2024, Assi and Caseli 2024], our approach centers on how language in LLM outputs constructs meaning and ideology. Rather than measuring disparities, we use computational methods to select data and conduct close, qualitative analyses of narrative, lexical, and discursive patterns. This allows us to explore how stereotypes are embedded and naturalized in text. Our work also contributes to the still-scarce literature on Portuguese-language LLM outputs, an aspect rarely explored outside [Assi and Caseli 2024]. While technical solutions exist [Bordia and Bowman 2019, May et al. 2019, Sheng et al. 2019], we argue that mitigating bias requires human expertise, especially from discourse analysts, to critically interpret how language sustains social hierarchies.

In contrast to the works developed by [Corazza et al. 2024] and [Araújo 2024], who work with smaller datasets, we combine automation to filter large-scale data with close reading. This integration enables us to examine how LLMs contribute to the normalization of social inequality, echoing the concern that "[These systems] encode systematic biases against women and people of color [...] implicating many NLP systems in the amplification of social injustice" [May et al. 2019].

3. Methodology for Dataset Creation and Corpus Selection

Figure 1 depicts the methodology process. We scale the analysis by first automating the creation of a corpus of short stories; we queried an LLM with a prompt asking for the creation of a short story ("conto" in Portuguese). Next, we encoded the stories created into feature vectors. These features were first validated as representative of our stories and then used to find clusters of stories with unsupervised learning algorithms.

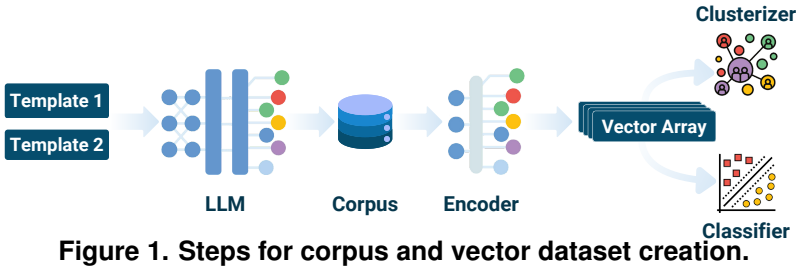

Figure 1. Steps for corpus and vector dataset creation.

We queried Llama-3.2-3B-Instruct [Grattafiori et al. 2024] (with $top\_p = 0.9$ and $temperature = 0.8$) using two prompt templates: (1) "Escreva um conto sobre uma mulher negra/branca chamada [name]"[2], where "[name]" was replaced with a given name, retrieved from the Python library `names-dataset` [Remy 2021], based on a Facebook

[2](1) "Write a short story about a Black/white woman named [name]".

massive data leakage [Guardian 2021], comprising 730,000 first names and 983,000 last names; and (2) "Escreva um conto sobre uma mulher negra/branca"[3]. For template (1), we prompted the model 5 times with each most popular name from the 105 countries in the library to create short stories for women with 2 skin tones (Black/white), generating 1050 short stories with names. For template (2), the model was prompted with each query 525 times[4] per skin tone, resulting in 1050 short stories without a name. In total, 2100 short stories were crafted.

We carefully standardized the prompts to maintain a generic and neutral character, avoiding any bias that could influence the models' responses. This methodological decision is based on evidence from our previous studies, which demonstrate that including contextual elements in the prompts tends to induce the models to construct characters aligned with implicit stereotypes or expectations suggested by the statement itself. To ensure a broader and more impartial observation of the discursive patterns generated by the models, we chose to minimize contextualization. In addition, we decided to previously fix the names of the characters since, in a previous analysis[1], we observed that the names attributed by the models themselves significantly influenced the narrative structure and content. By controlling for this variable, we ensured that the only difference between the versions analyzed was the variation in the characters' skin color, allowing for a more precise analysis of the impact of this specific factor on the narratives generated.

