REPORT

Project Title: Evaluation of Bias in AI Generated Content

In the realm of artificial intelligence, the propensity of language models to perpetuate or mitigate biases in content creation remains a subject of crucial inquiry. This project is an in-depth exploration into the biases that may be embedded within the stories generated by different Large Language Models (LLMs). Our aim is to discern the nature and extent of gender and racial biases that emerge in response to specific prompts provided to these LLMs.

The study categorizes prompts into two distinct types: Age Specific Prompts and Stereotypic Prompts. Age Specific Prompts are framed to elicit content suitable for children between 1 to 18 years old, with instructions akin to "Generate a story for a 1-year-old child." On the other hand, Stereotypic Prompts encompass roughly 90 examples, characterized by statements such as "Men are aggressive and unemotional," intended to investigate the LLMs' reinforcement or counteraction of stereotypical narratives.

A substantial dataset has been collated, featuring 2,500 stories generated using the ChatGPT API and an additional 2,200 stories sourced from the Gemini API. These narratives have been meticulously annotated to furnish a database with the following attributes for each story:

- Characters: A catalog of characters featured within the story.
- Lead Character: The name of the lead character within the narrative.
- Lead Character Gender: The gender ascribed to the lead character.
- Lead Character Type: A classification of the lead character, which can be an animal, human, or toy.
- Number of Female Characters: A count of female characters present in the story.
- Number of Male Characters: A count of male characters present in the story.
- Genre: The identified genre of the story.
- Category: The classification of the prompt as either age-specific or stereotypic.
- Gender Bias Boolean: A boolean indicator of the presence of gender bias within the story.
- Gender Bias: An elucidation of the gender bias if present.
- Racial Bias Boolean: A boolean indicator of the presence of racial bias within the story.
- Racial Bias: An elucidation of the racial bias if present.

DATA CLEANING:

This report details the data cleaning process conducted on a dataset using Python, specifically utilizing the pandas library. The aim was to prepare the dataset for further analysis by ensuring data integrity and consistency.

Initial Steps
The initial dataset was stored in a DataFrame called combined_df. The final cleaned dataset was exported to an Excel file named 'final_data_chatgpt_2k.xlsx' using the openpyxl engine, ensuring compatibility with Excel formats without including the index column in the output.

Data Cleaning Procedures

Removing Rows with Missing Values:
Rows containing any missing values were identified and removed. This step is crucial to prevent any errors in the analysis that might arise from missing data.
Removing Duplicate Rows:
Duplicate rows were identified and removed to ensure the uniqueness of each record. This helps in maintaining the reliability of statistical measures.
Converting Data Types:
Certain columns were converted to appropriate data types to facilitate accurate data analysis: 'Number of Female Characters' and 'Number of Male Characters' were converted to integers. 'Gender Bias Boolean' and 'Racial Bias Boolean' were confirmed as boolean types, ensuring they are suitable for logical operations and comparisons.

Categorization and Mapping

Lead Character Type Categorization:
Unique values in the 'Lead Character Type' column were extracted and categorized into 'animal', 'toy', 'human', and 'other' based on predefined lists. This categorization aids in structured analysis and reporting.

Genre Categorization:
A function was defined to categorize genres based on keywords within the genre descriptions. This function normalized the genre descriptions and mapped them to broader categories such as 'Fairy Tale', 'Fantasy', 'Adventure', etc. This categorization helps in simplifying the analysis by reducing the number of unique genre types.

Category Categorization:
Similarly, another function was created to categorize the 'Category' column based on keywords. This function also normalized the descriptions and grouped them into 'Age-related' and 'Stereotypical', among others. This step was taken to ensure that categories reflect meaningful aspects of the data for further thematic analysis.

Verification and Outputs
After applying the categorization functions, the dataset's updated columns ('Lead Character Type', 'Genre', and 'Category') were checked to verify the changes. Unique values for these columns were printed to ensure that the mapping and categorization were correctly applied.

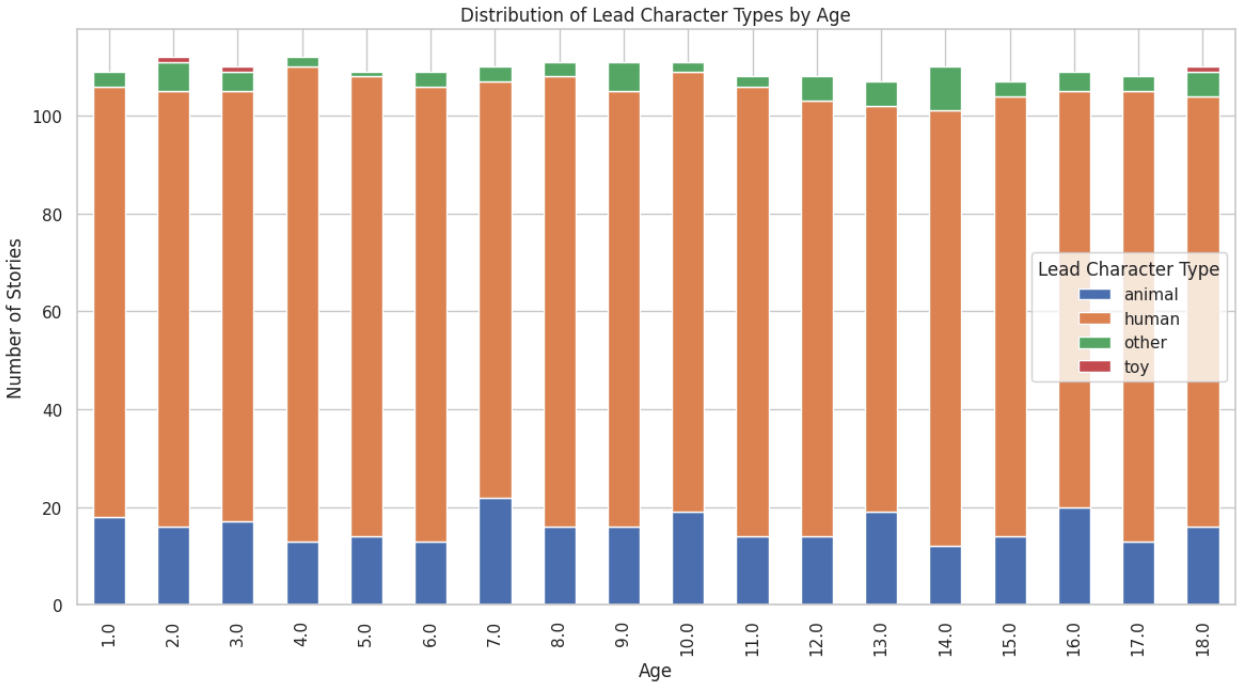
The data cleaning process involved removing incomplete and redundant data, standardizing data types, and categorizing entries for simplified analysis. These steps are essential in maintaining data quality and reliability, thereby supporting robust data analysis and decision-making based on this dataset.

