Exploring Large Language Models for Bias Mitigation and Fairness

Ibrahim MOHAMED SEROUIS[1] , **Florence SÈDES**[2]
[1]IRIT, Université Toulouse 3, Toulouse, France
[2]IRIT, Université Toulouse 3, Toulouse, France
{ibrahim.mohamed-serouis, florence.sedes}@irit.fr

Abstract

With the increasing integration of Artificial Intelligence (AI) in various applications, concerns about fairness and bias have become paramount. While numerous strategies have been proposed to mitigate bias, there is a significant gap in the literature regarding the use of Large Language Models (LLMs) in these techniques. This paper aims to bridge this gap by presenting an innovative approache that incorporates LLMs for bias mitigation and fairness in AI systems. Our proposed method, built on previous research, is designed to be model and system-agnostic, while keeping humans in the loop. We envision these approaches to foster trust between AI developers and end-users/stakeholders, contributing to the discourse on responsible AI.

1 Introduction

Artificial Intelligence (AI) has become an integral part of our daily lives, powering various applications from virtual assistants to recommendation systems. However, as AI systems increasingly interact with and make decisions about people, concerns about fairness and bias have come to the forefront. These systems, often trained on large datasets with not always clear guidelines for ensuring data quality during data collection, can inadvertently learn and perpetuate the biases present in these datasets, leading to unfair outcomes. One well-known example is the COMPAS system used in the United States criminal justice system, which predicts the likelihood of a defendant reoffending. A study by ProPublica found that the system was biased against African-American defendants, as they were more likely to be labeled as high-risk even if they had no prior convictions.

In response to these challenges, a part of the AI research community has been exploring various strategies to mitigate bias and ensure fairness. A plethora of methods have emerged, some involving reprocessing the data used for training such systems, others involving selecting the best models according to predefined fairness metrics, or others involving providing more transparency into the decision-making process. While LLMs, one of the most prominent breakthroughs in AI, have demonstrated effectiveness in a wide range of tasks, there exists a notable gap in the literature concerning studies advocating for their inclusion in existing techniques for bias mitigation. Furthermore, existing research has largely focused on specific dataset modalities (visual or textual), particular type of fairness, or the adaptation of optimization problems for specific AI models, with limited efforts towards establishing standards or frameworks that can be generalized to diverse problems, models, and modalities.

In this paper, we present and discuss an innovative approach that incorporate LLMs for mitigating bias and ensuring fairness in AI systems while keeping humans in the loop, and metrics to quantify its effectiveness. Our proposed approach, building on previous iterations in the literature, aims to be model, modality, and system-agnostic. The proposed metrics offer a unified measure to assess both the quality of the training data and the bias in AI models during decision-making. In its current form, our method has been evaluated using a task on a human-centric situations dataset, yielding promising outcomes and highlighting its model-agnostic capabilities. We foresee the application of our technique to further foster trust between AI system developers and end-users or stakeholders.

Paper structure. In Section 2, we present and discuss an overview of prior research, highlighting the gap for studies leveraging LLMs for mitigating bias. In Section 3, we introduce key concepts necessary for understanding the types of bias and fairness we refer to, and how the two concepts overlap and where they differ. In Section 4, we present and discuss strategies for mitigating bias using LLMs while highlighting their relation to types of bias they aim to solve, and the communication strategy that needs to be implemented for our approach to work, in Section 5. In Section 6, we present the specifics and results of an experimental application of our proposed methodology on a sample dataset. Finally, in Section 7, we engage in a discussion involving the challenges in leveraging LLMs for mitigating bias, and the limitations of our study.

2 Related work

2.1 FAIR principles

The FAIR principles, which emphasize the importance of making data Findable, Accessible, Interoperable, and

Reusable, have become a cornerstone for data management and stewardship in scientific research. These principles were first introduced by Wilkinson et al. in 2016 and have since been widely adopted across various disciplines to enhance data sharing and reuse [Wilkinson *et al.*, 2016]. The adoption of FAIR principles has been particularly significant in the life sciences, where large volumes of complex data are generated. The European Open Science Cloud (EOSC) is an example of a major project aimed at implementing FAIR data practices across Europe, promoting open science and facilitating access to research data [Collins *et al.*, 2018]. However, their implementation can be challenging and requires a collaborative effort from all stakeholders, as highlighted by [Collins *et al.*, 2018].