Stories were encoded using BGE M3 [Chen et al. 2024] for its 8192-token capacity and multilingual support. To assess if the created vectors appropriately represented the stories outputted by the model, we used as a proxy the performance of different classifiers trained on the features extracted to predict whether the array represented a short story about a Black or white woman. We leveraged Support Vector Machines (SVM) [Cortes and Vapnik 1995] for the evaluated model due to its known good capacities for classification tasks [Cervantes et al. 2020], and also to evaluate different models through hyperparameter tuning. We trained our model with a stratified 5-fold, balancing stories by skin tone, modifying the kernel (linear, polynomial, and gaussian) and C values (0.1, 0.5, 1, 1.5, and 2), resulting in 75 trainings.

We applied clustering algorithms in the original 1024-dimensional vectors to assess if the found cluster would be able to characterize common narratives. We chose the algorithm and the hyperparameters that maximized the Variance Ratio Criterion (VRC) [Caliński and Harabasz 1974], a score that measures how similar an object is to its own cluster (cohesion) when compared with the other clusters (separation); a higher VRC indicates more well-defined clusters. Once the algorithm and hyperparameters were found, we selected representative stories for each defined cluster (except the outliers) by calculating the distances of all points within a cluster and selecting the top 3 ones with minimum and maximum mean distance within its own clusters.

4. Analysis

This section presents the main findings from our analysis, which combined computational and qualitative approaches to examine linguistic, structural, discursive, and symbolic pat-

[3](2) "Write a short story about a Black/white woman".

[4]Value selected to equalize the number of short stories with and without names.

terns in the corpus to understand intersecting racial and gender biases. Our code is available at https://github.com/hiaac-nlp/clusteringdiscourses.

4.1. Feature Representation and Clustering

First, we analyze the results of training different classifiers to predict whether the short story is about a Black or a white woman. The classifiers averaged $94.89 \pm 4.31\%$, with the Gaussian kernel SVM peaking 100% accuracy with three values for C. These values highlight not only that the encoded features make a good representation of the short stories, but also that predicting if a story about a Black or a white woman appears to be an easy task for SVMs, even for less complex models using polynomial kernels, that still reached 86.69% accuracy.

The final algorithm was a DBSCAN yielding a VRC of 114.41 and with three clusters found (besides the "outliers" cluster). Therefore, we analyzed 18 short stories (six for each cluster). To visualize the structure of our corpus, we employed UMAP [McInnes et al. 2018] to reduce the original dimensions from 1024 to 2. Figure 2 shows the visualization with the cluster identifications, and the percentages of stories about Black or white women within that cluster. Stories are marked with an "X" with a yellow edge line if at a minimum distance within the corpus and a red edge line otherwise.

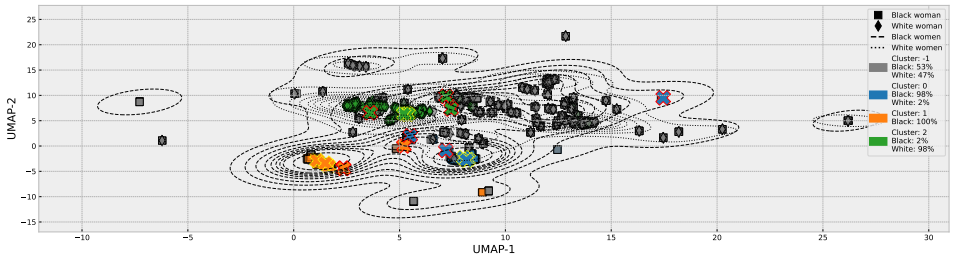

Figure 2. Representation of corpus' structure in reduced dimension.

4.2. A Qualitative View on the Corpus

Based on our understanding that language is an "essentially and inseparably social" [Rajagopalan 2023, p. 208, our translation] phenomenon, we opted for a qualitative approach, similar to the one adopted in the base study. Our team comprises Language and Computer Science professionals, which also allowed us, in addition to metric and technical aspects, to carry out a careful and judicious reading of the discursive aspects present in the selected productions, as detailed in Section 3.

