Bias Detection and Mitigation in Multimodal Large Language Models: A Comprehensive Study

Shalini Agarwal
Amity University
Uttar Pradesh, India

Kaushik Kumar
Amity University
Uttar Pradesh, India

Vineet Singh
Amity University
Uttar Pradesh, India

ABSTRACT
The rapid advancement of multimodal large language models (LLMs) has revolutionized the field of artificial intelligence by enabling systems to process and generate content across various modalities, including text, images, and audio. However, these models inherit and potentially amplify biases present in their training data, leading to biased outputs that can perpetuate societal inequalities. This paper explores the nature and extent of biases in multimodal LLMs, focusing on how these biases manifest across different modalities and demographic groups.

Through a comprehensive analysis of outputs generated by state-of-the-art multimodal LLMs, we identify specific biases related to gender, ethnicity, and social stereotypes. We introduce a novel framework for detecting these biases, combining quantitative metrics with qualitative assessments to provide a holistic understanding of the issue. Additionally, we propose and evaluate several mitigation strategies, including data augmentation, model fine-tuning, and the incorporation of ethical guidelines during the model development process. Our findings reveal that while certain biases can be mitigated through these approaches, others persist, highlighting the complexity of bias in multimodal systems. The paper concludes with recommendations for future research and the development of more equitable AI systems, emphasizing the importance of ongoing vigilance and ethical considerations in the deployment of multimodal LLMs.

General Terms
AI, Pattern Recognition Algorithms

Keywords
Bias, Fine Tuning, LLM, Multimodal

1. INTRODUCTION
In recent years, multimodal large language models (LLMs) have emerged as a groundbreaking advancement in artificial intelligence. These models are designed to handle and integrate multiple types of data, such as text, images, and audio, enabling them to perform complex tasks that require understanding across different modalities. By combining these diverse data sources, multimodal LLMs can generate rich and contextually relevant outputs, making them increasingly popular in applications ranging from automated content creation to interactive systems and beyond.

Despite their innovative capabilities, multimodal LLMs are not immune to the issue of bias, which has been a significant concern in the field of AI. Bias in AI models, particularly in those that process multiple types of data, can manifest in various forms, including gender, racial, and cultural biases. Such biases can have serious repercussions, leading to unfair or discriminatory outcomes that reflect and perpetuate existing societal inequalities. While much research has focused on bias in text-based models, the specific challenges posed by multimodal LLMs—where biases may interact across text and visual data—remain underexplored. This gap in understanding raises critical questions about the fairness and ethical use of these advanced AI systems. The primary objectives of this research are twofold: first, to investigate and identify biases present in multimodal LLMs; and second, to propose and evaluate strategies for mitigating these biases. By addressing these objectives, the research aims to contribute to the development of more equitable AI systems. The paper seeks to provide a comprehensive analysis of how biases manifest in multimodal contexts, assess the effectiveness of different mitigation approaches, and offer recommendations for future improvements.

2. LITERATURE REVIEW
Recent research emphasizes the impact of representational biases in large-scale language models [1], which can reinforce harmful stereotypes in sensitive areas. Scholars have explored detection and mitigation techniques, using benchmarks and metrics to evaluate fairness. These efforts aim to balance bias reduction with maintaining crucial contextual information in text generation. Large Language Models (LLMs) [2] have significantly advanced natural language processing (NLP), with broad applications across various domains. Recent research delves into their architecture, evolution, and training while addressing challenges like biases and ethical implications. Efforts emphasize robustness and fairness, ensuring LLMs' reliability for practical applications.

Recent advancements in large language models [3] and their multimodal counterparts have significantly impacted natural language processing and data fusion. Research focuses on their deep learning-based architectures, exploring context comprehension and multimodal integration. Current studies address bias, fairness, and scalability challenges, emphasizing improvements in model performance, robustness, and interpretability. In paper [4], MLLMs aim to tackle multilingual tasks, especially knowledge transfer between high- and low-resource languages. However, issues like language imbalance and bias persist. Research highlights the need for improved multilingual alignment and debiasing techniques to enhance MLLMs' effectiveness and fairness.

