ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 8, Issue 4, April 2025

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI:10.15680/IJMRSET.2025.0804288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Detecting and Mitigating Bias in Large Language Models

Gogineni Lekhana Chowdary, Dr. Nimmagadda Sreeram

B. Tech, Department of [AIML], NIMS University, Jaipur, India

Professor, RVR & JC College of Engineering, Chowdavaram, Guntur, India

ABSTRACT: Large Language Models (LLMs) are incredibly powerful, but they're not free from flaws—especially when it comes to bias. These models often reflect stereotypes and unfair patterns present in their training data. This project takes a closer look at how such biases show up, why they matter, and what we can do to reduce them. We explore ways to spot bias, like analyzing outputs, and test solutions including fine-tuning, better data handling, and prompt tweaks. The aim is to make LLMs more fair and reliable, helping build AI that works better for everyone, not just a few.

KEYWORDS: large language models, sterotypes, training data, prompt enginnering, Model Reliability.

I. INTRODUCTION

Large Language Models (LLM) are the recent advancements in Artificial Intelligence (AI) technology, and are likely to determine the future of technology and society. You can find tools like ChatGPT and BERT everywhere now, whether in classrooms, offices, mobile apps, and customer support. These models can write essays, answer questions, translate languages, and help people code. On the surface, they seem like magic If we take a deeper look at it, it becomes clear that they don't seem as neutral or objective as they are made out to be. That's where this project comes in As powerful as these models are, they're not perfect It has been observed that bias is a major issue by many researchers. Models are trained on huge datasets that are taken from internet which includes websites, social media, books, and etc. The dataset fails to provide representative samples of different communities and often includes racist and sexist stereotypes. Because they learn from that data, they reproduce and amplify the unfair patterns present in the data.For example, if a model constantly sees text where doctors are men and nurses are women, it might start suggesting or assuming that kind of stereotype in its responses. In more serious cases, biased models can produce harmful or offensive content, or make unfair decisions when used in things like hiring tools or legal tech. The risk isn't just academic—it's real, and it can impact people's lives in subtle but important ways.

That's why our project focuses on **detecting and mitigating bias** in large language models. We want to better understand where the bias comes from, how it appears in model behavior, and what can be done to reduce its effects. This isn't just about making AI "look good." It's about building systems that are trustworthy, respectful, and inclusive—AI that reflects the kind of fairness we expect in real life.

To do this, we explore several key areas. First, we look at how bias shows up in LLMs—what it looks like, and how to measure it. Next, we dive into different techniques to reduce that bias, such as improving the quality of training data, adjusting prompts to be more neutral, and fine-tuning models with targeted feedback. We also explore some of the ethical questions surrounding AI: who decides what counts as "fair"? How do we balance fairness with accuracy? And how can we ensure these tools work equally well for people of all backgrounds?

This project isn't about solving everything overnight. Bias in AI is a complex and evolving issue. But by studying it closely and trying different solutions, we're taking steps in the right direction. Our hope is that this work can help create a future where AI supports everyone fairly—regardless of gender, race, language, or background.

In short, this project is about making powerful technology more responsible. As AI continues to grow and become part of our daily lives, we believe it's essential to make sure it treats all users with respect. That starts by understanding its flaws and working actively to fix them—not just for the sake of better software, but for the sake of better outcomes for real people.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI:10.15680/IJMRSET.2025.0804288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