2.2 Critiques of data collection methods in Machine Learning

Early efforts in Machine Learning (ML) primarily focused on amassing vast amounts of data and optimizing the computing resources. However, recent years have witnessed a growing scrutiny of prevailing data collection practices within the research domain. Notable studies such as [Paullada *et al.*, 2021] and [Polyzotis *et al.*, 2019] have shed light on deficiencies inherent in current methodologies. Moreover, [Raji, 2020], advocating from the perspective of a black woman, offered a critical review, underscoring the imperative for heightened fairness within the current data collection practices. This sentiment aligns with the findings of [Paullada *et al.*, 2021], reiterated a year later, emphasizing the pressing need for equitable representation in research endeavors.

2.3 Bias and fairness mitigation studies

The increasing focus on equity and fairness in artificial intelligence (AI) has led to the development of numerous applications and methodologies aimed at addressing bias in machine learning (ML) models. [Dwork *et al.*, 2012] were among the pioneers in this field, providing a mathematical framework to define and measure fairness and bias in ML models. Their work laid the foundation for subsequent research by offering a structured approach to understand and mitigate bias. Building on this foundation, [Zemel *et al.*, 2013] introduced a method for learning data representations that ensure fairness with respect to sensitive attributes such as race and gender. Their approach involved transforming original data into a new representation space where the predictability of sensitive attributes is minimized while maintaining task-relevant information. Few years later, [Feldman *et al.*, 2015] introduced Disparate Impact, a metric aimed to quantify group fairness in AI decisions, and proposed an iterative method using convex optimization to find the optimal transformation that minimizes the disparate impact, while maintaining the accuracy of the trained model.

In the years that followed, researchers have explored various strategies to integrate fairness constraints and develop bias reduction techniques. For example, [Zafar *et al.*, 2017] proposed methods to incorporate fairness constraints directly into the optimization process of ML models. These efforts have been complemented by practical applications in society, such as the work by [Jang *et al.*, 2019], who developed a method to quantify biases in visual datasets with a focus on gender representation in commercials. This was further advanced by [Wang *et al.*, 2022] with the introduction of *REVISE*, a versatile tool designed to measure and mitigate bias in visual datasets, highlighting the ongoing commitment to creating fairer AI systems.

However, the aforementioned bias mitigation techniques have primarily concentrated on specific dataset modalities (e.g., visual or textual), a particular type of fairness, modifying the input data representation space, or adjusting the optimization constraints of AI models to achieve fairer representations. The incorporation of LLMs, one of the most groundbreaking innovations in AI, remains relatively underexplored in bias mitigation studies, as we can observe in recent surveys covering the strategies for mitigating bias in AI systems [Ferrara, 2024]. By addressing this limitation and exploring new avenues for fairness in AI, the field can continue to evolve and develop more robust, fair, and ethical AI systems.

2.4 Large Language Models

At present, LLMs stand as one of the most prominent breakthroughs in AI. There are many variations of LLMs in the market such as the GPT versions (GPT 1[Radford *et al.*, 2018], GPT2 [Radford *et al.*, 2019], GPT3 [Raji, 2020]), XLNET [Yang *et al.*, 2019], Llama from Meta [Touvron *et al.*, 2023], or the Gemma from Google [Team *et al.*, 2024], with Transformers [Vaswani *et al.*, 2017] and BERT [Devlin *et al.*, 2019] and being pioneers of these advancements.

Beyond their prowess in natural language processing tasks, LLMs have showcased their versatility across domains such as education, dataset generation, healthcare [Sallam, 2023], and scientific research [Jungherr, 2023]. However, despite their widespread adoption and multifaceted applications, there exist a notable lack of studies that leverage this innovation for reducing bias and ensuring fairness in AI systems. It is within this gap that our study positions itself, advocating for the integration of LLMs into practices aimed at mitigating bias and promoting fairness in AI systems.

3 Bias and fairness in the literature

In this section, we delineate key concepts: bias (Section 3.1), fairness (Section 3.2), and elucidate the difference between both concepts (Section 3.3). It serves as an entry point for presenting our solution, as subsequent sections draw extensively upon the provided definitions.

3.1 Bias in the literature

As defined by [Ferrara, 2024], bias in the literature are defined as the systematic errors that occur in decision-making processes, leading to unfair outcomes. Distinct forms of bias include algorithm bias, arising from algorithm design and implementation favoring specific attributes, confirmation bias, occurring as systems reinforce pre-existing biases, sampling bias, occurring when training data inadequately represent served populations, and measurement bias, emerging from data collection or measurement systematically over or under-representing certain groups. Recently, with the advent

of Generative AI [Feuerriegel *et al.*, 2024], generative bias has emerged [Ferrara, 2024], wherein generative model outputs disproportionately reflect specific attributes or patterns within training data, like an image generation model consistently producing male images when generating CEO images.