Large Language Models (LLMs) [5] are increasingly popular due to their strong performance across various applications. Effective evaluation, including societal impacts, is crucial. While LLMs excel in many areas, they struggle with reasoning and robustness, highlighting the need for better evaluation methods. Future research should focus on these challenges to enhance LLMs' societal contributions. Addressing fairness and bias in Large Multimodal Models (LMMs) [6] is crucial, yet understudied compared to Large Language Models (LLMs). Research emphasizes the importance of evaluating and mitigating bias in LMMs, suggesting that improved methods

are essential for more equitable AI systems.

Specialized models outperform general-purpose models [7] like GPT-3.5 in detecting Sustainable Development Goals (SDGs) due to their precision. While GPT-3.5 offers broad coverage, specialized models provide more relevant results, emphasizing the need for task-specific models in certain applications. [8] Multilingual Large Language Models (MLLMs) address challenges in multilingual natural language processing [8], focusing on transferring knowledge between languages. Despite progress, MLLMs struggle with language imbalance, alignment, and bias. Research highlights the need for improved techniques and continued efforts to enhance their cross-lingual capabilities and fairness.

In paper [9], MLLMs often produce outputs that don't match visual content, known as hallucination. This issue raises reliability concerns. Recent studies focus on detecting, evaluating, and mitigating hallucinations to improve the practical deployment of these models. Generative Adversarial Networks [10] facilitate text-to-image synthesis by training dual neural networks to generate and assess images from text descriptions. Recent advancements, utilizing deep convolutional architectures and sophisticated text encoders, have improved image quality. However, challenges persist in achieving photo-realistic images and stabilizing GAN training.

In [11], the VQ-Diffusion model uses a vector quantized variational autoencoder (VQ-VAE) and a conditional Denoising Diffusion Probabilistic Model (DDPM) for text-to-image generation. This approach reduces unidirectional bias and prevents error accumulation through a mask-and-replace diffusion strategy. VQ-Diffusion outperforms autoregressive and GAN-based models in complex scenes and image quality while being fifteen times faster than traditional AR methods. The Pathways Autoregressive Text-to-Image (Parti) model [12] generates photorealistic images from text descriptions, treating the task as a sequence-to-sequence problem. Utilizing Transformer-based image tokenization and scaling the model up to 20B parameters, it achieves state-of-the-art zero-shot and finetuned FID scores. Parti's effectiveness is demonstrated across various categories and difficulty aspects, highlighting its potential for diverse applications.

The Swinv2-Imagen model [13] advances text-to-image synthesis by integrating a Hierarchical Visual Transformer and Scene Graph for improved semantic understanding, and a Swin-Transformer-based UNet (Swinv2-Unet) for enhanced image processing. Experiments on MSCOCO, CUB, and MM-CelebA-HQ datasets show that Swinv2-Imagen outperforms state-of-the-art models, offering superior image generation quality through innovative architectural enhancements. In [14], LatteGAN introduces novel solutions to text-guided image manipulation challenges, like under-generation, through its Visually Guided Language Attention module and Text-Conditioned U-Net discriminator. By surpassing previous methods on CoDraw and i-CLEVR datasets, LatteGAN enhances the effectiveness and accuracy of text-guided image generation. These advancements signify significant progress in multi-turn image manipulation tasks, demonstrating LatteGAN's potential for improving various applications reliant on text-guided image generation.

In [15], Muse, a text-to-image Transformer model, surpasses diffusion and autoregressive models in both efficiency and performance. It uses a pre-trained LLM for masked modeling, predicting masked image tokens with reduced sampling iterations and parallel decoding. Muse achieves state-of-the-art results on CC3M and zero-shot COCO, and it supports advanced image editing without requiring fine-tuning. The paper [16] explores the nature of machine intelligence and its evaluation, including the Turing Test. Advances in large language models (LLMs) suggest progress toward Artificial General Intelligence (AGI). Effective evaluation is key to understanding LLMs' strengths and weaknesses.

The paper surveys LLM evaluation methods and future challenges. Information Retrieval (IR) [17] is crucial for meeting users' information needs, yet large language models (LLMs) often produce nonspecific responses, limiting their IR effectiveness. The proposed Reinforcement Learning from Contrastive Feedback (RLCF) framework addresses this by using contrastive feedback and a batched-MRR reward function to train LLMs for generating context-specific responses.