1. **Language models reflect societal data**
Large Language Models (LLMs) are trained on massive datasets sourced primarily from the internet, which inevitably carry the same societal biases found in human-generated content.
2. **Gender bias is well-documented**
Studies (e.g., Bolukbasi et al., 2016) show that LLMs often associate professions like "engineer" with males and "nurse" with females, perpetuating harmful stereotypes.
3.**Racial and ethnic bias affects model outputs**
Research reveals that LLMs respond differently to inputs based on racial or ethnic indicators, such as names or dialect, contributing to skewed or discriminatory responses.
4. **Training data plays a pivotal role**
The quality and diversity of training data significantly influence how fairly a model behaves. Biased inputs often lead to biased outputs.
5.**Bias can be both explicit and implicit**
While some bias is clearly reflected in word associations, others manifest subtly in tone, context selection, or likelihood of generating certain responses.
6.**Detection frameworks have emerged**
Tools like **StereoSet**, **CrowS-Pairs**, and custom probing tasks are commonly used to evaluate models for gender, racial, and ideological bias.
7.**Prompt engineering as a mitigation method**
Researchers have shown that rephrasing inputs can reduce bias in outputs, though this approach does not fundamentally solve the issue.
8. **Data-level interventions show promise**
Cleaning, balancing, and curating training data helps prevent models from learning biased relationships in the first place.
9. **Fine-tuning improves fairness**
Models can be fine-tuned on more inclusive datasets or tailored objectives to help reduce biased predictions and improve representational fairness.
10.**Human feedback strengthens model behavior**
Reinforcement Learning from Human Feedback (RLHF) allows LLMs to align better with ethical and socially responsible behavior, especially when fairness is prioritized.
11.**Post-processing mechanisms offer control**
Output filtering or rewriting techniques help catch biased or inappropriate outputs after generation, although they cannot modify the model's internal behavior.
12. **Fairness-performance trade-off is common**
Some mitigation strategies, while reducing bias, may compromise fluency, creativity, or task-specific accuracy—requiring careful balancing.
13.**Bias in LLMs is not intentional**
Models do not have beliefs or intentions; they reproduce patterns found in their training data without understanding social implications.
14.**Multilingual and cross-cultural bias is underexplored**
While English-language bias is well studied, biases across different languages and cultures remain a less-addressed but crucial area of research.
15. **Ongoing research and interdisciplinary efforts are essential**
Addressing bias requires continuous development, transparency, and collaboration across fields like NLP, ethics, sociology, and law.

III. APPROACH

Bias in large language models (LLMs) is a real concern that affects many aspects of our lives, from job applications to social media interactions. As these models become more integrated into our daily routines, it's crucial to ensure they operate fairly and equitably for everyone.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI:10.15680/IJMRSET.2025.0804288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Bias can show up in different ways within LLMs. For example, **representation bias** happens when certain groups are underrepresented in the training data. This can lead to models that don't accurately reflect the diversity of our society. **Measurement bias** can occur due to flaws in how data is collected or labeled, while **algorithmic bias** might be introduced through the model's design or training process. Recognizing these biases is the first step toward addressing them.

Detecting bias involves a mix of methods. On the quantitative side, researchers use tools like the Word Embedding Association Test (WEAT) to measure biases in word associations. There are also fairness metrics that assess how different demographic groups are treated by the model. On the qualitative side, human evaluations are essential. Diverse groups of people can review the model's outputs to spot biases that numbers alone might miss.

Once we identify bias, there are several strategies we can use to tackle it. **Preprocessing techniques** involve adjusting the training data before the model learns from it. This might mean adding more examples from underrepresented groups or removing biased instances. During the training phase, we can use **in-training techniques** like adversarial training, which challenges the model to reduce bias as it learns. Regularization techniques can also help keep bias in check. After the model generates outputs, **post-processing techniques** allow us to make adjustments to correct any biases, using algorithms designed to ensure fairness and representation.

Ethics play a vital role in this conversation. It's important to be transparent about how we detect and mitigate bias. Establishing accountability for biased outputs is crucial, as is involving diverse voices in the development process to create more equitable AI systems.

Looking ahead, collaborating with experts from various fields, such as social sciences and ethics, can deepen our understanding of bias. Exploring new AI architectures that might inherently reduce bias is another promising direction. Engaging with policymakers to create guidelines for fairness in AI will be essential for the future

3.1 Understanding Bias Detection

Detecting bias is the first step in addressing the negative impacts of bias in large language models (LLMs). Researchers have developed various techniques to identify and measure biases within these models. One of the most well-known methods is the **Word Embedding Association Test (WEAT)**, which looks at how different demographic groups are associated with certain stereotypes. By comparing the similarities between biased and unbiased word pairs, WEAT helps quantify bias in language models.