3.2 Fairness in the literature

Often confused with biases, fairness is a concept that focuses on the ethical and equitable treatment of individuals or groups within automated decision-making systems [Ferrara, 2024]. It involves ensuring that the outcomes and processes of AI systems do not discriminate against or disadvantage any particular demographic group based on characteristics such as race, gender, age, or socioeconomic status.

The main types of fairness studied in the literature are group fairness, that ensures that groups are treated equally or proportionally in AI systems, individual fairness, ensuring that similar individuals are treated similarly, procedural fairness, that ensures the process used to make decisions is fair and transparent, and, more recently, counterfactual fairness, aiming to ensure that the system would make the same decision for an individual, regardless of its demographic group.

3.3 Difference between bias and fairness

While interconnected, since bias introduces unfairness into AI systems, bias and fairness are two completely distinct concepts. While bias represents the presence of unfairness within AI systems, fairness encompasses the principles and practices aimed at mitigating bias and ensuring ethical and equitable outcomes for all stakeholders. Fairness emphasizes the need for transparency, accountability, and non-discrimination in decision-making processes. Recognizing and addressing biases are essential steps towards achieving fairness in automated decision-making processes.

4 Mitigating bias using LLMs

LLMs offer unique capabilities that enable the generation of coherent data and enhance transparency and explainability, crucial factors in addressing bias and ensuring fairness. In this section, we delve into our approach harnessing the power of LLMs to mitigate bias in AI systems, consisting of two independent modules: the data reprocessing module, detailed in Section 4.1, and the explanation module, delineated in Section 4.2. The complete architecture of our framework is illustrated in Figure 1.

4.1 LLMs for data (re)processing

Various techniques exist to mitigate sampling bias and measurement bias in the literature, often involving data pre or post-processing. These techniques may include resampling classes by generating synthetic yet coherent examples, or augmenting existing data with modified versions of coherent examples. They contribute in parallel to ensure group, individual, and counterfactual fairness by preventing over and under-representation, and providing alternative scenarios for training. Traditionally, addressing these biases demands extensive data observation and exploration of potential outcomes across different groups, such as demographic groups. Resampling with alternative versions and controlling the data quality of each row can be laborious and time-consuming. However, LLMs are recognized for their ability to generate coherent data, whether in the form of counterfactual versions [Chen *et al.*, 2023] or modified versions based on examples, sometimes surpassing traditional data augmentation methods [Yoo *et al.*, 2021]. We propose an iterative approach that involves human oversight while leveraging the power of LLMs.

In our approach, before the training phase, a dataset report is submitted to an LLM, and an expert determines the subsequent action: either issuing a prompt (command) to an LLM for generating new training data or initiating the training process. If a prompt is sent to an LLM to generate an alternative version of the training data, there is iterative feedback between the expert and the LLM until predefined conditions are met and a decision to proceed with model training is reached. Regardless of the chosen model, training and experiments are conducted, and detailed insights regarding the outcomes of the protected groups are provided either to an LLM for summarization (especially in cases involving multiple groups and subgroups) or directly to the expert. This is an automated process that can be achieved programatically without the input of a LLM. Finally, armed with comprehensive reports on input data and outcomes concerning protected groups, the expert can then opt to generate an alternative dataset version using an LLM to mitigate potential biases.

The expert can assess the quality of the data in terms of representation by computing Demographic Parity, that measures whether the generated data has equal representation of different demographic groups, in terms of uniqueness by computing the duplication rate, that measures the percentage of duplicates within generated data, precision, that proportion of data points that are relevant or important for a specific task or analysis, and in terms of consistency and timeliness by providing a rating on random batches of data. The potential summarization of the outcomes can be evaluated by the expert through a rating of accuracy of explanations. The fairness of the predictions of the model can be assessed using fairness metrics such as demographic parity, which measures the probability of a positive outcome being the same for all groups, regardless of their sensitive attributes (e.g., race, gender), equal opportunity, which measures the true positive rate (TPR) (proportion of actual positive instances that are correctly identified as positive by the model) being the same for all groups, and equalized odds, which requires both the true positive rate and the false positive rate (FPR) (proportion of actual negative instances that are incorrectly identified as positive by the model) to be the same for all groups.

In our proposed methodology, the calculation of the dataset's quality, denoted as DQ, is performed as follows:

$$DQ = \frac{1}{n}(\sum_{i=1}^{n} DQm_i * \alpha_i) + \frac{1}{m}(\sum_{j=1}^{m} DQr_j * \beta_j) \qquad (1)$$

Irrespective of the selected data quality metrics and ratings, we determine the sum of the data quality metrics (Dqm) multiplied by their corresponding normalization factor (α) which scales the values between 0 and 1, considering that different metrics may have varying scales. Subsequently, we add this

outcome with the sum of the data quality ratings multiplied by their respective normalization factor (β) for the same rationale.

