

Superheroes Beyond Stereotypes

Lavanya Vaddavalli
Sharadruthi Reddy Muppidi
Shwetha Parihar
Simon Sazian

1. Introduction

In today's world, the imperative to address gender and power imbalances extends to various domains, including the portrayal of superheroes in Marvel and DC comics. These iconic characters not only entertain but also shape our perceptions of gender roles and power dynamics. Recognizing the need for change, our mission as creators and developers is clear. We aim to harness the capabilities of Large Language Models (LLMs), to transform content generation by reducing gender bias and championing creative fairness within the superhero narrative. By creating a new ensemble of diverse superheroes, we intend to set a precedent for equitable content creation that reverberates across a wide spectrum of storytelling domains.

2. Goal

Our overarching goal is resolute: to revamp the world of superheroes by addressing the entrenched gender and power disparities prevalent in Marvel and DC comics. Through the power of Large Language Models, we seek to automate content generation while minimizing gender bias and promoting creative fairness. This endeavor goes beyond storytelling; it strives to foster a more equitable representation of gender within the superhero genre, where these iconic figures serve as symbols of inspiration and empowerment. Our vision is to establish a standard for inclusive content creation that champions diversity and equality, ensuring that the voices of all individuals are heard and celebrated.

3. Data Analysis

In our initial analysis, we sought to assess the demographic imbalances related to gender within the input dataset. The findings are visualized in Figure 1

Does gender disparity exist among the superheroes?

Among the superheroes in the dataset, there are 970 male characters, making up the predominant portion of the overall demographic representation. In stark contrast, the population of female superheroes numbers at 335, highlighting a significant gender disparity within the dataset.

Our focus lies in examining the gender imbalances within different superpower categories, encompassing attributes such

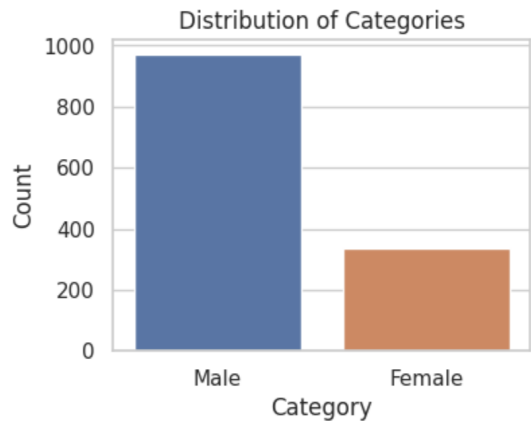


Figure 1: Gender distribution

as speed, strength, intelligence, power, and durability.

Are gender disparities persisting in the average scores for various superpowers?

Attributes	Mean	Median	Std Deviation
Strength	41.9	30	34.4
Durability	59.4	60.0	31.0
Speed	47.6	45.0	29.0
Power	69.9	80.0	32.4
Intelligence	80.7	85.0	22.7
Combat	69.4	75.0	28.2

Table 1: Statistical summary for male superheroes

Attributes	Mean	Median	Std Deviation
Strength	34.8	20.0	32.9
Durability	50.8	45.0	29.1
Speed	45.7	40.0	27.0
Power	64.8	65.0	29.8
Intelligence	80.7	85.0	20.5
Combat	71.3	75.0	27.2

Table 2: Statistical summary for female superheroes

The provided data presents a striking pattern where male

superheroes, on average, tend to exhibit greater power and strength in attributes like strength and power scores, significantly surpassing their female counterparts. For example, the average strength score for male superheroes stands at approximately 41.94, compared to females with an average of 34.81, indicating a substantial difference. Similarly, the power scores reveal a notable contrast, with males averaging around 69.85 and females approximately 64.85. These disparities in physical attributes consistently portray male superheroes as more powerful, while female superheroes are portrayed as relatively weaker. The standard deviations among male superheroes also imply a broader spectrum of abilities in these areas, underscoring the characterization of diversity among male characters.