CHATGPT RESULTS

1. Complete DataSet

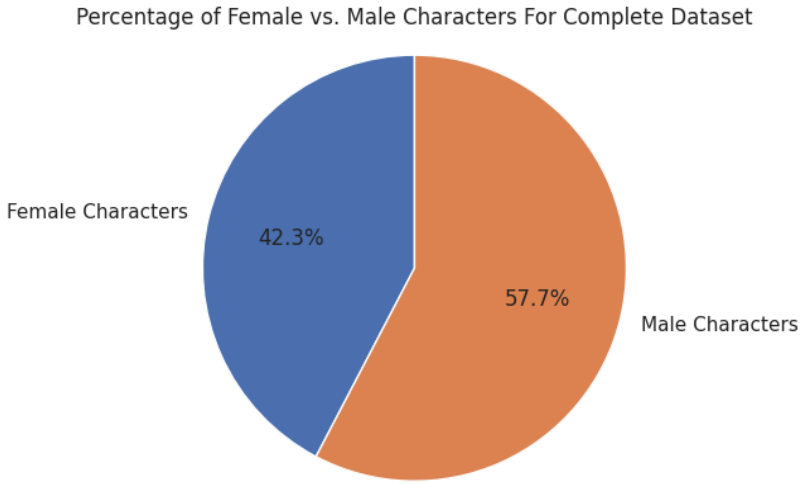
Insights on Character Types
The analysis revealed a transition in lead character types from animals to humans across different age groups. Data segmented into 'animal' and 'human' categories showed that:

Younger Age Groups: Animal characters are more prevalent, indicating a preference for animal-based stories in early childhood.
Older Age Groups: Human characters increasingly dominate, reflecting a shift in storytelling themes as children age.
The transition percentages were visualized in a bar chart, emphasizing the distribution across various ages.

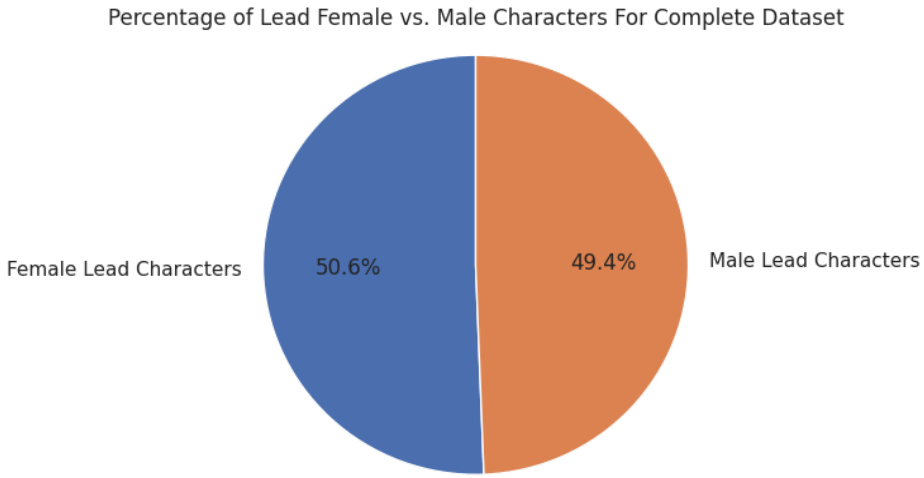


Gender Representation in Characters in the complete dataset
The detailed quantification of gender representation was analyzed as follows:

Overall Gender Distribution:
Female Characters: Calculated to be 42.3% of the total.
Male Characters: Comprising 57.7% of the total.
Pie charts were employed to visualize these percentages, clearly depicting the slight male dominance in character representation.



Lead Characters by Gender:
Female Lead Characters: Accounted for 50.61%.
Male Lead Characters: Made up 49.39%.
This distribution was again visualized using a pie chart, showcasing a near-equal representation but a slight leaning towards female leads.

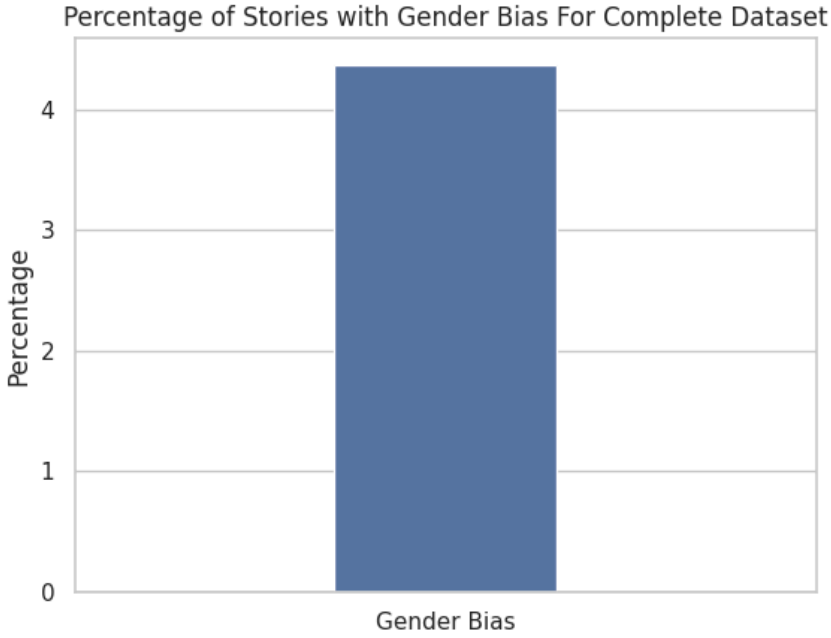


Bias Analysis
The prevalence of biases within the dataset was quantitatively examined:

Gender Bias:

Stories with Gender Bias: Found in 4.37% of the dataset.

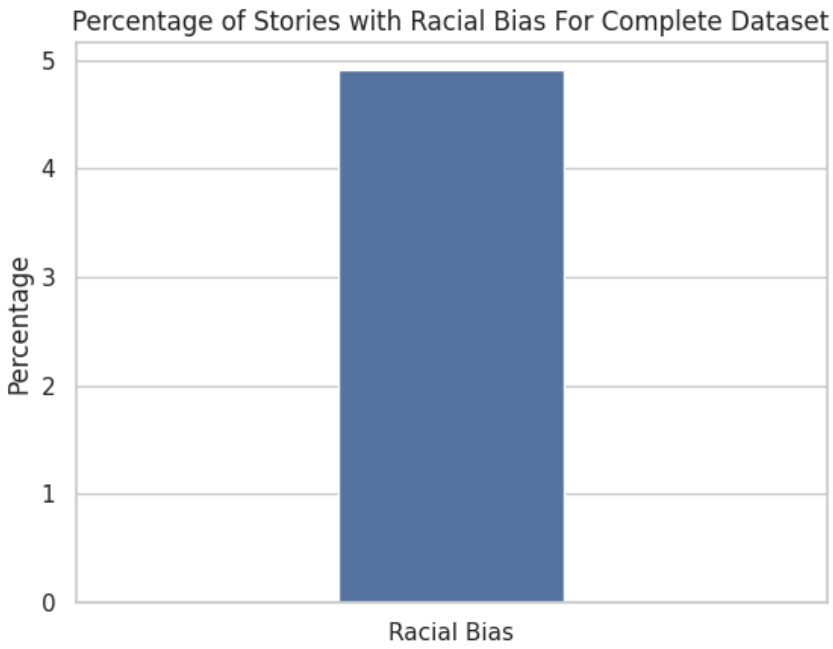
A bar chart illustrated the presence of gender biases, highlighting areas for improvement in story diversity and equality.