The total bias B in the proposed approach is assessed as:

$$B = \frac{1}{N} * \sum_{i=1}^{N} (bfm_i * \alpha_i) \tag{2}$$

Regardless of the chosen bias metrics, we establish the sum of the bias metrics (bfm) multiplied by their corresponding normalization factor (α). The overall effectiveness, E, of the proposed approach can be determined through the delta (difference) between the data quality of the generated dataset, DQ_{gen} and the data quality metric of the original dataset, DQ_{orig}, subtracted by the delta between the bias metric B of the generated dataset, B_{gen}, and the original dataset, B_{orig}. This can be formally expressed as:

$$E = (DQ_{gen} - DQ_{orig}) - (B_{gen} - B_{orig}) \tag{3}$$

Considering that the delta in the bias metric should be negative for a less biased dataset, the subtraction in this context is logical for enhancing the overall value of E, for indicating a superior effectiveness of the process. Our proposed metric can then assess both the quality of the generated data and the bias from the model in one unified metric.

4.2 LLMs for transparency, accountability, explainability

Transparency refers to the openness and clarity of AI systems regarding their operation, decision-making processes, and underlying algorithms. Explainability, on the other hand, specifically pertains to the ability of an AI system to provide understandable explanations for its decisions and actions. Both transparency and explainability are intricately contribute to ensuring procedural fairness (Section 3.2).

In many cases, ensuring transparency in AI systems involves incorporating explainability mechanisms that enable users to understand why and how certain decisions were made by the system. Various explainability techniques are available in the literature, including GRAD-CAM [Selvaraju *et al.*, 2017] mostly used for image data, SHAP [Lundberg and Lee, 2017] for tree-based models and neural networks, Integrated Gradients [Sundararajan *et al.*, 2017] and LIME [Ribeiro *et al.*, 2016] for image and text data, and GNN-Explainer for knowledge graphs [Ying *et al.*, 2019]. These methods produce diverse raw values, such as feature importance matrices or coefficients (SHAP, LIME), sub-graph and node importance scores (GNN Explainer), heatmaps (GRAD-CAM), and feature attribution vectors (Integrated Gradients). While an image juxtaposed to its heatmap, such as in the implementation of GRAD-CAM, can be fairly understood by a non-expert with a generic description that does not necessarily need the input of an LLM, explanations based on feature importance or attribution matrices can be challenging for non-image data to understand without expert interpretation. However, AI systems are not only destined to the experts, in most cases, but to the end-users and stakeholders.

For text-based tasks like Neural Machine Translation or Sequence Classification, leveraging LLMs can elucidate to the user that specific parts (one or many words) of the text contributed to a certain percentage of the translation or attributed class to the sentence. The LLM can take as input the feature attribution matrices, the context, and the task, generating a natural language explanation. Similarly, this approach can be applied to graph-based tasks, where LLMs can generate subgraphs indicating the nodes that contributed the most to the prediction of the model, along with a natural language explanation for their importance.

The model explanations undergo iterative refinement using a two-step feedback approach, involving randomly sampled batches of validation data. Initially, the explanations are fine-tuned on batches with an expert to ensure coherence between the raw values and the provided explanation. Subsequently, the refined explanation is presented to a non-expert, initiating a back-and-forth exchange between the expert and the non-expert to achieve a coherent and interpretable feedback from both parties. Finally, the LLM is fine-tuned with the new explanation generated based on the agreement between the expert and non-expert feedback. This process aims to produce explanations that are understandable by both novices and experts, thereby facilitating communication between stakeholders/users and scientists.

For evaluation, we propose a metric named **IFP** (**I**terative **F**eedback **I**mprovement), which would compare the distance between the ratings of clarity, understandability, trust, and usefulness provided by the non-expert for the initial explanations and for the explanations after the rounds of feedback. This can be formally expressed as:

$$IFP = \sum_{i=1}^{n} \alpha_i * (R_{gen(i)} - Ri_{orig(i)}) \tag{4}$$

Where α_i is the normalization coefficient for the rating, $R_{gen(i)}$ the ratings for the generated dataset, and $Ri_{orig(i)}$ the ratings for the generated dataset. A higher, positive distance would indicate significant improvement, a nil distance would signify no improvement, and a negative distance would suggest a decrease in the quality of explanation. However, if the initial ratings were already high enough, the absence of improvement might be considered normal. We can also anticipate such systems not improving significantly after a standard for explanations has been reached.

It is noteworthy to highlight that leveraging these quantifiable metrics, the potential bias introduced by the expert or the LLM are included but not segregated, when evaluating the effectiveness of the process in terms of E or IFP. We further discuss this aspect of our approach in Section 7.3.

