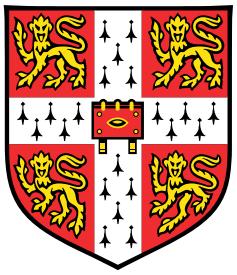
Mitigating and Assessing Bias and Fairness in Large Language Model-Generated Synthetic Tabular Data



Faria Zarin Subah

Supervisor: Prof. Mihaela van der Schaar

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy in Machine Learning and Machine Intelligence

Newnham College October 2024

I would like to dedicate this thesis to my loving parents, my brother, my inspiring teachers and Mahnaf, whose unwavering support made this journey possible. ...

Declaration

I, Faria Zarin Subah of Newnham College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described within are entirely my own, except where explicitly stated otherwise. This report does not contain material that has been previously submitted, in whole or in part, for any other degree or diploma. The word count, excluding declarations, bibliography, and images, but including tables, footnotes, figure captions, and appendices, is 14693.

All software used in this thesis was developed in Python. For all experimental analyses, the Synthcity[1] library was used to assess the quality (fidelity, diversity) of the synthetic data, and the Fairlearn[2] library was employed to evaluate its fairness. Scikit-learn[3] was utilized for training and testing the classification models. The GPT models used for evaluation were accessed via OpenAI[4][5].

Faria Zarin Subah
October 2024

[1] https://github.com/vanderschaarlab/synthcity
[2] https://github.com/fairlearn/fairlearn
[3] https://scikit-learn.org/stable/
[4] https://platform.openai.com/
[5] https://chat.openai.com/

Acknowledgements

This thesis would not have been possible without the unwavering support of several remarkable individuals. I am deeply grateful to my supervisor, Mihaela van der Schaar, for her invaluable guidance. A heartfelt thank you to Nabeel Seedat for always being there, helping me through the smallest hurdles. Your assistance was invaluable, and I am fortunate to have had your help during this journey.

A special note of thanks goes to John Dudley, my course director, whose exceptional guidance, mentoring and patience have profoundly influenced my academic development. John, your availability and encouragement made all the difference in navigating this challenging path. I deeply appreciate the hopeful and uplifting environment you fostered throughout this journey.

I would like to acknowledge my undergraduate thesis supervisor, Kaushik Deb, along with all my teachers for setting me on this path I find myself today. I am immensely grateful to the Prime Minister's Fellowship for their generous financial support, which made it possible for me to study at the University of Cambridge.

Above all, I would like to thank my family and friends from home for their unwavering support, even being miles apart. To my best friend, Mahnaf, your tremendous mental and emotional support has been my anchor, and I could not have come this far without you. Thank you for always believing in me and supporting my decisions.

Lastly, I want to thank my peers at Cambridge for their camaraderie and the cherished moments we shared, turning this chapter of my life into a truly remarkable journey.

Abstract

The rapid expansion of Large Language Models (LLMs) has opened new avenues for generating synthetic tabular data. However, these models often inherit societal biases from their training datasets, potentially leading to harmful outcomes for marginalized groups. This study examines whether LLMs, specifically GPT-4o, can generate fair synthetic tabular data through in-context learning (ICL) when guided by fairness constraints in prompts. Using the COMPAS dataset, the study is structured around three core research questions: the impact of data characteristics on synthetic data generation, the potential for bias mitigation through fairness-oriented prompting strategies, and the interaction of biases when LLM-generated synthetic data is used in downstream machine learning models.

We designed six fairness prompts and compared them with a general prompt lacking explicit fairness considerations. Our findings show that using few in-context samples (20-40) optimize the realism and fairness of synthetic data, with balanced sampling effectively reducing biases. However, mitigating biases through fairness-oriented prompts often leads to a trade-off with predictive accuracy, highlighting the challenges of aligning synthetic data with biased real-world benchmarks. This research contributes to the growing field of fair synthetic data generation by demonstrating that LLMs, when appropriately guided, can generate data that aligns with fairness goals, although challenges remain in maintaining both fairness and accuracy in downstream applications.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

> *"An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside."*
>
> — A.M. Turing, *Computing Machinery and Intelligence* (1950)

Alan Turing's observation highlights a fundamental challenge in the development of intelligent systems: the internal mechanisms of learning machines often remain opaque, even to their creators. This enigmatic nature is particularly relevant in the context of modern Large Language Models (LLMs), where their decision-making processes and inherent biases are difficult to fully comprehend or control. As these models become increasingly integrated into critical applications such as healthcare, education, finance, criminal justice, etc (Jungherr, 2023; Sallam, 2023), understanding and addressing the implications of their hidden workings, especially in terms of fairness, becomes crucial.

During the past few years, Large Language Models (LLMs) (Achiam et al., 2023; Brown et al., 2020; Ouyang et al., 2022; Team et al., 2023; Touvron et al., 2023) have witnessed a rapid expansion in their user base, drawing significant interest from both domain experts and the general public. Since the launch of ChatGPT (Achiam et al., 2023; Ouyang et al., 2022) by OpenAI in November 2022, these models have been widely utilized for various tasks including text generation, completion, summarization, translation, sentiment analysis, and conversational agents (Wolf, 2019). Notably, recent studies have leveraged LLMs for the tabular data classifications (Hegselmann et al., 2023; Liu et al., 2023b; Slack and Singh, 2023), where tabular information is transformed into natural language and presented to LLMs, along with a brief task description, to perform predictions. However, studies have revealed that LLMs may propagate societal biases inherent in the extensive datasets on which they were trained, potentially leading to adverse outcomes for marginalized groups (Abid et al., 2021; Askell et al., 2021; Basta et al., 2019; Ganguli et al., 2022a,b; Hutchinson et al.,

2020). With the growing adoption of LLMs across various sectors, understanding, addressing and mitigating these biases have become a critical concern. Although existing research has identified bias and unfairness in LLMs (Bi et al., 2023; Bordia and Bowman, 2019; Ferrara, 2023; Freiberger and Buchmann, 2024; Huang et al., 2023; Kotek et al., 2023; Nadeem et al., 2020; Zhang et al., 2023b; Zheng et al., 2023), to the best of our knowledge, no study has explored methods for achieving fairness in synthetic tabular data generation from a biased original dataset through in-context learning and effective prompting strategies in LLMs.

Synthetic data is artificially generated data, i.e., not obtained by direct measurement or collection from real-world events, but is created algorithmically to mimic the properties of real data (Wikipedia contributors, 2023). Currently, such data serves as a valuable tool for data augmentation and privacy preservation across various contexts (El Emam et al., 2020; Jordon et al., 2022). Recent research has highlighted the potential of utilizing LLMs (Li et al., 2023; Long et al., 2024) for generating synthetic data apart from statistical models (Fonseca and Bacao, 2023; Liu and Hsieh, 2019; Mooney, 1997; Raghunathan, 2021; Tang and He, 2015) and deep learning-based generative models (Goodfellow et al., 2020; Ho et al., 2020; Kingma, 2013)

Despite the widespread use of tabular data the fairness of LLM-generated synthetic data remain relatively unexplored (Borisov et al.; Grinsztajn et al., 2022). Current research on synthetic data generation has predominantly focused on the performance or utility aspects of the generated data and its impact on downstream model performance (Ghorbani and Zou, 2019; Just et al., 2023; Nohyun et al., 2022; Seedat et al., 2023) or on the potential of synthetic data in privacy protection (Wang et al., 2024; Wiest et al., 2024).

This largely overlooks a crucial aspect of reliable AI: assessing whether the synthetic data generated by LLMs are biased or unfair, and how such issues can be mitigated.

On the other hand, studies that address fairness issues primarily investigate LLMs as fair predictors, analyzing whether these models can perform fair classification on biased real tabular data (Chhikara et al., 2024; Liu et al., 2023b, 2024b), largely ignoring the challenge of fair synthetic data generation. Furthermore, most research on fair synthetic data generation has been limited to data modalities such as: text (Qureshi et al., 2024; Wang et al., 2023) and image (Li et al., 2023; Mishra et al., 2024), with relatively little attention given to tabular data which is arguably the most prevalent data type in both business and scientific contexts (Xu et al., 2024). Given these gaps, it is vital to thoroughly examine the fairness implications of using LLMs for generating synthetic tabular data.

In this research, our objective is to determine whether LLMs can grasp and implement the principles of fairness effectively. To achieve this, we focus on GPT-4o, a state-of-the-art variant of GPT-4 (Achiam et al., 2023), due to its advanced capabilities and superior alignment with human instructions (Islam and Moushi, 2024). We chose GPT-4o over other LLMs because it is designed with enhanced understanding and adherence to nuanced prompts (Shahriar et al., 2024), which is essential for enforcing specific fairness criteria. By conducting a rigorous study into how GPT-4o responds to prompts aimed at achieving fairness, we explore whether LLMs can effectively incorporate and apply these criteria when appropriately guided. Additionally, we assess whether biases in synthetic data generated by GPT-4o become more pronounced when utilized in downstream predictive tasks. Through this exploration, we aim to deepen our understanding of the fairness-related challenges associated with deploying LLMs like GPT-4o for synthetic data generation and their subsequent use in model predictions.

1.1 Research Questions

The rapid advancement of LLMs has opened new avenues for generating synthetic data (Deng et al., 2024; Hagos et al., 2024; Nazi and Peng, 2024; Nie et al., 2024; Raiaan et al., 2024; Ramos et al., 2024; Su et al., 2024). However, the deployment of LLMs in data synthesis raises significant concerns related to fairness and bias. Understanding and addressing these concerns is crucial for ensuring the ethical use of AI technologies (Jiao et al., 2024; Liyanage and Ranaweera, 2023; Serouis and Sèdes, 2024). This study aims to explore the extent to which LLM-generated synthetic tabular data are influenced by social biases and stereotypes. The following research questions (RQs) guide this exploration:

1. **Impact of Data:** How do the characteristics of data used in prompts influence LLM-driven synthetic data generation, specifically considering (i) the impact of the number of in-context samples provided, and (ii) the effects of the sampling method employed, such as random versus biased sampling?

2. **Bias mitigation:** Can biases in synthetic data be mitigated by using effective prompting strategies that incorporate fairness rules or constraints while maintaining the real data distribution and feature correlation intact? This question investigates whether LLMs can comprehend and implement fairness criteria when guided by such prompts.

3. **Interaction with a downstream model:** Do the biases present in synthetic tabular data exacerbate when classified using downstream machine learning models? This question

assesses the fairness-related challenges associated with deploying LLM-generated synthetic tabular data and utilizing it for downstream model prediction.

1.2 Contributions

The contributions of this thesis are as follow:

1. To our knowledge, this is the first study which investigates how various fairness criteria can be integrated into prompts to guide LLMs like GPT-4o in producing fair synthetic data.
2. We rigorously evaluate the generated data and analyze the amplification of biases when this synthetic data is utilized in downstream prediction tasks.
3. We thoroughly assess the accuracy-fairness tradeoff across various few-shot setups, carefully selecting in-context samples based on their number and sampling strategies.

1.3 Outline

This thesis is structured into five main chapters, each focusing on a specific aspect of our study.

- Chapter 1 introduces the research questions, contributions, and provides an overview of the thesis structure.
- Chapter 2 presents the background necessary to understand the context of the study. It begins with a discussion on Large Language Models (LLMs), exploring their emergent abilities, and then moves on to an overview of tabular data, discussing its defining characteristics, challenges, and the rationale for focusing on synthetic tabular data. It also reviews existing research on LLMs for tabular data generation, and finally, delves into bias and fairness in AI and the existing research of LLMs in addressing such fairness issues.
- Chapter 3 describes the methodology used in this research which sets the foundation for the experimental setup and subsequent analysis.
- Chapter 4 provides the evaluation of our experiments. It begins with the experimental setup, including details on the dataset, LLM used, prompt design, and classification models. The chapter then discusses the analysis of the COMPAS data, comparing

original and synthetic data across various prompt scenarios, and evaluating the fairness of the outputs using different fairness constraints.

- Chapter 5 concludes the thesis by summarizing key takeaways, discussing limitations, and highlighting areas for future work.

Chapter 2

Background

Large language models (LLMs), trained on extensive datasets, have demonstrated versatility beyond traditional NLP tasks (Fu et al., 2022). Recent research shows their emergent abilities, such as improved performance in few-shot learning (Wei et al., 2022a), sparking interest in their potential role in developing Artificial General Intelligence (Chang et al., 2024; Zhao et al., 2023b). LLMs are now seen as major AI breakthroughs, with notable models like GPT (Radford et al., 2019), XLNET (Yang, 2019), Llama (Touvron et al., 2023), and Gemini (Team et al., 2023), building on early advancements like Transformers (Vaswani, 2017)

ChatGPT, for example, showcases LLMs' ability to generate and comprehend human language (Liu et al., 2023c). Tabular data, essential in fields like finance, medicine, and education (Rundo et al., 2019; Sahakyan et al., 2021), has become a key focus, with researchers now exploring LLMs' potential in tasks involving prediction, table understanding, and data generation (Borisov et al., 2022; Hegselmann et al., 2023; Sui et al., 2023). However, there remains a significant gap in utilizing LLMs to reduce bias and enhance fairness in synthetic tabular data generation using effective prompting strategy.

2.1 Large Language Models (LLM)

(Fang et al., 2024) defines LLM as:

> *"A Large Language Model (LLM), denoted as M and parameterized by θ, is a Transformer-based model that may have an autoregressive, autoencoding, or encoder-decoder architecture. It is trained on an extensive corpus containing hundreds of millions to trillions of tokens, encompassing a range of pre-trained models"*

A language model predicts the likelihood of future or missing tokens in a word sequence. (Zhao et al., 2023b) categorize development of language models into four stages. The journey began with Statistical Language Models (SLMs) like N-Gram models, which struggled with dimensionality issues (Bengio et al., 2000; Saul and Pereira, 1997). Neural Language Models (NLMs) followed, using neural networks such as RNNs to generate word embeddings and improve generalization (Kim et al., 2016). Context-aware models like ELMo introduced bidirectional LSTMs, enhancing performance across NLP tasks (Peters et al., 2018; Wang et al., 2022). Pretrained Language Models (PLMs) like BERT and GPT-2 then leveraged transformer architectures and self-attention mechanisms to achieve remarkable results through pre-training and fine-tuning (Ding et al., 2023). The current focus is on Large Language Models (LLMs) like ChatGPT, which, due to their scale, demonstrate advanced capabilities beyond traditional tasks (Brown et al., 2020).

2.1.1 Emergent Abilities

Large Language Models (LLMs) have demonstrated several critical emergent abilities that highlights the advanced capabilities of LLMs, distinguishing them from smaller models and enabling them to tackle a broader range of tasks:

1. **In-context learning** where models solve tasks using examples in prompts without further training (Brown et al., 2020; Wei et al., 2022a)

2. **Instruction following** allows LLMs to perform new tasks based on natural language instructions, a skill enhanced through instruction tuning, particularly in larger models (Ouyang et al., 2022; Sanh et al., 2021)

3. **Multi-step reasoning** involves solving complex tasks by guiding the model through intermediate steps using chain-of-thought (CoT) prompting, which significantly improves performance, especially in models trained on code and exceeding 100B parameters (Wei et al., 2022b)

2.2 Overview of Tabular Data

Tabular data can be defined as - structured data organized into a grid format with rows and columns, where each column corresponds to a particular attribute or feature. It can be mathematically represented as a matrix $\mathbf{X}$ with dimensions $n \times m$, where $\mathbf{X} \in \mathbb{R}^{n \times m}$ represents the entire dataset. The parameter n corresponds to the number of rows, which are the individual records or instances, and m corresponds to the number of columns, which

are the features or attributes. Each entry x_{ij} in the matrix $\mathbf{X}$ represents the value of the j-th feature for the i-th record, where $x_{ij} \in \mathbb{R}$ (or another appropriate set depending on the data type, such as $\mathbb{Z}$ for integers or $\mathscr{C}$ for categorical values).

2.2.1 Defining Characteristics & Challenges

The typical characteristics of tabular data pose significant challenges in generative modeling (Fang et al., 2024; Manousakas and Aydöre, 2023)

1. There is often no prior knowledge about the structure of tabular data, making it difficult for models to grasp the inherent relationships between features. Features in tabular data can be correlated, requiring careful handling to avoid biases in model predictions.

2. Tabular data consists of a variety of feature types—categorical, numerical, binary, and textual—ranging from dense numerical features to sparse, high-cardinality categorical features. The mix of categorical and numerical features in tabular data complicates the process of learning a joint distribution over all features.

3. Missing values in tabular data challenge generative models in learning a complete distribution.

4. Often, there is insufficient tabular data available for training generative models, leading to difficulties in model development.

5. Unlike image or text data, the order of samples and features in tabular data is not inherently meaningful, limiting the applicability of position-based modeling techniques like CNNs.

2.2.2 Why Focus on Synthetic Tabular Data?

Focusing on synthetic tabular data is essential due to its significant influence in critical domains like healthcare, finance, and cybersecurity, where it drives research and policy decisions (Borisov et al., 2022; Dastile et al., 2020; Shwartz-Ziv and Armon, 2022). Despite its importance, progress in machine learning for tabular data remains slow compared to other modalities (van Breugel and van der Schaar, 2024) (see Figure 2.1). Synthetic data addresses issues such as noisy, imbalanced datasets, and the exclusion of marginalized communities, which can lead to biased models (van Breugel and van der Schaar, 2024). Additionally, synthetic data offers solutions to challenges like high costs, lengthy processes, and privacy concerns tied to real data collection and labeling (Manager, 2023; Porter, 2023). It can

be generated rapidly in large quantities, facilitating more controlled and precise AI model training and testing (Dilmegani, 2023; Savage, 2023). Gartner forecasts that synthetic data will dominate AI models by 2030, with 89% of tech executives considering it crucial for maintaining competitiveness (VentureBeat, 2021). Thus, given AI's potential to excel in the tabular domain, particularly in reasoning about real-world distributions and generalizing across variables (Borisov et al., 2022), focusing on synthetic tabular data becomes a critical research priority.

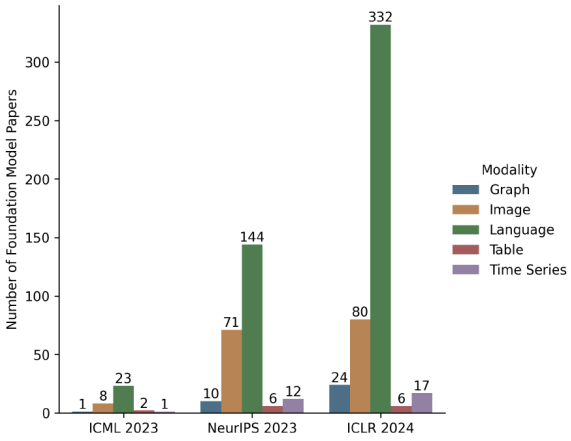


Fig. 2.1 The data, derived from (van Breugel and van der Schaar, 2024), highlights the rapid growth of language models (LLMs) in the field, while modalities such as tabular data remain significantly underrepresented (based on foundation model research across recent ML conferences).

2.2.3 Existing Research on Synthetic Tabular Data

Data synthesis is a critical task in tabular data modeling, essential for developing robust models. Synthetic data generation is used for augmenting sparse datasets, imputing missing values, and rebalancing imbalanced classes (Jolicoeur-Martineau et al., 2024; Onishi and Meguro, 2023; Sauber-Cole and Khoshgoftaar, 2022). Traditionally, methods like Copulas (Li et al., 2020; Patki et al., 2016) and Bayesian networks (Madl et al., 2023; Zhang et al., 2017) have been employed for this purpose. However, recent advances in generative models, including Variational Autoencoders (VAEs) (Darabi and Elor, 2021; Liu et al., 2023a; Ma et al., 2020; Vardhan and Kok, 2020; Xu et al., 2023b), Generative Adversarial Networks (GANs) (Baowaly et al., 2019; Choi et al., 2017; Park et al., 2018; Xu et al., 2019b), diffusion models (Kotelnikov et al., 2023; Lee et al., 2020; Xu et al., 2023a; Yang et al., 2024; Zhang et al., 2023a), and Large Language Models (LLMs), have significantly outperformed classical

methods like Bayesian networks (Xu et al., 2019b), opening new opportunities in data synthesis.

2.3 LLMs for Tabular Data Generation

Table 2.1 exhibits notable studies that leveraged the capabilities of LLMs in synthetic tabular data generation.

Table 2.1 Summary of LLM-based Tabular Data Generation Methods

Method	Used LLM	Fine-tuned or not
CLLM (Seedat et al., 2023)	GPT4	Non Fine-tuned
GReaT (Borisov et al., 2022)	GPT2/DistilGPT2	Fine-tuned
REaLTabFormer (Solatorio and Dupriez, 2023)	GPT2	Fine-tuned
TabMT (Gulati and Roysdon, 2024)	Masked Transformers - 24layer	Fine-tuned
TabuLa Zhao et al. (2023c)	DistilGPT2	Fine-tuned
TAPTAP (Zhang et al., 2023c)	GPT2/DistilGPT2	Fine-tuned

2.3.1 Application and Opportunities

Although language models have shown remarkable success in NLP tasks, their application to tabular data has been limited due to structural differences between text and tabular data. Nonetheless, there are growing opportunities to apply LLMs to tabular data modeling, potentially enhancing tasks like data synthesis (Fang et al., 2024) -

- Deep learning models frequently underperform on datasets that differ from those they were originally trained on, highlighting the potential of transfer learning through the pre-training and fine-tuning approach as a promising solution (Shwartz-Ziv and Armon, 2022).
- Converting tabular data into natural language interpretable by LLMs mitigates the curse of dimensionality often encountered with one-hot encoding in the preprocessing of high-dimensional categorical data.
- The development of emergent capabilities, such as chain-of-thought (CoT) prompting for sequential reasoning, has expanded language models beyond traditional language processing into broader task-solving roles. Further research is needed to explore the boundaries of these emergent abilities in LLMs when applied to tabular data modeling.

2.3.2 Converting Tabular Data to Text: Serialization

To input tabular data into LLMs, the structured data must be converted into a text format since LLMs operate as sequence-to-sequence models (Jaitly et al., 2023; Sui et al., 2024). A straightforward method is to convert it into a programming-readable format such as a Pandas DataFrame, JSON, or HTML. Alternatively, tables can be transformed into delimited text using commas or tabs. Some approaches convert tables into human-readable sentences based on column headers and cell values (Fang et al., 2024). In this study, we perform such text based serialization where individual rows are separated using curly braces and columns are comma-separated (,). Besides, embedding-based (Chen et al., 2023; Deng et al., 2022; Iida et al., 2021) and a less commonly used graph-based approach (Zhao et al., 2023a) are also employed across studies to serialize tabular data.

2.3.3 Strategies for Prompt Engineering

A prompt is input text fed into an LLM, and designing effective prompts is a complex task that has led to extensive research in prompt engineering.

Prompt Format: The simplest method involves concatenating a task description with a serialized table as a string, allowing the LLM to perform the described task and return a text-based response. Well-defined and properly formatted task descriptions have proven to be effective prompts (Marvin et al., 2023).

In-Context Learning Involves incorporating similar examples to guide the LLM in producing the desired output. (Sui et al., 2024) noted a significant drop in performance, with an overall accuracy decrease of 30.38%, when shifting from a 1-shot to a 0-shot setting. (Narayan et al., 2022) found that manually curated examples outperformed randomly selected ones by an average of 14.7 F1 points. (Chen, 2022) observed that while increasing from 1-shot to 2-shot often benefits the model, further increases do not necessarily lead to additional performance gains.

Chain-of-Thought and Self-Consistency Technique: encourages LLMs to break down tasks into step-by-step processes, improving reasoning abilities (Wei et al., 2022b). Program-of-Thoughts (PoT) (Chen et al., 2022) uses code-related comments, such as "Let's write a program step-by-step...," to guide the LLM.