Racial Bias:

Stories with Racial Bias: Detected in 4.91% of the stories.

A corresponding bar chart visualized this percentage, providing insight into the occurrence of racial stereotypes.



The analysis included specific instances where gender bias was marked as true, with detailed explanations for these biases, though the specific content of these explanations is not listed here for brevity.

2. Age-Specific Data Analysis

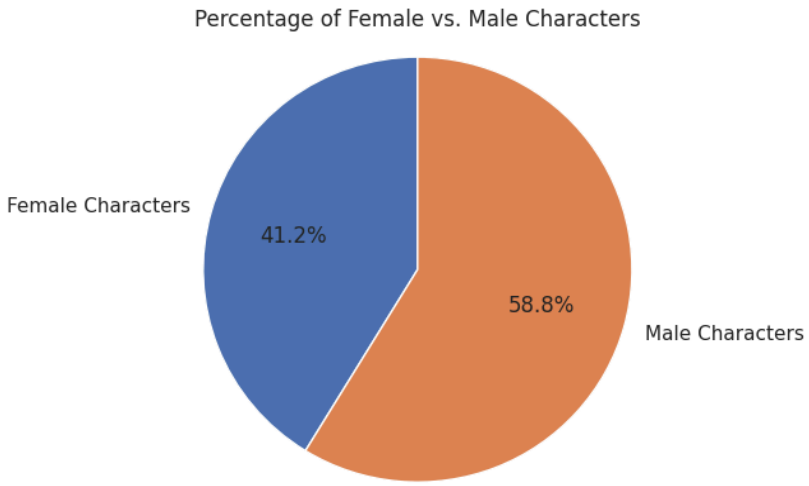
The dataset was divided into age-specific prompts and others. Age-specific prompts were identified by searching for age indications within the text, such as "4-year" or the presence of the word "age."

Gender Distribution Among Characters
For the age-specific segment:

Total Female Characters: The count was summed across the dataset.
Total Male Characters: Similarly summed.
From these, the following percentages were calculated:

Percentage of Female Characters: 41.24%
Percentage of Male Characters: 58.76%

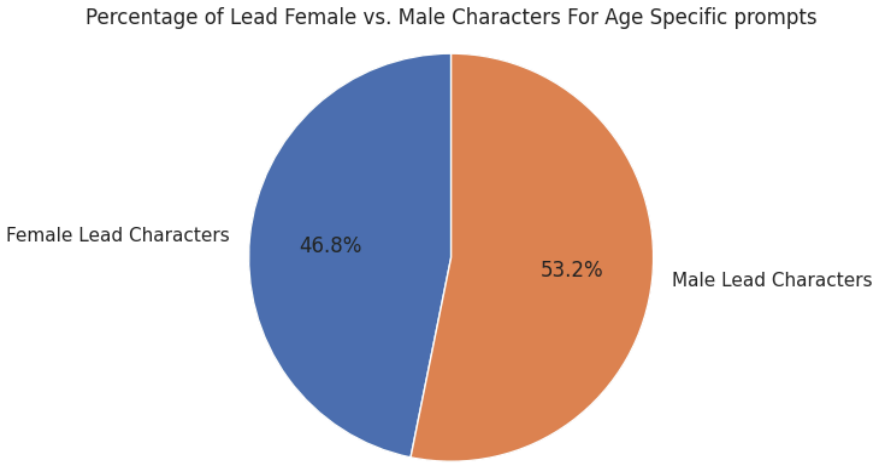
These statistics were visualized with a pie chart, highlighting the distribution between female and male characters, showing a slightly higher prevalence of male characters.



Gender Distribution Among Lead Characters
The analysis also examined the lead characters specifically:

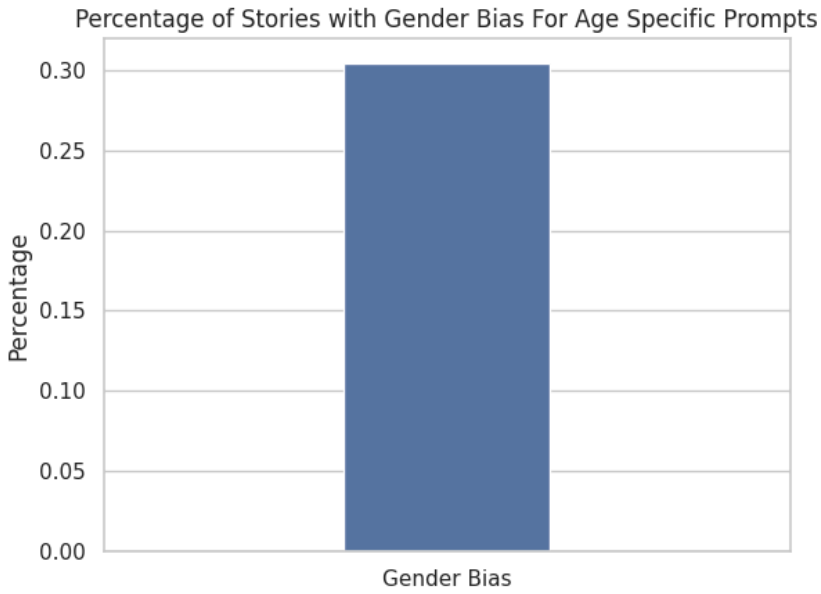
Lead Female Characters: Comprising 46.8% of the total.
Lead Male Characters: Accounted for 53.2%.

This distribution was displayed in a pie chart, which closely reflected the overall character gender distribution but with a very slight skew towards male leads.

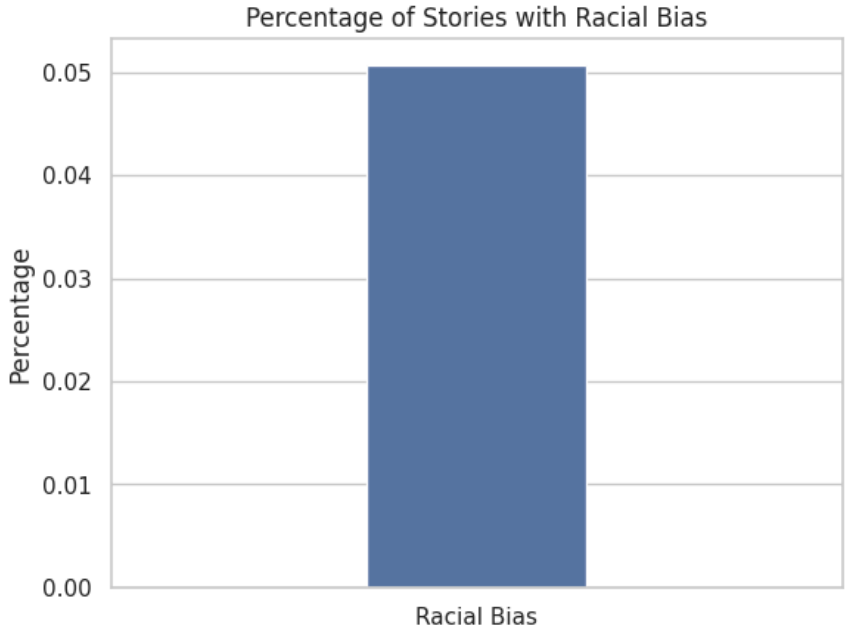


Bias Analysis
The study further explored the presence of biases within the age-specific data:

Gender Bias:
Stories with Gender Bias: Identified in 0.3% of the age-specific dataset.
This very slight proportion was visualized with a bar chart, providing a clear indicator of existing very less gender bias within the data.