5 Ensuring an optimized communication

Effective communication is pivotal in both methodologies, whether between experts and non-experts or between experts and LLMs. In the latter methodology, where experts engage with non-experts, communication revolves around non-formal concepts that are subject to interpretation. Consequently, these dialogues can potentially extend beyond necessary durations, diminishing process efficiency. To streamline communication, it is imperative to simplify guidelines and

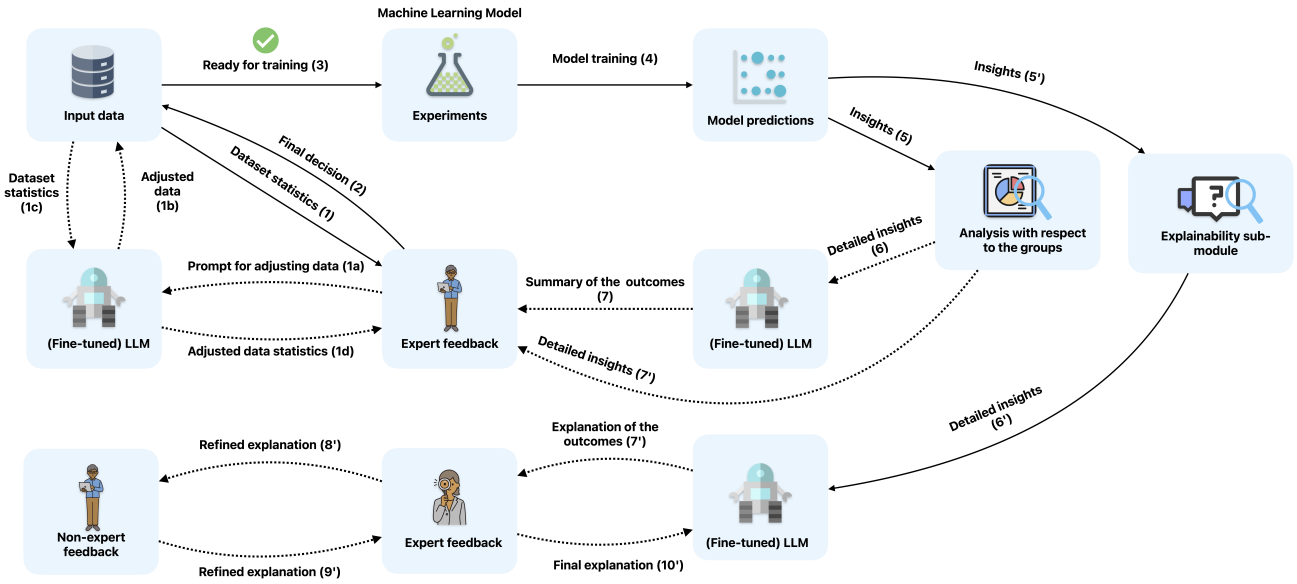

Figure 1: Integrating LLMs into bias mitigation practices.

evaluation metrics for non-experts while ensuring experts adhere to structured interrogation formats. Moreover, expertise in prompt engineering (practice of crafting specific instructions or queries given to language models, particularly LLMs, to elicit desired responses) is essential for experts to minimize unnecessary iterations with LLMs. This ensures that the communication process remains efficient and conducive to achieving desired outcomes.

Leveraging the aforementioned metrics, we can prompt the expert in both methodologies to offer a final explanation or dataset once a sufficient threshold has been met, either in terms of Effectiveness (as defined in Section 4.1), or IFP (Section 4.2). A key challenge lies in determining a specific threshold, as it would be highly domain-sensitive. Certain applications, such as hiring, education, or criminal justice, would derive greater benefits from fairness metrics than others, potentially necessitating a higher threshold in their context, while others may suffice with a lower one. Establishing a threshold or a set of thresholds based on domain-specific insights would be the responsibility of the expert and the potential AI governance team, with potential refinement based on user feedback. Failure to address this carefully could render even the most advanced LLMs in the literature of limited real-world benefit, only providing a higher computational cost rather than a safer, more trustworthy system.

6 Experiment on a sample dataset

6.1 Dataset description, task definition, and model selection

Our proposed approach was evaluated on the ObjectifyGaze dataset [Tores *et al.*, 2024]. The objective of the task is to predict whether a scene contains objectification towards a character or not, where objectification is defined as portraying a character as a mere object through dialogues, narratives, or camera angles that focus on specific body parts or physical attributes. The dataset comprises annotated scene graphs that include scenes, places, contexts, characters along with their attributes, and the objectification labels. Due to the human-centric nature of the task, fairness is of paramount importance.