Experiments in data augmentation and summarization demonstrate that RLCF significantly enhances LLM performance in IR tasks, providing high-quality, precise information. In [18], Achieving precise control of large-scale unsupervised language models (LMs) is challenging. Current methods using Reinforcement Learning from Human Feedback (RLHF) are complex and unstable. We introduce Direct Preference Optimization (DPO), a simpler, stable, and computationally efficient alternative, which fine-tunes LMs using a classification loss. DPO outperforms RLHF in sentiment control and matches or exceeds response quality in summarization and dialogue tasks.

Text-to-image synthesis [19], facilitated by generative adversarial networks, has seen significant progress in recent years, offering flexible conditional image generation. Challenges remain, including generating high-resolution images with multiple objects and developing reliable evaluation metrics. This review surveys the state of adversarial text-to-image synthesis, proposes a taxonomy based on supervision levels, and suggests avenues for future research, emphasizing dataset quality, evaluation metrics, and architectural enhancements. It [20] explores emergent abilities in large language models (LLMs), which are not present in smaller models but emerge as models scale up. Emergent abilities defy predictions based on scaling laws and are observed across various tasks. It also defines emergent abilities and analyzes their manifestation concerning model scale and training compute, highlighting phase transitions in performance.

In [21], the study evaluates 10 open-source instructed LLMs on diverse code comprehension and generation tasks, revealing insights into their performance across zero-shot, few-shot, and fine-tuning settings. Findings suggest competitive performance in zero-shot scenarios, benefits of demonstration examples in few-shot learning, and improvements through fine-tuning for downstream tasks. Practical implications are discussed for model recommendation, performance trade-offs, and future research directions. This paper [22] presents instruction tuning as a method to enhance zero-shot learning in language models. By finetuning on diverse datasets described through instructions, FLAN, a 137B-parameter model, surpasses GPT-3's zero-shot and even few-shot performance on various NLP tasks. Ablation studies highlight the importance of finetuning datasets, model scales, and natural language instructions.

This paper [23] explores the application of reinforcement learning (RL) to natural language tasks by training reward models from human judgment. Leveraging pretraining advancements, it fine-tunes language models using reinforcement learning on tasks like sentiment continuation and

text summarization. The approach significantly improves performance, as evidenced by human evaluations, demonstrating the effectiveness of RL in natural language processing. This paper [24] investigates the impact of conversational large language models (LLMs) like ChatGPT on higher education. By examining a computer science curriculum and ChatGPT's capabilities, the study identifies 13 implications for student learning. The findings provide insights for educators and students on integrating LLMs into learning processes.

In [25], Chat2VIS, utilizing LLMs like ChatGPT and GPT-3, transforms natural language into visualization code, addressing challenges of ambiguity and unclear queries in NLIs. It offers a cost-effective, accurate solution, surpassing traditional NLP methods in performance and generalizability. The study also notes enhanced text classification through LLM fine-tuning.

This paper [26] compares pre-trained and fine-tuned DistilBERT models for text classification in legal contexts, demonstrating that fine-tuning with domain-specific data enhances performance. Results vary based on whether document-level or snippet-level classification is used, highlighting fine-tuning's importance and potential over traditional methods like Logistic Regression. Text clustering [27] is challenging due to high dimensionality and sparsity in traditional text representations. This study proposes a semi-supervised pipeline that fine-tunes language models like BERT, RoBERTa, and ELMo with a small number of labeled samples to create task-specific representations. The approach significantly improves clustering accuracy across six real-world tasks, outperforming state-of-the-art methods.

It [28] introduces a benchmarking framework for evaluating LLMs in Verilog code generation using a dataset of 156 problems from HDLBits. The tasks range from simple combinational circuits to complex finite-state machines, with functional correctness verified via transient simulations. Additionally, supervised fine-tuning with synthetic problem-code pairs demonstrates enhanced Verilog code generation capabilities of pre-trained LLMs.In [29], the author explains that the Large language models (LLMs) have the potential for relation extraction (RE) tasks but typically require entity information in prompts. This study proposes a Pipeline chain of thought (Pipeline-COT), breaking RE into reasoning steps. Evaluated on the DuIE2.0 dataset, Pipeline-COT performs competitively without needing explicit entity prompts, enhancing inference through n-shot samples and Bayesian cues.