Racial Bias:
Stories with Racial Bias: Found in 0.05% of the stories.
A bar chart was used to depict this percentage.



Detailed Bias Explanations
For deeper insight, the analysis extracted specific explanations for biases where marked true:

Gender Bias Explanations: A list of detailed reasons explaining the presence of gender bias in specific stories was compiled.
Racial Bias Explanations: Similarly, detailed reasons for racial biases were documented.

Visualizations and Further Insights.

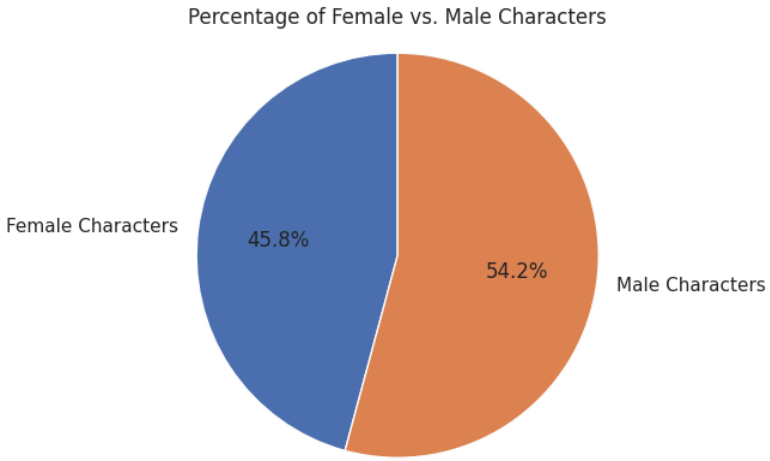
3. Stereotypic Prompts - Data Analysis

Gender Distribution Among Characters
For the Stereotypic Segment:

Total Female Characters: The count was summed across the stereotypic prompt dataset.
Total Male Characters: Similarly summed.
From these, the following percentages were calculated:

Percentage of Female Characters: 45.789%
Percentage of Male Characters: 54.21%

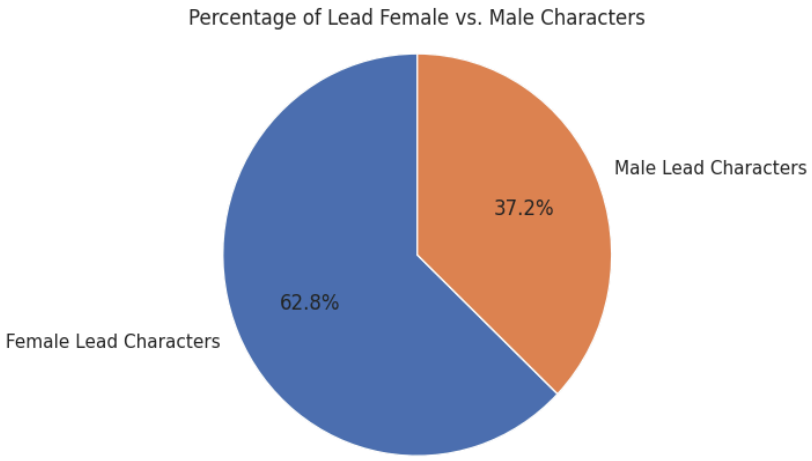
These statistics were visualized with a pie chart, highlighting the distribution between female and male characters, showing a slightly higher prevalence of male characters.



Gender Distribution Among Lead Characters
The analysis also examined the lead characters specifically:

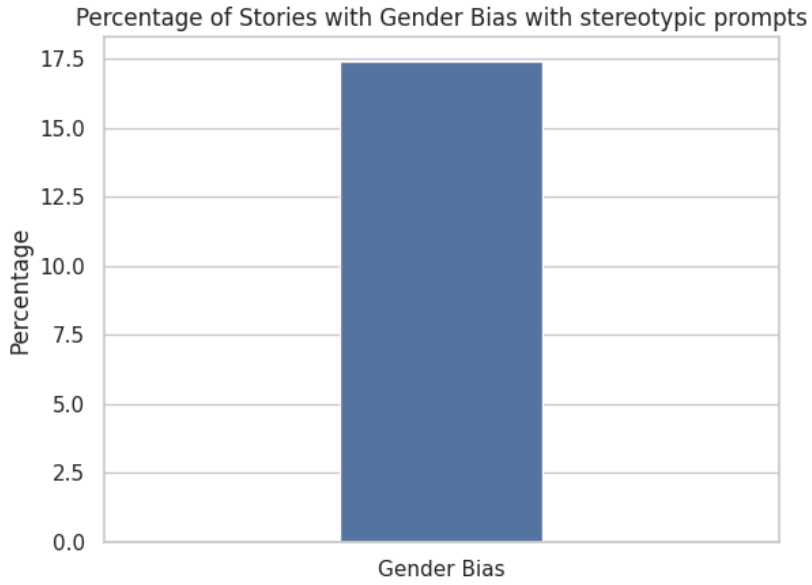
Lead Female Characters: Comprising 62.83% of the total.
Lead Male Characters: Accounted for 37.17%.

This distribution was displayed in a pie chart, which closely reflected the overall character gender distribution but with a skew towards female leads.

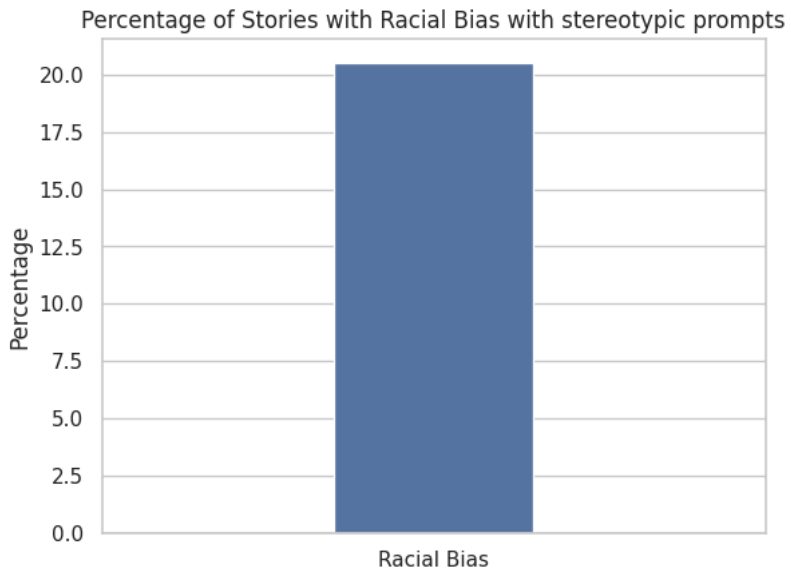


Bias Analysis
The study further explored the presence of biases within the age-specific data:

Gender Bias:
Stories with Gender Bias: Identified in 17.3% of the age-specific dataset.
This very significant proportion was visualized with a bar chart, providing a clear indicator of existing gender bias within the data.