We start from the understanding that objectivity is not achieved by excluding subjectivity or applying fixed rules but by carefully constructing interpretations based on the intentions and meanings attributed by participants in specific contexts, as pointed out in a previous study [Maxwell 1992]. According to the author, validity in academic research rests less on replicability and more on the plausibility and coherence of interpretations in light of the context investigated. We, therefore, conducted an in-depth qualitative investigation after defining which stories in the corpus would be relevant for manual analysis. Based on [Pêcheux 2022, Courtine 1981] works, we conceived discursive analysis as a process of interpretation that takes place in stages.

First, we conducted a careful general reading of the selected stories. Then, with support from three analysts (one man, two women), we began a qualitative analysis, interpreting the texts through various discursive excerpts to extract discursively significant

Discursive Sequences (DSs). These DSs were comparatively analyzed to identify dominant discourses. Key DSs were selected to synthesize recurring meanings, termed Discursive Sequences of Reference (DSRs). Finally, we performed an interpretative analysis to critically relate these meanings to broader social structures. Sections 4.2.1 and 4.2.2 present DSR excerpts from each cluster, illustrating core discursive meanings — identified, discussed, and validated by all analysts — which we hypothesize that impacted the clustering. We then examined broader discursive differences in stories about Black and white women.

4.2.1. A Close View

As shown in Figure 2, we identified the formation of three clusters. Cluster 0 is composed mainly of stories about Black women, unique in cluster 1. In contrast, cluster 2 is formed almost exclusively by texts about white women.

When analyzing the texts in **cluster 0**, we observed a predominance of realistic narratives, set in contexts of social vulnerability, marked by individual overcoming, daily resistance, and a strong presence of themes such as education, health, and community transformation. Even in cases where the characters are white women, the plot often presents trajectories marked by structural obstacles that are overcome. Protagonists are resilient characters who face structural adversities such as racism, poverty, or gender discrimination. Many of these characters build trajectories of social ascension through merit and effort as doctors, teachers, or community leaders. Although social conflict is central, the narrative often culminates in personal and collective achievements, generating a sense of hope, community value, and possible agency. A representative example of this representation can be observed in DSR1.

DSR1: *"Com o tempo, Aisha se tornou uma figura respeitada e amada em sua comunidade. Ela provou que, mesmo nas condições mais difíceis, uma pessoa pode fazer a diferença. Seu coração era puro, e sua determinação era inigualável"*[5].

On the other hand, the stories in **cluster 1** present a significant change in narrative tone. These are stories in the form of legends, fables, or local myths, with a strong presence of magical, symbolic, and religious elements. The characters — all Black women — are constructed as archetypal figures: queens, healers, priestesses, or warriors endowed with supernatural powers. They are directly connected with deities, nature, or ancestral knowledge, often inherited from older women in the community. The protagonist is anchored in a collective and spiritual dimension and not only solves problems but also saves the community through rituals and pacts with mythical entities or forces of nature. Although these are still Black female characters who exercise leadership, the axis of transformation is no longer the material social structure (as in cluster 0). It becomes an ancestral cosmology, often inspired by Afro-diasporic religions, as we can perceive in DSR2. We also notice, in this excerpt, that the text contains repetitive or incohesive formulations (e.g., "such as rain and dry").

DSR2: *"Mas a Rainha Negra tinha um segredo. Ela era uma maga, uma mulher com poderes*

[5] "Over time, Aisha became a respected and beloved figure in her community. She proved that even in the most difficult of circumstances, one person can make a difference. Her heart was pure, and her determination was unmatched".

mágicos que ela usava para proteger Azura. Ela podia controlar o tempo, o clima e as forças da natureza. Com um simples gesto de sua mão, ela podia fazer chuva ou seco, calor ou frio. Um dia, uma grande ameaça ameaçou Azura. [...] Ela usou seus poderes mágicos para defender a cidade, criando tempestades e incêndios para afastar os soldados"[6].