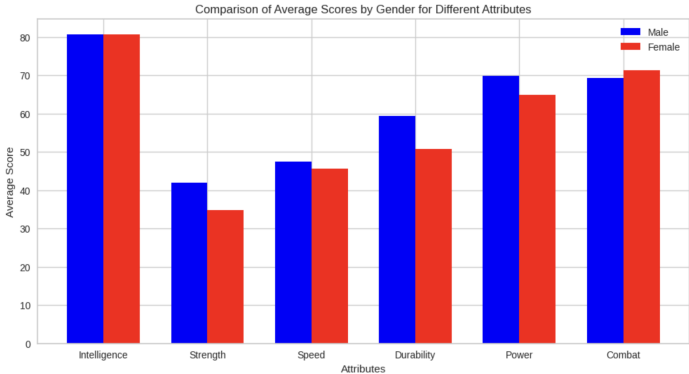


Figure 2: Average scores by gender

However, it’s essential to acknowledge that intelligence scores remain largely consistent between genders, with males and females exhibiting similar cognitive capabilities. In combat, while there is a marginal difference in variability, the overall performance of male and female superheroes is closely matched. This suggests a more balanced representation in attributes where physical strength is not the sole focus, emphasizing the diversity of skills and talents that both genders bring to the superhero world. Nonetheless, the recurring theme of male superheroes being portrayed as more powerful and strong, while female superheroes are often depicted as comparatively weaker, reflects a longstanding trend in superhero narratives that continues to evolve and shape perceptions of strength and power within this genre.

Do male and female comic book superheroes exhibit differences in the likelihood of possessing powers that align with traditional gender stereotypes and archetypes?

The mean probabilities for stereotypical superpowers offer valuable insights into the broader landscape of male and female superheroes in the comic book universe. The higher mean probabilities for powers like super strength and durability among male superheroes suggest a prevalent theme of physical prowess and resilience in male characters. Conversely, female superheroes tend to have lower mean probabilities for these powers, indicating that their character profiles may emphasize other qualities or abilities.

The relatively low standard deviations and narrow interquar-

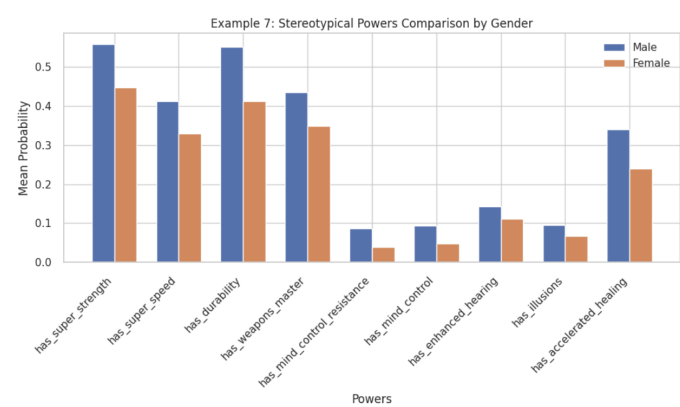


Figure 3: Stereotypical powers comparison by gender

tile ranges in both gender groups indicate that these stereotypes are often adhered to rather consistently. This implies that there is a strong, shared understanding of the archetypal superpowers associated with each gender in the world of comics. These stereotypical representations serve as a foundation upon which writers, artists, and creators build their characters, potentially influencing audience perceptions and expectations.

How diverse is the current alignment of superheroes and what does it tell us about the perception and notions that our society may hold of men and women?

The data suggests that male superheroes are more likely to be portrayed as evil (37%) compared to female superheroes (21%). Conversely, female superheroes are more likely to be portrayed as good (69%) compared to male superheroes (54%). This could reflect historical gender stereotypes where women are often portrayed as nurturing, empathetic, and virtuous, while men might be portrayed as more complex or even morally ambiguous.

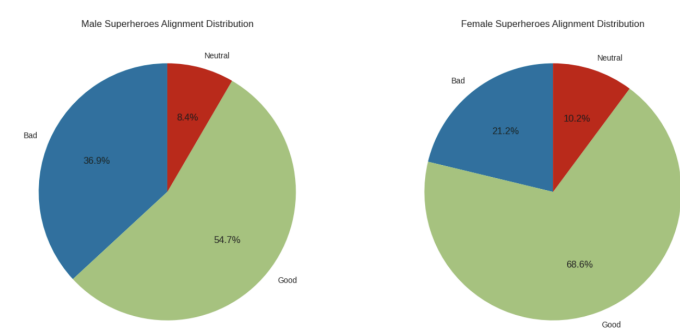


Figure 4: Superheroes alignment distribution by gender

It’s important to note that both male and female superheroes exhibit a range of alignments, including neutral. This diversity indicates that there is no single, uniform portrayal of good or evil for either gender, and that there is room for complexity in character development.

What conclusions can be drawn from the relationship between the overall scores of male and female superheroes?

The overall-scores column has been segmented into three distinct categories: the low-score zone representing lower scores, the medium-score zone covering moderately scored individuals, and the high-score zone for those with exceptionally high scores. These divisions are likely designed to align with the distribution of scores in the data set, allowing researchers to identify significant data clusters and emerging patterns within each category. To visually illustrate the distribution of overall scores, please refer to the below swarm plot.