Another important advancement is the development of **context-aware bias detection methods**. For instance, Zhao and colleagues introduced techniques that consider the context in which words are used, moving beyond static tests. This approach captures the nuances of language, making it more sensitive to biases that might be missed in simpler evaluations. Additionally, analyzing the training data itself is crucial. Researchers like Sheng et al. have shown that examining the datasets used to train LLMs reveals significant gender and racial biases. Their findings highlight the need for more balanced and representative datasets to ensure fairness in AI models.

Strategies for Mitigating Bias

Once biases are identified, it's essential to take steps to mitigate them. Several strategies have been proposed, including data augmentation, adversarial training, and algorithmic adjustments.

Data augmentation involves expanding the training dataset with diverse examples to reduce the model's exposure to biased data. For example, Zhao et al. demonstrated that adding counterfactual examples—where specific demographic attributes are altered—can significantly reduce gender bias in model outputs.

Adversarial training is another effective technique. This method incorporates adversarial objectives to help train models that produce unbiased representations. Ravfogel and colleagues introduced methods to neutralize unwanted biases, ensuring that the model's outputs remain consistent regardless of sensitive attributes.

Other algorithmic adjustments, such as **fairness constraints** and **regularization techniques**, guide LLMs to make unbiased decisions. For instance, Hardt et al. proposed equalized odds constraints, which ensure that model predictions are independent of sensitive attributes, leading to fairer classifications.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Post-processing techniques modify the model's outputs to achieve fairness without altering the underlying model. Kamiran and Calders developed methods to adjust decision thresholds after training, ensuring equitable outcomes across different demographic groups.

The Role of Knowledge Graphs in Enhancing Fairness

Knowledge Graphs (KGs) have emerged as powerful tools for integrating structured knowledge into AI models. Their structured nature allows for clear representation of entities and relationships, which can help counteract biases in LLMs. One key application of KGs in bias mitigation is enhancing training data with structured knowledge. Wang et al. explored how KGs can provide additional context and factual information, reducing reliance on biased associations learned from unstructured text. By incorporating KGs, models gain access to verified and balanced information, leading to more accurate and unbiased predictions.

KGs also facilitate the implementation of fairness constraints by providing a framework to define and enforce fairness criteria. Yu et al. utilized KGs to encode fairness-related attributes, enabling models to recognize and adhere to fairness guidelines during training and inference.

Moreover, KGs contribute to **explainable AI (XAI)**, enhancing the transparency and interpretability of LLMs. By integrating KGs, models can generate more understandable outputs, allowing users to see the reasoning behind decisions. This transparency is vital for identifying and correcting biases.

To effectively integrate KGs with LLMs, various techniques have been proposed. One approach is **embedding alignment**, where entities and relationships from KGs are embedded in the same vector space as the LLM's word embeddings. This alignment allows for seamless integration of structured knowledge. Yao et al. introduced a method to jointly train KG embeddings with LLMs, ensuring effective utilization of both information types.

Another method involves using **Graph Neural Networks (GNNs)** to encode the structural information of KGs before incorporating them into LLMs. GNNs can extract rich relational features that enhance the model's contextual understanding. Wu et al. demonstrated that using GNNs to enhance LLMs with KG-derived features improves both fairness and overall performance.

Attention mechanisms have also been adapted to incorporate KG information, allowing models to focus on relevant parts of the KG during processing. For example, Li et al. proposed a knowledge-aware attention mechanism that enables LLMs to attend to specific entities and relationships within a KG, improving the generation of unbiased and contextually appropriate responses.

3.2 Real-World Applications and Impact

The integration of KGs with LLMs for bias detection and mitigation has been applied across various domains, each benefiting uniquely from this approach. In healthcare, knowledge graph-augmented LLMs have helped reduce biases in diagnostic recommendations, ensuring that diverse patient populations receive equitable treatment.

In the financial sector, these integrated models have improved fairness in credit scoring and fraud detection, preventing biased decisions that could disproportionately affect marginalized groups. In the legal domain, knowledge graph integration has enhanced the fairness of case outcome predictions by providing balanced

3.3 Challenges and Future Directions

One of the main challenges is the complexity of language. Language is nuanced and context-dependent, making it difficult to pinpoint and quantify biases accurately. A word or phrase can have different meanings based on its context, complicating bias detection. For example, a term that seems neutral in one situation might be biased in another, leading to potential misunderstandings.