In our study, we leverage these strategies of prompt engineering and create prompts aimed to generate fair synthetic tabular data by GPT (Figure 3.3 contains a sample prompt)

2.4 Bias and Fairness

Large Language Models (LLMs) and machine learning systems have achieved significant success in various domains, but the risk of perpetuating societal harm shadows this success. Trained on vast, uncurated Internet data, LLMs often inherit and amplify stereotypes, misrepresentations, and exclusionary language, disproportionately impacting vulnerable and marginalized communities (Bender et al., 2021; Dodge et al., 2021; Sheng et al., 2021). These issues, broadly called "social bias," stem from deep-rooted historical and structural inequalities, leading to disparate treatment or outcomes among social groups.

Despite their widespread success, studies have found that "Robots are racist and sexist just like the people who created them"[1] with claims that include "Higher crime rates in black people"[2], "Dark skins are unattractive"[3], recommending less qualified male candidates over more qualified female candidates on job portals (Lahoti et al., 2019), and facial recognition software in digital cameras incorrectly detecting Asians as blinking more often than other groups[4]. Since machine learning models are trained on human-generated data, they are prone to reflecting and amplifying human biases and societal stereotypes in their decision-making processes (Wang et al., 2019b). The fairness of these models can be compromised when their outcomes vary based on protected characteristics such as race, religion, economic status, sexual orientation or gender (Mehrabi et al., 2021).

In machine learning, fairness and bias are closely related, as both concern how a model's predictions may advantage or disadvantage certain groups. Bias in a model leads to unfair outcomes, so mitigating bias is crucial to achieving fairness. However, it's important to recognize that absolute fairness is difficult to attain due to the varying definitions and criteria of fairness. As a result, there is currently no universal solution that can eliminate all forms of bias and render a model completely fair (Kheya et al., 2024).

2.4.1 Types of Bias

(Mehrabi et al., 2021) presents an exploratory survey comprising the different types of biases existing in ML systems, among which measurement bias is particularly relevant. Measurement bias occurs due to the selection, usage, and measurement of specific features (Suresh and Guttag, 2021). For example, in the COMPAS recidivism risk prediction tool, variables like prior arrests and the arrests of friends or family were used as proxies for "riskiness" or "criminal behavior." Besides, race of individuals is correlated to the risk of

[1] Robots are racist and sexist just like the people who created them - The Guardian
[2] Machine Bias: Risk Assessments in Criminal Sentencing - ProPublica
[3] Artificial Intelligence Beauty Contest Doesn't Like Black People - The Guardian
[4] Are Face Detection Cameras Racist? - Time

recidivism highly favoring Caucasians with a comparatively lower risk of reoffending. This approach is flawed because minority communities are often subjected to more frequent policing, leading to higher arrest rates. This does not inherently indicate that individuals from these communities are more dangerous but rather reflects biases in how they are monitored and assessed (Suresh and Guttag, 2019). In this study, we focus on addressing measurement bias within the COMPAS dataset by exploring whether large language models (LLMs) can mitigate this bias and generate fairer data through effective prompting strategies.

2.4.2 Fairness Assessment Tools and Metrics

There is no universal standard for measuring fairness, nor a definitive guideline on which metrics are most appropriate. Studies like - (Caton and Haas, 2020), provide an overview of various fairness measures aiming to provide a straightforward interpretation to assist in decision-making. Based on it, we utilize, parity-based metrics (statistical parity, disparate impact), confusion matrix-based metrics (equalized odds, equal opportunity, accuracy equality), and counterfactual fairness-based measures in our study (detailed in Chapter 4)

Researchers have introduced several tools to assess fairness in machine learning systems. Aequitas, for instance, allows users to evaluate models against various bias and fairness metrics across different population subgroups (Saleiro et al., 2018). AI Fairness 360 (AIF360) by IBM (Bellamy et al., 2018) aims to move fairness research into industrial settings, offering a benchmark for evaluating fairness algorithms. Other noteworthy tools for tabular datasets include FairDo (Duong and Conrad, 2024), Fairlearn (Bird et al., 2020), and FairX (Sikder et al., 2024), which also support the development of fair machine learning applications. We leveraged the Fairlearn library (Bird et al., 2020) to assess fairness.

2.4.3 Fair Synthetic Tabular Data

Reducing bias in synthetic tabular data, especially under differential privacy (DP), remains a critical challenge. Most studies have focused on GAN or diffusion-based frameworks for fair data generation, utilizing bias-penalized loss functions (Abroshan et al., 2022; Rajabi and Garibay, 2022; Xu et al., 2019b), or debiasing datasets prior to training (Chaudhari et al., 2022). Methods like DECAF and PreFair target bias reduction by eliminating undesirable causal links and limiting connections in underlying graphical models (Pujol et al., 2022; Van Breugel et al., 2021). Recent advancements include frameworks like MCRAGE (Behal et al., 2023) and TabDDPM (Kotelnikov et al., 2023), which leverage diffusion models to improve fairness and data fidelity in healthcare. The Bt-GAN framework (Ramachandra et al., 2024) enhances EHR utility by generating realistic, de-identified data that ensures

fairness in downstream tasks, while CuTS (Vero et al.) offers customizable synthetic tabular data generation with specified constraints, focusing on utility, privacy, and fairness. However, in comparison to these GAN or diffusion-based studies, assessing the fairness of synthetic tabular data generated by Large Language Models (LLMs) remains relatively unexplored. Most studies focus on the performance (Espinosa and Figueira, 2023; Seedat et al., 2023) and privacy aspects (Liu et al., 2024a) of such data overlooking fairness. Studies such as (Tiwald et al., 2021) demonstrates that synthetic data can be generated to be both representative and fair by incorporating fairness constraints in the generative model but no such work has been done leveraging emergent abilities of LLMs.

2.4.4 Fairness in LLM-Generated Tabular Data and Classification Tasks

Large Language Models (LLMs) trained on vast datasets, often inherit and amplify biases, leading to potentially harmful outcomes for underprivileged groups (Blodgett et al., 2020; Kumar et al., 2022). Research has focused on mitigating these biases through methods like RLHF (Ouyang et al., 2022) and RLAIF (Bai et al., 2022), which train LLMs to avoid reinforcing stereotypes and generating offensive content. These methods focus on training LLMs to produce fair and neutral outputs, but they may not be feasible for typical users who do not have the resource or expertise to fine-tune the models themselves. Additionally, benchmarks like CrowS-Pairs (Nangia et al., 2020), RealToxicityPrompts (Gehman et al., 2020), RedTeamingData (Perez et al., 2022), and HELM (Liang et al., 2022) have been developed to evaluate and assess bias in LLMs.

On the other hand, recent studies have started to specifically address fairness in LLM-based classification tasks. For instance, the work by (Liu et al., 2024b) explores how LLMs, such as GPT-3.5, inherit social biases from their training data, significantly impacting their fairness in tabular classifications. (Gupta et al., 2021) quantifies the bias present when transitioning from real to synthetic data in model training, highlighting the challenges in maintaining fairness. Moreover, a study by (Chhikara et al., 2024) examines few-shot learning techniques and their effectiveness in mitigating bias in LLMs, emphasizing the ongoing efforts to ensure equitable outcomes in these systems. These studies collectively underscore the growing focus on assessing and improving fairness in LLM-driven classification tasks. Despite these advancements, studies specifically addressing fairness in LLM-generated tabular data and its impact on bias when such synthetic data are used for downstream prediction remain limited.

LLMs have also shown the ability to perform tasks with minimal training data by leveraging contextual information (Brown et al., 2020; Radford et al., 2019). However, the effectiveness of LLMs heavily depends on prompt design, including the format, selection, and order of examples (Li and Qiu, 2023; Lu et al., 2021; Zhao et al., 2021a). Incorporating contextual information and fairness criteria within prompts can significantly improve the fairness of LLM outputs. These developments underscore the importance of incorporating fairness considerations directly into LLM prompts and exploring their impact on synthetic data generation, particularly in the tabular data domain. Additionally, it is crucial to analyze how bias may propagate in classification tasks trained on such synthetic data, a key focus of our study.

Chapter 3

Methodology

Over recent years, synthetic data has gained traction for its ability to surpass real data in applications like fairness enhancement (Rajabi and Garibay, 2022, 2023; Van Breugel et al., 2021; Wen et al., 2022; Xu et al., 2018, 2019a), data augmentation (Antoniou et al., 2017; Bing et al., 2022; Das et al., 2022; Dina et al., 2022), and artificial data generation (Wang et al., 2019a; Yoon et al., 2018). Despite significant attention to tools like DALL-E (Marcus et al., 2022) and ChatGPT (Roumeliotis and Tselikas, 2023; Wu et al., 2023) (van Breugel and van der Schaar, 2023), research on GPT-4's potential for generating fair synthetic data - an area with promising implications given its advanced capabilities (Islam and Moushi, 2024) remains scarce (Islam and Moushi, 2024). Building on this context, our research examines GPT-4's ability to apply fairness constraints in generating synthetic data from biased datasets while maintaining realism. We focus on a detailed analysis rather than benchmarking against other models. While quantitative metrics are crucial, further comparisons with other LLMs were deemed unnecessary, as existing research consistently ranks GPT-4 at the forefront of generative capabilities (Shahriar et al., 2024). Models like LLaMa-70b by Meta (Touvron et al., 2023), Gemini by Google DeepMind (Anil et al., 2023), and Mistral-7b by Mistral AI (Jiang et al., 2023) typically perform at levels comparable to GPT-3.5, falling short of GPT-4 (Chhikara et al., 2024; Seedat et al., 2023). Moreover, open-source models of similar quality, like LLaMA (Touvron et al., 2023) and FLAN-T5 (Chung et al., 2024), require substantial GPU VRAM, making broad analysis difficult. Additionally, many such models are trained on data from advanced models like GPT-4 (Gudibande et al., 2023), reducing the potential for valuable insights from additional comparisons.

This chapter details the methodology for evaluating GPT-4 (Wu et al., 2023) as a synthetic tabular data generator, with a focus on performance and fairness. Section 3.1 provides an overview of the approach, while Section 3.2 formulates the problem, detailing data partitioning, in-context sample selection strategy, and the training of classification models.

Section 3.3 explores the synthetic data generation process, explaining the rationale behind our approach. Lastly, Section 3.4 outlines the evaluation metrics employed in Chapter 4 to analyze the utility and fairness of the generated data.

3.1 Overarching Approach

As depicted in Figure 3.1, the framework involves dividing the real dataset (D_{real}) into training (D_{train}) and test sets (D_{test}). A subset of D_{train} is used as in-context examples for GPT-4 to generate a synthetic dataset (D_{syn}). Classification models are trained on both D_{train} and D_{syn}, and evaluated on D_{test}. A thorough comparison between D_{real} and D_{syn} is conducted using standardized metrics to assess the quality and potential biases in the synthetic data.

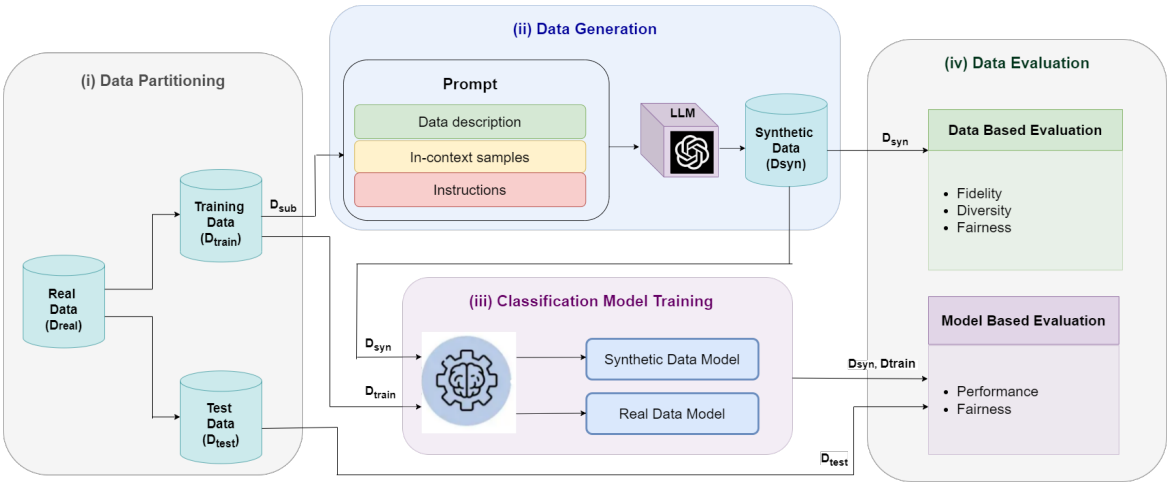


Fig. 3.1 Overview of the proposed methodology. (i) **Data Partitioning:** The real dataset is divided into a training set (D_{train}) and a test set (D_{test}) (ii) **Data Generation:** Synthetic tabular data (D_{syn}) is generated by LLM using a subset of (D_{train}) as in-context samples. (iii) **Classification model training:** Various classification models are separately trained on (D_{train}) and (D_{syn}) and evaluated on (D_{test}) following the TS-TR (Train on synthetic, Test on real) approach. (iv) **Data Evaluation:** The generated (D_{syn}) is separately evaluated for their statistical fidelity, diversity, and fairness using both data-based and model-based approaches.

3.2 Problem Formulation

Let $D_{\text{real}} = \{(x_i, y_i)\}_{i=1}^{M}$ be the original labeled dataset, where $x_i \in X$ and $y_i \in Y$. Here,

$X = \{1, \ldots, k\}$ represents the given feature vector,
$Y = \{0, 1\}$ represents the binary labels, and
$M = |D_{\text{real}}|$ is the total number of samples in the original dataset.

The goal is to generate a synthetic dataset, D_{syn}, and then train a classification model f on D_{syn}. To achieve this, the data partitioning process is carried out as follows:

1. **Data Partitioning:** The dataset D_{real} is divided into a training set D_{train} and a test set D_{test}, each containing 50% of the data while preserving the original class distribution. Let
$$D_{\text{real},0} = \{(x_i, y_i) \in D_{\text{real}} \mid y_i = 0\}, \quad M_0 = |D_{\text{real},0}|$$
and
$$D_{\text{real},1} = \{(x_i, y_i) \in D_{\text{real}} \mid y_i = 1\}, \quad M_1 = |D_{\text{real},1}|$$
denote the subsets with $y = 0$ and $y = 1$, respectively. The training and test sets are then defined by:

$$|D_{\text{train},0}| = \alpha M_0, \quad |D_{\text{train},1}| = \alpha M_1, \quad |D_{\text{test},0}| = (1-\alpha) M_0, \quad |D_{\text{test},1}| = (1-\alpha) M_1$$

with $\alpha = 0.5$, ensuring that both subsets reflect the original class imbalance.

2. **In-Context Examples and Sampling Strategy:** We define a subset $D_{\text{sub}} \subset D_{\text{train}}$ containing n samples, where $n \in \{20, 40, 100, 200, 500\}$. D_{sub} is selected using a sampling method that accounts for the label distribution y_i and the value of the sensitive attribute S, where $S \in \{s_1, s_2\}$. Here, $S = s_1$ represents the privileged group, and $S = s_2$ represents the unprivileged group. This carefully selected subset D_{sub} is then provided as in-context examples to the LLM in its prompt, as illustrated in Figure 3.1. In-context learning leverages the LLM's ability to interpret and generalize from a small number of examples provided in the prompt (Brown et al., 2020; Wei et al., 2022b). Given the LLM's limited context window (Kaplan et al., 2020), only a constrained number of examples D_{sub} can be included in the prompt. Therefore, the selection of these representative samples is crucial; it ensures that the LLM effectively grasps the task

at hand and generates high-quality synthetic data that accurately reflects the intended fairness constraints.

- **Random Sampling:** D_{sub} is selected by randomly sampling n examples from D_{train}, ensuring equal representation of each class y, without considering the sensitive attribute S. This results in:

  $$D_{\text{sub}} = \{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0\} \cup \{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1\},$$

  where the number of samples from each class is balanced:

  $$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0\}| = |\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1\}| = \frac{n}{2}.$$

  Here, the sampling is uniform across D_{train} with equal representation of both classes $y = 0$ and $y = 1$.

- **Balanced Sampling:** D_{sub} is drawn based on a balanced distribution of y with respect to the sensitive attribute, S.

  If n is the total length of D_{sub}, then D_{sub} will be drawn such that:

  $$\begin{aligned} D_{\text{sub}} = &\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \,\text{and}\, S = s_1\} \cup \{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \,\text{and}\, S = s_1\} \\ &\cup \{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \,\text{and}\, S = s_2\} \cup \{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \,\text{and}\, S = s_2\} \end{aligned}$$

  where each quarter of D_{sub} is sampled as follows:

  $$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \,\text{and}\, S = s_1\}| = |\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \,\text{and}\, S = s_1\}| =$$

  $$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \,\text{and}\, S = s_2\}| = |\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \,\text{and}\, S = s_2\}| = \frac{n}{4}.$$

- **Biased Sampling:** D_{sub} is drawn with a 70-30 split based on y and the sensitive attribute S. For the privileged group ($S = s_1$), where the original data has more $y = 0$, D_{sub} is adjusted to have 70% $y = 1$ and 30% $y = 0$. Conversely, for the unprivileged group ($S = s_2$), where the original data has more $y = 1$, D_{sub} is adjusted to have 70% $y = 0$ and 30% $y = 1$. This adjustment aims to counteract the original bias to ensure that the generated subset of data, D_{sub} has a distribution that corrects for the skewed representation of classes in the original dataset.

For the privileged group $S = s_1$:

$$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \text{ and } S = s_1\}| = 0.7 \times \frac{n}{2},$$
$$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \text{ and } S = s_1\}| = 0.3 \times \frac{n}{2}.$$

For the unprivileged group $S = s_2$:

$$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 0 \text{ and } S = s_2\}| = 0.7 \times \frac{n}{2},$$
$$|\{(x_i, y_i) \in D_{\text{train}} \mid y_i = 1 \text{ and } S = s_2\}| = 0.3 \times \frac{n}{2}.$$

Given D_{sub} along with the prompt, a generative model G (represented by the LLM in Figure 3.1) generates a synthetic dataset D_{syn} as follows:

$$D_{\text{syn}} = G(D_{\text{sub}})$$

where G is conditioned on D_{sub} as in-context samples.

3. **Training Classification Models on Real and Synthetic Data**

   We define two machine learning models:

   - **Real Data Model:** $f_{\text{real}} : X \to Y$ trained on D_{train}.
   - **Synthetic Data Model:** $f_{\text{syn}} : X \to Y$ trained on D_{syn}.

   The training process for both models involves minimizing a loss function $\mathscr{L}$:

$$f_{\text{real}} = \arg\min \mathscr{L}(f(D_{\text{train}}))$$

$$f_{\text{syn}} = \arg\min \mathscr{L}(f(D_{\text{syn}}))$$

   To ensure comparability, both models f_{real} and f_{syn} are evaluated on the same test set D_{test} using various performance and fairness metrics. The goal is to develop key insights by comparing:

$$\mathscr{C}_{\text{fair}}(f_{\text{real}}, D_{\text{test}}) \quad \text{vs.} \quad \mathscr{C}_{\text{fair}}(f_{\text{syn}}, D_{\text{test}})$$

$$\mathscr{C}_{\text{perf}}(f_{\text{real}}, D_{\text{test}}) \quad \text{vs.} \quad \mathscr{C}_{\text{perf}}(f_{\text{syn}}, D_{\text{test}})$$

Here, $\mathscr{C}_{\text{fair}}(f, D)$ represents the fairness metric(s) calculated for a given model f on the test dataset D. Similarly, $\mathscr{C}_{\text{perf}}(f, D)$ denotes the performance metric(s) for the model f on the same test dataset D. By comparing $\mathscr{C}_{\text{fair}}$ and $\mathscr{C}_{\text{perf}}$ across models trained on real data (f_{real}) and synthetic data (f_{syn}), we aim to assess how well the synthetic data preserves the fairness and performance characteristics observed in models trained on real data. This comparison is crucial for evaluating the utility of synthetic data, particularly in contexts where fairness and accuracy are both critical considerations.

3.3 Data Generation

Synthetic data generation aims to train a generative model G_θ on real data D_{real} to produce synthetic samples D_{syn} that replicate the statistical properties of D_{real} (Raghunathan, 2021). Traditional models like CTGAN (Xu et al., 2019b), TVAE (Xu et al., 2019b), NFLOW (Durkan et al., 2019), TabDDPM (Kotelnikov et al., 2023), SMOTE (Chawla et al., 2002), and GReaT (Borisov et al., 2022) often face limitations when the training dataset D_{train} is small, leading to the generation of synthetic data that may lack diversity and accuracy (Seedat et al., 2023). LLMs, with their vast pretraining, can overcome these limitations, generating high-quality synthetic data even from limited samples and incorporating fairness constraints through prompting (Chisca et al., 2024; Long et al., 2024; Zhao et al., 2023c). This study leverages GPT-4o mini for data generation task due to its advanced capabilities and cost-effectiveness (Achiam et al., 2023), offering superior intelligence over GPT-3.5-turbo while maintaining similar speed (Kalyan, 2023; Ye et al., 2023).

In this approach, we use the LLM in its frozen black box state choosing not to fine-tune based on the following key considerations.

1. **Baseline Evaluation:** Using the pre-trained model provides a baseline to assess its inherent capabilities and biases, leveraging in-context learning (Radford et al., 2019). This would help in understanding its natural tendencies in generating synthetic data (Brown et al., 2020).

2. **Generalization:** It reflects the model's generalization ability across tasks, crucial for generating realistic synthetic data without domain-specific modifications (Perez et al., 2021).

3. **Resource Efficiency:** Fine-tuning requires substantial resources; by using the pre-trained model, we focus on prompt engineering instead (Dodge et al., 2020; Strubell et al., 2020).

4. **Real-World Applicability:** Many users lack the expertise or resources for fine-tuning, making our approach more applicable to real-world scenarios where pre-trained models are used with fairness prompts (Bender et al., 2021).

To harness the emergent abilities of LLMs—such as leveraging prior knowledge (Chowdhery et al., 2023; Petroni et al., 2019; Singhal et al., 2023), processing contextual information (Raffel et al., 2020), and few-shot learning (Brown et al., 2020; Wei et al., 2023)—we designed our LLM prompt in three parts (Figure 3.3 provides an example prompt, with details for each prompt outlined in Section 4.1.3). :

- **Data Description:** A textual overview of the dataset, task, and feature meanings, utilizing the LLM's prior knowledge.

- **In-context samples:** Serialized samples from the training data (D_{train}) with features and labels provided in text format.

- **Instructions:** Directives for generating a synthetic dataset (D_{syn}) by identifying structural and feature-label relationships, ensuring the data mirrors the examples with and without fairness constraints.

Sample Prompt for Generating Synthetic Data with Demographic Parity

Instructions: You are a synthetic tabular data generator. Leverage your knowledge of criminal recidivism to generate 1000 realistic yet diverse and fair samples.
Ensure Demographic Parity on the basis of the 'race' feature. Demographic Parity implies that both `race_African-American` and `race_Caucasian` should have an equal probability of being assigned to the y=1 class.
I will provide real examples first. The synthetic data must maintain the original label distribution while applying a uniform distribution to the sensitive 'race' feature. Mirror the given examples in causal structure and feature-label distributions while also incorporating the specified fairness constraint, adjusting the influence of 'race'.

In-context samples: [{'sex': 1.0, 'age': 28.0, ... , 'c_charge_degree_M': 0.0, 'y': 1}...
...{'sex': 1.0, 'age': ..., 'c_charge_degree_M': 0.0, 'y': 0}]

Data Description: The output should be a markdown code snippet formatted in the following schema, including the leading and trailing ‘‘‘`{json}` and ‘‘‘`}`:

```
‘‘‘json
{
 "sex": string  // feature column
 "age": string  // feature column
 "juv_fel_count": string  // feature column
 "juv_misd_count": string  // feature column
 "juv_other_count": string  // feature column
 "priors_count": string  // feature column
 "age_cat_25-45": string  // feature column
 "age_cat_Greaterthan45": string  // feature column
 "age_cat_Lessthan25": string  // feature column
 "race_African-American": string  // feature column
 "race_Caucasian": string  // feature column
 "c_charge_degree_F": string  // feature column
 "c_charge_degree_M": string  // feature column
 "y": string  // binary label, y
}
```

DO NOT COPY THE EXAMPLES. Ensure the generated data is realistic, diverse, fair, and correctly conditioned on the features.