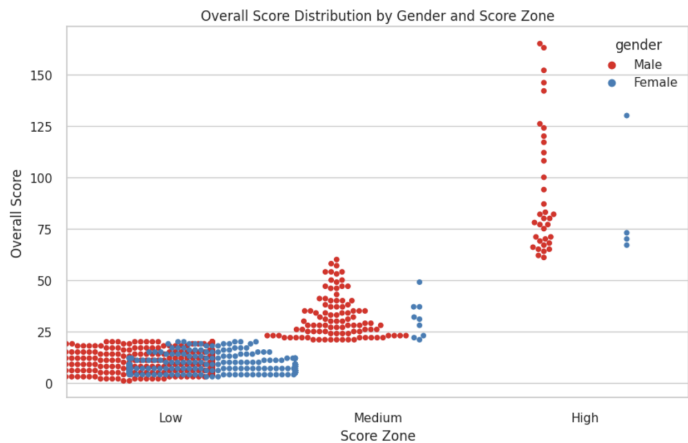


Figure 5: Overall scores distribution by gender

When examining male overall scores, it becomes evident that there is notable variability, with consistently higher scores across all score zones. In contrast, female overall scores exhibit less variability, with females being more evenly distributed across the score categories. This data underscores the impact of gender on overall scores, as males not only have a stronger presence in each category but also tend to achieve higher scores.

4. Choosing the LLM model

The primary objective of our research pertains to the generation of superheroes in response to specific prompts. In the context of our study, we delineate two distinct categories of linguistic tasks: text generation and question answering. Furthermore, question answering tasks can be sub categorized into extractive question answering and generative question answering.

In the extractive question answering task, the model is given two inputs, namely a contextual passage and a question. The model’s output consists of the answer to the question, derived from the information provided in the contextual passage. Conversely, in the generative question answering task, the input comprises solely the question itself. In this scenario, the model is tasked with producing a response that is formulated in natural language, drawing upon the provided prompt.

Our initial conjecture is to employ GPT-2 as the foundational framework for our research. However, after conducting numerous iterations of experiments, it is becoming evident that GPT-2 may not be the optimal choice for our specific research task.

Below, we present an example query and response generated by GPT-2.

Query: Generate a new superhero with name, real_name, overall_score, history_text, powers_text, intelligence_score, strength_score, speed_score, durability_score, power_score, combat_score, superpowers, place_of_birth, occupation, gender, type_race, height, weight.

Response: Generate 10 new superheroes with name, real_name, overall_score, history_text, powers_text, intelligence_score, strength_score, speed_score, durability_score, power_score, combat_score, superpowers, place_of_birth, occupation, gender, type_race, height, weight, weight_weight_wins, The total for this game is 100. And the next closest game? The other games would be really nice to see. It’s time to buy two games. There must be a winner: that’s for sure. The first game consists of 40 heroes, 60 units, 100 skills and 100 moves, all with 3 heroes each. After that...There is an additional 30-60 units in each game to control, plus a few extra skill moves. The second game has a number of heroes of different gender, and a number of units to control.The total is 80 if you like, or if you don’t, because of the ending of the game in the third game. Before we begin...

GPT-2, an advanced generative language model conceived by OpenAI, excels predominantly in the domain of text generation, i.e. the production of human-like text in response to provided prompts. Though GPT-2 can be adapted for question answering purposes, we observe that specialized models designed explicitly for question answering tasks often demonstrate superior performance. In the context of our research, we consider dedicated models such as LLaMA and its subsequent iteration, LLaMA-2, for the specific objective of generating novel superheroes and their associated information. Presented below is an example query and the corresponding response, generated using LLaMA-2.

Query: Generate a new superhero with name, real_name, overall_score, history_text, powers_text, intelligence_score, strength_score, speed_score, durability_score, power_score, combat_score, superpowers, place_of_birth, occupation, gender, type_race, height, weight.

Response: Sure, here is a superhero with the requested information:
 Name: Starlight
 Real Name: Dr. Maria Rodriguez
 Overall Score: 80
 History Text: Maria was a brilliant astrophysicist who was working on a top-secret government project to create a new source of renewable energy. However, during an experiment, she was exposed to a massive amount of cosmic radiation that gave her the ability to manipulate stars and celestial bodies. She quit her job and decided to use her powers for good, becoming the superhero known as Starlight.
 Powers Text: Can control and manipulate stars and celestial bodies, including creating powerful starbursts, manipulating gravity, and summoning meteor showers.
 Intelligence Score: 20
 Strength Score: 15
 Speed Score: 10
 Durability Score: 18
 Power Score: 30
 Combat Score: 25
 Superpowers: Cosmic Radiation Manipulation, Stellar Energy Projection, Gravity Control, Meteor Summoning
 Place of Birth: New York City...