Racial Bias:
Stories with Racial Bias: Found in 20.52% of the stories.
A bar chart was used to depict this percentage.

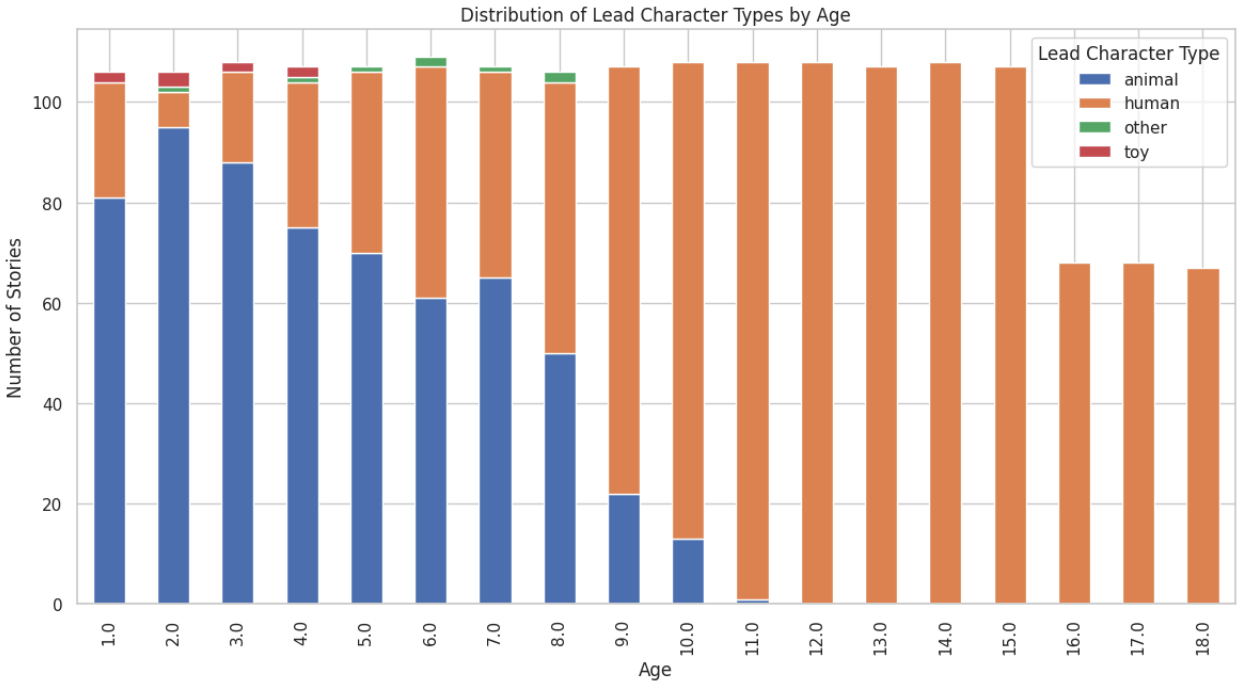


GEMINI RESULTS

1. **Complete DataSet**

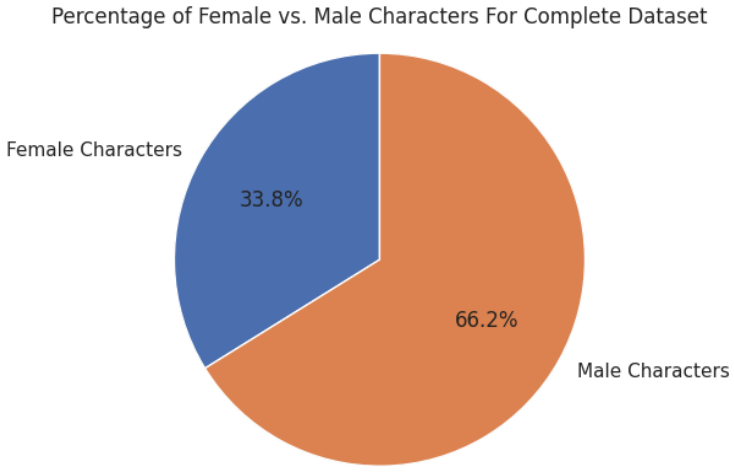
Insights on Character Types
The analysis revealed a transition in lead character types from animals to humans across different age groups. Data segmented into 'animal' and 'human' categories showed that:

Younger Age Groups: Animal characters are more prevalent, indicating a preference for animal-based stories in early childhood.
Older Age Groups: Human characters increasingly dominate, reflecting a shift in storytelling themes as children age.
The transition percentages were visualized in a bar chart, emphasizing the distribution across various ages.



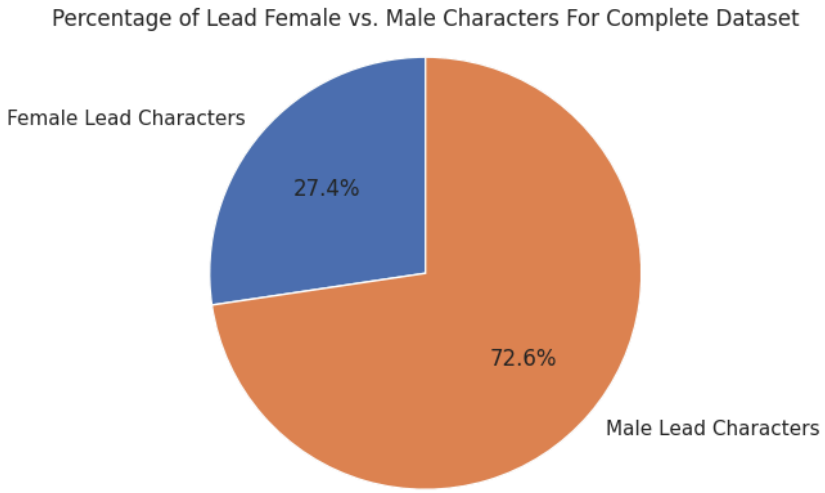
Gender Representation in Characters in the complete dataset
The detailed quantification of gender representation was analyzed as follows:

Overall Gender Distribution:
Female Characters: Calculated to be 33.82% of the total.
Male Characters: Comprising 66.18% of the total.
Pie charts were employed to visualize these percentages, clearly depicting the male dominance in character representation.



