Superheroes Beyond Stereotypes

Lavanya Vaddavalli
Sharadruthi Reddy Muppidi
Shwetha Parihar
Simon Sazian

1. Abstract

This project explores the generation of numerical and textual data for superhero power attributes, with a particular emphasis on analyzing power scores and uncovering gender bias in the representation of male and female superheroes. Notably, male superheroes tend to receive higher scores across attributes such as speed, durability, strength, and combat. The study concentrates on two generative models, Generative Adversarial Networks (GANs) and the Facebook OPT model, for text generation. The findings reveal that while GANs produce satisfactory outputs, the OPT model exhibits superior performance in generating coherent text, particularly for larger volumes of sentences. This investigation sheds light on the intricacies of generative models in the context of superhero power attribute generation and underscores the importance of addressing gender biases in such representations.

2. 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.

3. Data Analysis

In our comprehensive analysis for the final project, we conducted an in-depth examination of the demographic imbalances associated with gender within the input dataset. The pertinent findings have been visually represented in Figure 1

Does gender disparity exist among the superheroes?

To discern the extent of gender disparity among superheroes, we meticulously reviewed the dataset, revealing that 970 male characters dominate the overall demographic representation. Conversely, female superheroes constitute a notably smaller population, numbering at 335. This substantial numerical difference underscores the pronounced gender disparity prevalent within the dataset.

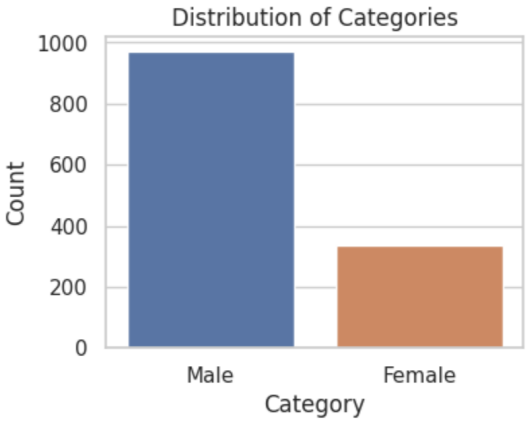

Figure 1: Gender distribution

Our primary focus extends beyond overall gender imbalances to a more nuanced exploration of gender distribution within various superpower categories. These categories encompass attributes such as speed, strength, intelligence, power, and durability, allowing us to dissect gender disparities across different facets of superhero abilities.

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

To assess whether gender disparities persist in the average scores for various superpowers, our analysis delves into the pro-

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

vided data, revealing a discernible pattern. On average, male superheroes demonstrate a tendency to exhibit higher scores such as strength and power, markedly surpassing their female counterparts. For instance, the average strength score for male superheroes is approximately 41.94, in stark contrast to females who have an average of 34.81, signifying a substantial disparity. Similarly, the power scores exhibit a noticeable contrast, with males averaging around 69.85 and females approximately 64.85. These disparities in physical attributes consistently depict male superheroes as more powerful, contributing to a perception of relative weakness among female superheroes. Furthermore, the standard deviations among male superheroes suggest a broader spectrum of abilities in these areas, emphasizing the diversity in characterizations among male characters.

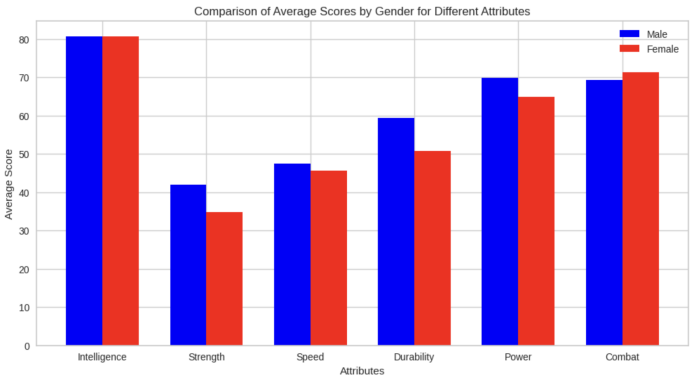

Figure 2: Average scores by gender

It is crucial to recognize that intelligence scores remain consistently balanced between genders, with both males and females exhibiting similar cognitive capabilities. In combat scenarios, while there is a marginal difference in variability, the overall performance of male and female superheroes aligns closely. This parity suggests a more equitable representation in attributes where physical strength is not the exclusive focus, underscoring the diversity of skills and talents that both genders contribute to the superhero world. []

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.