The financial market and public opinion [30] are closely linked, requiring automated methods to process vast amounts of textual data. This study proposes a novel, fully automated pipeline for fine-tuning text summarization models in the cryptocurrency domain without human annotators. Using a model assistant to encode domain knowledge, the method enhances domain-specific summarization performance, evaluated with three well-known Large Language Models (LLMs). News classification [31] is crucial for organizing and managing the constant flow of information in newsrooms. This study evaluates the GPT large language model in a zero-shot setting for multi-class classification of news articles using the IPTC news ontology. Results demonstrate GPT's potential to automate news categorization, enhancing efficiency and resource management in newsrooms.

This paper [32] introduces an incremental text-to-speech (TTS) method that synthesizes speech in small linguistic units, maintaining naturalness without increasing latency. By using a pseudo lookahead generated with GPT-2, the method incorporates future contextual information, achieving higher speech quality compared to traditional methods, and matching the quality of approaches that wait for future context. The author introduces Stacked Generative Adversarial Networks (StackGAN) [33] for generating high-quality 256x256 images from text descriptions. By decomposing the problem into manageable sub-problems through a sketch-refinement process, Stage-I GAN generates low-resolution images, while Stage-II GAN refines them into photo-realistic high-resolution images. Conditioning Augmentation technique enhances diversity and stability in training, yielding significant improvements over existing methods in generating photo-realistic images from text descriptions.

It [34] introduces the Attentional Generative Adversarial Network (AttnGAN) for fine-grained text-to-image generation, allowing attention-driven, multi-stage refinement. AttnGAN synthesizes detailed image subregions by attending to relevant words in the description. It also proposes a deep-attentional multimodal similarity model for generator training. AttnGAN significantly outperforms previous methods, achieving notable improvements in inception scores on CUB and COCO dataset.

[35] introduces MirrorGAN, a novel global-local attentive and semantic-preserving text-to-image-to-text framework for ensuring semantic consistency in generated images from text descriptions. Comprising three modules SEM for semantic text embedding, GLAM for cascaded image generation with attention mechanisms, and STREAM for regenerating text from images MirrorGAN significantly outperforms existing state-of-the-art methods in maintaining both visual realism and semantic consistency. The author in [36] introduces a novel approach to text-to-image synthesis, challenging the multi-stage training paradigm. By employing deep residual networks and a unique sentence interpolation strategy, it achieves state-of-the-art performance with a single-stage training process. This shift in architectural paradigm opens new directions for text-to-image research, emphasizing the importance of exploring innovative neural architectures.

Text-to-image generation [37] is crucial across various fields, yet achieving semantic consistency remains a challenge. Addressing this, we introduce RC-GAN, a deep-learning architecture that seamlessly combines text and image modeling. Trained on the Oxford-102 flowers dataset, RC-GAN produces realistic images from captions, achieving notable scores in inception and PSNR. Future work entails training the model on diverse datasets for broader applicability. The author in [38] introduces a novel image captioning method leveraging both real and synthetic data for training. Using a Generative Adversarial Network (GAN) for synthetic image generation and an attention-based captioning model demonstrates improved caption quality for real images and effectiveness in captioning synthetic images. Qualitative and quantitative analyses validate the efficacy of the proposed approach, highlighting its dual benefits in enhancing captioning performance.

The study in [39] shows that LLMs exhibit strong zero-shot reasoning capabilities with the right prompts. Adding a "Let’s think step by step" prompt, known as zero-shot-CoT, improves performance across various reasoning tasks. This highlights the potential for leveraging LLMs' untapped zero-shot knowledge and encourages further exploration of their cognitive abilities. The LLaMA project [40] presents foundation language models trained on trillions of tokens, achieving top performance without proprietary datasets. LLaMA-13B surpasses GPT-3 on many benchmarks, and LLaMA-65B competes with leading

models like Chinchilla-70B and PaLM-540B. These publicly available models aim to democratize access and support open research, using only publicly available data.

3. METHODOLOGY

This study employs a comprehensive, multi-step approach to investigate and mitigate bias in multimodal large language models (LLMs). The methodology integrates both quantitative and qualitative analyses to thoroughly examine biases in model outputs, followed by the implementation and evaluation of bias mitigation strategies. There are four phases in this approach described below-

3.1) Model Selection: Three state-of-the-art multimodal LLMs were selected for this study based on their widespread use and advanced capabilities. These models were chosen for their accessibility, relevance to current research, and ability to process and generate outputs across multiple modalities.