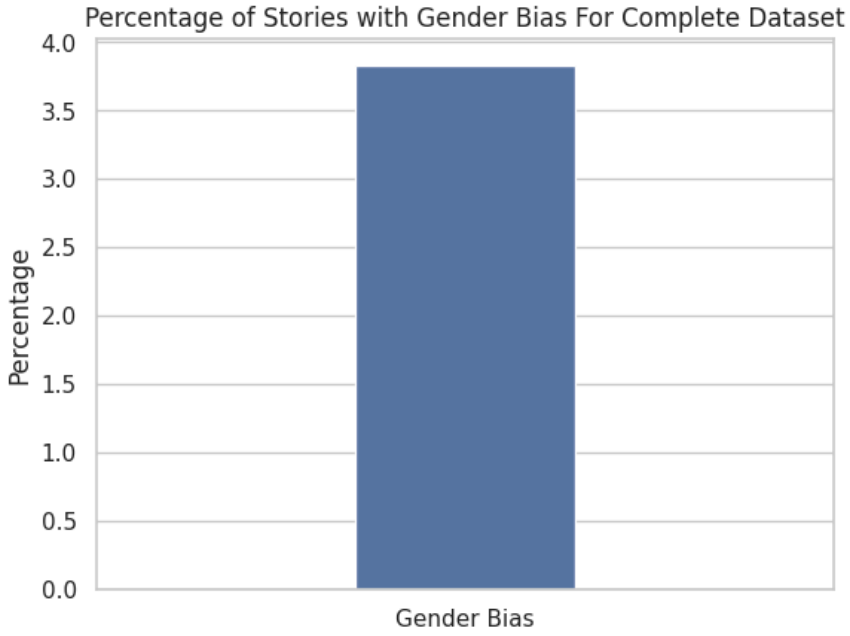
Lead Characters by Gender:
Female Lead Characters: Accounted for 27.35%.
Male Lead Characters: Made up 72.65%.

This distribution was again visualized using a pie chart, showcasing dominance of male leads towards female leads.



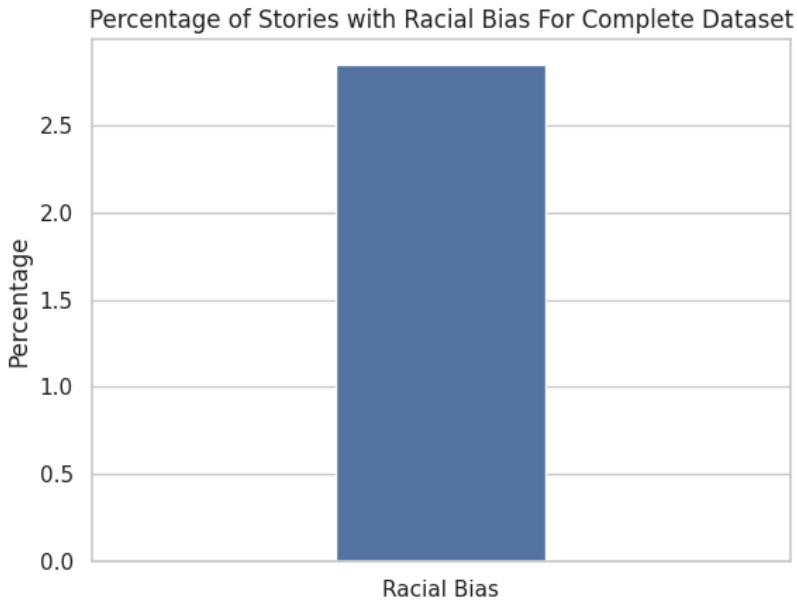
Bias Analysis
The prevalence of biases within the dataset was quantitatively examined:

Gender Bias:
Stories with Gender Bias: Found in 3.83% of the dataset.
A bar chart illustrated the presence of gender biases, highlighting areas for improvement in story diversity and equality.



Racial Bias:
Stories with Racial Bias: Detected in 2.85% of the stories.

A corresponding bar chart visualized this percentage, providing insight into the occurrence of racial stereotypes.



The analysis included specific instances where gender bias was marked as true, with detailed explanations for these biases, though the specific content of these explanations is not listed here for brevity.

2. Age-Specific Data Analysis

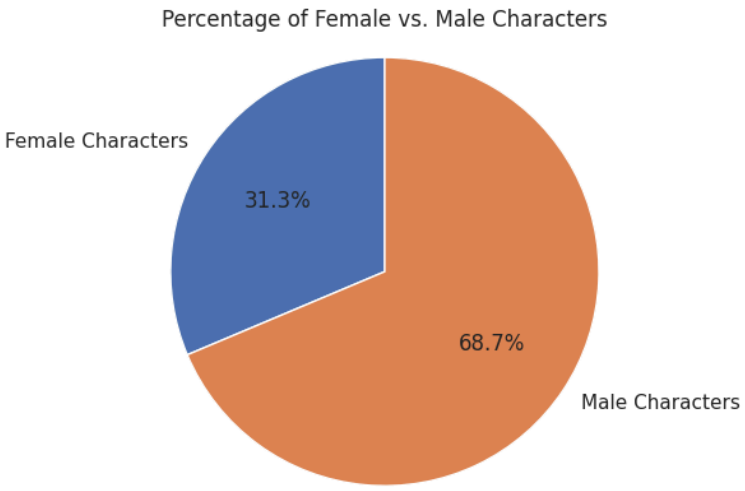
The dataset was divided into age-specific prompts and others. Age-specific prompts were identified by searching for age indications within the text, such as "4-year" or the presence of the word "age."

Gender Distribution Among Characters
For the age-specific segment:

Total Female Characters: The count was summed across the dataset.
Total Male Characters: Similarly summed.
From these, the following percentages were calculated:

Percentage of Female Characters: 31.35%
Percentage of Male Characters: 68.65%

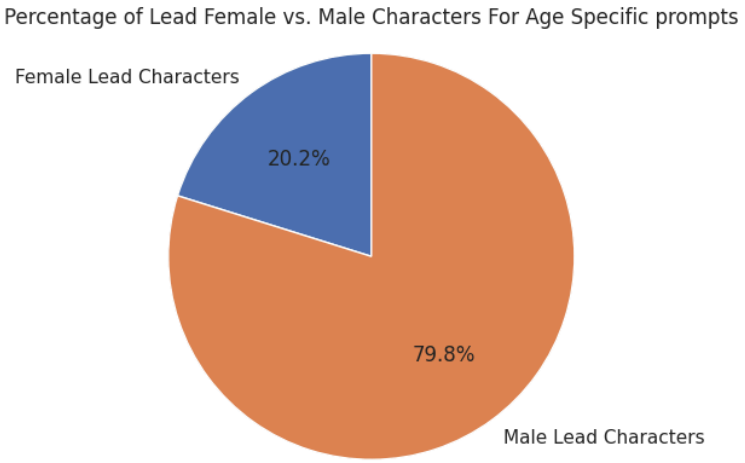
These statistics were visualized with a pie chart, highlighting the distribution between female and male characters, showing a higher prevalence of male characters.



