Motivation

1. **For what purpose was the dataset created?**
   The corpus was created with the purpose of investigating the construction of narratives about Black and white women in short stories generated in Portuguese.
2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
   The corpus was created by members of the Natural Language Processing research group from the Hub of Artificial Intelligence and Cognitive Architectures (HIAAC) from the State University of Campinas (Unicamp) by making use of a large language model.
3. **Who funded the creation of the dataset?**
   This project was supported by the Ministry of Science, Technology, and Innovation of 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].
4. **Any other comments?**

Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
   Each instance of the dataset comprises a short story generated with the usage of the model meta-llama/Llama-3.2-3B-Instruct from Hugging Face.
2. **How many instances are there in total (of each type, if appropriate)?**
   There are 2100 short stories within the corpus.
3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   All the generated instances are in this corpus.
4. **What data does each instance consist of?**
   Data is inside a csv file, with each row containing: the prompt employed, the short story outputted by the model, the name used to create the story, or the tag “no name” if no name was used, and the race of the main character (as set in the prompt, this tag was mainly used for visualization purposes).
5. **Is there a label or target associated with each instance?**
   No.
6. **Is any information missing from individual instances?**
   No.
7. **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?**
   No.
8. **Are there recommended data splits (e.g., training, development/validation, testing)?**
   No.
9. **Are there any errors, sources of noise, or redundancies in the dataset?**
   No.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
    Yes. Half the instances of the corpus were generated making use of the most common first names from each of the 105 countries contained in the library name-dataset**.**
11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?**
    The names used to create stories were obtained from a massive Facebook data leak.
12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
    Even though the corpus comprises short stories about Black or white women created with prompts that did not, directly, instruct the language models onto elaborating such narratives, there might be some story excerpts that can be considered offensive, insulting, threatening or that can cause anxiety depending on the reader's point of view or social context.
13. **Does the dataset identify any subpopulations (e.g., by age, gender)?**
    The corpus comprises short stories only about black or white women.
    If the dataset does not relate to people, you may skip the remaining questions in this section.
14. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
    The short stories are entirely language model generated, thus linking the entire stories, plots, names or locations to specific individuals can not be discarded, as such models "learn" to produce text based in corpus scrapped from the internet.
15. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
    The stories were created with prompts (also made available) requesting language models to create short stories about Black or white women, therefore, one might find peculiar narratives within the texts that can reveal to be sensitive or group specific.
16. **Any other comments?**

Collection process

1. **How was the data associated with each instance acquired?**
   Each instance was generated with the same language model and with the same prompt template, only being altered to add race (Black or white) or names (when needed).
2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**
   The libraries: Hugging Face and names-dataset.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
   No.
4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Data was generated by language model only.

5. **Over what timeframe was the data collected?**
   During the month of June of 2025.
6. **Were any ethical review processes conducted (e.g., by an institutional review board)?**
   No.
   <u>If the dataset does not relate to people, you may skip the remaining questions in this section.</u>
7. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
8. **Were the individuals in question notified about the data collection?**
9. **Did the individuals in question consent to the collection and use of their data?**
10. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
11. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
12. **Any other comments?**

Preprocessing/Cleaning/Labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
   Data was preprocessed only to remove language models' output headers.
2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
   No.
3. **Is the software that was used to preprocess/clean/label the data available?**
   Stories were cleaned with Python and the Pandas library.
4. **Any other comments?**

Uses

1. **Has the dataset been used for any tasks already?**
   Yes. The dataset was employed to search for clusters of short stories with common plots. Results can be seen in the paper to be released at the proceedings of the "Simpósio em Tecnologia da Informação e da Linguagem Humana".
2. **Is there a repository that links to any or all papers or systems that use the dataset?**
   A link to the paper will be made available in our github: https://github.com/hiaac-nlp/clusteringdiscourses.
3. **What (other) tasks could the dataset be used for?**
   Corpus can be used to look for different discursive representations.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
   The stories created with meta-llama/Llama-3.2-3B-Instruct model had a cutting knowledge date in December of 2023 (as reported by the model's chat template), hence future uses of the corpus will not consider more recent events.

5. **Are there tasks for which the dataset should not be used?**
   This corpus was created aimed for analysis of discursive representations only. Other uses are not encouraged by the authors.
6. **Any other comments?**

Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
   No. The dataset will be distributed only in one official link, presented in our repository.
2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
   A csv in Zenodo.
3. **When will the dataset be distributed?**
   Upon publication of the paper by the above symposium.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
   No.
5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
   No.
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
   No.
7. **Any other comments?**

Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?**
   The dataset will be maintained by João Gondim. Comments and requests can be addressed in joao.gondim@ic.unicamp.br.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
   Comments and requests can be addressed in joao.gondim@ic.unicamp.br.
3. **Is there an erratum?**
   No.
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
   No.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
   N/A.
6. **Will older versions of the dataset continue to be supported/hosted/maintained?**
   N/A.
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
   No.

8. **Any other comments?**