3.4 Data Evaluation

Evaluating synthetic data is essential to assess whether it maintains the statistical properties of the original dataset while adhering to fairness principles. This evaluation focuses on two key aspects: data quality (Section 3.4.1) and fairness (Section 3.4.2).

3.4.1 Synthetic Data Quality Measure

The quality of synthetic data is assessed by how well it replicates the characteristics of real data. Evaluation methods fall into two categories: the data-based approach (Section 3.4.1) directly compares synthetic data with real data (Manousakas and Aydöre, 2023; Yang et al., 2024), and the model-based approach (Section 3.4.2) involves training models on synthetic data and evaluating them on real test data.

3.4.1.1 Data Based Evaluation

1. **Fidelity (Resemblance or Similarity) Measure**

   Measures how closely synthetic data resembles real data. A high-fidelity synthetic dataset should consist of samples that appear "realistic" (El Emam et al., 2020; Zhao et al., 2021b). We assess univariate fidelity using Jensen-Shannon Divergence (JSD) (Lin, 1991), bivariate fidelity through pairwise correlation visualized via heat maps (Hittmeir et al., 2019), and multivariate fidelity using Wasserstein Distance (WD) (Arjovsky et al., 2017). These metrics align with standard approaches for evaluating statistical fidelity in synthetic datasets (Dankar et al., 2022b; El Emam et al., 2020).

   - **Jensen-Shannon Divergence (JSD):** JSD measures the similarity between two probability distributions (Dorodchi et al., 2019). It is a symmetric and finite variant of Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951), defined for discrete distributions as:

   $$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}, \tag{3.1}$$

   and for continuous distributions as:

   $$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx. \tag{3.2}$$

   p_i and q_i are the probabilities in distributions P and Q. KLD can be infinite if the distributions have non-matching supports, as Q may assign zero probability

where P does not. JSD (Lin, 1991) addresses this by smoothing the differences between distributions.

- **Wasserstein Distance (WD):** WD is used in population fidelity assessments, effectively addressing challenges posed by Kullback-Leibler Divergence (KLD) (Dandekar et al., 2017) and Total Variation distance (TVD) (McKenna et al., 2021), particularly with discontinuous mappings (Cheng et al., 2020).

  Given two distributions μ and ν over a space M:

$$W_1(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{M\times M} d(x,y)\, d\gamma(x,y) \tag{3.3}$$

  where γ is a transport plan with marginals μ and ν, and $d(x,y)$ is the distance between points x and y in M.

  WD is robust for comparing distributions over continuous or mixed variables, mitigating numerical instability associated with JSD for continuous variables (Zhao et al., 2021b).

Fidelity Metrics Interpretation

A lower JSD and WD value reflects greater fidelity, while higher values suggest increased divergence from the original data (Daniels, 2014).

2. **Diversity Measure**

Assesses whether the generated sample capture the variability of the real data. A robust generative model should produce diverse, high-quality samples (Alaa et al., 2022). We evaluate these aspects using the precision recall score from (Sajjadi et al., 2018) which provides a comprehensive assessment of both fidelity (precision) and diversity (recall) in synthetic data generation, surpassing single metrics like FID (Heusel et al., 2017). Precision measures how well generated samples align with the real data, while Recall evaluates how comprehensively the generated data covers the real distribution.

Diversity Metrics Interpretation

Higher precision indicates better fidelity, while higher recall indicates better diversity (Sajjadi et al., 2018).

3.4.1.2 Model Based Evaluation

While data-based evaluation allows for direct comparison between synthetic and real tabular data, challenges remain. A synthetic dataset that appears fair may not guarantee fair predictions in downstream tasks (van Breugel and van der Schaar, 2023), as a model's fairness can be compromised when applied to real data due to shifts in feature distribution (Jordon et al., 2022). This highlights the importance of model-based evaluation in assessing the fairness and reliability of models trained on synthetic data.

1. **Utility Measure** The application fidelity (Dankar et al., 2022a) or utility of synthetic data is primarily assessed through its performance in machine learning models (Dankar et al., 2022b; Hutt et al., 2022). The widely recognized train-on-real, test-on-real (TRTR) and train-on-synthetic, test-on-real (TSTR) methods involve training models on real and synthetic datasets separately and evaluating their predictive performance on a held-out real test set (Hutt et al., 2022; Manousakas and Aydöre, 2023). Using an independent test dataset reduces the risk of data leakage from the original training set (Hansen et al., 2023) and allows for a fair comparison between real and synthetic data models. We employ the following utility measures to analyze synthetic data performance relative to real data.

   - **Accuracy:** Measures the proportion of correct predictions, ensuring consistent performance across all groups. It is defined as:

   $$Acc = P(\hat{Y} = Y) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

   where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

   - **F1-Score:** The harmonic mean of precision and recall, particularly useful for imbalanced datasets. It is defined as:

   $$\text{Precision} = P(\hat{Y} = 1 \mid Y = 1) = \frac{TP}{TP + FP} \tag{3.5}$$

   $$\text{Recall} = P(Y = 1 \mid \hat{Y} = 1) = \frac{TP}{TP + FN} \tag{3.6}$$

   $$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{3.7}$$

   - **AUC Score:** Evaluates the model's ability to distinguish between positive and negative classes, defined as:

$$\text{AUC} = P(\hat{Y}_{\text{positive}} > \hat{Y}_{\text{negative}}) \tag{3.8}$$

where $\hat{Y}_{\text{positive}}$ and $\hat{Y}_{\text{negative}}$ are the model's predicted scores for positive and negative instances, respectively.

Utility Metrics Interpretation

Higher values of accuracy, F1-score, and AUC indicate better utility.

These metrics comprehensively assess model performance in downstream prediction tasks and also directly compares synthetic data with real data to evaluate its realism.

3.4.2 Synthetic Data Fairness Measure

Previous research shows that LLMs like GPT-3 (Brown et al., 2020), GPT-3.5, and GPT-4 (Achiam et al., 2023) exhibit significant social biases (Abid et al., 2021; Basta et al., 2019), which tend to worsen as models scale up (Askell et al., 2021; Ganguli et al., 2022a). These biases stem from the human-generated training data that reflects societal stereotypes and inequalities (Bolukbasi et al., 2016; Zhao et al., 2017), potentially leading to the reinforcement of harmful patterns when generating synthetic data from biased datasets (Bender et al., 2021). While traditional synthetic data evaluations focus on fidelity, diversity, and privacy, fairness is often neglected (Dankar et al., 2022b; Espinosa and Figueira, 2023; Manousakas and Aydöre, 2023; van Breugel and van der Schaar, 2023). Given the potential for biases in synthetic data, this research prioritizes fairness in its evaluation. While defining fairness is essential, there is no universally accepted metric or guideline for measuring it (Caton and Haas, 2024; Mehrabi et al., 2021).

> *"Broadly, fairness involves making decisions without prejudice or favoritism based on inherent or acquired traits" - (Saxena et al., 2019).*

3.4.2.1 Fairness Definition

This study focuses on five widely recognized fairness notions. For all fairness definitions, we assume, Y as the actual outcome, $\hat{Y}$ as the predicted outcome, and S as the sensitive attribute, where $S \in \{s_1, s_2\}$, with s_1 representing the privileged group and s_2 the unprivileged group, while X denotes the set of attributes excluding S.

Predictive Outcome based Definition focuses on ensuring that the model's predictions $\hat{Y}$, are fair and consistent across different demographic groups, regardless of the input data distribution.

1. **Demographic / Statistical Parity:**
   Demographic Parity requires that the probability of receiving a positive prediction ($\hat{Y} = 1$) is equal across different demographic groups ($S = s_1$ and $S = s_2$), ensuring fairness irrespective of group membership (Corbett-Davies et al., 2017; Feldman et al., 2015; Kamishima et al., 2012; Zemel et al., 2013).

$$P(\hat{Y} = 1 \mid S = s_1) = P(\hat{Y} = 1 \mid S = s_2) \tag{3.9}$$

Predictive Outcome and Actual Outcome based Definition considers both the actual outcome Y and the predicted outcome $\hat{Y}$ across different groups.

1. **Equal Opportunity:**
   This notion requires that the True Positive Rate (TPR) is equal across demographic groups ($S = s_1$ and $S = s_2$), ensuring that individuals in the positive class are treated fairly across different groups (Hardt et al., 2016; Pleiss et al., 2017):

$$P(\hat{Y} = 1 \mid Y = 1, S = s_1) = P(\hat{Y} = 1 \mid Y = 1, S = s_2) \tag{3.10}$$

2. **Equalized Odds:**
   Extending Equal Opportunity, Equalized Odds requires both TPR and False Positive Rate (FPR) to be equal across groups, ensuring consistent positive outcomes across classes (Berk et al., 2021; Verma and Rubin, 2018b):

$$P(\hat{Y} = 1 \mid Y = 1, S = s_1) = P(\hat{Y} = 1 \mid Y = 1, S = s_2) \tag{3.11}$$

$$P(\hat{Y} = 1 \mid Y = 0, S = s_1) = P(\hat{Y} = 1 \mid Y = 0, S = s_2) \tag{3.12}$$

Similarity based Definition asserts that individuals with similar attributes, except for the sensitive attribute S should receive consistent outcomes, ensuring fairness across demographic groups defined by S

1. **Fairness through Awareness / Causal Discrimination:** This criterion requires that individuals with identical attributes X, except for the sensitive attribute S, receive the

same predictions, ensuring equal treatment across demographic groups (Dwork et al., 2012; Galhotra et al., 2017).

$$\text{If } X_i = X_j, \text{ then } \hat{Y}_i = \hat{Y}_j \text{ regardless of } S.$$

2. **Fairness through Unawareness:** Asserts that a model is fair if it does not use sensitive/protected attributes like race or gender in its decision-making process (Grgic-Hlaca et al., 2016; Kusner et al., 2017).

$$f(X,S) = f(X) \tag{3.13}$$

This equation ensures that the model's predictions are independent of the sensitive attribute S.

General Fairness Definition: Beyond the five specific fairness definitions, we include a broader and more general definition of fairness referring it as Generic Fairness. This definition emphasizes overall fairness, ensuring the LLM's outputs are fair, unbiased, and equitable, without adhering to a specific fairness criterion.

3.4.2.2 Fairness Metrics

Fairness evaluations often use metrics derived from the confusion matrix (Barocas et al., 2023; Zafar et al., 2017). In this study, we apply the corresponding adaptations of the fairness definitions outlined in Section 3.4.2.1 to evaluate fairness metrics.

Data Based Evaluation: To evaluate fairness, we analyze class imbalances across sensitive attributes in both real and synthetic datasets (Yang et al., 2024). We use parity-based group fairness metrics—Statistical/Demographic Parity (Verma and Rubin, 2018a) and Disparate Impact (Caton and Haas, 2020)—commonly used to assess how equitably downstream models trained on synthetic data treat different demographic groups (Gupta et al., 2021; Liu et al., 2024b; Vero et al.; Yang et al., 2024). Here, we adapt these metrics to directly compare fairness between synthetic and real data, without relying on any intermediary classifier. This approach is crucial for identifying inherent biases in synthetic data. We also apply these metrics during model-based evaluations to assess how machine learning models affect fairness. This addresses our core research question (Section 1.1):

"Do the biases in synthetic tabular data exacerbate when classified using downstream machine learning models?"

1. **Demographic Parity Difference (DPD):** Measures the absolute difference in the probability of a positive outcome between the privileged and unprivileged group within the sensitive attribute. It can be calculated directly from the synthetic dataset as:

$$\text{DPD} = \left| \frac{\sum_{i=1}^{N_{s_2}} \mathbb{1}(Y_i = 1 \,|\, S = s_2)}{N_{s_2}} - \frac{\sum_{i=1}^{N_{s_1}} \mathbb{1}(Y_i = 1 \,|\, S = s_1)}{N_{s_1}} \right| \tag{3.14}$$

   Where N_{s_1} and N_{s_2} represent the number of samples in the privileged group $(S = s_1)$ and the unprivileged group $(S = s_2)$, respectively.

2. **Disparate Impact (DI):**

   Assesses the ratio of the probability of a positive outcome between the unprivileged and privileged groups. It can be calculated directly from the synthetic dataset as:

$$\text{DI} = \frac{\frac{\sum_{i=1}^{N_{s_2}} \mathbb{1}(Y_i=1 \,|\, S=s_2)}{N_{s_2}}}{\frac{\sum_{i=1}^{N_{s_1}} \mathbb{1}(Y_i=1 \,|\, S=s_1)}{N_{s_1}}}$$

   Where:

   $\mathbb{1}(Y_i = 1 \,|\, S = s_2)$ is an indicator function that equals 1 if the outcome $Y_i = 1$ for an individual in the unprivileged group, and 0 otherwise.

   $\mathbb{1}(Y_i = 1 \,|\, S = s_1)$ is an indicator function that equals 1 if the outcome $Y_i = 1$ for an individual in the privileged group, and 0 otherwise.

Model Based Evaluation:

Our approach evaluates the fairness of models trained on synthetic data using two key principles: equal allocation and equal performance (Agarwal et al., 2018).

- Equal allocation ensures predicted outcomes are distributed fairly across groups, measured by demographic parity difference (DPD).
- Equal performance requires consistent metrics across all groups, assessed via equalized odds difference (EOD) and equal opportunity difference (EOP).

1. **Demographic Parity Difference(DPD):** Demographic/Statistical parity (in Section 3.4.2.1) can be quantified using DPD which requires PPV (positive predicted value) to be equal across sensitive subgroups.

$$\Delta_{DPD} = \left| P(\hat{Y} = 1 \mid S = s_1) - P(\hat{Y} = 1 \mid S = s_2) \right|, \tag{3.15}$$

2. **Equalized Odds Difference (EOD):** Equalized odds (in Section 3.4.2.1) is measured via EOD which requires equal TPR (true positive rate) and FPR (false positive rate) across sensitive subgroups.

$$\begin{aligned}\Delta_{EOD} = \max\Big\{ &\big|P(\hat{Y}=1 \mid Y=0, S=s_1) - P(\hat{Y}=1 \mid Y=0, S=s_2)\big|, \\ &\big|P(\hat{Y}=1 \mid Y=1, S=s_1) - P(\hat{Y}=1 \mid Y=1, S=s_2)\big|\Big\},\end{aligned} \tag{3.16}$$

3. **Equality of Opportunity (EOP):** To quantify equality of opportunity, we calculate the difference in EOP which requires TPR across sensitive subgroups to be equal:

$$\Delta_{EoO} = \big|P(\hat{Y}=1 \mid Y=1, S=s_1) - P(\hat{Y}=1 \mid Y=1, S=s_2)\big|, \tag{3.17}$$

Furthermore for **subgroup level fairness** analysis we compute the differences of the following measures across the privileged and unprivileged groups:

1. **True Positive Rate (TPR)/Recall/Sensitivity:** measures the proportion of actual positive cases correctly identified by the model:

$$\Delta\text{TPR} = P(\hat{Y}=1 \mid Y=1, S=s_1) - P(\hat{Y}=1 \mid Y=1, S=s_2) \tag{3.18}$$

$$= \frac{\text{TP}_{s_1}}{\text{TP}_{s_1} + \text{FN}_{s_1}} - \frac{\text{TP}_{s_2}}{\text{TP}_{s_2} + \text{FN}_{s_2}} \tag{3.19}$$

2. **False Positive Rate (FPR):** indicates the proportion of negative cases incorrectly classified as positive by the model.

$$\Delta\text{FPR} = P(\hat{Y}=1 \mid Y=0, S=s_1) - P(\hat{Y}=1 \mid Y=0, S=s_2) \tag{3.20}$$

$$= \frac{\text{FP}_{s_1}}{\text{FP}_{s_1} + \text{TN}_{s_1}} - \frac{\text{FP}_{s_2}}{\text{FP}_{s_2} + \text{TN}_{s_2}} \tag{3.21}$$

3. **Positive Predictive Value (PPV)/Precision:** assesses the probability of correct predictions among all positive predictions

$$\Delta\text{PPV} = P(Y=1 \mid \hat{Y}=1, S=s_1) - P(Y=1 \mid \hat{Y}=1, S=s_2) \tag{3.22}$$

$$= \frac{\text{TP}_{s_1}}{\text{TP}_{s_1} + \text{FP}_{s_1}} - \frac{\text{TP}_{s_2}}{\text{TP}_{s_2} + \text{FP}_{s_2}} \tag{3.23}$$

4. **Subgroup Accuracy:** Two sensitive subgroups are considered fair if their accuracy rates are equal.

$$\Delta\text{Acc} = \frac{\text{TP}_{s_1} + \text{TN}_{s_1}}{\text{TP}_{s_1} + \text{TN}_{s_1} + \text{FP}_{s_1} + \text{FN}_{s_1}} - \frac{\text{TP}_{s_2} + \text{TN}_{s_2}}{\text{TP}_{s_2} + \text{TN}_{s_2} + \text{FP}_{s_2} + \text{FN}_{s_2}} \tag{3.24}$$

Lower values in DPD, EOD, and EOP indicate better fairness, with 0 representing perfect fairness. For DI, a value of 1 signifies absolute fairness, though the 80% rule is often applied (Alessandra, 1988; van Breugel and van der Schaar, 2023). According to this rule, a predictor may have a disparate impact if the positive outcome rate for a disadvantaged group ($s_2 = 1$) is less than 80% of that for a privileged group ($s_1 = 0$) (Feldman et al., 2015).

Fairness Metrics Interpretation

- For DI, the model is deemed fair if the ratio falls between 0.8 and 1.25.
- For DPD, EOD, and EOP, fairness is achieved if the absolute value is smaller than 0.1.
- For ΔTPR, ΔFPR, ΔPPV, and ΔAcc, the closer to zero, the better. Values less than 0.1 are considered fair.
- Applying these thresholds ensures that while striving to reduce bias, we also maintain a balance to minimize significant utility loss that might occur from over-correcting the predictor (Saxena et al., 2020).

In this study, we aligned our evaluation measures with the key properties outlined by (Saxena et al., 2020) for better synthetic data metrics. Our metrics provide insights into fidelity, and diversity, additionally, fairness, are interpretable for balancing fairness and utility, and support granular evaluation to identify underrepresented or high-risk groups. This ensures a robust and comprehensive assessment of synthetic data quality.

Chapter 4

Evaluation

In this chapter, the research questions (Section 1.1) are addressed through the structure of Sections 4.1 to 4.3. Section 4.1 details the experimental setup, providing the foundation for the analyses that follow, sensitive attributes, LLM configuration, prompt design, and the classification models used.

Section 4.3 directly addresses the research questions detailed in (Section 1.1) through its subsections: Section 4.3.1 corresponds to Research Question 1(i), exploring how different in-context samples affect LLM-driven synthetic data generation. Section 4.3.2 specifically addresses Research Question 1(ii), investigating the effects of sampling methods (random vs. biased) on synthetic data generation. Section 4.3.3 relates to Research Question 2, examining whether fairness can be maintained in synthetic data through carefully designed prompts.

Each subsection in Section 4.3 is further divided into data-based and model-based evaluations. Data-based evaluations assess the quality and fairness of the synthetic data immediately after generation. Model-based evaluations then measure the performance and fairness of the synthetic data when used for downstream prediction tasks, providing an indirect measure of any biases introduced during the data generation process. This dual approach ensures a comprehensive understanding of how bias manifest and potentially exacerbate across different stages of data generation and modeling.

4.1 Experimental Setup

4.1.1 Dataset and Sensitive Attribute

For our experimental analysis, we utilized the widely used COMPAS dataset, particularly fitting for fairness studies due to its relevance in assessing biases in recidivism prediction (Brennan et al., 2009; Larson et al., 2016). Despite similar accuracy across races, COMPAS

often assigns higher risk scores to Black defendants, favoring White defendants (Angwin et al., 2022). Detailed descriptions of the 13 features used in our study, including sex, age, charge degree, priors count, risk score, and the two-year recidivism outcome, are provided in Table 4.1. The race attribute, categorized as "African-American" and "Not African-American," is used as the sensitive attribute, with the dataset focusing on "Caucasian" and "African-American" individuals.

Table 4.1 Features in the COMPAS Recidivism Dataset (Preprocessed).

Feature	Type	Description
Sex	Categorical	The gender of the individual.
Race	Categorical	The race of the individual, grouped into African-American and Not African-American.
Age	Continuous	The age of the individual.
Juv Fel Count	Continuous	The number of juvenile felony counts.
Juv Misd Count	Continuous	The number of juvenile misdemeanor counts.
Juv Other Count	Continuous	The number of other juvenile offenses.
Priors Count	Continuous	The number of prior convictions or charges.
Age Cat 25-45	Categorical	Age category indicator: 25-45 years.
Age Cat Greater than 45	Categorical	Age category indicator: Greater than 45 years.
Age Cat Less than 25	Categorical	Age category indicator: Less than 25 years.
Race African-American	Categorical	Race indicator: African-American.
Race Caucasian	Categorical	Race indicator: Caucasian.
Charge Degree F	Categorical	Charge degree indicator: Felony (F).
Charge Degree M	Categorical	Charge degree indicator: Misdemeanor (M).
Two-Year Recid (Target)	Binary	The target variable indicating whether an individual recidivated within two years.

We worked with a preprocessed version of the COMPAS dataset[1], available from the OpenML repository[2], which includes 5,278 samples and a binary label predicting whether an individual will recidivate within two years, based on demographic and criminal history information. The original dataset contains over 50 attributes (Angwin et al., 2016).

4.1.2 Large Language Model: GPT4o

Large Language Models (LLMs), known for their vast parameters and learning capabilities, are key tools in modern AI (Chang et al., 2024; Zhao et al., 2023b). We utilized OpenAI's "gpt-4o-mini" model, released in March 2023 (Achiam et al., 2023), for its high-quality output with reduced computational overhead (OpenAI, 2024). This model, pre-trained on next-word prediction (Vaswani, 2017) and fine-tuned with RLHF (Christiano et al., 2017; Ouyang et al., 2022), features a 128K token context window, valuable for fitting multiple examples in prompts for synthetic data generation. Each experiment was conducted across 5 random seeds for robustness. The GPT configuration included a temperature of 0.9, a max token limit of 4000, top-p at 0.95, and penalties for frequency (0) and presence (1).

[1] mlr3fairness developers. *COMPAS Recidivism Risk Score Data Set*. 2023. https://mlr3fairness.mlr-org.com/reference/compas.html

[2] https://www.openml.org/search?type=data&status=active&id=42192

4.1.3 Prompt Design

In a traditional in-context learning framework, an LLM L generates output y from a prompt p as $y = L(p)$. Typically, prompts include task context, in-context examples, and specific instructions. In our setup, we construct the prompt p by combining instructions (I), in-context examples (n), data description (d), and fairness rules (f), formally defined as $p = Q(I, f, n, d)$, where $Q(.)$ denotes their concatenation. Our analysis focuses on three primary prompt categories (see Table 4.2), each designed to explore different aspects of synthetic data generation.

Table 4.2 Overview of Prompt Categories and Sampling Methods.

Prompt Category	**Description**	**Sampling Method**	**No of IC Samples n**
General Prompt	Prompt with context, without fairness constraint.	Random	20, 40, 100, 200, 500
Zero-Prior Prompt	Prompt without context.	Random Balanced Biased	Optimal n Optimal n Optimal n
Prompt with Fairness Constraints	Prompt with context and fairness constraint.	Balanced	Optimal n

- **General Prompt:** This prompt includes contextual information like data description and instructions but omits explicit fairness constraints. It primarily evaluates how varying the number of in-context (IC) samples ($n = 20, 40, 100, 200, 500$) affects synthetic data realism, predictor performance, and fairness. This links to research question 1(i), focusing on how the quantity of IC samples influences LLM-driven synthetic data generation. Here, $p = Q(I, n, d)$, with $f = \emptyset$ (no fairness rules).