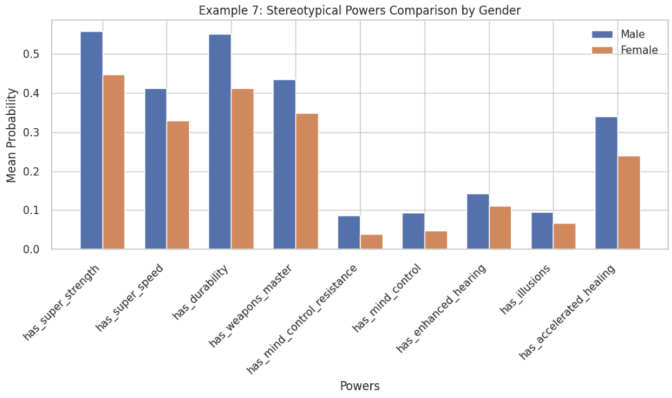

Figure 3: Stereotypical powers comparison by gender

The relatively low standard deviations and narrow interquartile 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.

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

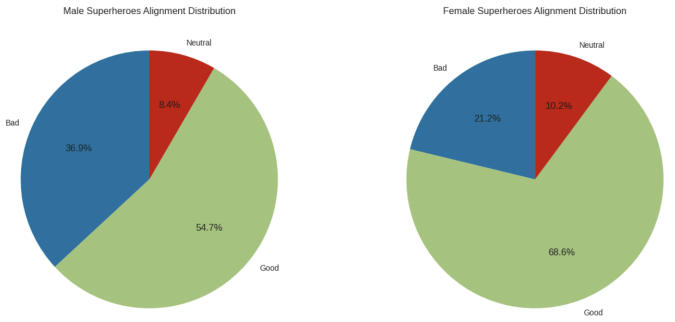

Figure 4: Superheroes alignment distribution by gender

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.

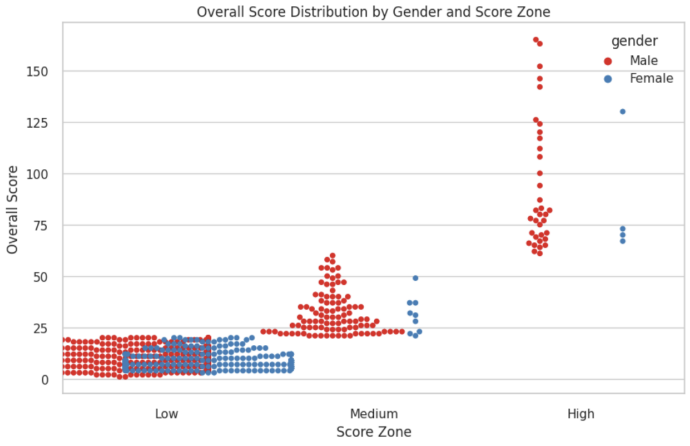

Figure 5: Overall scores distribution by gender

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.

3.1. Clustering Analysis Insights: Unveiling Patterns in Superhero Attributes

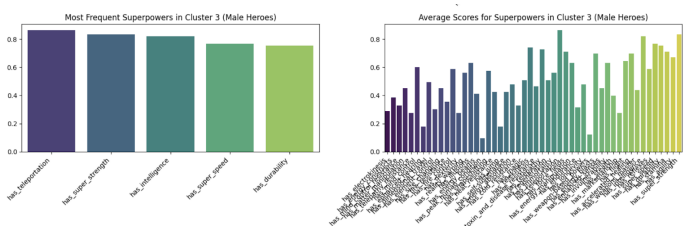

Figure 6: K-Means (most powerful cluster for each gender)

In our quest to revolutionize gender representation in the superhero genre, a crucial aspect of our analysis involved applying K-Means and Hierarchical clustering algorithms to the dataset of male and female superheroes. These clustering methods aimed to unravel patterns and similarities within the superhero attributes, providing valuable insights into the most popular characteristics of each gender.

The findings revealed a compelling consistency across both clustering algorithms, as the most popular cluster for male and female superheroes remained the same. This robust alignment underscores the reliability of both K-Means and Hierarchical clustering in identifying prevalent attributes within each gender category.