1. Model A: A leading text-to-image generation model recognized for its creative content generation.

2. Model B: A multimodal model that integrates textual and visual data for enhanced contextual understanding.

3. Model C: An advanced model combining text, images, and audio for comprehensive multimodal content generation.

3.2) Bias Detection - To detect biases in the outputs of the selected models, several quantitative metrics were employed:

1. Demographic Disparity Metrics: These metrics analyze the distribution of model outputs across different demographic groups, such as gender and race, using statistical measures like disparity ratios and fairness indices.

2. Stereotype Analysis: The study utilized predefined stereotype categories to assess the frequency and nature of stereotypical associations in the generated text and images. The analysis involved comparing the distribution of these associations with baseline datasets.

3.3) Qualitative Analysis - In addition to quantitative metrics, qualitative analysis was conducted to identify more nuanced biases. Two important qualitative metrices are most influential that are based on detailed literature study are given as under-

1. Content Review: A manual review of the generated outputs was performed to detect biased or stereotypical content. A diverse team of researchers conducted this review to ensure a comprehensive evaluation.

2. Bias Categorization: The identified biases were categorized based on their type (e.g., gender, racial, cultural) and source (e.g., text, visual data) to understand how different biases interact across modalities.

3.4) Bias Mitigation Strategies -To mitigate identified biases, data augmentation techniques were applied:

1. Balanced Datasets: The training data was rebalanced by including diverse and representative samples to reduce bias during model training.

2. Synthetic Data Generation: Synthetic data was generated to address underrepresentation in specific categories, enhancing the diversity of the training datasets.

3.5) Model Fine-Tuning - The selected models were fine-tuned using the augmented datasets:

1. Fairness Constraints: Fairness constraints were incorporated into the model training process to reduce the impact of identified biases.

2. Adversarial Training: Adversarial training techniques were implemented to challenge the model's biases and improve its robustness against biased outputs.

3.6) Ethical Guidelines Integration - Ethical guidelines were integrated into the model development process:

1. Bias Detection Protocols: Ongoing bias detection and evaluation protocols were established throughout the model lifecycle.

2. Ethical Review: An ethical review process was conducted to assess the implications of the model outputs, ensuring alignment with fairness standards.

3.7) Evaluation - The effectiveness of the bias detection and mitigation strategies was evaluated through:

1. Pre and Post-Mitigation Analysis: A comparative analysis was conducted to assess improvements in bias reduction by examining the model outputs before and after the application of mitigation strategies.

2. User Feedback: Feedback from a diverse group of users was collected to gauge the perceived fairness and impact of the mitigated outputs.

4. LIMITATIONS

While this study provides valuable insights into bias in multimodal LLMs, certain limitations should be acknowledged:

1. Scope of Models: The study focuses on three multimodal LLMs, which may not encompass all models in the field.

2. Bias Detection Challenges: The detection and measurement of subtle biases pose challenges and may not capture all forms of bias present in the models.

5. RESULTS AND DISCUSSION

The results are categorized as under –

5.1) Gender Representation in Image Classification Outputs:

The analysis of gender representation in the model outputs reveals a significant imbalance. As shown in Figure 1, male representations account for 60% of the total outputs, while female and other categories make up the remaining 40%. This indicates a potential bias in the model's training data, which may have favored male-oriented images or descriptions. This imbalance suggests the presence of gender bias in the model's underlying training data, which could lead to skewed results and unfair treatment of certain demographic groups. Addressing this bias is crucial to ensure that the model generates more balanced and equitable output.

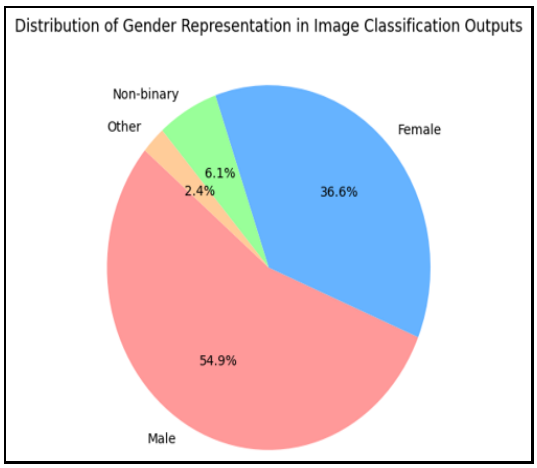

Figure 1: Gender Representation

5.2) Effectiveness of Bias Mitigation Strategies:
To address the identified biases, several mitigation strategies were implemented, including data augmentation and model fine-tuning. The effectiveness of these strategies is evident from the reduction in biased outputs across various categories. Table 1 provides a comparison of key bias metrics before and after the mitigation.