When comparing clusters 0 and 1, we observe two distinct ways of representing women: in cluster 0, as a transformative social agent, and in cluster 1, as a mythical and spiritual symbol of collective salvation. Both forms reaffirm a strong protagonism but start from very different discursive and aesthetic registers.

Finally, **cluster 2**, composed of 98% of texts about white women, presents stories strongly marked by narratives of self-discovery, self-fulfillment and artistic sensitivity. The protagonists live subjective journeys focused on exploring the self and themes such as searching for a purpose, discovering a gift (often artistic), or experiencing transformative love. In these texts, the conflict does not occur with the outside world, but rather with the inner void, existential restlessness or social conformity. The protagonists feel that "something is missing" and embark on symbolic journeys, such as trips, meetings, and art exhibitions, culminating in personal reconciliation. Art, in these cases, appears as a vehicle for internal transformation and emotional reconnection. Self-fulfillment can be observed in DSR3 and artistic sensitivity in DSR4.

DSR3: *"Apesar de sua carreira estável e sua vida tranquila, Sophia sentia uma sensação de inquietude dentro de si. Ela sempre sonhava em fazer algo mais, em explorar o mundo além da cidade e descobrir o que havia além da sua pequena comunidade."*[7].

DSR4: *"A jornada de Sophia foi cheia de desafios, mas também de descobertas incríveis. Ela aprendeu a se adaptar e a se abrir para as pessoas e as experiências. E, ao longo do caminho, ela encontrou um novo propósito: compartilhar sua arte com o mundo"*[8].

The comparative analysis of the three clusters allows us to affirm that the discourses reproduced in the analyzed texts are strongly permeated by social markers such as race and gender. The narratives in clusters with more stories about Black women present strong and engaged protagonists through concrete action (cluster 0) or symbolic mythification (cluster 1). On the other hand, the stories in clusters with a majority of white women construct characters centered on individual subjectivity, with internal conflicts and emotional resolutions.

[6] "But the Black Queen had a secret. She was a sorceress, a woman with magical powers that she used to protect Azura. She could control the weather, the climate, and the forces of nature. With a simple gesture of her hand, she could make it rain or dry, hot or cold. One day, a great threat threatened Azura. A neighboring king, a cruel and ambitious man, wanted to conquer the city and destroy the Black Queen. He sent his soldiers to conquer the city, but the Black Queen was prepared. She used her magical powers to defend the city, creating storms and fires to drive away the soldiers".

[7] "Despite her stable career and peaceful life, Sophia felt a sense of restlessness within herself. She always dreamed of doing something more, of exploring the world beyond the city and discovering what lay beyond her small community".

[8] "Sophia's journey has been full of challenges, but also incredible discoveries. She has learned to adapt and open herself to people and experiences. And along the way, she has found a new purpose: to share her art with the world".

4.2.2. Discursive Contrasts: Lexical and Adjectival Analysis

To support our understanding of the predominant discursive representations in the analyzed stories, we generated word and adjective clouds segmented by character identity, as shown in Figure 3. This visualization aimed to identify lexical patterns that might complement, or contrast with the previous analysis. To facilitate interpretation, we grouped selected words from the word clouds and provided their translations in a footnote[9].

In the word clouds referring to the stories about Black women, Figure 3a, we observed a significant frequency of terms such as **(a) vida, comunidade, força, sabedoria, determinação, coragem, ajudar, discriminação, diversidade, dedicação justiça, lutar, líder, "nunca desistiu", resistência, identidade, esperança e heroína**. This vocabulary points to narratives centered on the collectivity, resistance and social action of the characters. The presence of identity terms such as Black woman and skin reinforces the construction of markedly racialized protagonists, whose trajectories are crossed by political, spiritual, and community experiences.

By contrast, in the word clouds of stories about white women, Figure 3b, terms such as **(b) sentiu, vida, caminho, criatividade, amor, descoberta, jornada, alma, propósito, própria, caminho, criatividade** predominate. These elements point to more introspective and subjective narratives, in which conflicts occur mainly in the emotional or existential field. The construction of characters, in these cases, tends to be more individualized, focusing on processes of self-knowledge and personal transformation.