- **Zero-Prior Prompt:** This prompt excludes contextual information to assess the LLM's inherent knowledge, focusing on how different IC sampling methods (see Section 3.2) affect performance, fairness, and fidelity of the synthetic data. This approach directly tests research question 1(ii), examining the effects of sampling methods. Here, $p = Q(n)$, with $I = d = f = \emptyset$, isolating the impact of IC samples.

- **Prompt with Fairness Constraints:** This prompt incorporates fairness constraints along with contextual instructions and data description $p = Q(I, f, n, d)$. It directly addresses research question 2, exploring whether LLMs can understand and apply fairness rules while maintaining the real data distribution and feature correlations.

4.1.3.1 Framework for Prompts with and without fairness

Prompt w/o fairness

You are a synthetic tabular data generator. Your objective is to create data that mirrors the given examples in terms of causal structure and feature-label distributions while also ensuring the samples are as diverse as possible.
I will provide real examples first. Leverage your knowledge of criminal recidivism to generate 1000 realistic yet diverse samples.
example data: {in-context examples, n}
{data_description, d}
Avoid copying the examples directly. Ensure new and diverse data such that the labels are correctly conditioned on the features.

Prompt with fairness

You are a synthetic tabular data generator. Leverage your knowledge of criminal recidivism to generate 1000 realistic yet diverse and fair samples.
{fairness_rule, f}
I will provide real examples first. The synthetic data must maintain the original label distribution while applying a uniform distribution to the sensitive 'race' feature. Mirror the given examples in causal structure and feature-label distributions while also incorporating the specified fairness constraint adjusting the influence of 'race'.
example data: {in-context examples, n}
{data_description, d}
DO NOT COPY THE EXAMPLES. Ensure the generated data is realistic, diverse, fair and correctly conditioned on the features.

Note: Text highlighted in blue indicates the differences between the two prompts.

In our setup for generating synthetic data using COMPAS, the general prompt or prompt without fairness is $p = Q(I,n,d)$, where I is shown in Figure 4.1.3.1, d in Figure 4.1.3.1, and $n \in \{20,40,100,200,500\}$.

For generating data with fairness notions, the prompt is $p = Q(I,f,n,d)$, with I in Figure 4.1.3.1, f in Table 4.11, d in Figure 4.1.3.1, and n being the optimal value from the general prompt analysis. The fairness definitions in Table 4.11 are discussed in Section 3.4.2.1.

Data description *d* in prompt

The output should be a markdown code snippet formatted in the following schema, including the leading and trailing ```{json} and ```}:

```
```json
{
 "sex": string  // feature column
 "age": string  // feature column
 "juv_fel_count": string  // feature column
 "juv_misd_count": string  // feature column
 "juv_other_count": string  // feature column
 "priors_count": string  // feature column
 "age_cat_25-45": string  // feature column
 "age_cat_Greaterthan45": string  // feature column
 "age_cat_Lessthan25": string  // feature column
 "race_African-American": string  // feature column
 "race_Caucasian": string  // feature column
 "c_charge_degree_F": string  // feature column
 "c_charge_degree_M": string  // feature column
 "y": string  // binary label, y
}
```

4.1.4 Classification Models

LLMs like GPT-4, despite their data generation strengths, are prone to inherent biases and even with fine-tuning, often fail to achieve the fairness and generalization that traditional ML models offer (Chhikara et al., 2024). Fine-tuning is also resource-intensive, making traditional models more practical and reliable. Therefore, we used established ML models—logistic regression (Cox, 1958), decision tree (Loh, 2011), random forest (Breiman, 2001), SVM (Cortes, 1995), and XGBoost (Chen and Guestrin, 2016)—to evaluate the synthetic data generated by GPT-4. These models are well-studied, generalize effectively to unseen data, and offer a consistent benchmark across various tabular datasets (Kotelnikov et al., 2023).
```

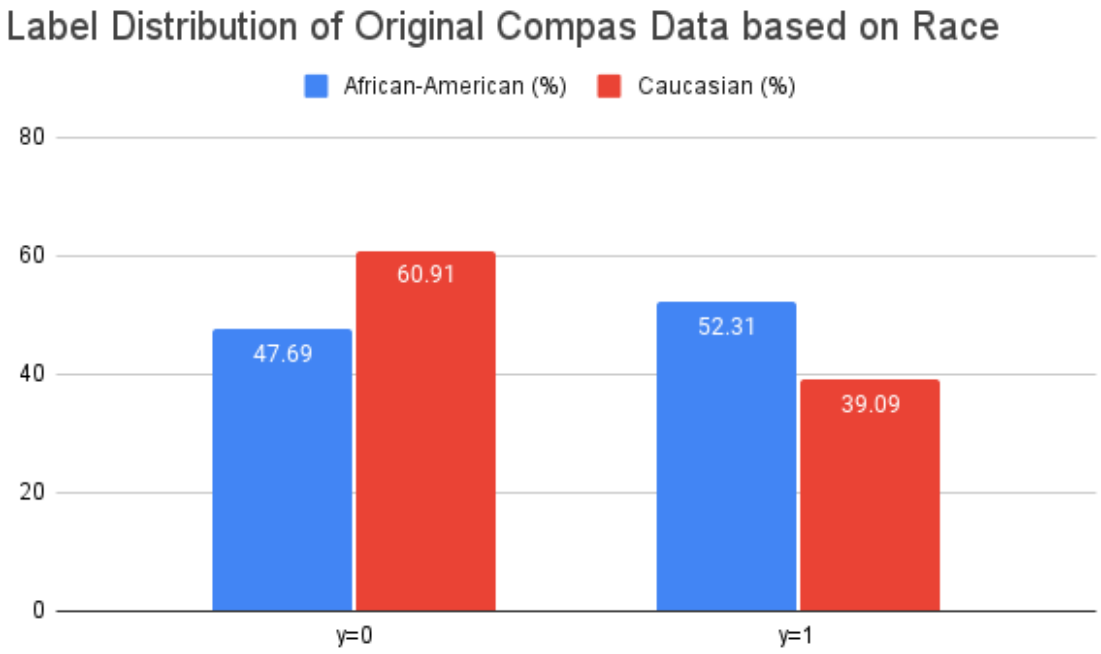
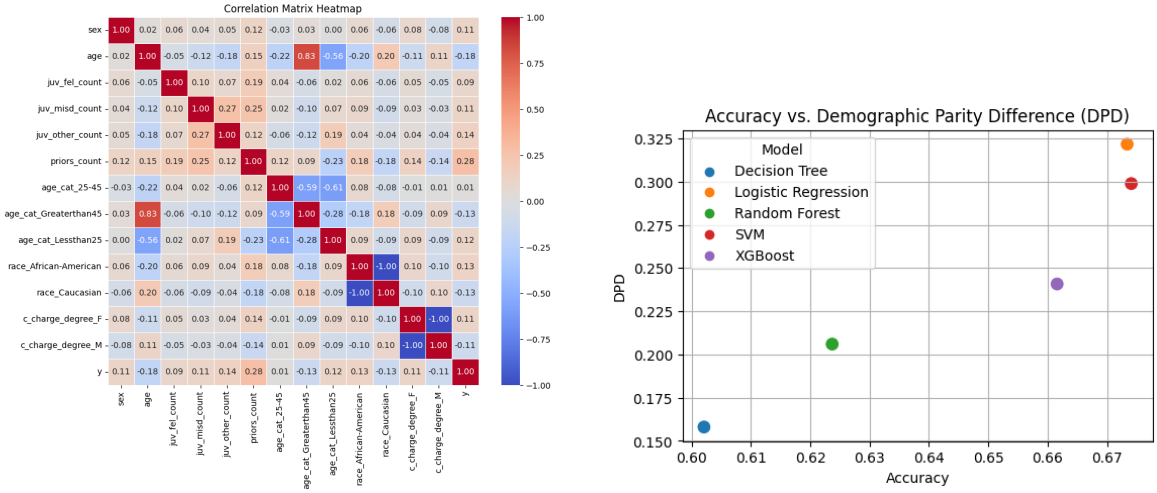


Fig. 4.1 Class distribution based on sensitive attribute "race"



(a) Correlation matrix

(b) Performance vs fairness tradeoff

Fig. 4.2 (a) Correlation heatmap showing a positive correlation between African-American and recidivism, and a negative correlation between Caucasian and non-recidivism (b) Scatter plot showing a trade-off between model accuracy and fairness, with higher accuracy models often having greater fairness disparities

Table 4.3 Performance and Fairness Metrics for Different Classifiers on Real COMPAS data.

	Performance Measure			Fairness Measure		
Classifier	**Acc ↑**	**F1 ↑**	**AUC ↑**	**DPD ↓**	**EOD ↓**	**EOP ↓**
Decision Tree	0.6021	0.549	0.6019	**0.158**	**0.1422**	**0.1531**
Logistic Regression	**0.6734**	**0.6351**	**0.7273**	0.3217	0.3561	0.3561
Random Forest	0.6237	0.5958	0.6589	0.206	0.2362	0.1914
SVM	**0.6741**	**0.6365**	**0.7302**	0.2989	0.3197	0.3197
XGBoost	0.6616	0.6182	0.7018	0.2408	0.2617	0.2617

Note: Upward Arrow (↑) indicates that higher values are better for the performance metrics (Acc, F1, AUC). Downward Arrow (↓) indicates that lower values are better for the fairness metrics (DPD, EOD, EOP). The best values for each metric are highlighted in bold.

4.2 Original COMPAS Data Analysis

The preprocessed COMPAS dataset from OpenML[3] contains around 5,000 samples with a label imbalance: 53% for $y = 0$ (no risk of reoffending) and 47% for $y = 1$ (risk of reoffending within two years). The label distribution shows racial disparity, with African-Americans more likely to be classified as high-risk compared to Caucasians, as depicted in Figure 4.1. The correlation heatmap in Figure 4.2 supports this, showing a slight positive correlation (0.11) between race_African-American and $y = 1$, and a slight negative correlation (-0.13) between race_Caucasian and $y = 0$.

Direct application of Disparate Impact (DI) and Demographic Parity Difference (DPD) yields unfair results with DPD = 0.1345 and DI = 1.3452, where fairness is generally indicated by DI in the range of 0.8-1.25 and DPD, EOD, and EOP values below 0.1 (see 3.4.2.2). Table 4.3 shows that while logistic regression and SVM deliver strong predictive performance, they do so at the expense of fairness. Decision trees, though offering lower predictive accuracy, exhibit better fairness. The fairness issues in logistic regression may arise due to its linear nature, which can amplify biases by overemphasizing features like race. Simpler models like decision trees may avoid these biases by not capturing complex feature interactions, while more complex models like Random Forest and XGBoost are prone to capturing and potentially amplifying subtle biases in the data.

The scatter plot in Figure 4.2 highlights a trade-off between fairness and predictive accuracy, where models prioritizing accuracy often exhibit greater fairness disparities. An interesting observation is that applying the DPD formula directly to the dataset yields a value of 0.1345, but after using a Decision Tree classifier, it increases to 0.158, indicating reduced fairness. This shift occurs because, initially, fairness is measured on raw data without

[3] https://www.openml.org/search?type=data&status=active&id=42192

model influence. Once a classifier is introduced, it trains on the data, accounting for inter-feature interactions, which can exacerbate or introduce biases as it optimizes for accuracy. Consequently, the increase in DPD and reduction in fairness are tied to the classifier's handling of complex feature interactions, potentially amplifying existing biases.

Key Takeaways

- Logistic regression and SVM provide strong predictive performance but compromise fairness, whereas decision trees offer better fairness at the cost of accuracy demonstrating an accuracy-fairness tradeoff (Table 4.3)

4.3 Synthetic COMPAS Data Analysis

This analysis is structured into three key subsections, each corresponding to a different prompt strategy (discussed in detail in Section 4.1.3):

4.3.1 Prompt without fairness/General Prompt

When generating synthetic tabular data using LLMs like GPT4, there isn't any universally agreed-upon "golden standard" for the number of in-context examples (IC samples) to provide in a few-shot learning scenario. It largely depends on the specific task, the complexity of the data, and the desired level of accuracy. This prompt assesses synthetic data fidelity without imposing fairness constraints, focusing on the impact of varying IC sample sizes n where $n \in \{20, 40, 100, 200, 500\}$ on data realism and fairness. IC samples are randomly selected with balanced classes, generating 1000 synthetic samples across 5 seeds for robustness. Figure 4.3 shows that while synthetic data mirrors real data distribution, there is a slight increase in racial bias, particularly with African-Americans classified as high-risk ($y = 1$). This bias arises from initial imbalances in the IC samples, highlighting the risk of bias transfer in synthetic data generation. To address this, balanced and biased sampling strategies are explored in subsequent experiments.

4.3.1.1 Data-based Evaluation

Table 4.4 shows the performance of GPT-4-generated synthetic data with varying IC sample sizes, evaluated on fidelity and diversity metrics.

The results indicate that 40 IC samples strike a "sweet spot" scross all metrics where GPT-4 effectively captures the necessary patterns and relationships without overfitting or underfitting. Additionally, the low standard deviations for 40 IC samples suggest stable

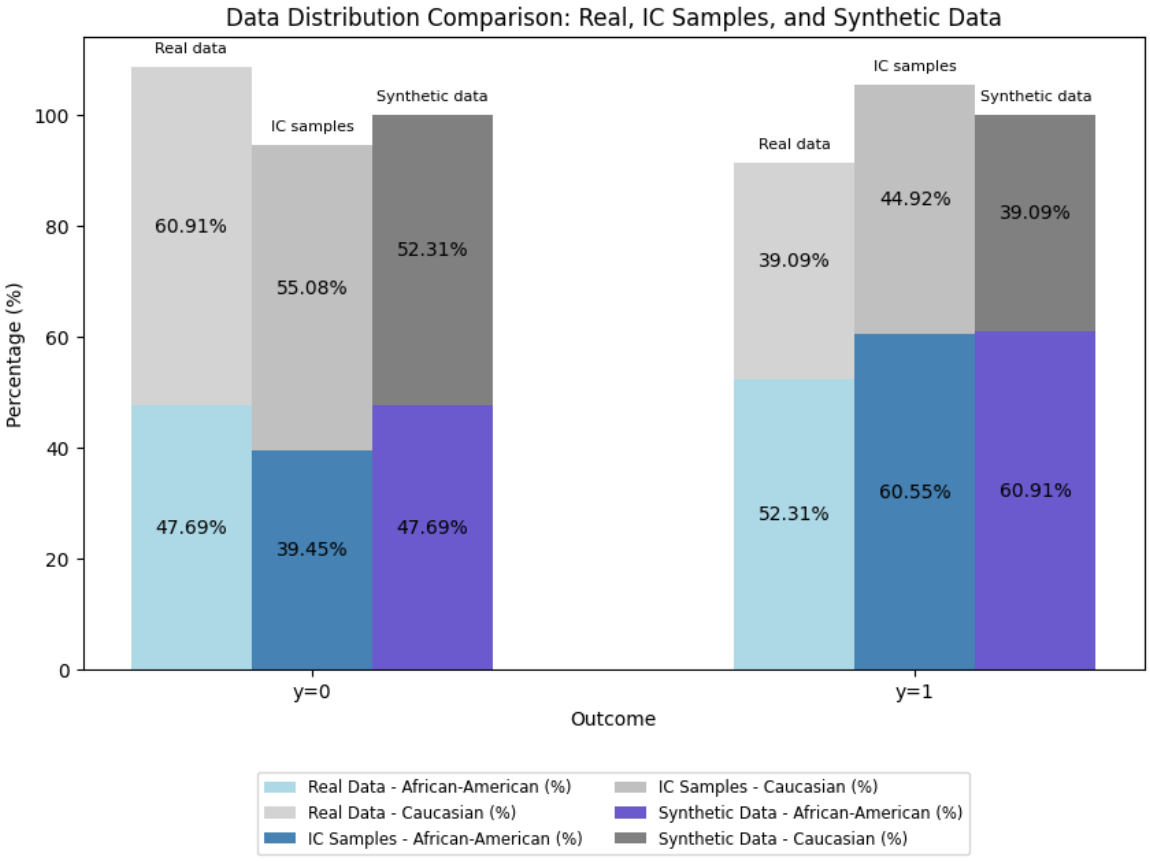


Fig. 4.3 Label distribution based on race in real, in-context samples and synthetic COMPAS data

Table 4.4 Effect of quantity of IC samples in synthetic data fidelity and diversity

No of IC Samples	**JSD (↓)**	**WD (↓)**	**Precision (↑)**	**Recall (↑)**
20	0.0228 (± 0.0022)	0.1458 (± 0.0182)	0.6307 (± 0.0336)	0.8708 (± 0.0335)
40	**0.0209 (± 0.0046)**	**0.1322 (± 0.0451)**	**0.6519 (± 0.0215)**	**0.8751 (± 0.0453)**
100	0.0217 (± 0.0035)	0.153 (± 0.0545)	0.6472 (± 0.0504)	0.7975 (± 0.0394)
200	0.0223 (± 0.0026)	0.1547 (± 0.0376)	0.6431 (± 0.0339)	0.8493 (± 0.0316)
500	0.0225 (± 0.0025)	0.1432 (± 0.0256)	0.6085 (± 0.0624)	0.8713 (± 0.0148)

results. For example, the standard deviations for JSD and WD at 40 IC samples are among the lowest, reinforcing the notion that 40 IC samples are a robust choice for synthetic data generation. The optimal performance with 40 IC samples is likely due to GPT-4's context window limits, where adding more samples does not necessarily improve outcomes.

Table 4.5 Data based fairness evaluation on synthetic data varying number of IC samples

No of IC Samples	**DPD (↓)**	**DI (≈ 1)**
20	**0.0743 (± 0.0368)**	**1.0923 (± 0.2459)**
40	0.1334 (± 0.0987)	1.3501 (± 0.3036)
100	0.144 (± 0.0735)	1.3015 (± 0.1397)
200	0.1351 (± 0.1081)	1.3093 (± 0.2942)
500	0.1576 (± 0.0787)	1.3394 (± 0.163)
Real Data	0.1345	1.3452

Table 4.5 shows that 20 IC samples generate the least biased synthetic data, even fairer than the real data, while 40 IC samples align closely with the real data's fairness. The improved fairness with 20 IC samples, despite the lack of fairness constraints, might be due to the model having less data to overfit on existing biases, inadvertently leading to fairer outcomes but at the cost of lower fidelity and diversity metrics like Precision and Recall.

4.3.1.2 Model-based Evaluation

To assess the impact of varying IC sample sizes on performance and fairness across different classifiers, we conducted a comprehensive analysis summarized in Table 4.3. Logistic Regression consistently provided the best accuracy and AUC, followed by SVM, while Decision Tree performed worst in these metrics across all sample sizes $n \in \{20, 40, 100, 200, 500\}$. However, Decision Tree showed less bias, highlighting the performance-fairness trade-off observed in real COMPAS data in Table 4.5. Regarding the size of ic samples n, $n = 40$ yielded the best accuracy and AUC across classifiers, while fairness was optimal with $n = 20$ for Decision Tree and $n = 100$ for other classifiers, though with reduced performance. This highlights that the performance-fairness trade-off still holds true for synthetic data (see Table 4.7).

Interestingly, for all sample sizes, DPD increases if calculated based on model prediction, in comparison to the DPD that was computed directly on synthetic data (see Table 4.6), indicating that bias exacerbates when synthetic data is used for prediction tasks. This finding reinforces the idea that inter-feature interactions within the dataset can amplify existing biases.

Table 4.6 Comparison of DPD Metrics Across Different IC Sample Sizes

No of IC Samples	DPD_data (↓)	DPD_model (↓)
20	0.0743 (± 0.0368)	0.1174 (± 0.0737)
40	0.1334 (± 0.0987)	0.2120 (± 0.0617)
100	0.1440 (± 0.0735)	0.1846 (± 0.1102)
200	0.1351 (± 0.1081)	0.1775 (± 0.0644)
500	0.1576 (± 0.0787)	0.1464 (± 0.0499)
Real Data	0.1345	0.158

DPD_data: DPD directly computed on synthetic data. **DPD_model**: DPD based on predictive performance using a Decision Tree classifier, which showed the best fairness compared to other classifiers (as shown in Table 4.9).

Key Takeaways

1. Best outcomes in synthetic data realism are generally observed within the 20-40 sample range, highlighting that more data isn't always better for in-context learning-based generation. This insight is particularly valuable for data efficiency, demonstrating that high-quality results can be achieved without the need for large datasets—a crucial advantage in scenarios where data collection is challenging or resource-intensive (see Table 4.4).

2. The initial racial bias against African-Americans in the synthetic data underscores the importance of careful IC sample selection, revealing the risk of bias transfer and the need to explore balanced and biased sampling strategies in further experiments (see Figure 4.3).

3. The fairness vs fidelity trade-off in synthetic data is consistent with real COMPAS data (see Table 4.7).

4. DPD increased when calculated based on model predictions compared to direct computation on synthetic data, suggesting that prediction tasks can exacerbate bias (see Table 4.6). This highlights the potential for inter-feature interactions to amplify existing biases.

5. Despite achieving accuracy and AUC close to real data, synthetic data generally increases bias compared to real data (see Table 4.7). This result is expected, as the LLM was not explicitly guided to prioritize fairness in the data generation process. Upcoming sections will explore fairness-constrained prompts using $n = 40$ IC samples.