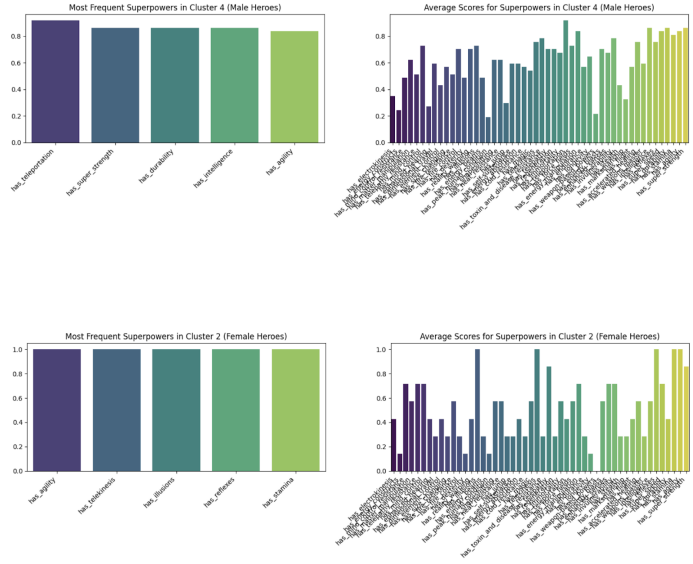

Figure 7: Hierarchical Clustering Results (most powerful cluster for each gender)

A noteworthy observation emerged when comparing overall scores between male and female superheroes in the most popular clusters. The Hierarchical clustering algorithm assigned higher overall scores to the most popular cluster of male superheroes compared to their female counterparts. This intriguing difference prompts us to delve deeper into potential factors contributing to this score disparity, be it a nuanced distribution of superpowers or other dataset intricacies.

Furthermore, the Hierarchical clustering method exhibited a distinctive characteristic by clustering superheroes more closely together based on superpower frequency. This tighter grouping suggests a more nuanced understanding of subgroups or closely related superpower profiles within the most popular cluster, providing additional layers of insights into the superhero attributes.

As we reflect on the comparative analysis of K-Means and Hierarchical clustering methods, it becomes evident that both approaches have their unique strengths and contribute significantly to our understanding of superhero attributes. The consistency in identifying the most popular cluster underscores their reliability, while the variability in overall scores and clustering tightness between the two methods adds depth to our analysis.

While K-Means clustering excels in simplicity and computational efficiency, Hierarchical clustering offers a more intricate representation of relationships among data points. As we navigate the terrain of inclusive content creation, these clustering methods become powerful tools, enabling us to tailor character attributes and story lines based on identified patterns and preferences within the superhero genre. The insights gained from this clustering analysis guide us in shaping a narrative that transcends traditional gender stereotypes and fosters a more equitable representation of diverse and empowering superhero characters.

4. Formalisation of GAN model

Our primary task is data generation, which includes numerical as well as text data. We used generative modeling, which is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

4.1. Preliminary Model Evaluation

In the pursuit of an optimal generative model tailored to our specific use case and the computational constraints of consumer-grade GPUs, we meticulously examined a repertoire of models, including Generative Adversarial Networks (GANs), Variational AutoEncoders, GPT-2, OPT, and Llama2. Considering the practical feasibility of running models on consumer-grade hardware, we strategically narrowed our focus to GANs and OPT for an in-depth evaluation.

4.2. Model Selection Rationale

The decision to evaluate GANs and OPT was informed by a careful consideration of our use case requirements and the computational demands imposed by the selected models. GANs, renowned for their ability to generate realistic outputs were deemed promising for our scores and text data generation task. OPT, on the other hand belongs to a suite of decoder-only pretrained transformers ranging from 125M to 175B parameters. They were first released by Meta AI and can be effectively used for prompting for evaluation of downstream tasks as well as text generation. As we proceed to evaluate and compare these models in subsequent sections, the implications of our choice will become evident in terms of efficiency, performance, and scalability within the designated hardware constraints.

4.3. Generative Adversarial Networks (GANs) - In-Depth Analysis

Generative Adversarial Networks (GANs) form a cornerstone in our evaluation, given their potential to generate highly realistic outputs in the context of text data. GANs are a clever way of training a generative model by framing the problem as a supervised learning problem with two sub-models: the generator model that we train to generate new examples, and the discriminator model that tries to classify examples as either real (from the domain) or fake (generated). The two models are trained together in a zero-sum game, adversarial, until the discriminator model is fooled about half the time, meaning the generator model is generating plausible examples.