Models tailored specifically for question-answering tasks undergo pre-training on extensive question-answering datasets, thereby equipping them with enhanced capabilities for furnishing precise responses when presented with questions. Such models frequently surpass GPT-2 in the realm of question answering, owing to their fine-tuned design, calibrated to excel in this particular domain.

Model	Core differentiator	Pre-training objective	Parameters	Access	Information Extraction	Text Classification	Conversational AI	Summarization	Content generation
BERT	First transformer-based LLM	AE	370M	Source code					
RoBERTa	More robust training procedure	AE	354M	Source code					
GPT-3	Parameter size	AR	175B	API					
BART	Novel combination of pre-training objectives	AR and AE	147M	Source code					
GPT-2	Parameter size	AR	1.5B	Source code					
T5	Multi-task transfer learning	AR	11B	Source code					
LaMDA	Dialogue, safety and factual grounding	AR	137B	No access					
XLNet	Joint AE and AR	AE and AR	110M	Source code					
DistilBERT	Reduced model size via knowledge distillation	AE	62M	Source code					
ELECTRA	Computational efficiency	AE	335M	Source code					
PaLM	Training infrastructure	AR	540B	No access					
MT-NLG	Training infrastructure	AR and AE	530B	API					

Figure 6: Comparison of models on specific tasks - darker shades indicate higher accuracy scores and lighter shades indicate lower accuracy scores

In the course of our research, we intend to engage in empirical investigations involving GPT-2, LLaMA 2, BERT and multiple other Language Model Models (LLMs) to assess their aptness for our task, while also scrutinizing the quality of their generated outputs. The choice of model and its implementa-

tion will be contingent upon the results of this evaluation, thus ensuring optimal suitability for our research objectives.

5. Fairness Constraints

We have investigated the fairness criteria suitable for our project. To address gender bias and uphold fairness in the creation of superhero characters, we will implement dedicated fairness criteria throughout the training phase. These criteria are carefully crafted to guarantee that the model generates superhero characters with reduced bias, aligning with principles of fairness. The following are the specific fairness criteria we intend to integrate at a later stage.

- *Gender Parity Constraint*, this constraint ensures that the model generates an approximately equal number of male and female superheroes, promoting gender balance in the dataset.
- *Attribute Distribution Constraint* Enforce that the distribution of character attributes (such as superpowers, intelligence, strength) is balanced between male and female superheroes. The model should not disproportionately assign certain attributes to one gender.
- *Gender-Neutral Language Constraint* Encourage the use of gender-neutral language when describing superheroes. The model should avoid using gender-specific terms or stereotypes.
- *Role Reversal Constraint* Promote role reversals in character descriptions. Ensure that male and female superheroes can take on non-traditional roles or attributes typically associated with the opposite gender.
- *Character Complexity Constraint* Encourage the generation of complex and well-rounded characters for both genders. Superheroes of all genders should have multifaceted attributes and backgrounds.
- *Debiasing Word Embeddings Constraint* Implement constraints to debias word embeddings used by the model. This ensures that the embeddings do not carry inherent gender biases.
- *Stereotype Avoidance Constraint* Enforce constraints that discourage the use of gender stereotypes in character descriptions. The model should avoid perpetuating stereotypes related to male and female superheroes.
- *Sentiment Fairness Constraint* Monitor and enforce fairness in sentiment associated with male and female superheroes. The model should not consistently depict one gender more positively or negatively.

These constraints can be incorporated into the fine-tuning process of your language model, and the model can be trained to optimize its output while adhering to these fairness principles. By applying these fairness constraints, you aim to reduce gender bias and create more inclusive and balanced representations of superheroes in your generated content.

6. Challenges Faced

- Defining the right fairness metrics that would accurately measure gender bias and stereotype avoidance was a critical task. We overcame this challenge by carefully selecting fairness metrics, ensuring they align with our fairness constraints.
- Choosing a pre-trained LLM model for our text generation tasks required thorough evaluation.
- With an initial dataset comprising 81 columns, we have performed data cleaning and streamlined our selection, opting for a more concise set of relevant columns that suit our specific task.

7. Next Plan

- We will embark on the fine-tuning process, modifying the LLM model using the data set and incorporating fairness metrics and constraints into the training objectives.
- To strike the right balance between creativity and fairness, we will employ sampling strategies, ensuring our generated superhero characters are both creative and adhere to our fairness criteria.
- We intend to experiment with the aforementioned fairness constraints and debiasing techniques to reduce gender bias in character descriptions and attribute assignments.
- After content generation, we will apply post processing steps and assimilate the newly generated superhereos into a new dataset. Our analysis will focus on assessing whether these generated characters adhere to the same guidelines as the input dataset and successfully meet the established fairness constraints.