We trained a Graph Convolutional Network (GCN) [Kipf and Welling, 2017], a Vanilla Message Passing Neural Network (Vanilla MPNN) [Gilmer *et al.*, 2017] using Tensorflow-GNN [Ferludin *et al.*, 2022], and a Random Forests classifier [Breiman, 2001] on the aforementioned dataset for detecting objectification within scenes. To evaluate the performance of our model, we computed classic metrics for binary classification, including accuracy, F1-score, and ROC AUC score. Our model achieved decent efficiency in terms of these metrics, although out of the scope of this paper.

6.2 Protected attributes and fairness metrics

Protected attributes such as gender, race, and social characteristics are commonly considered in fairness evaluations. Although our model performed well in terms of standard binary classification metrics, we observed that it leaned towards "not objectifying" for scenes with non-caucasian characters, which account for only 0.08% of the dataset. To address this issue and avoid creating biases towards specific groups, we applied our methodology for evaluating and mitigating bias in our system. Specifically, we focused on the data (re)processing part due to insufficient resources for fine-tuning a large language model (LLM) for explanations.

We used Demographic parity difference (DP), Equalized odds difference (EOD), and Equal Opportunity Difference (EOP) as the bias and fairness metrics in our approach, with a

normalization value (alpha) of 1 for all the metrics since they possess the same scale (between 0 and 1). Their calculation methods are available in Appendix A.

We evaluated the quality of data using Uniqueness, that measures the degree to which data values are distinct and not duplicated, and Completeness, that measures the degree to which data values are uniform and free from contradictions. Other metrics such as Timeliness could have been used in a collaboration with domain experts, which we could not accomplish at the time of the experiments. Mistral [Jiang *et al.*, 2023] with 7 billion parameters was used as the LLM in our experiment, and the fairness metrics were computed using Fairlearn [Weerts *et al.*, 2023].

6.3 Instructions provided to the LLM

We provided Mistral with three essential pieces of information for generating counterfactual data:

- The ethnicity classes present in our current dataset
- The number of instances to generate per minority class (in our case, the non-caucasian examples): it is essential for maintaining a certain balance within the classes
- The precise data format to generate, along with a sample for reference

The generated data were subsequently combined with the original data, and the same set of prior experiments were conducted.

6.4 Results

We explored three approaches: oversampling by duplicating examples from the minority classes (1), generating counterfactual examples using a LLM (2) to alter the ethnicity attribute of characters and some scene details such as the identities of the characters, and combining the LLM-generated data with the duplicated data (3). The results in Table 1 show that involving LLMs for counterfactual data generation improved the fairness metrics, for the model that had the best results on the dataset (GCN).

Simply duplicating examples from the minorities did not improve the fairness metrics, although a slight increase in raw accuracy was observed. Our method showed a higher efficiency (E) than the other methods, as well as better results in terms of F1-score, although out of the scope of this paper. The decrease in data quality when duplicating examples is due to a decrease in the Uniqueness of data within the dataset. The Vanilla MPNN model and Decision Tree classifier exhibited similar trends, with all metrics showing significant improvement when using our proposed approach.

To evaluate the adaptability of our approach with different metrics, we incorporated Disparate Impact in our bias evaluation metrics, which measures the difference in selection rate between two groups defined by a sensitive attribute, such as race or gender [Feldman *et al.*, 2015], and used different combinations of metrics. All our metrics were still computable, and the results showed a similar trend, with B and E improving with our approach..

These findings demonstrate the versatility of our calculation method in adapting to different scales and fairness metrics, and its independence from the underlying model, opening up new possibilities for quantifying bias in model outcomes. It is noteworthy that our approach enables the assessment of both data quality and outcome bias with a single metric E, or separately with B or DQ.

However, our approach is more readily applicable to tabular data, textual data, and attributed graphs datasets, while visual datasets require more computing resources for generating sets of images. Nevertheless, with the emergence of smaller visual LLMs, there is potential for applying this technique to visual data, particularly for generating counterfactual images. Our aim is to apply our methodology on a larger scale once sufficient resources become available.

Data	B	DQ	E
Baseline	0.10	0.80	-
O + Duplicated examples	0.14	0.76	-0.08
O + LLM + Duplicated examples	0.06	0.84	0
O + LLM	**0.03**	**0.90**	**0.17**

Table 1: Results on the ObjectifyGaze dataset for the GCN model. (Baseline=O=Original dataset, B=total bias, E=Effectiveness, as defined in Section 4)

7 Discussion

In this section, we discuss about the use of traditional methods compared to LLMs, the challenges in leveraging LLMs for mitigating bias, and the shortcomings of our study.