Table 4.7 Performance and fairness evaluation for different classification models across varying number of IC samples

IC Samples	Performance Measure		Fairness Measure		
	Acc (↑)	AUC (↑)	DPD (↓)	EOD (↓)	EOP (↓)
Decision Tree					
20	0.5792 (± 0.0437)	0.5734 (± 0.0435)	**0.1174** (± 0.0737)	**0.1228** (± 0.0556)	**0.1069** (± 0.0873)
40	**0.6075** (± 0.0234)	**0.6024** (± 0.0269)	0.212 (± 0.0617)	0.2396 (± 0.0578)	0.2522 (± 0.0787)
100	0.5791 (± 0.0485)	0.5764 (± 0.0445)	0.1846 (± 0.1102)	0.2006 (± 0.1316)	0.11 (± 0.1602)
200	0.5894 (± 0.0203)	0.5814 (± 0.0227)	0.1775 (± 0.0644)	0.187 (± 0.0681)	0.1716 (± 0.0613)
500	0.5523 (± 0.0265)	0.547 (± 0.0324)	0.1464 (± 0.0499)	0.1778 (± 0.0544)	0.1431 (± 0.0594)
Real Data	0.6021	0.6019	0.158	0.1422	0.1531
Logistic Regression					
20	0.6353 (± 0.0321)	0.6864 (± 0.0351)	0.2489 (± 0.1518)	0.2809 (± 0.1423)	0.2481 (± 0.1946)
40	**0.6482 (± 0.0217)**	**0.7099 (± 0.0135)**	0.3027 (± 0.1152)	0.3490 (± 0.1139)	0.3490 (± 0.1139)
100	0.6000 (± 0.0722)	0.6464 (± 0.0896)	**0.2142 (± 0.0983)**	**0.2484 (± 0.0879)**	**0.1648 (± 0.2064)**
200	0.6387 (± 0.0297)	0.6687 (± 0.0552)	0.3092 (± 0.1354)	0.3352 (± 0.1600)	0.3352 (± 0.1600)
500	0.5920 (± 0.0844)	0.6272 (± 0.1327)	0.3065 (± 0.1746)	0.3397 (± 0.1616)	0.2975 (± 0.2203)
Real Data	0.6734	0.7273	0.3217	0.3561	0.3561
Random Forest					
20	0.602 (± 0.0431)	0.6303 (± 0.0554)	0.1972 (± 0.1171)	0.2258 (± 0.1039)	0.1769 (± 0.1662)
40	**0.6209 (± 0.0259)**	**0.6619 (± 0.0188)**	0.2242 (± 0.1017)	0.2411 (± 0.1153)	0.2409 (± 0.123)
100	0.5842 (± 0.0285)	0.6114 (± 0.0638)	**0.2277 (± 0.1184)**	**0.2375 (± 0.1262)**	**0.1304 (± 0.2342)**
200	0.6161 (± 0.0216)	0.6408 (± 0.0304)	0.2753 (± 0.0771)	0.3024 (± 0.0969)	0.3034 (± 0.1088)
500	0.577 (± 0.0644)	0.607 (± 0.0767)	0.2264 (± 0.1608)	0.263 (± 0.1642)	0.2591 (± 0.1932)
Real Data	0.6237	0.6589	0.206	0.2362	0.1914
SVM					
20	0.5968 (± 0.031)	0.6657 (± 0.0451)	0.1671 (± 0.0937)	0.1832 (± 0.0656)	0.1444 (± 0.0529)
40	**0.6416 (± 0.0232)**	**0.7123 (± 0.0195)**	0.2545 (± 0.0674)	0.2502 (± 0.0631)	0.2397 (± 0.0697)
100	0.5741 (± 0.0872)	0.6575 (± 0.0833)	**0.1472 (± 0.111)**	**0.1569 (± 0.1004)**	**0.117 (± 0.1174)**
200	0.6271 (± 0.0466)	0.6917 (± 0.0608)	0.2515 (± 0.1094)	0.2486 (± 0.108)	0.23 (± 0.1237)
500	0.596 (± 0.0782)	0.6517 (± 0.118)	0.1921 (± 0.105)	0.2073 (± 0.0707)	0.1557 (± 0.1208)
Real Data	0.6741	0.7302	0.2989	0.3197	0.3197
XGBoost					
20	0.6033 (± 0.0427)	0.6416 (± 0.0643)	0.2075 (± 0.1273)	0.2246 (± 0.1166)	0.1953 (± 0.1682)
40	**0.6233 (± 0.0216)**	**0.6679 (± 0.0257)**	0.2415 (± 0.1034)	0.2604 (± 0.1145)	0.2604 (± 0.1145)
100	0.5781 (± 0.0548)	0.6211 (± 0.0668)	**0.1826 (± 0.0978)**	**0.2057 (± 0.1051)**	**0.1209 (± 0.1939)**
200	0.6228 (± 0.0232)	0.6611 (± 0.0341)	0.2576 (± 0.0561)	0.288 (± 0.0774)	0.288 (± 0.0774)
500	0.5773 (± 0.0595)	0.6063 (± 0.079)	0.2333 (± 0.1413)	0.262 (± 0.1516)	0.2421 (± 0.1815)
Real Data	0.6616	0.7018	0.2408	0.2617	0.2617

4.3.2 Zero Prior Prompt (Prompt without context)

In this study, we assess the quality and fairness of synthetic data generated using zero-prior prompts, where only in-context samples $n = 40$ are provided, without any descriptive context or instructions. This experiment, discussed further in Section 4.1.3 aims to understand the influence of LLM's internal knowledge when generating data samples without explicit natural language instructions. This approach allows us to observe how the model's inherent understanding impacts the generated outputs in the absence of guided context.

4.3.2.1 Data Based Evaluation

The analysis in Table 4.8 underscores the importance of contextual guidance in GPT-4's synthetic data generation. Without context, the model relies on internal knowledge, leading to increased fidelity (higher precision) but reduced diversity (lower recall). Without context, the model's internal biases become more apparent as it replicates patterns it "knows" from the provided IC samples without being directed to explore new or varied patterns. This results in synthetic data that closely mirrors the training data but lacks broader exploration, as shown in the t-SNE plot (Figure 4.4). In contrast, adding context enhances diversity (wider coverage) while slightly decreasing fidelity, as the model is guided to explore beyond its internalized patterns.

Precision and recall values in Table 4.8 reflect this trade-off, indicating that zero-prior prompts closely replicates provided examples with almost no diversity with (Recall $\rightarrow 0$), whereas prompts with context better balance fidelity and diversity. This highlights the necessity of including contextual information in prompts to effectively harness the LLM's potential for generating diverse and representative synthetic data, emphasizing the critical role of thoughtful prompt design.

Table 4.8 Effect of sampling strategy on the quality of synthetic data

Prompt Category	Sampling Method	JSD (↓)	WD (↓)	Precision (↑)	Recall (↑)
Zero prior	Random	**0.0133 (± 0.0037)**	0.2609 (± 0.0485)	0.9209 (± 0.052)	0.04 (± 0.0)
	Balanced	0.015 (± 0.0024)	**0.2555 (± 0.0553)**	**0.93 (± 0.04)**	**0.110 (± 0.0023)**
	Biased	0.0143 (± 0.0028)	0.2673 (± 0.0465)	0.9185 (± 0.067)	0.061 (± 0.0012)
General prompt	Random	**0.0209 (± 0.0046)**	0.1322 (± 0.0451)	0.6519 (± 0.0215)	0.8751 (± 0.0453)
	Balanced	0.0289 (± 0.0034)	**0.1131 (± 0.0326)**	**0.6769 (± 0.0255)**	**0.8962 (± 0.0335)**
	Biased	0.0309 (± 0.0018)	0.1402 (± 0.0541)	0.6233 (± 0.0125)	0.8543 (± 0.0371)

When evaluating fairness in synthetic data generated with $n = 40$ IC samples (see Table 4.9) using sampling strategies discussed in Section 3.2, balanced sampling generally yields less biased results, although it doesn't fully comply with the 80% rule (Alessandra, 1988). Random sampling balances the outcome variable y, but may still skew the sensitive attribute S,

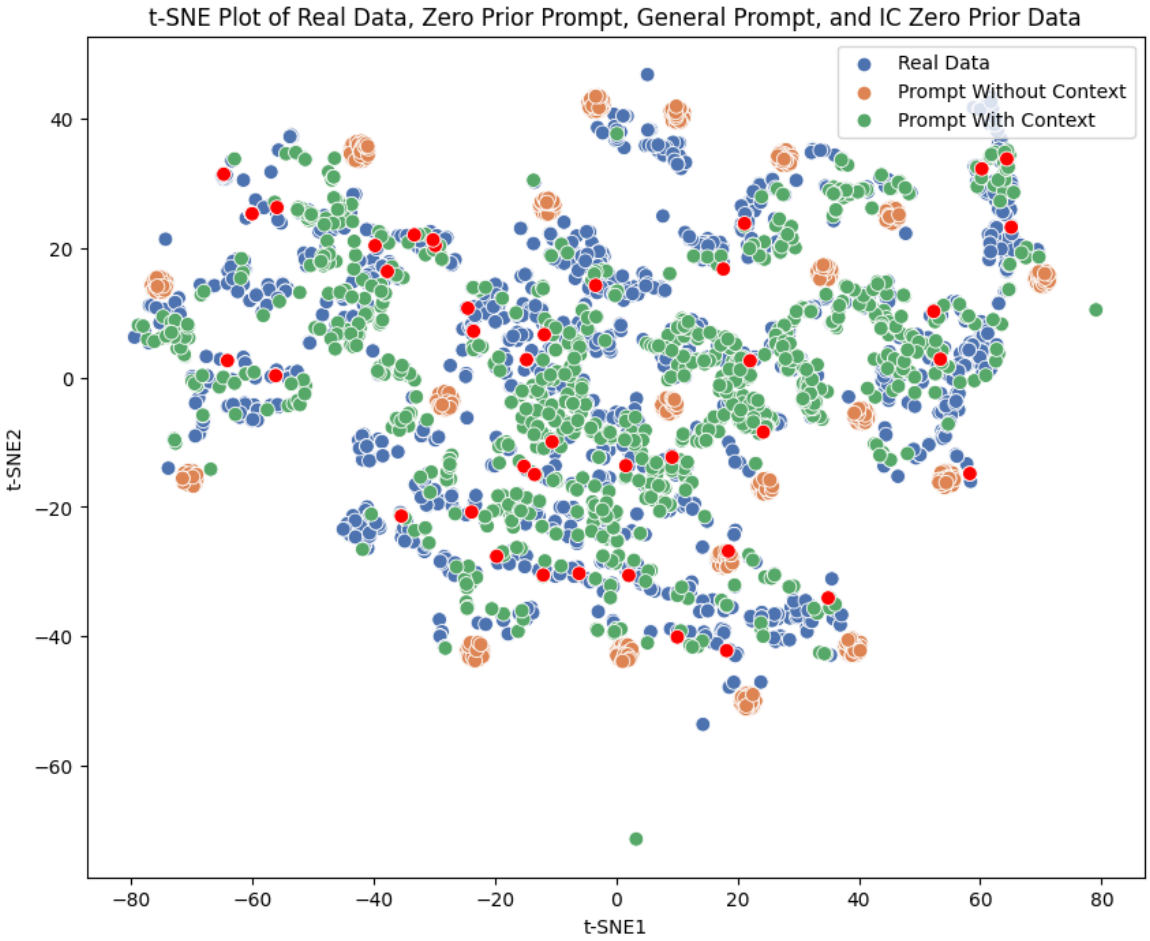


Fig. 4.4 Synthetic data generated using prompt with and without context

Table 4.9 Data based evaluation of fairness using zero prior and different sampling strategies

Prompt Category	**Sampling Method**	**DPD** (↓)	**DI** (≈ 1)
Zero prior/prompt without context	Random	0.2042 (± 0.1485)	1.7922 (± 0.9535)
	Balanced	0.1664 (± 0.0915)	1.0791 (± 0.4098)
	Biased	0.5229 (± 0.1155)	0.3375 (± 0.0994)
General prompt	Random	0.1334 (± 0.0987)	1.3501 (± 0.3036)
	Balanced	0.1302 (± 0.0152)	1.3015 (± 0.4121)
	Biased	0.3212 (± 0.0533)	0.4357 (± 0.0994)
Real Data	-	0.1345	1.3452

reflecting the original data's bias towards Caucasians.In biased sampling, data is intentionally skewed to reverse the original bias in the dataset. Although this method seeks to counteract existing biases but it may introduce other disparities. Balanced sampling ensures equal representation of both y and S, guiding GPT-4 towards equitable distributions.

4.3.2.2 Model Based Evaluation

Table 4.10 reveals that zero prior prompts with IC samples selected using a balanced sampling strategy consistently reduced bias, as indicated by DPD, EOP, and EOD scores, almost equivalent to models trained and tested on real data. This improvement in fairness likely arises from the LLM generating data that reflects the less biased subset of IC samples. Conversely, random and biased sampling strategies exacerbated bias. Hence, for subsequent experiments with fairness-constrained prompts, we use the balanced sampling strategy.

Table 4.10 Model based performance and fairness evaluation for zero prior prompt using different sampling techniques

Sampling Method	**Acc (↑)**	**F1 (↑)**	**AUC (↑)**	**DPD (↓)**	**EOD (↓)**	**EOP (↓)**
			Decision Tree			
Random	**0.5853** (± 0.0236)	**0.5224** (± 0.0227)	**0.5944** (± 0.0234)	0.2709 (± 0.1426)	0.2759 (± 0.1385)	0.2643 (± 0.1317)
Balanced	0.5253 (± 0.0481)	0.5182 (± 0.0508)	0.5108 (± 0.0624)	**0.1592** (± 0.1503)	**0.1821** (± 0.137)	**0.0936** (± 0.1572)
Biased	0.51 (± 0.0425)	0.457 (± 0.0872)	0.5102 (± 0.0419)	0.3325 (± 0.2094)	0.3946 (± 0.2064)	0.3213 (± 0.2354)
Real Data	0.6021	0.549	0.6019	0.158	0.1422	0.1531
			Logistic Regression			
Random	**0.5965** (± 0.0713)	**0.5861** (± 0.04)	**0.6418** (± 0.0906)	0.3388 (± 0.1595)	0.353 (± 0.159)	0.2943 (± 0.2512)
Balanced	0.5423 (± 0.0531)	0.537 (± 0.1208)	0.5098 (± 0.1243)	**0.283** (± 0.2034)	**0.3245** (± 0.2334)	**0.1798** (± 0.2424)
Biased	0.5348 (± 0.0395)	0.4904 (± 0.0805)	0.5578 (± 0.0482)	0.3512 (± 0.1532)	0.4078 (± 0.1779)	0.3518 (± 0.1805)
Real Data	0.6734	0.6351	0.7273	0.3217	0.3561	0.3561
			Random Forest			
Random	**0.6098** (± 0.0266)	**0.593** (± 0.0258)	**0.6436** (± 0.0353)	0.2607 (± 0.1297)	0.2814 (± 0.1541)	0.2601 (± 0.1542)
Balanced	0.5206 (± 0.048)	0.5094 (± 0.1173)	0.5103 (± 0.1163)	**0.1443** (± 0.1753)	**0.1626** (± 0.1642)	**0.0584** (± 0.1875)
Biased	0.5058 (± 0.0447)	0.4444 (± 0.072)	0.5102 (± 0.0527)	0.3035 (± 0.1088)	0.4088 (± 0.2631)	0.4802 (± 0.2747)
Real Data	0.6237	0.5958	0.6589	0.206	0.2362	0.1914
			SVM			
Random	**0.6174** (± 0.0495)	**0.584** (± 0.0784)	**0.6729** (± 0.0625)	0.1851 (± 0.0842)	0.2096 (± 0.0826)	0.1722 (± 0.1066)
Balanced	0.5180 (± 0.0807)	0.4833 (± 0.1301)	0.5473 (± 0.1224)	**0.1032** (± 0.1012)	**0.1395** (± 0.0962)	**0.0112** (± 0.1649)
Biased	0.5123 (± 0.0357)	0.4932 (± 0.1884)	0.5323 (± 0.0681)	0.1975 (± 0.0967)	0.2261 (± 0.0928)	0.1407 (± 0.1435)
Real Data	0.6741	0.6365	0.7302	0.2989	0.3197	0.3197
			XGBoost			
Random	**0.6053** (± 0.0395)	**0.599** (± 0.0225)	**0.649** (± 0.0441)	0.211 (± 0.112)	0.2144 (± 0.1181)	0.2045 (± 0.1294)
Balanced	0.5385 (± 0.0451)	0.5377 (± 0.0845)	0.5568 (± 0.0399)	**0.0954** (± 0.1439)	**0.1264** (± 0.1441)	**0.0999** (± 0.168)
Biased	0.5117 (± 0.0511)	0.4661 (± 0.0895)	0.4947 (± 0.0526)	0.2836 (± 0.1896)	0.3597 (± 0.1814)	0.277 (± 0.215)
Real Data	0.6616	0.6182	0.7018	0.2408	0.2617	0.2617

Key Takeaways

1. Explicit guidance is essential for LLMs to accurately capture real data distribution when generating synthetic data (see Table 4.8 and Figure 4.4).
2. Balanced sampling of IC samples consistently reduce bias in synthetic data, making it comparable to real data, whereas random and biased sampling exacerbate existing biases (Table 4.10).

4.3.3 Prompt with Fairness Constraints

In this section, we conduct experiments using six distinct prompts, each designed with a specific focus on fairness. Five of these prompts incorporate particular fairness constraints aligned with the definitions outlined in Section 3.4.2.1. The prompts are provided in Table 4.11. Additionally, we also use a sixth prompt, referred to as "Generic Fairness," which does not adhere to any specific fairness rule but instead provides general instructions to the LLM to generate fair synthetic data (Table 4.11). Using these prompts, we generate synthetic COMPAS datasets and evaluate them in terms of both quality and fairness. The results are then compared with those obtained using a general prompt that does not explicitly address fairness considerations and the original COMPAS data.

4.3.3.1 Data Based Evaluation

In the subsequent analysis, the different fairness notions are mapped to specific prompts designed to enforce distinct fairness criteria (Table 4.11). The "No Fairness Constraint" shown in Table 4.12 reflects the general prompt discussed in Section 4.1.3, which aims to generate synthetic data that closely mirrors real in-context examples without explicitly enforcing fairness. This provides a basis for comparison with synthetic data generated using specific fairness notions.

Table 4.12 shows that while EOD and Causal Fairness achieve the lowest JSD and WD values, the differences in these metrics across other fairness notions are minimal compared to the no fairness constraint data. Precision ranges from 0.60 to 0.65, with EOD achieving the highest precision, while recall ranges from 0.81 to 0.83, with Causal Fairness achieving the highest recall. Overall, the quality of synthetic data declines when fairness constraints are applied, particularly in recall, which decreases more than precision. This suggests that applying fairness constraints prioritizes fidelity over diversity, resulting in synthetic data that is less diverse but still relatively accurate. Such reduction in overall synthetic data quality is expected, as the no-fairness prompt focuses on closely replicating real in-context examples,

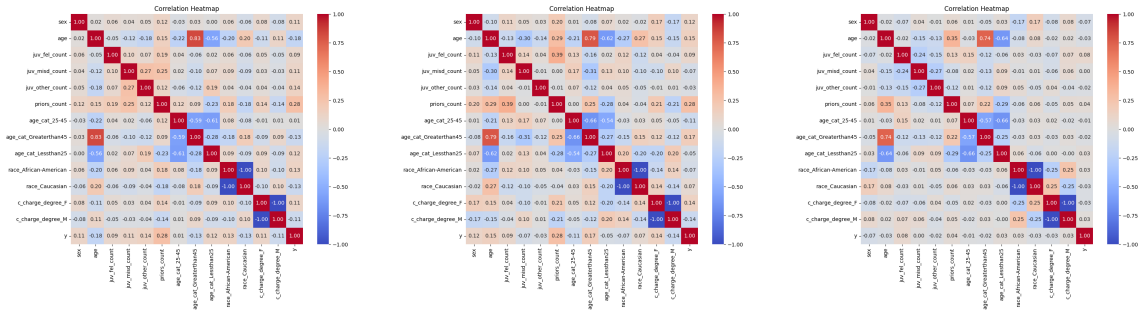
Table 4.11 Fairness Notions and Corresponding Fairness Rules, f

Fairness Notions	**Prompts adhering to respective fairness notions**
Demographic Parity (DP)	Ensure Demographic Parity on the basis of the 'race' feature. Demographic Parity implies that both `race_African-American` and `race_Caucasian` should have an equal probability of being assigned to y=1 class.
Equal Opportunity (EOP)	Ensure Equal Opportunity on the basis of the 'race' feature. True Positive Rate should be the same across both `race_African-American` and `race_Caucasian`. The probability of a person with a true label of y=1 being correctly classified as y=1 should be the same for both races.
Equalized Odds (EOD)	Ensure 'Equalized Odds' on the basis of the 'race' feature. True Positive Rate and False Positive Rate should be the same across both `race_African-American` and `race_Caucasian`. The probability of a person with a true label of y=1 being correctly classified as y=1 and the probability of a person with a true label of y=0 being incorrectly classified as y=1 should be the same for both races.
Causal Discrimination (Causal)	Ensure 'Causal Discrimination' on the basis of the 'race' feature. People having similar attributes should be assigned similar results. Both races, `race_African-American` and `race_Caucasian` with the same features should be assigned the same outcome, that is, either both will be assigned y=1 or both will be assigned y=0.
Fairness through Unawareness (FTU)	Ensure 'Fairness through Unawareness' on the basis of the 'race' feature. The outcomes or labels in the synthetic data should be generated independently of any 'race' features. For any two individuals with the same attributes, the outcome should be similar irrespective of race.
Generic Fairness (Generic)	Ensure that the generated data is unbiased with respect to 'race'. Fairness, in this context, means treating all racial groups equitably, ensuring that the synthetic data does not reflect, reinforce, or perpetuate any discrimination or prejudice. The generated synthetic data should be impartial and reasonable, providing morally right outcomes with no racial group being advantaged or disadvantaged. Maintain the overall distribution of labels and other non-sensitive features as they appear in the original data. However, adjust the influence of race so that it does not affect the distribution of labels or outcomes, ensuring equitable and fair results across all racial groups.

Table 4.12 Evaluation of synthetic data quality using different fairness notions

Fairness Notion	**JSD** (↓)	**WD** (↓)	**Precision** (↑)	**Recall** (↑)
DP	0.02 (± 0.0021)	0.1351 (± 0.0194)	0.6217 (± 0.034)	0.805 (± 0.0407)
EOP	0.0224 (± 0.0019)	0.1585 (± 0.0171)	0.6203 (± 0.0215)	0.8202 (± 0.0361)
EOD	0.0197 (± 0.0029)	0.1258 (± 0.0183)	**0.6536 (± 0.024)**	0.8191 (± 0.0235)
Causal	0.0188 (± 0.0021)	0.1048 (± 0.0092)	0.642 (± 0.0372)	**0.8358 (± 0.0425)**
FTU	0.0222 (± 0.0016)	0.1419 (± 0.0137)	0.6042 (± 0.0299)	0.8265 (± 0.0369)
Generic	0.0224 (± 0.0019)	0.1585 (± 0.0171)	0.6203 (± 0.0215)	0.8202 (± 0.0361)
No fairness constraint	0.0209 (± 0.0046)	0.1322 (± 0.0451)	0.6519 (± 0.0215)	0.8751 (± 0.0453)

while fairness-incorporated prompts must balance realism with specific fairness criteria, leading to slightly less faithful but fairer synthetic data across demographic groups. The comparison of correlation heatmaps, as presented in Figure 4.5a, demonstrates that the LLM effectively adhered to the fairness prompt, successfully reducing the correlation between the outcome variable y and the sensitive attributes S from 0.13 (in Figure 4.5a) to 0.03 (in Figure 4.5c). The prompt containing no fairness notion also reduces the correlation slightly than that of real data (in Figure 4.5b)



(a) Real Data (b) No fairness prompting (c) Generic fairness prompting

Fig. 4.5 Correlation heatmap for real and synthetic data

To further assess the effectiveness of fairness-incorporated prompts in generating synthetic data that is not only realistic but also accurately represents the underlying distribution of the sensitive race attribute, a subgroup-level analysis in Table 4.13 reveals that EOP, EOD, and Causal Fairness perform well in maintaining fidelity, with Causal Fairness also excelling in diversity. However, precision and recall dip slightly under fairness constraints compared to the no fairness notion, as the model without fairness constraints better replicates the COMPAS dataset's variety but risks perpetuating existing biases.

Table 4.14 shows that all fairness notions, except Demographic Parity (DP), achieve better fairness compared to the no fairness constraint prompt - which mirrors the bias in the real data - when DPD and DI are directly computed on synthetic data. DP underperforms in

Table 4.13 Sub-group level analysis of synthetic data quality based on sensitive attribute, race.

Fairness Notion	Precision (↑)		Recall (↑)	
	African-American	Caucasian	African-American	Caucasian
DP	0.6612 (± 0.0277)	0.655 (± 0.0528)	0.7834 (± 0.054)	0.854 (± 0.0352)
EOP	0.706 (± 0.033)	0.6763 (± 0.0305)	0.802 (± 0.0443)	0.8758 (± 0.0188)
EOD	0.6927 (± 0.0188)	0.6861 (± 0.0338)	0.8159 (± 0.0319)	0.8911 (± 0.0293)
Causal	0.6913 (± 0.0507)	0.6909 (± 0.0187)	0.8452 (± 0.0328)	0.8732 (± 0.031)
FTU	0.6732 (± 0.0309)	0.6578 (± 0.0538)	0.8023 (± 0.0478)	0.8883 (± 0.0189)
Generic	0.6661 (± 0.0299)	0.6456 (± 0.0426)	0.8006 (± 0.0388)	0.8669 (± 0.0472)
No fairness notion	0.7062 (± 0.0348)	0.728 (± 0.0368)	0.8774 (± 0.0278)	0.9095 (± 0.0431)

both fairness and subgroup fidelity and diversity (see Table 4.13), while other fairness notions successfully balance these objectives, albeit with some reduction in overall data quality due to the inherent trade-off.