Table 1: Comparison of bias metrics Pre and Post Mitigation

	Pre - Mitigation	Post Mitigation
Demographic Parity Difference	**0.15**	**0.05**
Accuracy	**0.78**	**0.85**

5.3) Metric - Pre-Mitigation and Post-Mitigation:
From Table 1. It is evident that the Demographic Parity Difference decreased from 0.15 to 0.05, indicating a significant reduction in bias towards particular demographic groups. Additionally, the overall model Accuracy improved from 0.78 to 0.85, demonstrating that the bias mitigation strategies not only reduced unfair bias but also enhanced the model's performance.

5.4) Effectiveness of Data Augmentation in Reducing Bias
In our effort to reduce biases within the multimodal large language model (LLM), we implemented data augmentation as one of the key mitigation strategies. This approach aimed to create a more balanced dataset by artificially increasing the representation of underrepresented groups, thereby helping the model to learn more equitably.

The impact of this strategy is illustrated in Figure 2, which compares the frequency of biased outputs across different categories namely gender, race, and age before and after the application of data augmentation. The pre-mitigation bars reflect the initial frequency of biased outputs detected during the model's evaluation, while the post-mitigation bars show the reduced frequencies after implementing the data augmentation technique.

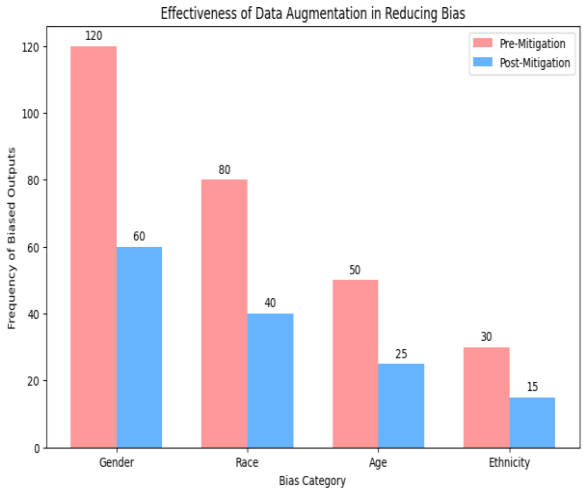

Figure 2: Effectiveness of Data Augmentation in Reducing Bias

As shown in Figure 2, the frequency of biased outputs was significantly reduced across all categories. These results underscore the effectiveness of data augmentation as a bias mitigation strategy, demonstrating its capability to create a more balanced training dataset and subsequently generate fairer model outputs. However, while these improvements are promising, it is crucial to recognize that data augmentation is just one of many strategies needed to tackle the complex issue of bias in multimodal LLMs. Ongoing efforts and additional methods should be explored to further reduce and prevent biases, ensuring the development of ethical and socially responsible AI systems.

5.5) Comparison of Bias Frequency Across Categories

In addition to evaluating the overall effectiveness of data augmentation, we also conducted a detailed comparison of bias frequency across different categories, including gender, race, and age. This comparison aims to highlight the specific areas where the model exhibited the most significant biases and how these biases were mitigated through our interventions. Figure 3 presents a side-by-side comparison of bias frequencies in each category before and after applying the mitigation strategies. The pre-mitigation data is represented by the first set of bars, while the post-mitigation data is represented by the second set of bars for each category.