7.1 Can traditional methods be more effective?

While LLMs offer a promising avenue for bias mitigation, they may not always surpass traditional methods in terms of data quality or efficiency. In scenarios requiring explanation generation, the manual crafting of explanations can be laborious, particularly considering the diverse nature of input data. In such cases, LLMs may hold an advantage due to their ability to automate the generation process. However, studies such as [Sobieszek and Price, 2022] and more recent work by [Kandpal *et al.*, 2023] have highlighted instances where LLMs struggle to capture context effectively, exhibiting limitations in originality and emotional nuance compared to human authors. Consequently, for tasks demanding high-level interpretative contexts and emotional intelligence, human input may yield higher quality synthetic and counterfactual data. Nonetheless, LLMs excel in generating large volumes of data at a rapid pace, surpassing human capabilities in terms of sheer output. In such scenarios, human involvement becomes essential for verifying the quality of the generated data. Thus, our proposed framework maintains human involvement alongside LLM utilization, recognizing the complementary strengths of both approaches and emphasizing the necessity of regular auditing for successful integration.

7.2 Challenges in leveraging LLMs for fairness

While LLMs can be foreseen as a promising direction towards fairer AI systems, certain characteristics hinder their widespread adoption. Firstly, defining appropriate metrics

for assessing the potential bias generated by the LLMs used in our approaches, especially the potential introduced algorithmic bias, poses a significant challenge. Traditional fairness metrics may not be directly applicable to language generation tasks, necessitating the development of novel evaluation methods tailored to LLMs. Moreover, the computing resources required for training and running inferences on LLMs are intensive. Small organizations and researchers with limited access to compute resources may face challenges in leveraging LLMs effectively for bias mitigation.

However, advancements in LLM quantization techniques, as demonstrated by Yao et al. [Yao *et al.*, 2024], offer a glimpse of a future where more researchers can access LLM inference and training capabilities. This democratization of LLM technologies has the potential to catalyze innovation and accelerate progress towards fairer AI systems in research.

7.3 Limitations of our study

While we attempted to cover the most prominent types of bias and fairness in the literature, we did not exhaustively cover every conceivable form of bias. For instance, some surveys on bias and fairness have highlighted interaction bias, wherein AI systems interact with humans in a biased manner based on a protected attribute (*e.g* race, gender), or causal fairness, which entails ensuring that the system does not perpetuate historical biases and inequalities. However, as elucidated in studies such as [Ferrara, 2024], individual fairness and causal fairness are significantly interconnected. Moreover, for interaction and algorithmic bias, we could not think of solutions for mitigating involving LLMs. From our perspective, traditional model selection techniques hold an advantage over LLMs in addressing algorithm bias, and concealing protected attributes from the system before querying appears to provide a superior approach for mitigating interaction bias.

Furthermore, to the best of our knowledge, there exists only few studies addressing the efficiency of LLMs in counterfactual or alternative data generation for specific topics and tasks. Our current evaluation metrics for our methodology do not take into account the inherent capabilities of the chosen LLM.

8 Conclusion

In this paper, we have delved into the critical intersection of artificial intelligence (AI), bias mitigation, and fairness. Our discourse has underscored the nuanced relationship between different types of bias and the corresponding strategies for mitigation. Through a thorough examination of existing literature and methodologies, we have identified a significant void pertaining to the integration of LLMs in bias mitigation strategies.

To bridge this gap, we have proposed innovative model and system-agnostic approaches that harness the power of LLMs to tackle biases across diverse AI applications. While our methodologies offer promising avenues for bias mitigation, it is essential to acknowledge its limitations, such as the exclusion of certain types of bias like interaction bias.

Looking ahead, an intriguing area for future research lies in the development of evaluation strategies for our framework that would include the efficacy of LLMs in data generation and explainability within the domain of application. As AI continues to evolve, it becomes increasingly imperative to prioritize fairness and inclusivity, ensuring that AI systems uphold ethical standards and contribute to the collective welfare. Through collaborative endeavors and interdisciplinary collaboration, we can collectively advance towards a more equitable and just AI-powered future.

Ethical Statement

As our methodology progresses, we anticipate its broad application across various AI and Machine Learning domains. Yet, the inclusion of human oversight in refining explanations or data generation processes introduces the potential for unintended influence from individuals or groups whose beliefs and intentions are unknown to us. As we could observe with Tay, the Microsoft chatbot unveiled on Twitter that started exhibiting racist, violent, and disrespectful behavior due to deliberate manipulation by users, giving so much power to a human or a single research group can be detrimental.