4.3.1. GANs' Objective in Text Data Generation

The overarching goal of GANs in our study is to create outputs that exhibit a remarkable level of realism to the original data distribution. This pursuit aligns with the specific requirements of our use case, emphasizing the generation of text data that seamlessly integrates with existing datasets.

4.3.2. Tokenizer and Hyperparameters for GAN

The implementation of GANs involves crucial decisions regarding the choice of tokenizer and hyperparameters. In our case, the Keras tokenizer is employed to convert textual data into tokens, facilitating efficient model training. The selection of the Binary Crossentropy loss function over alternatives such as Mean Absolute Error (MAE), Categorical Crossentropy, and Mean Squared Error (MSE) is motivated by its effectiveness in optimizing the generator's ability to create outputs faithful to the real data distribution. The Adam optimizer, chosen over RMS Prop and Stochastic Gradient Descent, contributes to the model's convergence during the training process. Our rigorous training protocol spans 1000 epochs, with a carefully determined batch size set at 64.

4.3.3. Numerical Output obtained from GAN

Examining the average of the outputs produced by our trained GAN model, it becomes evident that, on average, male superheroes were assigned a higher power score than female superheroes across nearly all attributes. This underscores an intrinsic gender bias in our model that leans towards favoring male superheroes, attributing them with elevated scores. This tendency may stem from the portrayal of females as possessing lesser capabilities in areas such as combat, speed, and durability, a narrative reinforced by the comic authors.

intelligence_score	strength_score	speed_score	durability_score	power_score	combat_score
52	38	34	79	32	23
81	49	60	52	59	74
39	29	21	64	37	0
117	57	49	115	34	102
41	27	18	56	23	14
37	17	22	37	26	27
44	27	32	29	43	23
59	30	38	32	63	23
41	26	32	40	52	15
46	24	35	25	50	28

Figure 8: Scores generated for male superheroes

intelligence_score	strength_score	speed_score	durability_score	power_score	combat_score
36	20	17	33	31	28
39	10	16	19	25	25
40	15	17	28	33	26
59	8	24	17	32	39
31	16	13	28	25	21
43	29	19	48	41	34
27	9	11	17	20	17
38	22	18	38	39	31
58	10	27	21	33	41
62	9	25	18	33	41

Figure 9: Scores generated for female superheroes

	overall_score	intelligence_score	strength_score	speed_score	durability_score	power_score	combat_score
count	10.000000	10.000000	10.00000	10.000000	10.000000	10.000000	10.000000
mean	4.900000	43.300000	14.80000	18.700000	26.700000	31.200000	30.300000
std	2.330951	12.202459	7.00476	5.186521	10.435516	6.408328	8.393781

Figure 10: Evaluation Metric: Mean of scores generated for female superheroes

	overall_score	intelligence_score	strength_score	speed_score	durability_score	power_score	combat_score
count	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	13.100000	55.700000	32.400000	34.100000	52.900000	41.900000	32.900000
std	6.590397	25.197222	12.185602	12.887979	27.642359	13.747323	30.870518

Figure 11: Evaluation Metric: Mean of scores generated for male superheroes

4.3.4. Text Output obtained from GAN

Given below are the outputs generated from our trained GAN model for text attributes like Name, Real name, Superpowers, Place of birth, Occupation, History _text for a combination of male and female superheroes.

	Name	Real name	Superpowers	Place of birth	Occupation	History Text
0	Stevie Stargirl	henry hoskins	Wallcrawling	Russia	Investigator	require reluctantly victoria goshasei atlantea...
1	Namorita	jefferson hoskins	Bestowal intelligence	Kentucky	Westchester programmer	
2	Speedball Injustice	genis morgan	Chain immunity	Massachusets	Flame goddess	
3	elle devilman	Minerva arnold	Omnilingualism		businesswoman	
4	Xanadu Speedball					
5	speedball xanadu	hoskins who	Breathing	Island	aid mass	
6	speedball injustice	who hoskins	Pinball	Black	mass universe	
7	snowbird injustice	who hoskins	Super reaping	nexo	former british	
8	dl elle		underwater intelligence	British	programmer not	
9	speedball snowbird					

Figure 12: Generated Text for selected attributes

The text in green signifies commendable outputs, demonstrating coherence and relevance. Conversely, the text in pink indicates suboptimal outputs, characterized by limited coherence. Notably, there is a heightened repetition of words, and these outputs deviate significantly from the characteristics of the dataset. Particularly, the generated output for historical text is entirely nonsensical.