Another significant hurdle is the need for diverse training data. Many LLMs are trained on datasets that do not adequately represent the full spectrum of human experiences. This lack of diversity can perpetuate existing biases and limit the model's ability to generalize across different demographics. When certain groups are underrepresented, the model may produce outputs that reinforce stereotypes rather than challenge them.

Measuring bias accurately is also a considerable challenge. Current metrics for assessing bias may not capture the full extent of bias present in model output.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI:10.15680/IJMRSET.2025.0804288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Enhancing data representation is another critical area. We should curate more balanced and representative datasets that reflect diverse perspectives. Techniques like data augmentation and counterfactual examples can help mitigate biases in training data, ensuring that models are more equitable.

Integrating ethical frameworks into AI development is essential for fairness and accountability. Collaborating with ethicists and community stakeholders can help shape AI systems that prioritize inclusivity. Establishing clear ethical guidelines will provide a foundation for responsible AI practices.

IV. METHODOLOGY AND RESULTS

Our methodology consists of several interconnected steps:

4.1 Dataset Selection and Preprocessing

The first step involves carefully selecting diverse and representative datasets that reflect a wide range of perspectives. We will preprocess these datasets to remove any explicit biases and ensure they are suitable for training. This step is crucial for laying a solid foundation for the subsequent phases of our research.

4.2 Knowledge Graph Integration

Next, we will integrate knowledge graphs into the training process. Knowledge graphs provide structured information that can enhance the model's understanding of relationships and context, helping to mitigate biases that arise from limited or skewed training data. This integration allows the model to draw on a broader knowledge base, improving its ability to generate fair and balanced outputs.

4.3 Model Training and Fine-Tuning

With the enriched dataset and knowledge graph in place, we will proceed to train and fine-tune the LLM. This phase involves adjusting the model's parameters to optimize its performance while being mindful of bias. We will employ techniques such as transfer learning to leverage pre-trained models, ensuring efficient training and better results.

4.4 Bias Detection and Mitigation Techniques

To identify biases within the model, we will implement various detection techniques, such as fairness metrics and adversarial testing. Once biases are detected, we will apply mitigation strategies, including re-weighting training samples and adjusting model outputs, to reduce the impact of identified biases.

4.5 Data Visualization

Data visualization plays a critical role in understanding the results of our bias detection and mitigation efforts. We will create visual representations of our findings, making it easier to communicate complex data and insights to a broader audience. This step will help stakeholders grasp the implications of our research and the effectiveness of our methodology.

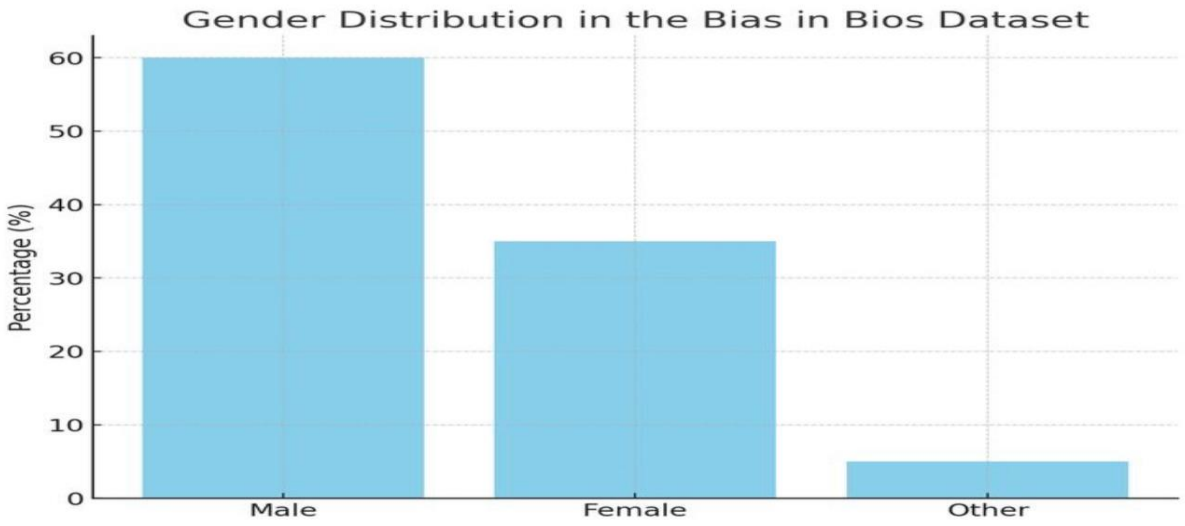

Figure 1:Gender Distribution in the Bias in Bios Dataset

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI:10.15680/IJMRSET.2025.0804288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