We assert that regular, independent, and external audits of these systems by specialized organizations in AI ethics are indispensable for their responsible evolution. Without diligent monitoring of the system outputs, the unchecked deployment of these technologies could exacerbate existing biases, depending on their intended applications. Therefore, we advocate for stringent oversight measures to safeguard against the potential misuse and ensure the ethical integrity of AI systems, especially on a method that relies heavily on prompt engineering.

Acknowledgements

This work has been supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38- 00012-01.

A Fairness metrics calculation

A.1 Demographic parity difference (DP)

$$DP = P(\hat{Y} = 1|A = a) - P(\hat{Y} = 1|A = b) \tag{5}$$

Where $\hat{Y}$ is the predicted outcome, A is the sensitive attribute, and a and b are two different values of the attribute.

A.2 Equalized odds difference (EOD)

$$EOD = |(TPR_a - FPR_a) - (TPR_b - FPR_b)| \tag{6}$$

Where TPR_a and FPR_a are the true positive rate and false positive rate for the privileged group, respectively, and TPR_b and $FPR_b$are the true positive rate and false positive rate for the unprivileged group, respectively.

A.3 Equal opportunity difference (EOP)

$$EOP = |TPR_a - TPR_b| \tag{7}$$

Where TPR_a and TPR_b are the true positive rates for the privileged and unprivileged groups, respectively.

References

[Breiman, 2001] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[Chen *et al.*, 2023] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. DISCO: Distilling Counterfactuals with Large Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Collins *et al.*, 2018] Sandra Collins, Françoise Genova, Natalie Harrower, Simon Hodson, Sarah Jones, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaitė, and Peter Wittenburg. Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. 2018. Publisher: Luxembourg: Publications Office of the European Union.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. Association for Computing Machinery. event-place: Sydney, NSW, Australia.

[Ferludin *et al.*, 2022] Oleksandr Ferludin, Arno Eigenwillig, Martin Blais, Dustin Zelle, Jan Pfeifer, Alvaro Sanchez-Gonzalez, Wai Lok Sibon Li, Sami Abu-El-Haija, Peter Battaglia, Neslihan Bulut, Jonathan Halcrow, Filipe Miguel Gonçalves de Almeida, Pedro Gonnet, Liangze Jiang, Parth Kothari, Silvio Lattanzi, André Linhares, Brandon Mayer, Vahab Mirrokni, John Palowitch, Mihir Paradkar, Jennifer She, Anton Tsitsulin, Kevin Villela, Lisa Wang, David Wong, and Bryan Perozzi. TF-GNN: Graph Neural Networks in TensorFlow, 2022. arXiv:2207.03522 [physics, stat].

[Ferrara, 2024] Emilio Ferrara. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), 2024.

[Feuerriegel *et al.*, 2024] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, February 2024.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017. ISSN: 2640-3498.

[Jang *et al.*, 2019] Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. _eprint: 2310.06825.

[Jungherr, 2023] Andreas Jungherr. Using ChatGPT and other large language model (LLM) applications for academic paper assignments. 2023. Publisher: Otto-Friedrich-Universität.

[Kandpal *et al.*, 2023] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Paullada *et al.*, 2021] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[Polyzotis *et al.*, 2019] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. Data validation for machine learning. *Proceedings of machine learning and systems*, 1:334–347, 2019.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and others. Improving language understanding by generative pre-training. 2018. Publisher: OpenAI.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raji, 2020] Inioluwa Deborah Raji. Handle with Care: Lessons for Data Science from Black Female Scholars. *Patterns*, 1(8):100150, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Sallam, 2023] Malik Sallam. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI, 2023. Issue: 6.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[Sobieszek and Price, 2022] Adam Sobieszek and Tadeusz Price. Playing games with AIs: the limits of GPT-3 and similar large language models. *Minds and Machines*, 32(2):341–364, 2022. Publisher: Springer.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Team *et al.*, 2024] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and others. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[Tores *et al.*, 2024] Julie Tores, Lucile Sassatelli, Hui-Yin Wu, Clement Bergman, Lea Andolfi, Victor Ecrement, Frederic Precioso, Thierry Devars, Magali Guaresi, Virginie Julliard, and Sarah Lecossais. Visual Objectification in Films: Towards a New AI Task for Video Interpretation, 2024. _eprint: 2401.13296.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and others. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2022] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision*, 130(7):1790–1810, July 2022.

[Weerts *et al.*, 2023] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems, 2023. _eprint: 2303.16626.

[Wilkinson *et al.*, 2016] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and others. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. Publisher: Nature Publishing Group.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[Yao *et al.*, 2024] Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19377–19385, 2024. Issue: 17.

[Ying *et al.*, 2019] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[Yoo *et al.*, 2021] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.