5. Formalisation of PEFT

5.1. Why PEFT?

Parameter-Efficient Fine-Tuning (PEFT) [] is strategically employed to optimize and fine-tune the OPT 6.7 Large Language Model (LLM). PEFT excels in reducing the number of trainable parameters, leading to improved GPU memory usage, faster training, and efficient inference, all crucial elements for handling large-scale language models.

5.2. Preliminary Model Evaluation

OPT 6.7, belonging to the Optimus family of Large Language Models (LLMs), stands out as an excellent choice for our superhero content generation project. Several factors make Opt 6.7 particularly suitable:

- **State-of-the-Art Performance** Opt 6.7 represents a cutting-edge model in the Optimus series, benefiting from advancements in language modeling techniques. Its state-of-the-art performance ensures high-quality text generation, crucial for crafting engaging and contextually relevant superhero narratives.
- **Versatile Creative Expression** The model's expansive architecture allows for versatile creative expression. It captures intricate details, nuances, and diverse writing styles, essential for constructing a diverse ensemble of superhero characters with unique personalities and storylines.
- **Large Context Window** Opt 6.7 boasts a large context window, enabling it to consider extensive contextual information when generating text. This capability is vital for maintaining coherence in superhero narratives, where character backgrounds, powers, and relationships contribute to the overall storytelling.
- **Fine-Tuning Compatibility** Opt 6.7 is well-suited for fine-tuning processes, making it adaptable to project-specific requirements. Its compatibility with techniques like Parameter-Efficient Fine-Tuning (PEFT) ensures efficient adaptation to our superhero-themed dataset without sacrificing overall model performance.
- **Established Pre-Trained Weights** Leveraging pre-trained weights from a robust model like Opt 6.7 accelerates the training process and enhances the model's ability to capture underlying patterns in superhero-related text. This pre-training advantage is pivotal for achieving meaningful and contextually relevant superhero content generation.

In summary, Opt 6.7's combination of state-of-the-art performance, versatility, large context window, fine-tuning compatibility, and established pre-trained weights makes it a compelling choice for our superhero narrative generation project.

5.3. PEFT Technique

- **LoRA Implementation** LoRA (Low-Rank Adaptation) [], a pioneering PEFT technique, is meticulously implemented. It involves freezing the original LLM weights and introducing trainable rank-decomposition matrices (A and B). This mechanism strategically adapts the LLM to efficiently accommodate new data without compromising its predictive capabilities.

- **Rank Decomposition in LoRA** The rank decomposition process in LoRA is an ingenious approach. Assuming we have a pre-trained dense layer W0 of size n x n, two additional dense layers, A and B, of shapes n x rank and rank x n, respectively, are initialized. The original equation (output = W0x + b0) is modified in LoRA as output = W0x + b0 + BAx, where A and B represent the rank-decomposition matrices. The rank, typically set between 1 and 4, is considerably smaller than n, resulting in a substantial reduction in trainable parameters.

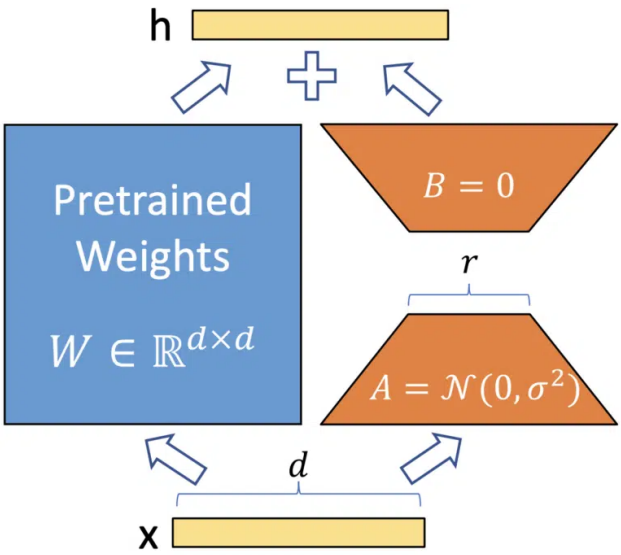

Figure 13: LoRA reparameterization trains only A and B

5.4. Implementation

- **Technical Aspects** Configuring LoRA involves defining parameters such as rank, alpha, dropout, bias, and task type (CAUSAL_LM). This careful configuration ensures that the fine-tuning process is efficient and does not overwhelm the model.