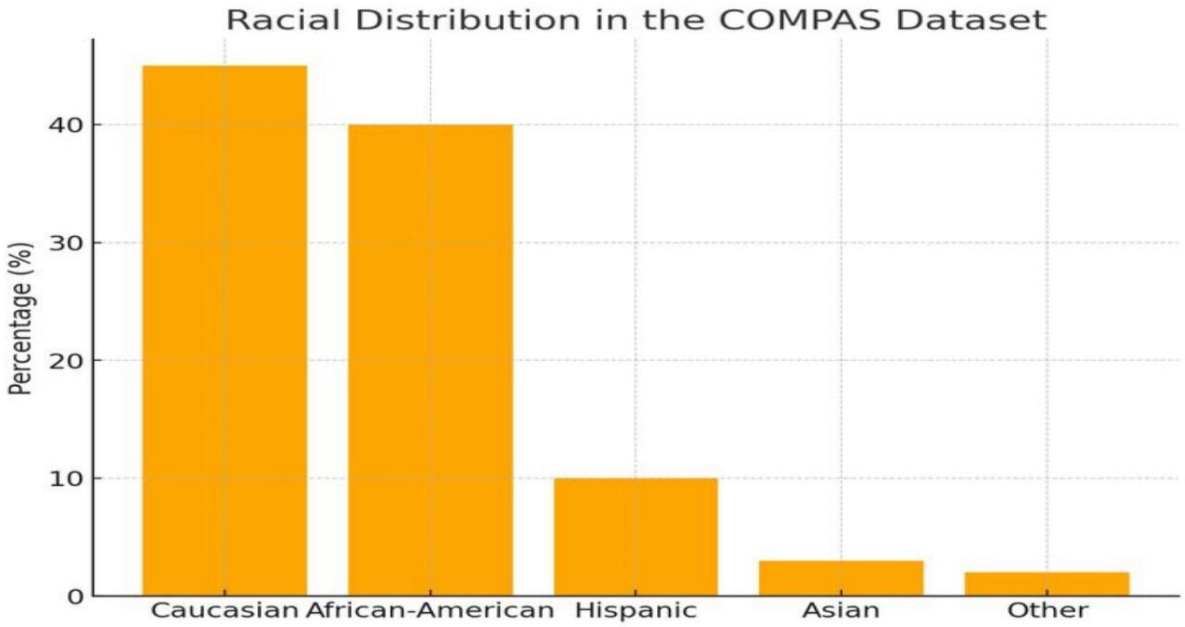

Figure 2:Racial Distribution in the COMPAS Dataset

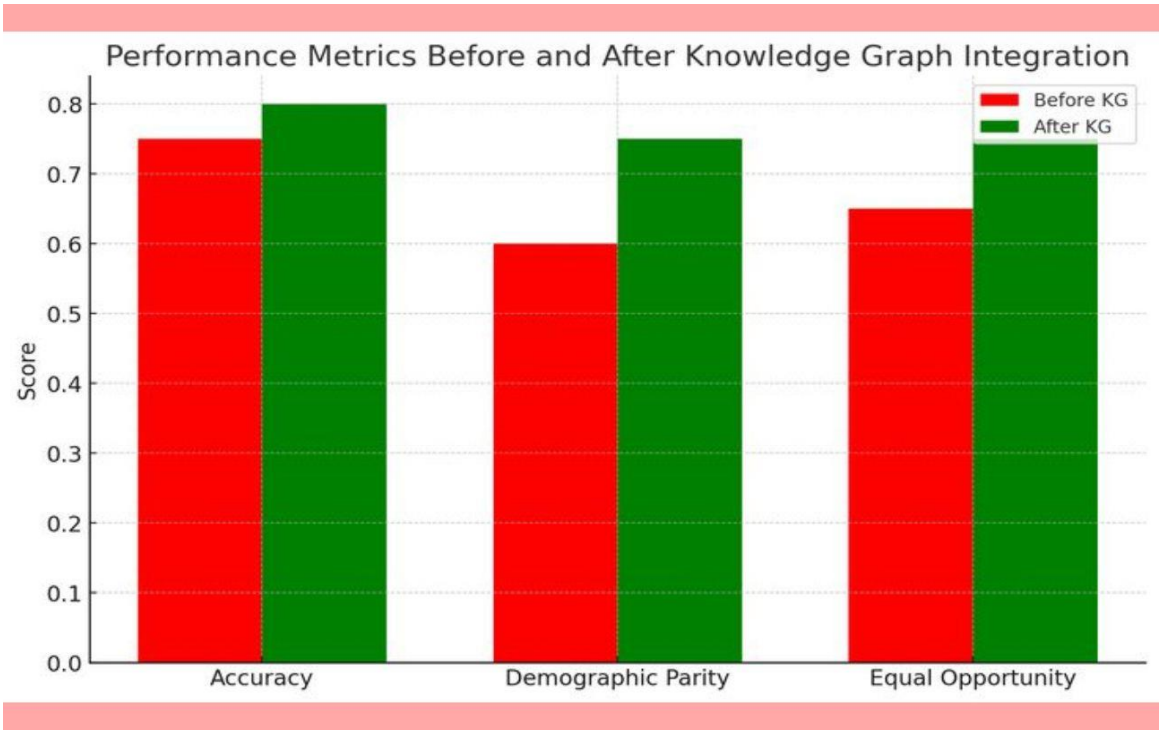

Figure 3:Performance Metrics Before and After Knowledge Graph Integration

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. RESULTS ANALYSIS

Finally, we will analyze the results of our experiments, comparing the performance of the bias-mitigated model against baseline models. This analysis will provide insights into the effectiveness of our KGAT approach and highlight areas for further improvement

VI. CONCLUSION

By employing a comprehensive methodology that integrates knowledge graphs into the training of Large Language Models, this research aims to significantly reduce bias and enhance the fairness of AI systems. Our findings will contribute to the ongoing discourse on ethical AI and provide practical solutions for developers and researchers working with LLMs

REFERENCES

1. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 149-158). ACM.
   Link: https://dl.acm.org/doi/10.1145/3287560.3287598
2. Dev, S., & Phillips, J. (2020). A Survey of Bias in Machine Learning through the Lens of the Law. *ACM Computing Surveys*, 53(6), 1-35
   Link: https://dl.acm.org/doi/10.1145/3397271
3. Gururangan, S., Marasović, A., Swayamdipta, S., et al. (2020). Don't Take the Easy Way Out: Ensemble Baselines for Natural Language Processing. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1891-1900).
   Link: https://aclanthology.org/2020.acl-main.171.pdf
4. Zhao, J., Wang, T., Yatskar, M., et al. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2979-2989).
   Link: https://aclanthology.org/D17-1302.pdf
5. Binns, R., & Van Kleek, M. (2020). The Role of Knowledge Graphs in Fairness and Bias Mitigation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 1-15).
   Link: https://dl.acm.org/doi/10.1145/3351095.3372830
6. Bertolini, M., & Gatti, L. (2021). Knowledge Graphs for Bias Mitigation in Natural Language Processing. *Journal of Artificial Intelligence Research*, 70, 1-30.
7. Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters*.
   Link: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/
8. Zhang, B., Lemoine, B., Mitchell, M., et al. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 335-344).
   Link: https://dl.acm.org/doi/10.1145/3287560.3287598
9. Wang, T., & Sennrich, R. (2020). Knowledge Graphs for Natural Language Processing: A Survey. *ACM Computing Surveys*, 53(6), 1-36.
   Link: https://dl.acm.org/doi/10.1145/3397271
10. Mitchell, M., et al. (2019). Model Cards for Model Reporting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (pp. 220-229).
    Link: https://dl.acm.org/doi/10.1145/3287560.3287596

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com