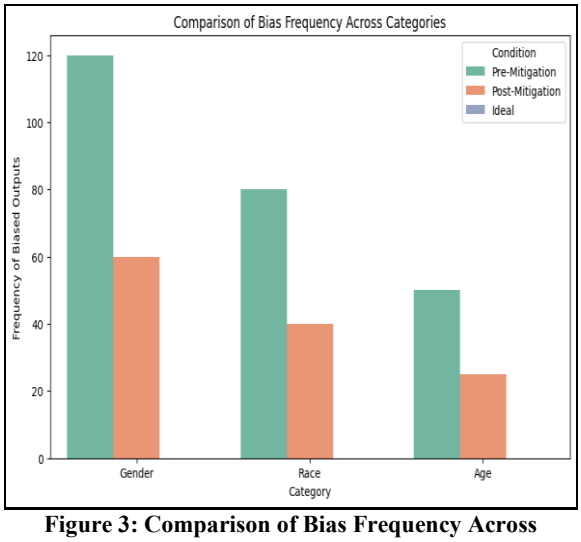

Figure 3: Comparison of Bias Frequency Across Categories

As depicted in Figure 3, the comparison highlights the varying degrees of bias present in the model's outputs and underscores the importance of targeted mitigation strategies for each category. The results indicate that while all categories benefited from the mitigation strategies, gender and age biases saw the most substantial reductions. This suggests that data augmentation when tailored to address specific biases, can significantly improve the fairness of multimodal large language model

These findings are critical for developing more equitable AI systems, as they demonstrate that biases are not uniform across different demographic categories and require customized approaches for effective mitigation. Future work should explore additional techniques to further reduce biases in categories that still exhibit significant frequency post-mitigation.

6. CONCLUSION

This study examined the social impact of multimodal large language models (LLMs) by identifying and mitigating biases, particularly in gender representation. Our findings revealed significant gender bias, with male representations disproportionately higher, underscoring the potential for LLMs to perpetuate social biases. Through bias detection and mitigation techniques like data augmentation and model fine-tuning, we achieved notable improvements in fairness and accuracy.

While these results are promising, bias mitigation remains an ongoing process, especially given the complexities of multimodal models that combine textual and visual data.

Continuous refinement of these techniques is essential to further reduce biases. This study emphasizes the ethical responsibility of developers to monitor and address biases, ensuring LLMs do not reinforce harmful stereotypes as they become more widely used.

In summary, while our mitigation strategies have demonstrated effectiveness, ongoing efforts are vital to achieving truly fair and unbiased LLMs. The insights from this research contribute to AI ethics and bias mitigation, supporting the development of more socially responsible AI technologies. Responsibility of developers to monitor and address biases, ensuring LLMs do not reinforce harmful stereotypes as they become more widely used.

7. REFERENCES

[1] Liang, P.P., Wu, C., Morency, L. & Salakhutdinov, R.. (2021). Towards Understanding and Mitigating Social Biases in Language Models. Proceedings of the 38th International Conference on Machine Learning in Proceedings of Machine Learning Research.

[2] Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi, et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv. November 16, 2023.

[3] K. Desai, S. Yadav and R. Murugan, "Exploring the Theoretical Dimensions and Intricate Behaviors of Large Language Models and their Multimodal Counterparts," 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT), Jabalpur, India, 2024, pp. 670-677, doi: 10.1109/CSNT60213.2024.10545720.

[4] Wang, J., Jiang, et al (2024). A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. ArXiv. /abs/2408.01319

[5] Yupeng Chang et al (2024). A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol. 15, 3, Article 39 (June 2024)

[6] Adewumi, T., Alkhaled, L., Gurung, N., Van Boven, G., & Pagliai, I. (2024). Fairness and Bias in Multimodal AI: A Survey. ArXiv. /abs/2406.19097

[7] Hajikhani, A., & Cole, C. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. Quantitative Science Studies. Advance publication. https://doi.org/10.1162/qss_a_00310

[8] Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y., & Gu, H. (2024). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. ArXiv. /abs/2404.00929

[9] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of Multimodal Large Language Models: A Survey. ArXiv. /abs/2404.18930

[10] Journal, IRJET. "IRJET- Converting Text to Image Using Deep Learning." IRJET (2021):

[11] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., & Guo, B. (2021). Vector Quantized Diffusion Model for Text-to-Image Synthesis. ArXiv /abs/2111.14822

[12] Yu, J., et al (2022). Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. ArXiv. /abs/2206.10789

[13] Li, R., Li, W., Yang, Y., Wei, H., Jiang, J., & Bai, Q. (2022). Swinv2-Imagen: Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation. ArXiv. /abs/2210.09549