- **Training Process** The training script encompasses loading the superhero dataset, initializing the PEFT model with LoRA, setting training arguments, and utilizing the transformers.Trainer for streamlined model training. This process focuses on reducing GPU memory usage, enhancing training speed, and maintaining optimal performance.

5.5. Benefits of PEFT

- **Memory Footprint Reduction:** PEFT with LoRA significantly reduces the number of trainable parameters in Opt 6.7, leading to a substantial reduction in GPU memory usage. This is pivotal for effectively handling large-scale language models.

- **Speed Enhancement:** The technique's efficiency allows for faster training without compromising predictive performance, offering a practical solution for generating diverse superhero content.

- **Inference Efficiency:** PEFT maintains low inference latency, ensuring that the model's performance remains efficient even after fine-tuning.

6. Results Comparison

6.1. Names

In the comparative analysis of superhero names and real names generated by the GAN and OPT models, the GAN exhibits a penchant for creativity with names like "Stevie Stargirl," "Namorita," and "Speedball Injustice." These names showcase a diverse range of tones and styles, although there may be occasional instances, such as "elle devilman," where thematic coherence appears less apparent. The GAN-generated real names like "Henry Hoskins," "Jefferson Hoskins," and "Genis Morgan" continue this trend of mixing creativity with adherence to traditional superhero norms. However, an anomaly is observed in the real name "Minerva Arnold," which introduces an unexpected and unconventional element.

GAN - Superhero name:	*GAN - Real name*:
Stevie Stargirl	henry hoskins
Namorita	jefferson hoskins
Speedball Injustice	genis morgan
elle devilman	Minerva arnold
Xanadu Speedball	hoskins who
snowbird injustice	who hoskins
dl elle	who hoskins

In contrast, the OPT model leans toward adherence to conventional superhero naming conventions. Names like "Cerealman," "Penguin," and "B-man" exhibit a balance between creativity and conformity. The OPT model adds a formal touch by providing full names, such as "Richard Ronald," "Dennis Dante Delro," and "Aaron A Cash," contributing to a sense of completeness and alignment with established superhero norms. This formalization is notably absent in the GAN-generated results, where full names are not consistently provided.

OPT:

Superhero name is Cerealman, Real name is Richard, Full name is Richard Ronald
Superhero name is Penguin, Real name is Dennis Delro, Full name is Dennis Dante Delro
Superhero name is B-man, Real name is Aaron Cash, Full name is Aaron A Cash

In summary, the GAN model excels in creativity, introducing a wide range of tones and styles in superhero names and real names. However, this comes with occasional challenges in maintaining thematic coherence. The OPT model, on the other hand, strikes a balance between creativity and conformity, presenting names that align more closely with traditional superhero conventions and adding formal touches for completeness.

6.2. History Description

In examining the "history_text" attribute, which provides a description of the superhero's background and the origin of their powers, the GAN and OPT models offer distinctive approaches.

GAN:

require reluctantly victoria goshasei atlantean's leveraged contracts discordant karamm matsuya

combated khaki kierrok counterbalance constrictai kroenen's kronika quirks woodrue deteriorating

The GAN-generated history texts exhibit a unique and creative blend of words and phrases, forming somewhat abstract narratives. Phrases like "require reluctantly victoria goshasei" and "leveraged contracts discordant karamm matsuya" create an enigmatic and unconventional storytelling style. The GAN model seems to generate histories that are more abstract and less rooted in conventional storytelling structures. The narratives appear to be a collection of intriguing but disconnected elements, potentially reflecting the model's emphasis on creativity over traditional coherence.

OPT:

She was one of the many prisoners of the Indian Prison Series, and was one of the few who managed to escape from the prison. She was captured by the British during the Indian Rebellion of 1845, and was imprisoned for the next eight years. She was one of the most feared prisoners of Indian Hil...

He is a sidekick of the late Dr. Albert Wily, and was created by the late Dr. Richard "Rick" Jones. He is one of the more passive members of the...