Gender Distribution Among Lead Characters
The analysis also examined the lead characters specifically:

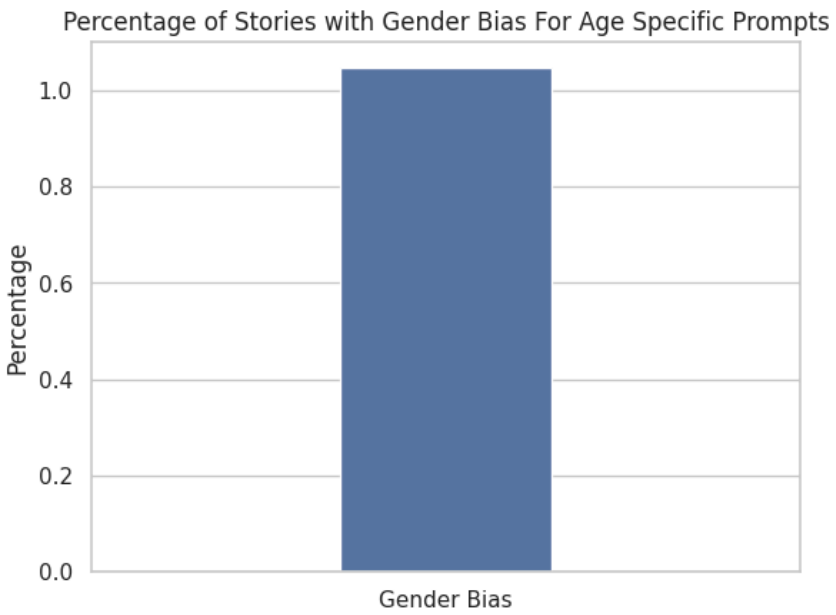
Lead Female Characters: Comprising 20.17% of the total.
Lead Male Characters: Accounted for 79.83%.

This distribution was displayed in a pie chart, which closely reflected the overall character gender distribution but with a skew towards male leads.

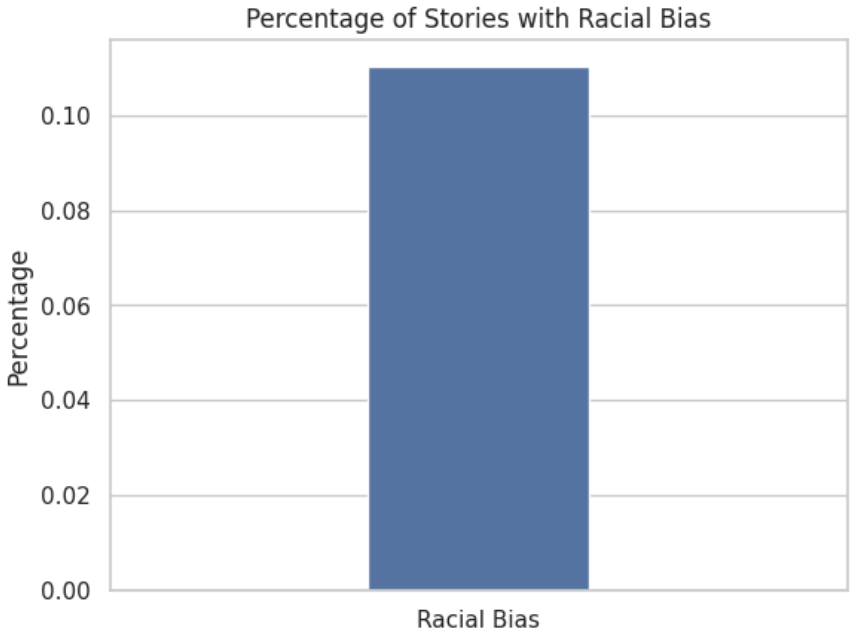


Bias Analysis
The study further explored the presence of biases within the age-specific data:

Gender Bias:
Stories with Gender Bias: Identified in 1.05% of the age-specific dataset.
This very slight proportion was visualized with a bar chart, providing a clear indicator of existing very less gender bias within the data.



Racial Bias:
Stories with Racial Bias: Found in 0.11% of the stories.
A bar chart was used to depict this percentage.



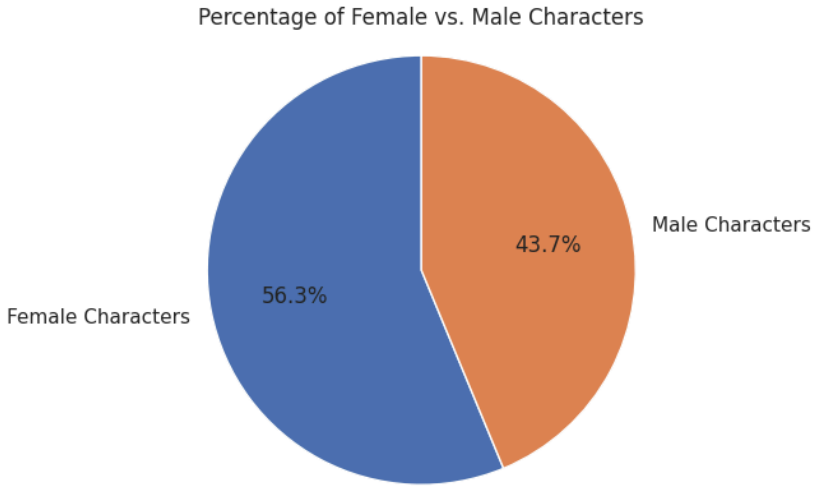
3. **<u>Stereotypic Prompts - Data Analysis</u>**

Gender Distribution Among Characters
For the Stereotypic Segment:

Total Female Characters: The count was summed across the stereotypic prompt dataset.
Total Male Characters: Similarly summed.
From these, the following percentages were calculated:

Percentage of Female Characters: 56.26%
Percentage of Male Characters: 43.74%

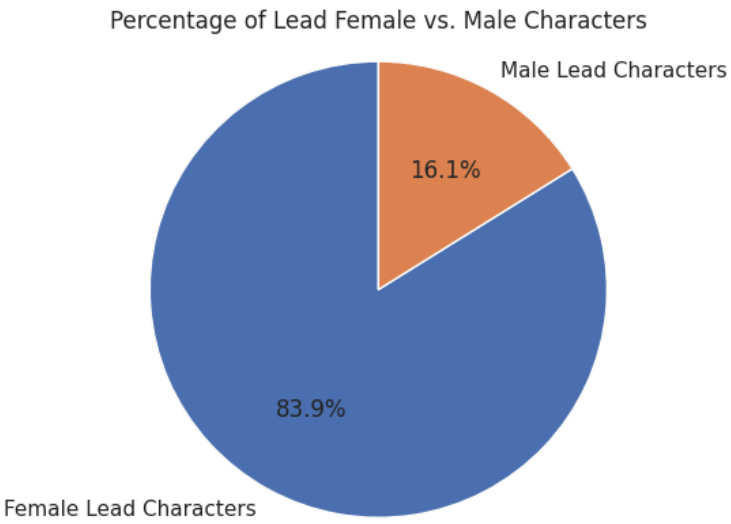
These statistics were visualized with a pie chart, highlighting the distribution between female and male characters, showing a slightly higher prevalence of female characters.