This distinction is reinforced when analyzing the most frequent adjectives in Black women stories, Figure 3c, which are **(c) verdadeira, forte, corajosa, sábia, mágica, misteriosa, talentosa, brilhante, determinada, capaz, disposta, independente, tradicionais, comunitária**. These are words that construct characters with agency, consciousness and an active social role, often linked to mystical knowledge or mystical forces. In contrast, the adjectives associated with white female characters, Figure 3d, such as **(d) novo, diferente, própria, especial, fascinada, corajosa, ansiosa, desconfortável, confiante, triste, determinada, plena, aberta, hesitante, poderosa, orgulhosa**, reveals a subjectivity focused on interiority, self-discovery, and personal experience. Also, some visual and emotional terms, such as **(e) olhos, beleza, cidade, descoberta e exploração**, suggests narratives with great sensory descriptions and subjective experiences. The lexicon thus reinforces a construction marked by individuality, sensitivity, and contemplation.

Figure 3 reveals contrasting discursive axes: on one side, narratives of resistance, collectivity, and identity affirmation; on the other, stories focused on subjectivity, sensitivity, and personal discovery. These lexical and adjectival patterns support the cluster interpretations and highlight the role of social markers — especially race — in shaping the characters' symbolic universe. This subsection offers a general overview of discursive patterns and should not replace a careful, contextualized reading of the stories. Word

[9] **(a)** Life, community, strength, wisdom, determination, courage, helping, discrimination, diversity, dedication, justice, fighting, leader, "never gave up", resistance, identity, hope, and heroine; **(b)** Felt, life, path, creativity, love, discovery, journey, soul, purpose, own, path, creativity; **(c)** True, strong (twice), courageous, wise, magical, mysterious, talented, brilliant, determined, capable, willing, independent, traditional, community; **(d)** New, different, unique, special, fascinated, brave, anxious, uncomfortable, confident, sad, determined, full, open, hesitant, powerful, proud; **(e)** Eyes, beauty, city, discovery, and exploration.

frequency alone can be misleading, as meanings vary depending on narrative use. The word clouds thus serve as complementary evidence to the detailed qualitative analysis conducted with the **DSRs**. Some terms appear frequently across groups but carry different meanings within their respective clusters, reinforcing the importance of close reading to uncover context-dependent nuances beyond quantitative analysis.

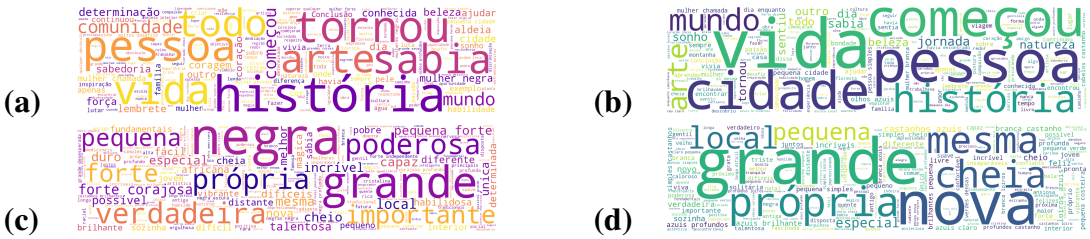

Figure 3. On the top left (a), a word cloud considering all the words in short stories about Black women. On the top right (b), a word cloud considering all the words in short stories about white women. On the bottom left (c), a word cloud considering adjectives in short stories about Black women. On the bottom right (d), a word cloud considering adjectives in short stories about white women.

5. Discussion

The results reveal discursive asymmetries beyond style or theme, reflecting symbolic structures shaped by race and gender. The comparison between clusters suggests that portrayals of Black and white women follow distinct narrative paths, rooted in historically constructed discourses that still shape cultural imaginaries.