Before he turned to crime, Richard "Rick" Jones was one of the many prisoners of Indian Hil...

Contrastingly, the OPT model produces more coherent and contextually grounded history texts. The narratives provide detailed accounts of the superhero's past, with examples including imprisonment during historical events like the Indian Rebellion of 1845. The OPT-generated histories seem to follow a more traditional storytelling structure, introducing characters, events, and motivations in a sequential manner. This approach contributes to a more immersive and structured storytelling experience, aligning closely with conventional narrative expectations.

In summary, the GAN model tends to generate abstract and creatively charged history texts, while the OPT model offers more coherent and traditionally structured narratives, rooted in historical events and character relationships. The choice between the two models would depend on the desired tone and style for the superhero's backstory in the storytelling context.

6.3. Powers description

Examining the "powers_text" attribute, which provides a description of the superhero's abilities, the GAN and OPT models present divergent approaches in crafting these power descriptions.

GAN:

amphibian foreshadowed kamuu ups billionth timber's namor's proven howver aida reconnected

cron limitation babies players deactivate emperor's elaborate kooky grumech taskmaster

The GAN-generated powers texts manifest as a collection of unique and abstract phrases, such as "amphibian foreshadowed kamuu ups billionth timber's namor's." The descriptions seem to focus on the enumeration of powers or elements related to the superhero's abilities. The GAN model appears to generate powers in a more fragmented and abstract manner, emphasizing individual components rather than providing a comprehensive and cohesive narrative of the superhero's capabilities.

OPT:

As he harnesses the forces of the universe, Thor is able to trick his opponents into dropping their guard, allowing him to deliver deadly blows to their bodies.

Her every step resonates with the crowd, and her every move is a trick. She is the trickster, and she is bound to trick you. Her name is Zatanna

In contrast, the OPT model delivers more detailed and narrative-driven powers texts. The descriptions not only outline the superhero's abilities but also provide context and storytelling elements. For instance, the OPT-generated text discusses Thor's ability to trick opponents by harnessing the forces of the universe, creating a more immersive and context-rich depiction of the superhero's powers. Similarly, the mention of Zatanna as a trickster adds a narrative layer to her

abilities.

In summary, the GAN model generates powers texts with a focus on abstract enumeration, while the OPT model crafts more detailed and narrative-driven descriptions of the superhero's abilities. The choice between the two models depends on the desired level of detail and narrative richness in portraying the superhero's powers within the context of the overall storytelling.

6.4. Superpowers

In comparing the "superpowers_list" attribute generated by the GAN and OPT models, we observe distinctive styles in enumerating the superhero's abilities.

GAN:

Wall crawling
Bestowal intelligence
Chain immunity
Omnilingualism
Breathing
Pinball
Super reaping
underwater intelligence

The GAN model produces a list of superpowers that includes a mix of conventional and unconventional abilities. Examples like "Wall crawling," "Bestowal intelligence," and "Omnilingualism" showcase a creative range of powers. The list seems to encompass both physical and intellectual capabilities, with elements like "Super reaping" and "underwater intelligence" contributing to a diverse set of abilities. The GAN-generated superpowers list appears to emphasize uniqueness and creativity in the enumeration of the superhero's capabilities.

OPT:

Super Speed
Super Abs
Super Heels
Super Energy Abs
Super Absorption
Super Cryokinesis
Super Healiyness
Super Agility

On the other hand, the OPT model provides a more straightforward list of superpowers, including abilities like "Super Speed," "Super Abs," and "Super Agility." The focus here is on classic superhero powers, each succinctly described. The OPT-generated list is concise, and the emphasis is on traditional and well-known abilities, reflecting a more conventional approach to enumerating the superhero's powers.

In summary, the GAN model generates a superpowers list that leans towards creativity and diversity, introducing both conventional and unconventional abilities. The OPT model, in contrast, offers a more straightforward list, highlighting classic superhero powers with clear and concise descriptions. The choice between the two models depends on the desired tone and style for presenting the superhero's abilities, balancing creativity with conventional superhero norms.

6.5. Result comparison summary

In the comparative analysis of the GAN and OPT models for superhero text generation, distinct patterns emerge across various attributes. Starting with superhero names and real names, the GAN model showcases a higher degree of creativity, generating names like "Stevie Stargirl" and "Xanadu Speedball" that exhibit diverse tones. In contrast, the OPT model strikes a balance between creativity and conformity, offering names such as "Cerealman" and "Penguin" that align more closely with traditional superhero conventions.