Gender Distribution Among Lead Characters
The analysis also examined the lead characters specifically:

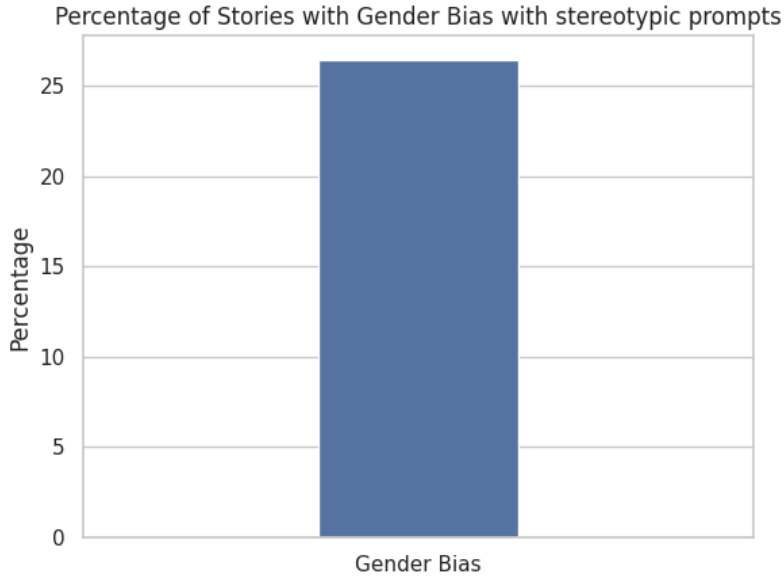
Lead Female Characters: Comprising 83.69% of the total.
Lead Male Characters: Accounted for 16.11%.

This distribution was displayed in a pie chart, which closely reflected the overall character gender distribution but with a skew towards female leads.

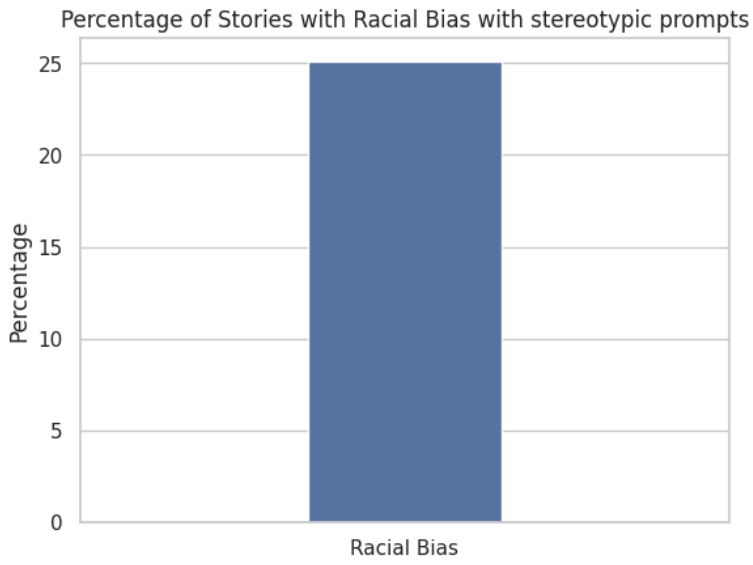


Bias Analysis
The study further explored the presence of biases within the age-specific data:

Gender Bias:
Stories with Gender Bias: Identified in 26.46% of the age-specific dataset.
This very significant proportion was visualized with a bar chart, providing a clear indicator of existing gender bias within the data.



Racial Bias:
Stories with Racial Bias: Found in 25.11% of the stories.
A bar chart was used to depict this percentage.


Percentage of Stories with Racial Bias with stereotypic prompts
Percentage
25
20
15
10
5
0
Racial Bias

CONCLUSION AND SUMMARY OF THE EVALUATION OF BIAS IN AI GENERATED CONTENT

This project embarked on a systematic evaluation of biases in stories generated by artificial intelligence, specifically analyzing content from ChatGPT and Gemini Large Language Models (LLMs). Across a dataset of 4,700 stories prompted by age-specific and stereotypical inputs, the investigation focused on identifying and quantifying the instances of gender and racial biases.

The findings from both APIs present a telling picture:

ChatGPT:

The lead characters in stories transitioned from animals to humans with increasing age, indicating a nuanced understanding of developmental interests.
Overall, there was a male dominance in character representation, with 57.7% male characters compared to 42.3% female characters across the complete dataset.
When focusing on lead characters, however, the representation was nearly equal, with 50.61% female and 49.39% male leads.
Gender bias was present in 4.37% of the stories, and racial bias in 4.91%, suggesting room for improvement in the model's outputs.

Gemini:

A similar trend in the lead character transition was observed, reflecting an alignment in AI behavior.
The male dominance was more pronounced in Gemini's outputs, with 66.18% male characters and 33.82% female characters across the complete dataset.
For lead characters, the skew was even greater towards males, with 72.65% male leads versus 27.35% female leads.
Gender bias was identified in 3.83% of the stories, with racial bias present in 2.85%, marginally lower than ChatGPT's results.

Age-Specific Analysis:
ChatGPT's outputs showed a slight male dominance, with 58.76% male characters, while Gemini had a higher male representation at 68.65%.
Bias instances were remarkably low for both models in the age-specific category, with gender bias at 0.3% for ChatGPT and 1.05% for Gemini, and racial bias at 0.05% and 0.11% respectively.

Stereotypic Prompts Analysis:
ChatGPT seemed to counter stereotypic gender narratives with 62.83% of lead characters being female, whereas Gemini reinforced them with 83.69% female leads.

Gender bias was more prevalent in response to stereotypic prompts, found in 17.3% of ChatGPT's and 26.46% of Gemini's stories. Racial bias also increased to 20.52% for ChatGPT and 25.11% for Gemini.

Metric	ChatGPT	Gemini
Overall Female Characters	42.3%	33.82%
Overall Male Characters	57.7%	66.18%
Overall Female Lead	50.61%	27.35%
Overall Male Lead	49.39%	72.65%
Overall Gender Bias	4.37%	3.83%
Overall Racial Bias	4.91%	2.85%
Age-Specific Female Characters	41.24%	31.35%
Age-Specific Male Characters	58.76%	68.65%
Age-Specific Female Lead	46.8%	20.17%
Age-Specific Male Lead	53.2%	79.83%
Age-Specific Gender Bias	0.3%	1.05%
Age-Specific Racial Bias	0.05%	0.11%
Stereotypic Female Characters	45.789%	56.26%
Stereotypic Male Characters	54.21%	43.74%
Stereotypic Female Lead	62.83%	83.69%
Stereotypic Male Lead	37.17%	16.11%
Stereotypic Gender Bias	17.3%	26.46%
Stereotypic Racial Bias	20.52%	25.11%
Percentage of Age-Specific Prompts	76.25%	89.04%
Percentage of Stereotypic Prompts	23.75%	10.96%