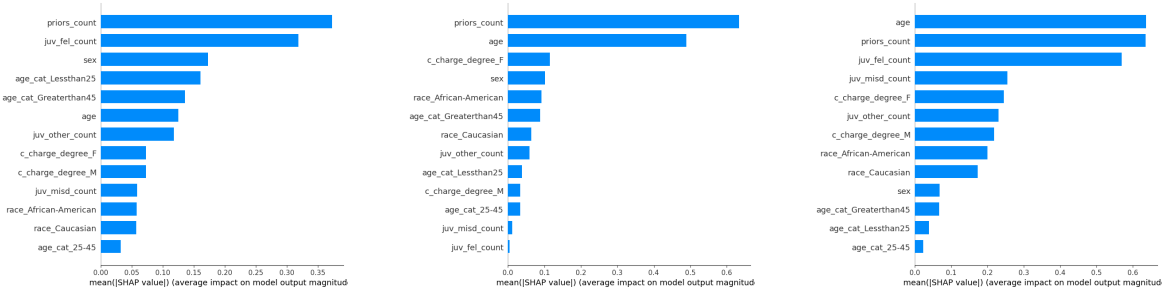
Table 4.14 Data based evaluation of synthetic data fairness using different fairness notions

Fairness Notion	**DPD** (↓)	**DI** (≈ 1)
DP	0.1411 (± 0.0873)	1.3488 (± 0.2532)
EOP	0.0785 (± 0.033)	1.1446 (± 0.1051)
EOD	0.0715 (± 0.0496)	1.1067 (± 0.1575)
Causal	0.0716 (± 0.0352)	1.1221 (± 0.1265)
FTU	0.0367 (± 0.016)	0.979 (± 0.0718)
Generic	**0.0322 (± 0.0278)**	**1.0104 (± 0.0788)**
No fairness constraint	0.1334 (± 0.0987)	1.3501 (± 0.3036)
Real Data	0.1345	1.3452

4.3.3.2 Model Based Evaluation

Table 4.15 presents performance (accuracy, F1, AUC) and fairness (DPD, EOD, EOP) metrics for six fairness prompts across five classifiers. Performance generally decreased while fairness improved compared to the no fairness prompt. Decision Tree consistently showed the best fairness but lower performance, while Logistic Regression excelled in performance but struggled with fairness. These findings align with earlier observations (in Table 4.3, 4.7), highlighting that simpler models like Decision Trees are fairer but less accurate, whereas more complex models like Logistic Regression prioritize performance,

often at the expense of fairness. Further, in Table 4.16, we represent the utility and fairness metrics of different fairness notions using only the decision tree classfier (refer to Table A.1-A.4 in the Appendix A for other classifiers) which revealed that, the model was successful in generating fair synthetic data in most cases where DPD, EOD and EOP reduced below 0.1.



(a) Fairness prompt (b) Real data (c) No fairness prompt

Fig. 4.6 SHAP value analysis for feature importance across different datasets. The plots display the impact of individual features on the model's output, with higher SHAP values indicating greater influence.

The SHAP analysis reveals that synthetic data generated using fairness prompt (Fairness through Unawareness (FTU) in this case) (Figure 4.6a) significantly reduces the influence of race on model predictions compared to the original biased data and synthetic data generated without fairness constraints demonstrating the prompt's effectiveness in mitigating bias. In contrast, the real biased data shows a strong correlation between race and model output (Figure 4.6b). The synthetic data generated without any fairness prompt still retains some bias (Figure 4.6c). This underscores the importance of integrating fairness prompts in synthetic data generation to achieve more equitable outcomes in downstream tasks.

Table 4.15 Performance Metrics for Different Fairness Prompts using different classifiers

Demographic Parity Prompt						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.5225 (± 0.0486)	0.4762 (± 0.0511)	0.5171 (± 0.0513)	**0.1476 (± 0.0894)**	**0.1621 (± 0.0835)**	**0.0665 (± 0.208)**
Logistic Regression	**0.5828 (± 0.0432)**	**0.533 (± 0.0606)**	**0.6176 (± 0.0591)**	0.384 (± 0.2757)	0.3394 (± 0.2737)	0.3876 (± 0.2737)
Random Forest	0.5537 (± 0.0453)	0.5128 (± 0.0577)	0.5676 (± 0.0565)	0.2589 (± 0.2455)	0.2794 (± 0.2525)	0.2211 (± 0.2482)
SVM	0.5505 (± 0.0693)	0.4865 (± 0.1294)	0.5781 (± 0.089)	0.1477 (± 0.1562)	0.1661 (± 0.1472)	0.0869 (± 0.1414)
XGBoost	0.5541 (± 0.0601)	0.5022 (± 0.0783)	0.5666 (± 0.0876)	0.1986 (± 0.2664)	0.2072 (± 0.257)	0.1916 (± 0.2602)
Equal Opportunity Prompt						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.541 (± 0.0406)	0.5223 (± 0.0547)	0.5416 (± 0.0431)	**0.0992 (± 0.1082)**	**0.1135 (± 0.0913)**	**0.0802 (± 0.0982)**
Logistic Regression	**0.6086 (± 0.0385)**	**0.6219 (± 0.0373)**	**0.6702 (± 0.038)**	0.3715 (± 0.1805)	0.3713 (± 0.172)	0.3497 (± 0.1642)
Random Forest	0.5743 (± 0.0384)	0.5483 (± 0.0494)	0.5945 (± 0.0371)	0.1517 (± 0.1279)	0.1702 (± 0.1107)	0.1628 (± 0.1465)
SVM	0.5739 (± 0.0445)	0.5712 (± 0.0743)	0.6413 (± 0.055)	0.1532 (± 0.0687)	0.1825 (± 0.0516)	0.1137 (± 0.0727)
XGBoost	0.561 (± 0.0405)	0.5615 (± 0.0345)	0.596 (± 0.0478)	0.2178 (± 0.0963)	0.2235 (± 0.1012)	0.1905 (± 0.0946)
Equalized Odd Prompt						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.5441 (± 0.0138)	0.5042 (± 0.0457)	0.543 (± 0.0153)	**0.0861 (± 0.0551)**	**0.096 (± 0.046)**	**0.0813 (± 0.0557)**
Logistic Regression	**0.596 (± 0.0327)**	**0.5981 (± 0.0447)**	**0.6519 (± 0.0257)**	0.275 (± 0.1792)	0.2773 (± 0.1835)	0.2453 (± 0.1565)
Random Forest	0.5858 (± 0.0294)	0.5365 (± 0.0432)	0.577 (± 0.0412)	0.1744 (± 0.0995)	0.1914 (± 0.0783)	0.1385 (± 0.1053)
SVM	0.5663 (± 0.0387)	0.5938 (± 0.0595)	0.6032 (± 0.0483)	0.1452 (± 0.0739)	0.1667 (± 0.051)	0.0978 (± 0.0621)
XGBoost	0.5679 (± 0.0333)	0.5559 (± 0.0383)	0.5966 (± 0.0403)	0.167 (± 0.1016)	0.1834 (± 0.1168)	0.1404 (± 0.0942)
Causal Fairness						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.5518 (± 0.0344)	0.4879 (± 0.0528)	0.5451 (± 0.032)	**0.1101 (± 0.0731)**	**0.139 (± 0.0648)**	**0.1075 (± 0.0658)**
Logistic Regression	**0.6275 (± 0.033)**	**0.5394 (± 0.0527)**	**0.6593 (± 0.036)**	0.2391 (± 1.408)	0.2787 (± 0.1504)	0.2663 (± 0.1738)
Random Forest	0.5821 (± 0.0311)	0.5296 (± 0.0672)	0.6076 (± 0.0479)	0.1978 (± 0.1274)	0.2003 (± 0.1041)	0.1897 (± 0.1233)
SVM	0.5773 (± 0.0603)	0.4906 (± 0.065)	0.6107 (± 0.0765)	0.1233 (± 0.0709)	0.1541 (± 0.037)	0.1083 (± 0.0933)
XGBoost	0.5781 (± 0.03)	0.5128 (± 0.0493)	0.6084 (± 0.0389)	0.1485 (± 0.1048)	0.1633 (± 0.0897)	0.156 (± 0.0988)
Fairness Through Unawareness						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.5413 (± 0.0514)	0.5022 (± 0.0462)	0.5376 (± 0.0486)	**0.086 (± 0.0974)**	**0.1012 (± 0.0943)**	**0.0735 (± 0.1343)**
Logistic Regression	**0.6128 (± 0.0596)**	**0.5829 (± 0.0664)**	**0.6464 (± 0.0319)**	0.1871 (± 0.0566)	0.2045 (± 0.1194)	0.1376 (± 0.2)
Random Forest	0.5692 (± 0.0689)	0.5287 (± 0.0291)	0.6164 (± 0.0695)	0.1354 (± 0.1064)	0.1663 (± 0.1511)	0.1216 (± 0.0934)
SVM	0.5782 (± 0.0698)	0.5711 (± 0.0758)	0.5912 (± 0.0847)	0.1093 (± 0.0841)	0.1031 (± 0.0765)	0.0824 (± 0.0721)
XGBoost	0.5751 (± 0.0379)	0.5246 (± 0.0298)	0.6037 (± 0.0642)	0.097 (± 0.1007)	0.1165 (± 0.0624)	0.0824 (± 0.1003)
Generic Fairness						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.5173 (± 0.0255)	0.4799 (± 0.0293)	0.5131 (± 0.0223)	**0.0288 (± 0.0186)**	**0.0596 (± 0.0288)**	**0.0034 (± 0.0372)**
Logistic Regression	**0.5682 (± 0.0748)**	**0.6057 (± 0.083)**	**0.6116 (± 0.1035)**	0.1861 (± 0.2143)	0.2023 (± 0.068)	0.1323 (± 0.1499)
Random Forest	0.5269 (± 0.04)	0.5128 (± 0.0464)	0.5364 (± 0.0514)	0.0921 (± 0.075)	0.1315 (± 0.0752)	0.016 (± 0.1663)
SVM	0.522 (± 0.0802)	0.5887 (± 0.0676)	0.5341 (± 0.137)	0.0747 (± 0.0722)	0.0862 (± 0.0826)	0.0356 (± 0.0627)
XGBoost	0.527 (± 0.0307)	0.5025 (± 0.0359)	0.5351 (± 0.0419)	0.0876 (± 0.0561)	0.1024 (± 0.09)	0.0073 (± 0.096)
No Fairness Notion						
Model	**Acc** (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
Decision Tree	0.6075 (± 0.0234)	0.5298 (± 0.0598)	0.6024 (± 0.0269)	0.212 (± 0.0617)	0.2396 (± 0.0578)	0.2522 (± 0.0787)
Logistic Regression	0.6482 (± 0.0217)	0.5553 (± 0.1022)	0.7099 (± 0.0135)	0.3027 (± 0.1152)	0.349 (± 0.1139)	0.349 (± 0.1139)
Random Forest	0.6209 (± 0.0259)	0.5251 (± 0.1082)	0.6619 (± 0.0188)	0.2242 (± 0.1017)	0.2411 (± 0.1153)	0.2409 (± 0.123)
SVM	0.6416 (± 0.0232)	0.6267 (± 0.0437)	0.7123 (± 0.0159)	0.2545 (± 0.0674)	0.2502 (± 0.0631)	0.2397 (± 0.0697)
XGBoost	0.6233 (± 0.0216)	0.5361 (± 0.0826)	0.6679 (± 0.0257)	0.2415 (± 0.1034)	0.2604 (± 0.1145)	0.2604 (± 0.1145)

Table 4.16 Performance and fairness evaluation using decision tree and prompts with various fairness notions

Fairness Notion	Performance Measure			Fairness Measure		
	Acc (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
DP	0.5225 (± 0.0486)	0.4762 (± 0.0511)	0.5171 (± 0.0513)	0.1477 (± 0.1562)	0.1661 (± 0.1472)	0.0665 (± 0.208)
EOP	0.541 (± 0.0406)	**0.5223 (± 0.0547)**	0.5416 (± 0.0431)	0.0992 (± 0.1082)	0.1135 (± 0.0913)	0.0802 (± 0.0982)
EOD	0.5441 (± 0.0138)	0.5042 (± 0.0457)	0.543 (± 0.0153)	0.0861 (± 0.0551)	0.096 (± 0.046)	0.0318 (± 0.057)
Causal	**0.5518 (± 0.0344)**	0.4879 (± 0.0528)	**0.5451 (± 0.032)**	0.1101 (± 0.0731)	0.139 (± 0.0648)	0.1075 (± 0.0658)
FTU	0.5413 (± 0.0514)	0.5022 (± 0.0462)	0.5376 (± 0.0486)	0.086 (± 0.0974)	0.1012 (± 0.0943)	0.0735 (± 0.1343)
Generic	0.5173 (± 0.0255)	0.4799 (± 0.0293)	0.5131 (± 0.0223)	**0.0288 (± 0.0186)**	**0.0596 (± 0.0288)**	**0.0334 (± 0.0372)**
No fairness notion	0.6075 (± 0.0234)	0.5298 (± 0.0598)	0.6024 (± 0.0269)	0.212 (± 0.0617)	0.2396 (± 0.0578)	0.2522 (± 0.0787)
Real Data	0.6021	0.549	0.6019	0.158	0.1422	0.1531

Key Takeaways

1. Fairness constraints improve fairness but reduce synthetic data quality, especially recall, prioritizing fidelity over diversity (Table 4.12, 4.13).
2. Decision Trees achieve better fairness with lower performance, while Logistic Regression excels in performance but struggles with fairness (Table 4.15)
3. DP based prompts underperforms, while EOD, EOP, and Causal Fairness comparatively balances fidelity and fairness, though with some loss in data quality (Table 4.16).
4. Decision Tree classifier effectively generates fair synthetic data, with DPD, EOD, and EOP values below 0.1 in most cases (Table 4.16).

4.3.4 Subgroup Level Analysis using Fairness Prompts

Subgroup analysis based on sensitive attributes is essential for evaluating the impact of fairness interventions on different demographic groups, especially in critical areas like recidivism prediction. Table 4.17, 4.18, 4.19 presents utility measures for the decision tree classifier which demonstrated the best fairness in Table 4.15, focusing on the African-American (AA) and Caucasian (C) racial subgroups for each fairness notion. Table A.5 - A.8 in the Appendix A displays such subgroup analysis for the other classifiers. Each fairness notion in the LLM prompt targets a specific aspect of fairness, such as Demographic Parity (DP) focusing on PPV, Equal Opportunity (EOP) on TPR, and Equalized Odds (EOD) on both TPR and FPR. While computing the primary metric for each fairness notion is crucial, relying solely on it may not fully capture fairness. Therefore, we also examine the impact

on other relevant metrics to gain a more comprehensive understanding of fairness dynamics. The primary metric for each fairness notion is highlighted in yellow.

We begin evaluating with the TS-TR method of train-test which provides an understanding of how well the synthetic data generalizes and whether the fairness constraints hold in a real-world setting. To compare with original data model, we also implement the TR-TR approach. Additionally, we perform a train-on-synthetic, test-on-synthetic (TS-TS) evaluation performing an 80-20 split of the synthetic data across 5 random seeds to understand its internal consistency. It determines whether the model is able to maintain fairness when applied to data from the same distribution (i.e., the synthetic data) and assess whether synthetic data has inherently enforced the fairness constraint as intended during generation.

Rationale for TS-TS (Train-on-Synthetic, Test-on-Synthetic) Evaluation:

In the TS-TR method of evaluation, the model was trained on a synthetic dataset that reflects fairness (implied by fairness constraint prompts). When such models are tested on a biased real-world dataset (e.g., COMPAS where predictions historically favored Caucasians over African-Americans), it might lead to discrepancies in predictions.

As the model was trained to be fair, it might produce predictions ($\hat{y}$) that attempt to correct the bias. For example, if the model has learned to treat African-American and Caucasian individuals equally in terms of predicting recidivism risk, assigning similar probabilities to both groups when predicting on the real dataset. However, if the real dataset's ground truth labels (y) are biased (e.g., higher rates of recidivism predictions for African-Americans), the model's fair predictions will not align with these biased labels. For instance, the model might predict a lower risk of recidivism for African-Americans compared to what is reflected in the biased real data. This could result in lower True Positive Rates (TPR) for the African-American group or higher False Positive Rates (FPR) when compared to the biased ground truth.

Such mismatch makes it challenging to interpret utility metrics. For example, a lower True Positive Rate (TPR) for African-Americans in the test set could indicate that the model is "correcting" for bias, but because the ground truth labels are biased, the model's fair predictions appear to perform poorly. This doesn't mean the model is unfair; rather, it indicates that the real dataset is not a suitable benchmark for evaluating fairness. Essentially, the biased nature of the real dataset undermines the evaluation of fairness-focused models, leading to potentially misleading conclusions. However, this discrepancy actually highlights the bias in the real data rather than any flaw in the model. To better quantify fairness in this context, synthetic-to-synthetic evaluations (TS-TS) or re-annotating and adjusting the real dataset to reduce inherent biases before using it for evaluation can be considered. However, re-annotations requires domain expertise and substantial manual effort which is not feasible

in our context. Moreover, re-annotation might introduce new biases or inconsistencies if not performed meticulously. Given these challenges, we opted for the TS-TS evaluation approach. This will allow us to assess the model's internal consistency and adherence to fairness without being confounded by the biases present in the real dataset.

4.3.4.1 Demographic Parity

Table 4.17 Subgroup level analysis on demographic parity based prompt using decision tree

Utility Measure	TS-TR with fairness	TS-TR without fairness	TR-TR	TS-TS with fairness	TS-TS without fairness
Acc_AA (↑)	0.5099 ± 0.0531	0.6016 ± 0.0308	0.5882	0.497 ± 0.0463	0.7077 ± 0.0645
Acc_C (↑)	0.5553 ± 0.0366	0.6097 ± 0.016	0.627	0.5679 ± 0.0323	0.6496 ± 0.0396
ΔAcc	-0.0454 ± 0.0645	-0.0081 ± 0.0347	-0.0388	-0.0709 ± 0.0554	0.0581 ± 0.0757
PPV_AA (↑)	0.5313 ± 0.0577	0.6333 ± 0.0326	0.6159	0.5697 ± 0.0641	0.7565 ± 0.0782
PPV_C (↑)	0.4326 ± 0.0332	0.5077 ± 0.0337	0.4957	0.4983 ± 0.1084	0.6339 ± 0.0401
ΔPPV	0.0987 ± 0.0668	0.1256 ± 0.0468	0.1189	0.0714 ± 0.1252	0.1226 ± 0.0876
TPR_AA (↑)	0.491 ± 0.1109	0.5689 ± 0.0935	0.5609	0.5227 ± 0.0626	0.7474 ± 0.0581
TPR_C (↑)	0.4044 ± 0.0859	0.3053 ± 0.0684	0.4296	0.463 ± 0.0638	0.5976 ± 0.0658
ΔTPR	0.0866 ± 0.1381	0.2636 ± 0.1164	0.1313	0.0597 ± 0.0891	0.1498 ± 0.0875
FPR_AA (↓)	0.4693 ± 0.0908	0.3626 ± 0.0786	0.3821	0.5454 ± 0.0674	0.376 ± 0.1254
FPR_C (↓)	0.3474 ± 0.0983	0.1941 ± 0.0555	0.2457	0.3519 ± 0.0519	0.3058 ± 0.0371
ΔFPR	0.1219 ± 0.1332	0.1685 ± 0.0965	0.1364	0.1935 ± 0.0857	0.0702 ± 0.1307

Table 4.17 provides us with the following key observations:

1. **Without Fairness Constraints:** The difference in PPV between African American and Caucasian groups increases significantly when fairness is not enforced, showing a higher discrepancy (ΔPPV = 0.1256 and 0.1226 for the two settings) compared to the original data (ΔPPV = 0.1189). This exacerbates the existing bias.

2. **With Fairness Constraints:** shows a reduction in the PPV disparity (ΔPPV = 0.0987 and 0.0714), i.e., (ΔPPV < 0.1), suggesting that fairness interventions are effective in generating fair data in both real world and controlled synthetic environment scenario.

3. **Accuracy Trade-offs:** Implementing fairness constraints leads to a decrease in overall accuracy across all settings. Because, the need to balance predictive outcomes across racial groups can dilute the model's ability to achieve the highest possible accuracy.

4. **TPR disparity in TS-TR approach:** DP-based constraints reduced TPR disparity among both racial groups. For settings like TS-TR, where testing occurs on biased real data, the decrease in TPR for African-Americans, although initially seeming negative, is justified because the real-world dataset exhibits bias that results in disproportionately higher positive outcomes for African Americans. When fairness constraints are

introduced, they attempt to equalize the predictive outcomes between different racial groups, which can result in a decrease in TPR for African Americans. This decrease is an intended effect of reducing the bias that causes inflated positive outcomes for this group.

5. **TPR disparity in TS-TS approach** In contrast, the controlled synthetic environment reveals that fairness constraints reduce disparities without the same level of pre-existing real-world bias like the former case. While the reduction in ΔTPR here also points to successful fairness interventions, the overall lower TPR for both groups (compared to the non-fairness scenario) suggests that the model is more conservative in predicting positives under fairness constraints. Such a decrease in disparity (ΔTPR) between the groups, while reflective of a more balanced model, also indicates that the model's sensitivity is adjusted down for the unprivileged group to bring it in line with the privileged group. This is an essential trade-off in fairness implementations, aiming to reduce preferential biases but also affecting the overall detection rate of true positives.

6. **Overcompensation Effects of Fairness Constraints on FPR:** Fairness constraints increased FPR for African Americans in both TS-TR and TS-TS scenarios compared to the no fairness prompting, suggesting overcompensation while balancing PPV. Thus, fairness interventions, while reducing one form of bias, might inadvertently introduce or fail to adequately address another.

4.3.4.2 Equal Opportunity

Table 4.18 Subgroup level analysis on equal opportunity based prompt using decision tree

Utility Measure	TS-TR with fairness	TS-TR without fairness	TR-TR	TS-TS with fairness	TS-TS without fairness
Acc_AA (↑)	0.5432 ± 0.0399	0.6016 ± 0.0308	0.5882	0.5576 ± 0.0811	0.7077 ± 0.0645
Acc_C (↑)	0.5389 ± 0.0517	0.6097 ± 0.016	0.627	0.5795 ± 0.0317	0.6496 ± 0.0396
ΔAcc	0.0043 ± 0.0651	-0.0081 ± 0.0347	-0.0388	-0.0219 ± 0.0871	0.0581 ± 0.0757
PPV_AA (↑)	0.5612 ± 0.0345	0.6333 ± 0.0326	0.6159	0.5982 ± 0.0885	0.7565 ± 0.0782
PPV_C (↑)	0.4235 ± 0.0544	0.5077 ± 0.0337	0.4957	0.5789 ± 0.0834	0.6339 ± 0.0401
ΔPPV	0.1377 ± 0.0644	0.1256 ± 0.0468	0.1189	0.0193 ± 0.1219	0.1226 ± 0.0876
TPR_AA (↑)	0.5590 ± 0.0998	0.5689 ± 0.0935	0.5609	0.6071 ± 0.0997	0.7474 ± 0.0581
TPR_C (↑)	0.4641 ± 0.0489	0.3053 ± 0.0684	0.4296	0.5643 ± 0.0689	0.5976 ± 0.0658
ΔTPR	0.0949 ± 0.1114	0.2636 ± 0.1164	0.1313	0.0428 ± 0.1213	0.1498 ± 0.0875
FPR_AA (↓)	0.4740 ± 0.0629	0.3626 ± 0.0786	0.3821	0.5099 ± 0.0862	0.376 ± 0.1254
FPR_C (↓)	0.4128 ± 0.0796	0.1941 ± 0.0555	0.2457	0.4108 ± 0.0827	0.3058 ± 0.0371
ΔFPR	0.0612 ± 0.1007	0.1685 ± 0.0965	0.1364	0.0991 ± 0.1194	0.0702 ± 0.1307

The observations from Table 4.18 are summarized as follows:

1. **Impact on TPR:** The TPR for African-Americans decreases in both TS-TR and TS-TS settings after applying fairness constraints, with the disparity in TPR (ΔTPR) narrowing to 0.0949 and 0.0428, respectively. This reduction shows that Equal Opportunity principles are effectively integrated into the synthetic data generation. In contrast, the ΔTPR in the TR-TR setting is 0.1313 but escalates to 0.2636 in TS-TR when synthetic data is generated without fairness guidelines. This increase highlights how the lack of fairness considerations can amplify existing biases in the original data.