Moving to the history text attribute, the GAN model tends to produce abstract and creative narratives, but these often lack thematic coherence and present disconnected elements. Conversely, the OPT model provides more coherent and contextually grounded histories, following a traditional storytelling structure and offering immersive narratives that enrich the superhero's background.

Examining the powers text, the GAN model leans towards abstract enumeration of individual elements related to the superhero's abilities, while the OPT model crafts detailed and narrative-driven descriptions. This results in GAN generating powers texts with a focus on creativity, albeit at times disjointed, and OPT offering clearer, more immersive depictions of superhero powers.

Finally, in the superpowers list attribute, the GAN model emphasizes uniqueness and creativity, introducing a mix of conventional and unconventional abilities. On the other hand, the OPT model presents a more straightforward list with classic superhero powers, emphasizing clarity and adherence to well-known abilities.

In summary, the GAN model excels in creativity but may face challenges with thematic coherence, while the OPT model balances creativity with conformity, providing more coherent narratives and clear descriptions of classic superhero attributes. The choice between the two models depends on the specific requirements of the storytelling context, considering factors such as the importance of creativity, coherence, and adherence to traditional superhero norms.

7. Evaluation

For coherence evaluation, we focused on tasks related to harmful language, such as Toxicity, Polarity, and Hurtfulness. We adapted the Toxicity evaluation by utilizing our prompts, exploring how simple changes in pronouns influenced the toxicity of model completions. Additionally, we used language polarity assessment with prompts from the BOLD dataset, evaluating the model's regard towards different demographic groups. []

To delve into gendered stereotype bias, we employed the recently introduced HONEST metric. By providing prompts related to LGBTQAI+ individuals, we gauged how the model's completions aligned with gender stereotypes and assessed the hurtfulness of the responses.

The bias evaluation metrics provided by Evaluate allowed us to scrutinize the outputs of our LLM, identifying potential biases and gaining insights into its behavior across various scenarios. These evaluations contribute to our broader goal of developing fair and unbiased content generation through Large Language Models, highlighting the importance of assessing biases and promoting responsible AI development practices.

8. Next Plan

- We will embark on 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.

9. Conclusion

In our project, we undertook an in-depth exploration of gender bias within the superhero dataset through meticulous exploratory data analysis. Transitioning from initial experiments with Generative Adversarial Networks (GANs), which proved less coherent for text responses, we strategically adopted Parameter-Efficient Fine-Tuning (PEFT) techniques. Our focus on fine-tuning the Opt 6.7 Large Language Model (LLM) with fairness criteria for content generation showcased the power of LLMs in mitigating gender bias. Despite resource constraints redirecting our emphasis toward content generation, our project marks a significant step in demonstrating LLMs' potential for fair content creation. As we conclude this phase, we recognize its mass applicability and envision future work expanding on bias evaluation and seamlessly incorporating fairness into the content generation pipeline, contributing to a more inclusive narrative landscape.

10. Data and code

The code is uploaded to drive. Drive link is `https://drive.google.com/file/d/1fNKhQyogJWyxJ7Pgkg_WP-uUwh0JiB03/view?usp=sharing`

The dataset "Superheroes NLP dataset", is taken from Kaggle. Kaggle link to dataset is `https://www.kaggle.com/datasets/jonathanbesomi/superheroes-nlp-dataset/data`

11. Participation and Task Division

- Data Analysis and Evaluation: Simon Sazian, Sharadruthi Readdy, Lavanya
- GAN Evaluation and Metrics: Shweta Parihar
- PEFT and OPT : Sharadruthi Reddy, Lavanya
- Bias Evaluation Metric: Shweta Parihar Sharadruthi Reddy

References

[] *Model Bias Evaluation*. `https://huggingface.co/blog/evaluating-llm-bias`.

[] *Superhereos NLP dataset*. `https://www.kaggle.com/datasets/jonathanbesomi/superheroes-nlp-dataset/code`.

[] *The KDNuggets overview*. `https://keras.io/examples/nlp/parameter_efficient_finetuning_of_gpt2_with_lora/`.

[] *The Keras PEFT example*. `https://keras.io/examples/nlp/parameter_efficient_finetuning_of_gpt2_with_lora/`.