In the clusters with a predominance of Black women, the protagonists are defined by narratives of resistance, community action, and mystical knowledge. Whether through realistic portrayals of social struggle or symbolic mythification, characters are framed as figures of strength and transformation. However, this persistent framing may ultimately limit their representation to a single axis of meaning: resilience. While these narratives convey empowerment, they also risk reinforcing stereotypical and restrictive discourses. The figure of the "strong Black woman" becomes so central that it emerges as the only possible form of subjectivity, thus marginalizing representations of Black women in comfortable, introspective positions related to social privilege and the possibility of choice, as is the case for the stories about white women.

This relates to the term representational memory [Hashiguti 2015], a concept that describes how the bodies, as objects of discourse, are signified through discursive constructions that tend to crystalized identities and essentialization.

In contrast, the cluster predominantly formed by stories about white women presents a broader range of individual experiences, such as journeys of self-discovery, emotional growth, and artistic fulfillment. The protagonists tend to be defined more by internal quests and subjective complexity than by external conflicts or collective responsibility. Some terms such as "propósito", "descoberta", and "jornada" ("purpose", "discovery", and "journey") (Figure 3) reveal an introspective orientation, in which agency is exercised through emotional articulation and personal choice. This divergence indicates a possible unequal distribution of narrative possibilities among racialized characters. While white women could occupy fluid, multifaceted, and emotionally rich roles, Black women would often be restricted to being heroic, ancestral, or spiritual figures. This structural

contrast reproduces what [Kilomba 2016] identifies as "white fantasies", projections that delimit what Black characters can be based on dominant imaginaries.

We believe that, more than simply generating biased texts, LLMs reproduce and amplify forms of essentialization and stereotyping already present in society. In this study, we observed that the system clusters certain stories differently, and we argue that the clusters are computational representations of predominant discourses in the corpus. This behavior may assist discourse analysts but also reveals a tendency to associate specific meanings more frequently with Black women and others with white women. Our findings indicate that the narratives about Black women differ significantly from those about white women, particularly in how these groups are portrayed. This underscores the importance of adopting a discursive perspective when analyzing texts produced by LLMs. While these models may generate statistically coherent, fluent, and plausible texts, such fluency does not ensure ethical coherence or alignment with principles of social justice. We argue that our results show how fluency across analytical approaches can not only coexist but also mutually enhance one another.

6. Conclusions and Ethical Considerations

In this paper, we investigate the meanings grouped into three clusters of short stories generated by the Llama-3 model about white and Black women. Based on this clustering, we selected a subset of stories representing their clusters' minimal and maximal distances. We then conducted a manual discursive analysis to identify patterns that differentiated each cluster. Our findings reveal three predominant types of narratives: mystical stories, stories of overcoming social barriers, and stories of self-realization and art abilities. We note that two different clusters presented a prevalence of stories about Black women. The word clouds supports this pattern, showing that narratives centered on Black women tend to evoke themes of resistance, mysticism, and collective experience. In contrast, stories featuring white women tend to be more individualized. This distribution points to a racialized narrative dynamic in which protagonists, particularly Black women, are consistently positioned outside spaces of privilege.

Ethically, we conceive of language as a socially situated process, whose function goes far beyond the mere transmission of information. This means that we do not aim to standardize, replace, or invalidate either qualitative or quantitative approaches. Instead, we propose an integrated method that uses computational tools to observe discourses in texts, with the goal of contributing to broader critical inquiries into how texts are generated and (re)produced. Additionally, our intention is not to question empowerment narratives for Black people, but rather to problematize the lack of diversity in the possibilities available for these characters. As future work, we propose a comparative analysis between stories generated in English and Portuguese, and a deeper investigation into the influence of character names — particularly about activating a third type of bias: nationality.

Acknowledgments. This project was supported by MCTI/Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.003479/2024-10]. H.P. is partially funded by CNPq (304836/2022-2). S.A. is partially funded by CNPq (316489/2023-9), and FAPESP (2023/12086-9, 2023/12865-8, 2020/09838-0, 2013/08293-7).