2. **Effect on FPR:** Fairness constraints slightly increase the FPR for both racial groups but effectively lowers the disparity compared to scenarios without fairness measures. This adjustment illustrates that fairness interventions can uniformly increase the rate of false positives, potentially leading to overcompensations, yet they manage to maintain a reduced disparity in FPR.

3. **Accuracy versus Fairness Trade-off:** The overall accuracy of the model decreases, which underscores the common trade-off between maximizing model accuracy and achieving fairness.

4.3.4.3 Equalized Odds

Table 4.19 Subgroup level analysis on equalized odds based prompt using decision tree

Utility Measure	TS-TR with fairness	TS-TR without fairness	TR-TR	TS-TS with fairness	TS-TS without fairness
Acc_AA (↑)	0.5399 ± 0.0174	0.6016 ± 0.0308	0.5882	0.5842 ± 0.0333	0.7077 ± 0.0645
Acc_C (↑)	0.5479 ± 0.0409	0.6097 ± 0.016	0.627	0.5515 ± 0.0418	0.6496 ± 0.0396
ΔAcc	-0.0080 ± 0.0444	-0.0081 ± 0.0347	-0.0388	0.0327 ± 0.0534	0.0581 ± 0.0757
PPV_AA (↑)	0.5660 ± 0.0201	0.6333 ± 0.0326	0.6159	0.6056 ± 0.0857	0.7565 ± 0.0782
PPV_C (↑)	0.4299 ± 0.0406	0.5077 ± 0.0337	0.4957	0.5369 ± 0.0431	0.6339 ± 0.0401
ΔPPV	0.1361 ± 0.0455	0.1256 ± 0.0468	0.1189	0.0687 ± 0.0955	0.1226 ± 0.0876
TPR_AA (↑)	0.5153 ± 0.0892	0.5689 ± 0.0935	0.5609	0.6004 ± 0.0444	0.7474 ± 0.0581
TPR_C (↑)	0.4471 ± 0.0844	0.3053 ± 0.0684	0.4296	0.5487 ± 0.0492	0.5976 ± 0.0658
ΔTPR	0.0682 ± 0.1228	0.2636 ± 0.1164	0.1313	0.0517 ± 0.0660	0.1498 ± 0.0875
FPR_AA (↓)	0.4332 ± 0.0822	0.3626 ± 0.0786	0.3821	0.4295 ± 0.0421	0.376 ± 0.1254
FPR_C (↓)	0.3872 ± 0.0961	0.1941 ± 0.0555	0.2457	0.4521 ± 0.0707	0.3058 ± 0.0371
ΔFPR	0.0460 ± 0.1259	0.1685 ± 0.0965	0.1364	-0.0226 ± 0.0825	0.0702 ± 0.1307

Table 4.19 reveals the following:

1. **Impact of TPR and FPR:** In TS-TR environment, both TPR and FPR for African-Americans and Caucasians show significantly reduced disparities compared to settings without fairness constraints indicating a successful application of Equalized Odds in the data generation process, aiming for equal true and false positive rates across groups.

2. **Accuracy and PPV:** Despite the consistent disparities in Accuracy and PPV between racial groups under the Equalized Odds framework, individual accuracy and PPV for each group decrease. This suggests that while Equalized Odds can promote fairness by balancing disparities, it may also compromise the model's overall performance and its ability to accurately predict positive outcomes.

4.3.4.4 Causal Fairness

Causal fairness is defined by the principle that predictions should remain consistent for individuals with similar relevant characteristics, regardless of their sensitive attributes. To assess causal fairness within racial subgroups, we adopted counterfactual fairness analysis in Table 4.20. This process involves generating counterfactual scenarios by altering the sensitive attribute (for instance, changing race from African American to Caucasian) while keeping all other features unchanged. We then compare the model's predictions across these scenarios. The key metric is the proportion of instances where the model's prediction shifts solely due to the change in the sensitive attribute. A lower proportion of such changes indicates a higher level of causal fairness.

Table 4.20 Counterfactual fairness analysis

Classifiers	**Causal Fairness**
Random Forest	0.1854 (± 0.044)
XGBoost	0.0 (± 0.0)
Logistic Regression	0.1276 (± 0.057)
Decision Tree	0.1594 (± 0.1285)
SVM	0.0163 (± 0.0081)

Table 4.20 reveals notable variations in their reliance on sensitive attributes across classifiers. Random Forest and Decision Tree exhibits the highest causal fairness score, indicating a significant dependence on sensitive features, leading to less fairness compared to other models. In contrast, XGBoost achieves perfect counterfactual fairness with a score of 0.0 (± 0.0) making it the fairest model in this context. SVM also demonstrates minimal dependence ranking as one of the fairer models, albeit not as fair as XGBoost while Logistic Regression shows a moderate level of dependence. This overall analysis underscores the trade-offs between fairness and model complexity, with simpler models like XGBoost and SVM generally exhibiting greater fairness.

To ensure that causal fairness does not lead to overall performance degradation, Table 4.21 displays accuracy anlaysis among subgroup levels revealing minimal accuracy trade-offs across subgroups compared to scenarios without fairness constraints. This indicates that

Table 4.21 Subgroup accuracy analysis with and without Causal Fairness notion across classifiers

Classifiers	Causal Fairness			No Fairness Notion		
	Acc_AA	Acc_C	ΔAcc	Acc_AA	Acc_C	ΔAcc
Random Forest	0.5829 ± 0.0255	0.5865 ± 0.0429	-0.0036 (± 0.0499)	0.613 ± 0.0396	0.6318 ± 0.0181	-0.0188 (± 0.0435)
XGBoost	0.5698 ± 0.0327	0.5833 ± 0.0622	-0.0135 (± 0.0703)	0.615 ± 0.0306	0.636 ± 0.0106	-0.0210 (± 0.0324)
Logistic Regression	0.6275 ± 0.0274	0.6344 ± 0.0246	-0.0069 (± 0.0368)	0.6475 ± 0.0259	0.6493 ± 0.017	-0.0018 (± 0.0310)
Decision Tree	0.5451 ± 0.0264	0.5676 ± 0.0404	-0.0225 (± 0.0483)	0.6016 ± 0.0308	0.6097 ± 0.016	-0.0081 (± 0.0347)
SVM	0.5777 ± 0.0436	0.5766 ± 0.0856	0.0011 (± 0.0961)	0.6392 ± 0.0226	0.6453 ± 0.0268	-0.0061 (± 0.0351)

among the examined fairness notions, causal fairness is particularly effective in balancing the trade-off between fairness and performance.

4.3.4.5 Fairness through Unawareness (FTU)

To assess FTU, each prediction model was evaluated by training it twice: once with the sensitive attribute (e.g., race) included as a feature, and once without it. The comparison focused on key utility metrics such as accuracy, PPV, TPR, FPR. By observing the changes in these metrics with and without the sensitive attribute S, we analyze the extent to which the model's predictions were influenced by that attribute (in Table 4.22). Smaller changes in these metrics indicate stronger adherence of the generated synthetic data to the FTU principle mentioned in the prompt suggesting that the model does not rely heavily on the sensitive attribute in making predictions.

Table 4.22 Subgroup level analysis on various classifiers with and without sensitive attributes

	Acc_AA (↑)	Acc_C (↑)	PPV_AA (↑)	PPV_C (↑)	TPR_AA (↑)	TPR_C (↑)	FPR_AA (↓)	FPR_C (↓)
Decision Tree								
with S	0.5414 ± 0.0499	0.5361 ± 0.0574	0.5687 ± 0.0560	0.4242 ± 0.0481	0.5170 ± 0.0882	0.4451 ± 0.0673	0.4319 ± 0.1008	0.4053 ± 0.1356
w/o S	0.5475 ± 0.0515	0.5355 ± 0.076	0.5768 ± 0.0618	0.4265 ± 0.066	0.5093 ± 0.0863	0.4311 ± 0.0562	0.4108 ± 0.0971	0.3972 ± 0.1607
diff (d)	−0.0061 ± 0.0718	0.0006 ± 0.0951	−0.0081 ± 0.0831	−0.0023 ± 0.0812	0.0077 ± 0.1236	0.0140 ± 0.0877	0.0211 ± 0.1394	0.0081 ± 0.2082
Logistic Regression								
with S	0.6229 ± 0.0391	0.6042 ± 0.0995	0.6565 ± 0.0436	0.5425 ± 0.0771	0.6347 ± 0.1677	0.5024 ± 0.2809	0.3805 ± 0.1841	0.3302 ± 0.3372
w/o S	0.6166 ± 0.0417	0.6171 ± 0.075	0.651 ± 0.061	0.5502 ± 0.0695	0.6432 ± 0.2072	0.4985 ± 0.2469	0.4124 ± 0.2748	0.3064 ± 0.2776
diff (d)	0.0063 ± 0.0562	−0.0129 ± 0.1236	0.0055 ± 0.0753	−0.0077 ± 0.1044	−0.0085 ± 0.2655	0.0039 ± 0.3704	−0.0319 ± 0.3367	0.0238 ± 0.4362
Random Forest								
with S	0.5490 ± 0.0435	0.5924 ± 0.0172	0.5754 ± 0.0434	0.4814 ± 0.0214	0.5551 ± 0.0910	0.4364 ± 0.0896	0.4577 ± 0.1463	0.3070 ± 0.0821
w/o S	0.58 ± 0.0141	0.5627 ± 0.0487	0.6076 ± 0.0257	0.449 ± 0.0744	0.6076 ± 0.0257	0.4505 ± 0.0446	0.4045 ± 0.0966	0.364 ± 0.1138
diff (d)	−0.0310 ± 0.0454	0.0297 ± 0.0514	−0.0322 ± 0.0503	0.0324 ± 0.0775	−0.0525 ± 0.0944	−0.0141 ± 0.0998	0.0532 ± 0.1752	−0.0570 ± 0.1409
SVM								
with S	0.5791 ± 0.0481	0.5768 ± 0.1091	0.6033 ± 0.0600	0.5024 ± 0.0782	0.6586 ± 0.2337	0.5689 ± 0.2693	0.5078 ± 0.2984	0.4182 ± 0.3396
w/o S	0.5798 ± 0.0492	0.5753 ± 0.11	0.6045 ± 0.0615	0.501 ± 0.0808	0.6586 ± 0.2325	0.501 ± 0.0808	0.5062 ± 0.2991	0.42 ± 0.3396
diff (d)	−0.0007 ± 0.0690	0.0015 ± 0.1543	−0.0012 ± 0.0847	0.0014 ± 0.1123	0.0000 ± 0.3301	0.0679 ± 0.2803	0.0016 ± 0.4224	−0.0018 ± 0.4803
XGBoost								
with S	0.5583 ± 0.0371	0.5652 ± 0.0490	0.5867 ± 0.0436	0.4576 ± 0.0462	0.5440 ± 0.0636	0.4617 ± 0.1042	0.4261 ± 0.1111	0.3681 ± 0.1399
w/o S	0.5712 ± 0.0301	0.5673 ± 0.0638	0.5987 ± 0.0408	0.4636 ± 0.0619	0.5602 ± 0.0857	0.4699 ± 0.0749	0.4169 ± 0.109	0.37 ± 0.1461
diff (d)	−0.0129 ± 0.0480	−0.0021 ± 0.0806	−0.0120 ± 0.0603	−0.0060 ± 0.0766	−0.0162 ± 0.1068	−0.0082 ± 0.1288	0.0092 ± 0.1565	−0.0019 ± 0.2023

Table 4.22 reveals the following observations across all classifiers.

1. The differences in accuracy and PPV values between models trained with and without sensitive attributes are relatively small across all classifiers.

2. TPR and FPR differences between models trained with and without sensitive attributes show some variability, particularly in models like Random Forest and Logistic Regression. However, this increase is not uniform across all classifiers, indicating some models' ability to correctly identify TPR and FPR is largely preserved even without sensitive attributes.

3. Overall, GPT appears to have successfully generated synthetic data that aligns with the FTU principle, as the models trained without the sensitive attribute still perform comparably to those trained with it.

4.3.4.6 Generic Fairness

When measuring fairness across the overall synthetic dataset, the Generic Fairness prompt yields the most favorable results, significantly reducing DPD, EOD, and EOP. However, this improvement in fairness comes at the cost of the synthetic data's quality. A more detailed subgroup-level analysis, as shown in Table 4.23, reveals a nuanced picture: while overall disparities in accuracy, PPV, TPR, and FPR are reduced under the Generic Fairness prompt—leading to optimistic results in the overall fairness analysis (Table 4.16)—there is a notable trade-off. Specifically across both TS-TR and TS-TS scenarios, utility measures like accuracy, PPV, and TPR decrease across individual racial subgroups, and FPR increases, indicating that the fairness intervention, while successful in reducing bias at a high level, compromises the model's performance within each subgroup. This trade-off highlights the complexity of fairness interventions, where improvements in overall fairness may come at the expense of utility in specific demographic groups.

Table 4.23 Subgroup level analysis based on generic fairness

Utility Measure	TS-TR with fairness	TS-TR without fairness	TR-TR	TS-TS with fairness	TS-TS without fairness
Acc_AA (↑)	0.5160 ± 0.0257	0.6016 ± 0.0308	0.5882	0.5469 ± 0.0641	0.7077 ± 0.0645
Acc_C (↑)	0.5271 ± 0.0321	0.6097 ± 0.016	0.627	0.5156 ± 0.0295	0.6496 ± 0.0396
ΔAcc	-0.0111 ± 0.0411	-0.0081 ± 0.0347	-0.0388	0.0313 ± 0.0936	0.0581 ± 0.0757
PPV_AA (↑)	0.5441 ± 0.0386	0.6333 ± 0.0326	0.6159	0.5772 ± 0.0368	0.7565 ± 0.0782
PPV_C (↑)	0.4098 ± 0.0225	0.5077 ± 0.0337	0.5299	0.5022 ± 0.0625	0.6339 ± 0.0401
ΔPPV	0.1343 ± 0.0447	0.1256 ± 0.0468	0.086	0.0750 ± 0.0993	0.1226 ± 0.0876
TPR_AA (↑)	0.4885 ± 0.0820	0.5689 ± 0.0935	0.5609	0.5539 ± 0.0946	0.7474 ± 0.0581
TPR_C (↑)	0.4515 ± 0.0552	0.3053 ± 0.0684	0.4296	0.5022 ± 0.0625	0.5976 ± 0.0658
ΔTPR	0.0370 ± 0.0988	0.2636 ± 0.1164	0.1313	0.0517 ± 0.1571	0.1498 ± 0.0875
FPR_AA (↓)	0.4540 ± 0.1094	0.3626 ± 0.0786	0.3821	0.4566 ± 0.0772	0.376 ± 0.1254
FPR_C (↓)	0.4241 ± 0.0856	0.1941 ± 0.0555	0.2457	0.4686 ± 0.0502	0.3058 ± 0.0371
ΔFPR	0.0299 ± 0.1389	0.1685 ± 0.0965	0.1364	-0.0120 ± 0.1274	0.0702 ± 0.1307

Key Takeaways

1. DP, EOP and EOD based fairness prompts effectively reduce TPR and FPR disparities between racial groups, promoting fairness but at the cost of reduced accuracy and PPV. Overcompensation, particularly in FPR for underrepresented groups, is a common downside, highlighting the classic trade-off between fairness and model performance (Table 4.17, 4.18, 4.19).

2. Causal and FTU based prompts minimize the impact of sensitive attributes on predictions, with Causal Fairness showing a strong balance between reducing bias and preserving accuracy (Table 4.20, 4.21, 4.22).

3. FTU maintains fairness without using sensitive attributes but shows some variability in TPR and FPR, making it a strong, albeit slightly less consistent, option (Table 4.22).

4. Generic approach broadly reduces bias but at the cost of significantly lowering accuracy, TPR, and PPV, making it less effective where high predictive performance is also required (Table 4.23).

Chapter 5

Discussion and Conclusion

In Section 5.1, we summarize the key takeaways that address each research question, highlighting the critical conclusions drawn from our analysis; Section 5.2 discusses the limitations and outlines directions for future work, while Section 5.3 provides concluding remarks.

5.1 Key Takeaways

RQ1: Impact of Data

How do the characteristics of data used in prompts influence LLM-driven synthetic data generation, specifically considering (i) the impact of the number of in-context samples provided, and (ii) the effects of the sampling method employed, such as random versus biased sampling?

Using 20-40 in-context (IC) samples is optimal for generating realistic and efficient synthetic data using GPT, while balanced sampling effectively reduces bias.

Higher IC sample counts don't necessarily improve data quality generated by GPT and can amplify biases without fairness constraints. Balanced sampling, which equally represents outcomes and sensitive attributes, consistently reduces bias, while random or biased sampling can either maintain or exacerbate existing disparities.

RQ2: Bias Mitigation

Can biases in synthetic data be mitigated by using effective prompting strategies that incorporate fairness rules or constraints while maintaining the real data distribution and feature correlation intact? This question investigates whether LLMs can comprehend and implement fairness criteria when guided by such prompts.

Yes, biases in synthetic data can be mitigated using fairness-oriented prompts, though this often results in a trade-off with predictive accuracy.

Our analysis shows that while most fairness measures achieve equitable outcomes, this improvement in fairness comes with a significant reduction in performance. Efforts to balance fairness across racial subgroups can worsen individual predictive metrics like TPR, FPR, and PPV. Such poor performance observed in the Train-on-Synthetic, Test-on-Real (TS-TR) setting is largely due to the misalignment between the fairness-focused synthetic data used for training and the biased real-world data used for testing. In such cases, classification models might produce fair predictions that conflict with the biased ground truth labels in the real data, resulting in seemingly poor performance metrics. This highlights the limitations of using biased real data as a benchmark for fairness, suggesting Synthetic-to-Synthetic (TS-TS) evaluation or reannotation of the real dataset before using it for evaluation.

RQ3: Interaction with a Downstream Model

Do the biases present in synthetic tabular data exacerbate when classified using downstream machine learning models? This question assesses the fairness-related challenges associated with deploying LLM-generated synthetic tabular data and utilizing it for downstream model prediction.

Yes, biases in synthetic tabular data are indeed exacerbated when classified using downstream machine learning models.

This amplification occurs because classifiers, particularly more complex ones like Random Forest and XGBoost, account for inter-feature interactions that can intensify existing biases as they optimize for accuracy. In contrast, simpler models like Decision Tree exhibit better fairness, as they are less likely to capture and amplify these complex biases.

Based on both TS-TR and TS-TS based evaluation it can be concluded -

1. Causal Fairness based prompt is the most promising, balancing fairness with performance.

2. FTU (Fairness through Unawareness) offers a good alternative by reducing reliance on sensitive attributes.

3. DP, EOP, and EOD are effective for fairness but often reduce accuracy

4. Generic Fairness significantly compromises predictive performance to achieve broader fairness improvements, ultimately resulting in superior fairness compared to other approaches.

5.2 Limitations and Future Work

While our study addresses several key aspects, we recognize certain limitations that could be explored in future research.

1. Our study focuses exclusively on the GPT-4 model. Therefore, our conclusions are specific to GPT-4 and cannot be generalized to other LLMs, which may exhibit varying behaviors in terms of data generation, bias mitigation, and interaction with downstream models. This model-specific focus limits the broader applicability of our findings.

2. The impact of data characteristics, like the number of in-context samples and sampling methods, may differ based on the specific dataaset, context or domain. Since we only used the COMPAS dataset, our results may not be applicable to other types of data or use cases.

3. Although we aimed to address five widely used and significant type of bias and fairness discussed in the literature, we did not cover every possible form of bias comprehensively.

5.3 Concluding Remarks

This study aims to mitigate biases in tabular datasets through equitable synthetic data generation, promoting fair decision-making leveraging GPT-4's emergent abilities. In summary, while GPT-4 shows potential in generating fair synthetic data, achieving an optimal balance between fairness and accuracy remains a challenge.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

Mahed Abroshan, Mohammad Mahdi Khalili, and Andrew Elliott. Counterfactual fairness in synthetic data generation. In *NeurIPS Workshop on Synthetic Data for Empowering ML Research*, 2022.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification, 2018. URL https://arxiv.org/abs/1803.02453.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

Anita M Alessandra. When doctrines collide: Disparate treatment, disparate impact, and watson v. fort worth bank & trust. *U. Pa. L. Rev.*, 137:1755, 1988.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1, 2023.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. URL https://arxiv.org/abs/1701.07875.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.

Keira Behal, Jiayi Chen, Caleb Fikes, and Sophia Xiao. Mcrage: Synthetic healthcare data for fairness. *arXiv preprint arXiv:2310.18430*, 2023.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arxiv preprint. *arXiv preprint arXiv:1810.01943*, 2018.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Guanqun Bi, Lei Shen, Yuqiang Xie, Yanan Cao, Tiangang Zhu, and Xiaodong He. A group fairness lens for large language models. *arXiv preprint arXiv:2312.15478*, 2023.

Simon Bing, Andrea Dittadi, Stefan Bauer, and Patrick Schwab. Conditional generation of medical time series for extrapolation to underrepresented populations. *PLOS Digital Health*, 1(7):e0000074, 2022.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

V Borisov, T Leemann, K Seßler, J Haug, M Pawelczyk, and G Kasneci. Deep neural networks and tabular data: A survey. arxiv 2021. *arXiv preprint arXiv:2110.01889*.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40, 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020. URL https://arxiv.org/abs/2010.04053.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Bhushan Chaudhari, Himanshu Chaudhary, Aakash Agarwal, Kamna Meena, and Tanmoy Bhowmik. Fairgen: Fair synthetic data generation. *arXiv preprint arXiv:2210.13023*, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Nuo Chen, Linjun Shou, Ming Gong, Jian Pei, Chenyu You, Jianhui Chang, Daxin Jiang, and Jia Li. Bridge the gap between language models and tabular understanding. *arXiv preprint arXiv:2302.09302*, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Wenhu Chen. Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv:2210.06710*, 2022.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Cheng Cheng, Beitong Zhou, Guijun Ma, Dongrui Wu, and Ye Yuan. Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data. *Neurocomputing*, 409:35–45, 2020.

Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. Few-shot fairness: Unveiling llm's potential for fairness-aware classification. *arXiv preprint arXiv:2402.18502*, 2024.

Andrei-Victor Chisca, Andrei-Cristian Rad, and Camelia Lemnaru. Prompting fairness: Learning prompts for debiasing large language models. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 52–62, 2024.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Ashish Dandekar, Remmy AM Zen, and Stéphane Bressan. Comparative evaluation of synthetic data generation methods. In *Proceedings of ACM Conference (Deep Learning Security Workshop)*, 2017.

Max Daniels. Statistical distances and their implications to gan training. https://qnkxsovc.gitlab.io/prob-vis/, 2014. Accessed: 2024-08-10.

Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022a. doi: 10.1109/ACCESS.2022.3144765.

Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022b.

Sajad Darabi and Yotam Elor. Synthesising multi-modal minority samples for tabular data. *arXiv preprint arXiv:2105.08204*, 2021.

Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-Vincentelli, and Costas J Spanos. Conditional synthetic data generation for robust machine learning applications with limited pandemic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11792–11800, 2022.

Xolani Dastile, Turgay Celik, and Moshe Potsane. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, 2020.

Heng Deng, Qian Zhou, Ziwei Zhang, Taohu Zhou, Xiaoqing Lin, Yi Xia, Li Fan, and Shiyuan Liu. The current status and prospects of large language models in medical application and research. *Chinese Journal of Academic Radiology*, pages 1–9, 2024.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

Cem Dilmegani. Synthetic data vs real data: Benefits, challenges in 2023. https://research.aimultiple.com/synthetic-data-vs-real-data/, 2023. Accessed: 2024-09-03.