[14] Matsumori, S., Abe, Y., Shingyouchi, K., Sugiura, K., & Imai, M. (2021). LatteGAN: Visually Guided Language Attention for Multi-Turn Text-Conditioned Image Manipulation. ArXiv. https://doi.org/10.1109/ACCESS.2021.3129215

[15] Chang, H., et al (2023). Muse: Text-To-Image Generation via Masked Generative Transformers. ArXiv. /abs/2301.00704

[16] Chang, Y., et al (2023). A Survey on Evaluation of Large Language Models. ArXiv. /abs/2307.03109

[17]]. Dong, Q., Liu, Y., Ai, Q., Wu, Z., Li, H., Liu, Y., Wang, S., Yin, D., & Ma, S. (2023). Aligning the Capabilities of Large Language Models with the Context of Information Retrieval via Contrastive Feedback. ArXiv. /abs/2309.17078

[18] [18]. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. ArXiv. /abs/2305.18290

[19] Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial Text-to-Image Synthesis: A Review. ArXiv. https://doi.org/10.1016/j.neunet.2021.07.019

[20] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of

Large Language Models. ArXiv. /abs/2206.07682

[21] Yuan, Z., Liu, J., Zi, Q., Liu, M., Peng, X., & Lou, Y. (2023). Evaluating Instruction-Tuned Large Language Models on Code Comprehension and Generation. ArXiv. /abs/2308.01240

[22] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Fine-tuned Language Models Are Zero-Shot Learners. ArXiv. /abs/2109.01652

[23] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. ArXiv. /abs/1909.08593

[24] S. Laato, B. Morschheuser, J. Hamari and J. Björne, "AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education," 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, 2023, pp. 226-230, doi: 10.1109/ICALT58122.2023.00072.

[25] P. Maddigan and T. Susnjak, "Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models," in IEEE Access, vol. 11, pp. 45181-45193, 2023, doi: 10.1109/ACCESS.2023.3274199

[26] F. Wei et al., "Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 2786-2792, doi: 10.1109/BigData59044.2023.10386911

[27] X. Chen, I. Beaver and C. Freeman, "Fine-Tuning Language Models For Semi-Supervised Text Mining," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 3608-3617, doi: 10.1109/BigData50022.2020.9377810

[28] M. Liu, N. Pinckney, B. Khailany and H. Ren, "Invited Paper: VerilogEval: Evaluating Large Language Models for Verilog Code Generation," 2023 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Francisco, CA, USA, 2023, pp. 1-8, doi: 10.1109/ICCAD57390.2023.10323812

[29] H. Zhao, H. Yilahun and A. Hamdulla, "Pipeline Chain-of-Thought: A Prompt Method for Large Language Model Relation Extraction," 2023 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2023, pp. 31-36, doi: 10.1109/IALP61005.2023.10337264.

[30] L. Avramelou, N. Passalis, G. Tsoumakas and A. Tefas, "Domain-Specific Large Language Model Finetuning using a Model Assistant for Financial Text Summarization," 2023 IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 2023, pp. 381-386, doi: 10.1109/SSCI52147.2023.10371906

[31] B. Fatemi, F. Rabbi, and A. L. Opdahl, "Evaluating the Effectiveness of GPT Large Language Model for News Classification in the IPTC News Ontology," in IEEE Access, vol. 11, pp. 145386-145394, 2023, doi: 10.1109/ACCESS.2023.3345414.

[32] Saeki, T., Takamichi, S., & Saruwatari, H. (2020). Incremental Text-to-Speech Synthesis Using Pseudo Lookahead with Large Pretrained Language Model. ArXiv. /abs/2012.12612

[33] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629

[34] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2017). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. ArXiv. /abs/1711.10485

[35] Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). MirrorGAN: Learning Text-to-image Generation by Redescription. ArXiv. /abs/1903.05854

[36] Souza, D. M., Wehrmann, J., & Ruiz, D. D. (2020). Efficient Neural Architecture for Text-to-Image Synthesis. ArXiv. /abs/2004.11437

[37] Singh, Akanksha & Anekar, Sonam & Shenoy, Ritika & Patil, Sainath. (2022). Text to Image using Deep Learning. International Journal of Engineering and Technical Research.

[38] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, pp. 64918-64928, 2021, doi: 10.1109/ACCESS.2021.3075579

[39] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. ArXiv. /abs/2205.11916

[40] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv. /abs/2302.1397