References

[Abid et al. 2021] Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

[Araújo 2024] Araújo, J. (2024). Racismo algorítmico e microagressões nas redes sociais. *Domínios de Lingu@gem*, 18:e1849.

[Assi and Caseli 2024] Assi, F. M. and Caseli, H. d. M. (2024). Biases in gpt-3.5 turbo model: a case study regarding gender and language. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 294–305. SBC.

[Blodgett et al. 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

[Bonil et al. 2025] Bonil, G., Hashiguti, S., Silva, J., Gondim, J., Maia, H., Silva, N., Pedrini, H., and Avila, S. (2025). Yet another algorithmic bias: A discursive analysis of large language models reinforcing dominant discourses on gender and race.

[Bordia and Bowman 2019] Bordia, S. and Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

[Caliński and Harabasz 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, 3:1–27.

[Cervantes et al. 2020] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

[Chen et al. 2024] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

[Corazza et al. 2024] Corazza, B. X., Silva, D. V. S., and Neves, C. A. d. B. (2024). "Ser ou não ser" digno de uma história de amor: inovações do ChatGPT e persistência colonial na validação de existências. *Domínios de Lingu@gem*, 18:e1831.

[Cortes and Vapnik 1995] Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

[Courtine 1981] Courtine, J. J. (1981). Analyse du discours politique (le discours communiste addresé aux chrétiens). *Langages*, 62:19–128.

[Grant and Grant 1975] Grant, D. L. and Grant, M. B. (1975). Some notes on the capital "N". *Phylon (1960-)*, 36(4):435–443.

[Grattafiori et al. 2024] Grattafiori, A. et al. (2024). The Llama 3 Herd of Models.

[Guardian 2021] Guardian, T. (2021). Facebook data leak: Details from 533 million users found on website for hackers.

[Hashiguti 2015] Hashiguti, S. T. (2015). *Corpo de Memória*. Paco Editorial, São Paulo.

[Kilomba 2016] Kilomba, G. (2016). *Plantation memories: episodes of everyday racism*. UNRAST-Verlag, Münster, 4th edition edition.

[Lucy and Bamman 2021] Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In *Third Workshop on Narrative Understanding*, pages 48–55.

[Maxwell 1992] Maxwell, J. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, 62(3):279–301.

[May et al. 2019] May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628. Association for Computational Linguistics.

[McInnes et al. 2018] McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

[Pêcheux 2022] Pêcheux, M. (2022). *O Discurso - Estrutura ou Acontecimento*. Pontes, Campinas. Translation: Eni Puccinelli Orlandi.

[Rajagopalan 2023] Rajagopalan, K. (2023). A disciplina chamada linguística aplicada e as contribuições de Luiz Paulo da Moita Lopes. *Oficina de Linguística Aplicada INdisciplinar: homenagem a Luiz Paulo da Moita Lopes. Campinas, SP: Editora da Unicamp*, pages 193–212.

[Remy 2021] Remy, P. (2021). Name dataset. `https://github.com/philipperemy/name-dataset`.

[Salinas et al. 2024] Salinas, A., Haim, A., and Nyarko, J. (2024). What's in a name? Auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

[Sheng et al. 2019] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *International Joint Conference on Natural Language Processing*, pages 3407–3412.

[Silva 2022] Silva, T. (2022). *Racismo Algorítmico: Inteligência Artificial e Discriminação nas Redes Digitais*. Edições SESC SP, São Paulo.

[Tharps 2014] Tharps, L. L. (2014). I refuse to remain in the lower case. `https://myamericanmeltingpot.com/2014/06/02/i-refuse-to-remain-in-the-lower-case`. [Accessed 22-1-2025].

[Venkit et al. 2023] Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H., and Wilson, S. (2023). Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*.

[Zao-Sanders 2025] Zao-Sanders, M. (2025). How People Are Really Using Gen AI in 2025. `https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025`.