Ayesha Siddiqua Dina, AB Siddique, and D Manivannan. Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. *IEEE Access*, 10:96731–96747, 2022.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Mohsen Dorodchi, Erfan Al-Hossami, Aileen Benedict, and Elise Demeter. Using synthetic data generators to promote open science in higher education learning analytics. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4672–4675. IEEE, 2019.

Manh Khoi Duong and Stefan Conrad. Measuring and mitigating bias for tabular datasets with multiple protected attributes. *arXiv preprint arXiv:2405.19300*, 2024.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

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.

Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

Erica Espinosa and Alvaro Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Jane Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, Christos Faloutsos, et al. Large language models (llms) on tabular data: Prediction, generation, and understanding-a survey. 2024.

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*, pages 259–268, 2015.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.

Vincent Freiberger and Erik Buchmann. Fairness certification for natural language processing and large language models. In *Intelligent Systems Conference*, pages 606–624. Springer, 2024.

Yao Fu, Hao Peng, and Tushar Khot. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*, 2022.

Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510, 2017.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022a.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022b.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

Manbir Gulati and Paul Roysdon. Tabmt: Generating tabular data with masked transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Aman Gupta, Deepak Bhatt, and Anubha Pandey. Transitioning from real to synthetic data: Quantifying the bias in model. *arXiv preprint arXiv:2105.04144*, 2021.

Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *arXiv preprint arXiv:2407.14962*, 2024.

Lasse Hansen, Nabeel Seedat, Mihaela van der Schaar, and Andrija Petrovic. Reimagining synthetic tabular data generation through data-centric ai: A comprehensive benchmark. *Advances in Neural Information Processing Systems*, 36:33781–33823, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. Utility and privacy assessments of synthetic data for regression tasks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5763–5772. IEEE, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*, 2023.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*, 2020.

Stephen Hutt, Ryan S Baker, Michael Mogessie Ashenafi, Juan Miguel Andres-Bray, and Christopher Brooks. Controlled outputs, full data: A privacy-protecting infrastructure for mooc data. *British Journal of Educational Technology*, 53(4):756–775, 2022.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. Tabbie: Pretrained representations of tabular data. *arXiv preprint arXiv:2105.02584*, 2021.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024.

Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. Towards better serialization of tabular data for few-shot classification. *arXiv preprint arXiv:2312.12464*, 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, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*, 2024.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International Conference on Artificial Intelligence and Statistics*, pages 1288–1296. PMLR, 2024.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Andreas Jungherr. Using chatgpt and other large language model (llm) applications for academic paper assignments. 2023.

Hoang Anh Just, Feiyang Kang, Jiachen T Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*, 2023.

Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. The pursuit of fairness in artificial intelligence models: A survey. *arXiv preprint arXiv:2403.17333*, 2024.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*, 2022.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE, 2019.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed: 2024-08-25.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Xiaonan Li and Xipeng Qiu. Finding support examples for in-context learning. *arXiv preprint arXiv:2302.13539*, 2023.

Zheng Li, Yue Zhao, and Jialin Fu. Sync: A copula based framework for generating synthetic data from aggregated sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 571–578. IEEE, 2020.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Chien-Liang Liu and Po-Yen Hsieh. Model-based synthetic sampling for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1543–1556, 2019.

Qinyi Liu, Mohammad Khalil, Ronas Shakya, and Jelena Jovanovic. Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics, 2024a. URL https://arxiv.org/abs/2401.06883.

Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. Goggle: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. Investigating the fairness of large language models for predictions on tabular data. *arXiv preprint arXiv:2310.14607*, 2023b.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. Confronting LLMs with traditional ML: Rethinking the fairness of large language models in tabular classifications. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3603–3620, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.198. URL https://aclanthology.org/2024.naacl-long.198.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023c.

Udara Piyasena Liyanage and Nimnaka Dilshan Ranaweera. Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts. *Journal of Computational Social Dynamics*, 8(11):15–25, 2023.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33:11237–11247, 2020.

Tamas Madl, Weijie Xu, Olivia Choudhury, and Matthew Howard. Approximate, adapt, anonymize (3a): A framework for privacy preserving training data release for machine learning. *arXiv preprint arXiv:2307.01875*, 2023.

Data Privacy Manager. Meta hit with record €1.2b gdpr fine – data privacy manager, 2023. URL https://dataprivacymanager.net/meta-hit-with-record-e1-2b-gdpr-fine/. Accessed: 2024-09-03.

Dionysis Manousakas and Sergül Aydöre. On the usefulness of synthetic tabular data generation. *arXiv preprint arXiv:2306.15636*, 2023.

Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer, 2023.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data, 2021. URL https://arxiv.org/abs/2108.04978.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1538–1545, 2024.

Christopher Z Mooney. *Monte carlo simulation*. Number 116. Sage, 1997.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, and Christopher Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.

Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI, 2024.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

Ki Nohyun, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2022.

Soma Onishi and Shoya Meguro. Rethinking data augmentation for tabular data in deep learning. *arXiv preprint arXiv:2305.10308*, 2023.

OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2024-08-28.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 399–410. IEEE, 2016.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070, 2021.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Alexis Porter. Lessons learned from gdpr fines in 2023. *CPO Magazine*, 2023. URL https://www.cpomagazine.com/data-protection/lessons-learned-from-gdpr-fines-in-2023/. Accessed: 2024-09-03.

David Pujol, Amir Gilad, and Ashwin Machanavajjhala. Prefair: Privately generating justifiably fair synthetic data. *arXiv preprint arXiv:2212.10310*, 2022.

M Atif Qureshi, Arjumand Younus, and Simon Caton. Inclusive counterfactual generation: Leveraging llms in identifying online hate. In *International Conference on Web Engineering*, pages 34–48. Springer, 2024.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Trivellore E Raghunathan. Synthetic data. *Annual review of statistics and its application*, 8 (1):129–140, 2021.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction*, 4(2): 488–501, 2022.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. Distance correlation gan: Fair tabular data generation with generative adversarial networks. In *International Conference on Human-Computer Interaction*, pages 431–445. Springer, 2023.

Resmi Ramachandra, Md Fahim Sikder, David Bergström, and Fredrik Heintz. Bt-gan: Generating fair synthetic healthdata via bias-transforming generative adversarial networks. *Journal of Artificial Intelligence Research*, 79:1313–1341, 2024.

Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. A review of large language models and autonomous agents in chemistry. *arXiv preprint arXiv:2407.01603*, 2024.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

Francesco Rundo, Francesca Trenta, Agatino Luigi Di Stallo, and Sebastiano Battiato. Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24): 5574, 2019.

Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE access*, 9:135392–135422, 2021.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Rick Sauber-Cole and Taghi M Khoshgoftaar. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1):98, 2022.

Lawrence Saul and Fernando Pereira. Aggregate and mixed-order markov models for statistical language processing. *arXiv preprint cmp-lg/9706007*, 1997.

Neil Savage. Synthetic data could be better than real data, Apr 2023. URL https://www.nature.com/articles/d41586-023-01445-8.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283:103238, 2020.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. Curated llm: Synergy of llms and data curation for tabular augmentation in ultra low-data regimes. *arXiv preprint arXiv:2312.12112*, 2023.

Ibrahim Mohamed Serouis and Florence Sèdes. Exploring large language models for bias mitigation and fairness. In *1st International Workshop on AI Governance (AIGOV) in conjunction with the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

Sakib Shahriar, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency, 2024. URL https://arxiv.org/abs/2407.09519.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*, 2021.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Md Fahim Sikder, Resmi Ramachandranpillai, Daniel de Leng, and Fredrik Heintz. Fairx: A comprehensive benchmarking tool for model analysis using fairness, utility, and explainability. *arXiv preprint arXiv:2406.14281*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Dylan Slack and Sameer Singh. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*, 2023.

Aivin V Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*, 2023.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*, 2024.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. Tap4llm: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *arXiv preprint arXiv:2312.09039*, 2023.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study, 2024. URL https://arxiv.org/abs/2305.13062.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. ACM, October 2021. doi: 10.1145/3465416.3483305. URL http://dx.doi.org/10.1145/3465416.3483305.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

Bo Tang and Haibo He. Kerneladasyn: Kernel based adaptive synthetic data generation for imbalanced learning. In *2015 IEEE congress on evolutionary computation (CEC)*, pages 664–671. IEEE, 2015.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Paul Tiwald, Alexandra Ebert, and Daniel T. Soukup. Representative fair synthetic data, 2021. URL https://arxiv.org/abs/2104.03007.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Boris van Breugel and Mihaela van der Schaar. Beyond privacy: Navigating the opportunities and challenges of synthetic data. *arXiv preprint arXiv:2304.03722*, 2023.

Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*, 2024.

Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

L Vivek Harsha Vardhan and Stanley Kok. Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders. In *Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37 th International Conference on Machine Learning*, 2020.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

VentureBeat. 89% of tech execs see synthetic data as a key to staying ahead, 2021. URL https://venturebeat.com/ai/89-of-tech-execs-see-synthetic-data-as-a-key-to-staying-ahead. Accessed: 2024-04-20.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7, 2018a. doi: 10.1145/3194770.3194776.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018b.

Mark Vero, Mislav Balunovic, and Martin Vechev. Cuts: Customizable tabular synthetic data generation. In *Forty-first International Conference on Machine Learning*.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 2022.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019a.

Rui Wang, Pengyu Cheng, and Ricardo Henao. Toward fairness in text generation via mutual information minimization based on importance sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 4473–4485. PMLR, 2023.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319, 2019b.

Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. Harmonic: Harnessing llms for tabular data synthesis and privacy protection. *arXiv preprint arXiv:2408.02927*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Bingyang Wen, Yupeng Cao, Fan Yang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Modeling tabular data using causally-aware gan. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Isabella C Wiest, Marie-Elisabeth Lessmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. Anonymizing medical documents with local, privacy preserving large language models: The llm-anonymizer. *medRxiv*, pages 2024–06, 2024.

Wikipedia contributors. Synthetic data — wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Synthetic_data, 2023. Accessed: 2024-06-26.

T Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pages 570–575. IEEE, 2018.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019a.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019b.

Shengzhe Xu, Cho-Ting Lee, Mandar Sharma, Raquib Bin Yousuf, Nikhil Muralidhar, and Naren Ramakrishnan. Are llms naturally good at synthetic tabular data generation?, 2024. URL https://arxiv.org/abs/2406.14541.

Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. Detime: Diffusion-enhanced topic modeling using encoder-decoder based llm. *arXiv preprint arXiv:2310.15296*, 2023a.

Weijie Xu, Jinjin Zhao, Francis Iannacci, and Bo Wang. Ffpdg: Fast, fair and private data generation. *arXiv preprint arXiv:2307.00161*, 2023b.

Zeyu Yang, Peikun Guo, Khadija Zanna, and Akane Sano. Balanced mixed-type tabular data synthesis with diffusion models. *arXiv preprint arXiv:2404.08254*, 2024.

Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In *International Conference on Machine Learning*, pages 5699–5707. PMLR, 2018.

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.

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.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023a.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, 2023b.

Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. Generative table pre-training empowers models for tabular prediction. *arXiv preprint arXiv:2305.09696*, 2023c.

Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. Large language models are complex table parsers. *arXiv preprint arXiv:2312.11521*, 2023a.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021a.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021b.

Zilong Zhao, Robert Birke, and Lydia Chen. Tabula: Harnessing language models for tabular data synthesis. *arXiv preprint arXiv:2310.12746*, 2023c.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

Appendix A

Appendix

Table A.1 Performance and Fairness Evaluation using Logistic Regression with Various Fairness Notions

Fairness Notion	**Performance Measure**			**Fairness Measure**		
	Acc (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
DP	0.5828 (± 0.0432)	0.533 (± 0.0606)	0.6176 (± 0.0591)	0.384 (± 0.2757)	0.3949 (± 0.2675)	0.3876 (± 0.2703)
EOP	0.6086 (± 0.0385)	**0.6219 (± 0.0373)**	**0.6702 (± 0.038)**	0.3715 (± 0.1805)	0.3713 (± 0.172)	0.3497 (± 0.1642)
EOD	0.596 (± 0.0327)	0.5981 (± 0.0447)	0.6519 (± 0.0257)	0.275 (± 0.1792)	0.2773 (± 0.1835)	0.2453 (± 0.1565)
Causal	**0.6275 (± 0.033)**	0.5394 (± 0.0527)	0.6593 (± 0.036)	0.2391 (± 1.408)	0.2787 (± 0.1504)	0.2663 (± 0.1738)
Unawareness	0.6182 (± 0.0509)	0.5829 (± 0.0668)	0.6464 (± 0.0634)	0.1871 (± 0.0566)	0.2043 (± 0.2194)	0.1376 (± 0.2)
Generic	0.5682 (± 0.0748)	0.5887 (± 0.0676)	0.6116 (± 0.1035)	**0.1861 (± 0.2143)**	**0.2023 (± 0.068)**	**0.1323 (± 0.1499)**
No fairness notion	0.6482 (± 0.0217)	0.5553 (± 0.1022)	0.7099 (± 0.0135)	0.3027 (± 0.1152)	0.349 (± 0.1139)	0.349 (± 0.1139)
Real Data	0.6734	0.6351	0.7273	0.3217	0.3561	0.3561

Table A.2 Performance and Fairness Evaluation using Random Forest with Various Fairness Notions

Fairness Notion	**Performance Measure**			**Fairness Measure**		
	Acc (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
DP	0.5537 (± 0.0453)	0.5128 (± 0.0577)	0.5676 (± 0.0655)	0.2589 (± 0.2455)	0.2794 (± 0.2525)	0.2007 (± 0.3357)
EOP	0.5743 (± 0.0384)	**0.5483 (± 0.0494)**	0.5945 (± 0.0371)	0.1517 (± 0.1279)	0.1702 (± 0.1107)	0.1628 (± 0.1465)
EOD	0.5585 (± 0.0294)	0.5365 (± 0.0352)	0.577 (± 0.0412)	0.1744 (± 0.0995)	0.1914 (± 0.0783)	0.1385 (± 0.1053)
Causal	**0.5821 (± 0.0311)**	0.5296 (± 0.0296)	**0.5972 (± 0.0499)**	0.1782 (± 0.1274)	0.2003 (± 0.1041)	0.1897 (± 0.1233)
Unawareness	0.5692 (± 0.0286)	0.5287 (± 0.0291)	0.5833 (± 0.0356)	0.1486 (± 0.124)	0.1665 (± 0.1511)	0.1216 (± 0.0934)
Generic	0.5269 (± 0.04)	0.5128 (± 0.0464)	0.5364 (± 0.0514)	**0.0921 (± 0.075)**	**0.1315 (± 0.0752)**	**0.016 (± 0.1663)**
No fairness notion	0.6209 (± 0.0259)	0.5251 (± 0.1082)	0.6619 (± 0.0188)	0.2242 (± 0.1017)	0.2411 (± 0.1153)	0.2409 (± 0.123)
Real Data	0.6237	0.5958	0.6589	0.206	0.2362	0.1914

Table A.3 Performance and Fairness Evaluation using SVM with Various Fairness Notions

Fairness Notion	**Performance Measure**			**Fairness Measure**		
	Acc (↑)	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
DP	0.5505 (± 0.0693)	0.4865 (± 0.1294)	0.5781 (± 0.089)	0.1476 (± 0.0894)	0.1621 (± 0.0835)	0.0869 (± 0.1104)
EOP	0.5739 (± 0.045)	0.5712 (± 0.0734)	**0.6413 (± 0.0555)**	0.1532 (± 0.0687)	0.1825 (± 0.0516)	0.1137 (± 0.0727)
EOD	0.5663 (± 0.0387)	0.5938 (± 0.0595)	0.6032 (± 0.0385)	0.1452 (± 0.0739)	0.1667 (± 0.051)	0.0978 (± 0.0621)
Causal	**0.5782 (± 0.0698)**	0.4906 (± 0.065)	0.6107 (± 0.0776)	0.1233 (± 0.0709)	0.1541 (± 0.037)	0.1083 (± 0.0933)
Unawareness	0.5773 (± 0.0603)	0.5711 (± 0.0758)	0.5912 (± 0.1058)	0.1093 (± 0.0841)	0.1013 (± 0.0765)	0.0897 (± 0.071)
Generic	0.522 (± 0.0802)	**0.6057 (± 0.083)**	0.5341 (± 0.137)	**0.0747 (± 0.0722)**	**0.0862 (± 0.0826)**	**0.0356 (± 0.0627)**
No fairness notion	0.6416 (± 0.0232)	0.6267 (± 0.0437)	0.7123 (± 0.0159)	0.2545 (± 0.0674)	0.2502 (± 0.0631)	0.2397 (± 0.0697)
Real Data	0.6741	0.6365	0.7302	0.2989	0.3197	0.3197

Table A.4 Performance and Fairness Evaluation using XGBoost with Various Fairness Notions

Fairness Notion	**Performance Measure**			**Fairness Measure**		
	Acc	**F1** (↑)	**AUC** (↑)	**DPD** (↓)	**EOD** (↓)	**EOP** (↓)
DP	0.5541 (± 0.0601)	0.5022 (± 0.0783)	0.5666 (± 0.0876)	0.1986 (± 0.2664)	0.2072 (± 0.257)	0.1916 (± 0.2602)
EOP	0.561 (± 0.0405)	**0.5615 (± 0.0345)**	0.596 (± 0.0478)	0.2178 (± 0.0963)	0.2235 (± 0.1012)	0.1905 (± 0.0946)
EOD	0.5679 (± 0.0333)	0.5559 (± 0.0383)	0.5966 (± 0.0403)	0.167 (± 0.1016)	0.1834 (± 0.1168)	0.1404 (± 0.0942)
Causal	**0.5781 (± 0.03)**	0.5128 (± 0.0493)	**0.6084 (± 0.0389)**	0.1485 (± 0.1048)	0.1633 (± 0.0897)	0.156 (± 0.0988)
Unawareness	0.5751 (± 0.0379)	0.5246 (± 0.0298)	0.5851 (± 0.0525)	0.097 (± 0.1007)	0.1165 (± 0.0624)	0.0824 (± 0.1003)
Generic	0.527 (± 0.0307)	0.5025 (± 0.0359)	0.5351 (± 0.0419)	**0.0876 (± 0.0561)**	**0.1024 (± 0.09)**	**0.0073 (± 0.096)**
No fairness notion	0.6233 (± 0.0216)	0.5361 (± 0.0826)	0.6679 (± 0.0257)	0.2415 (± 0.1034)	0.2604 (± 0.1145)	0.2604 (± 0.1145)
Real Data	0.6616	0.6182	0.7018	0.2408	0.2617	0.2617

Table A.5 Subgroup level analysis on a decision tree classifier trained on synthetic data generated using different fairness notion

Fairness Notion	**Acc_AA** (↑)	**Acc_C** (↑)	**Recall_AA** (↑)	**Recall_C** (↑)	**FPR_AA** (↓)	**FPR_C** (↓)	**TNR_AA** (↑)	**TNR_C** (↑)
DP	0.566 ± 0.0427	0.6082 ± 0.0459	0.6458 ± 0.2043	0.2583 ± 0.1125	0.5212 ± 0.2539	0.1662 ± 0.1306	0.4788 ± 0.2539	0.8338 ± 0.1306
EOP	0.6016 ± 0.0307	0.6192 ± 0.0535	**0.8065 ± 0.1444**	0.4568 ± 0.1200	0.6221 ± 0.1844	0.2761 ± 0.1624	0.3779 ± 0.1844	0.7239 ± 0.1624
EOD	0.5874 ± 0.0257	0.6089 ± 0.0538	0.7351 ± 0.1628	0.4898 ± 0.1686	0.5739 ± 0.2267	0.3142 ± 0.1941	0.4261 ± 0.2267	0.6858 ± 0.1941
Causal	**0.6275 ± 0.0274**	**0.6344 ± 0.0246**	0.5551 ± 0.1075	0.2888 ± 0.1022	**0.3030 ± 0.0695**	**0.1427 ± 0.0869**	**0.6970 ± 0.0695**	**0.8573 ± 0.0869**
Unawareness	0.6229 ± 0.0391	0.6042 ± 0.0995	0.6347 ± 0.1677	0.5024 ± 0.2809	0.3805 ± 0.1841	0.3302 ± 0.3372	0.6195 ± 0.1841	0.6698 ± 0.3372
Generic	0.5693 ± 0.0577	0.5667 ± 0.1063	0.7206 ± 0.2071	**0.5830 ± 0.2145**	0.5960 ± 0.2881	0.4438 ± 0.2942	0.4040 ± 0.2881	0.5562 ± 0.2942
Real Data	0.6795	0.6641	0.7226	0.3665	0.3676	0.144	0.6324	0.856

Table A.6 Subgroup level analysis on random forest classifier trained on synthetic data generated using different fairness notion

Fairness Notion	**Acc_AA** (↑)	**Acc_C** (↑)	**Recall_AA** (↑)	**Recall_C** (↑)	**FPR_AA** (↓)	**FPR_C** (↓)	**TNR_AA** (↑)	**TNR_C** (↑)
DP	0.5416 ± 0.0429	0.5549 ± 0.0724	0.5643 ± 0.2083	0.3578 ± 0.1298	0.4833 ± 0.1813	0.318 ± 0.1976	0.5167 ± 0.1813	0.682 ± 0.1976
EOP	0.5730 ± 0.0385	0.5749 ± 0.0475	**0.6780 ± 0.1282**	0.4442 ± 0.0682	0.4651 ± 0.1121	0.3408 ± 0.1028	0.5349 ± 0.1121	0.6592 ± 0.1028
EOD	0.5458 ± 0.0376	0.5699 ± 0.0402	0.5831 ± 0.1039	0.4451 ± 0.0625	0.4949 ± 0.1439	0.3496 ± 0.0924	0.5051 ± 0.1439	0.6504 ± 0.0924
Causal	**0.5829 ± 0.0255**	**0.5924 ± 0.0172**	0.5624 ± 0.0796	0.3602 ± 0.0760	**0.3947 ± 0.0605**	**0.2676 ± 0.1165**	**0.6053 ± 0.0605**	**0.7324 ± 0.1165**
Unawareness	0.5490 ± 0.0435	0.5865 ± 0.0429	0.5551 ± 0.0910	0.4364 ± 0.0896	0.4577 ± 0.1463	0.3070 ± 0.0821	0.5423 ± 0.1463	0.6930 ± 0.0821
Generic	0.5243 ± 0.0530	0.5336 ± 0.0323	0.5370 ± 0.1280	0.5175 ± 0.1100	0.4896 ± 0.1524	0.4560 ± 0.1010	0.5104 ± 0.1524	0.5440 ± 0.1010
Real Data	0.6203	0.6061	0.6598	0.4515	0.4229	0.2942	0.5771	0.7058

Table A.7 Subgroup level analysis on SVM classifier trained on synthetic data generated using different fairness notion

Fairness Notion	**Acc_AA (↑)**	**Acc_C (↑)**	**Recall_AA (↑)**	**Recall_C (↑)**	**FPR_AA (↓)**	**FPR_C (↓)**	**TNR_AA (↑)**	**TNR_C (↑)**
DP	0.5452 ± 0.0581	0.5585 ± 0.09	0.5122 ± 0.2455	0.4252 ± 0.1713	0.4187 ± 0.2164	0.3556 ± 0.182	0.5813 ± 0.2164	0.6444 ± 0.182
EOP	0.5719 ± 0.0276	0.5768 ± 0.1091	**0.8195 ± 0.2376**	0.5544 ± 0.2242	0.5331 ± 0.2332	0.4088 ± 0.232	0.4699 ± 0.2332	0.5912 ± 0.232
EOD	0.5681 ± 0.0226	0.5635 ± 0.0673	0.7293 ± 0.1956	0.6316 ± 0.2116	0.6079 ± 0.2438	0.4804 ± 0.2248	0.3921 ± 0.2438	0.5196 ± 0.2248
Causal	**0.5791 ± 0.0481**	**0.5768 ± 0.0747**	0.4700 ± 0.0877	0.3617 ± 0.0807	**0.3046 ± 0.0650**	**0.2848 ± 0.1629**	**0.6954 ± 0.0650**	**0.7152 ± 0.1629**
Unawareness	0.5777 ± 0.0436	0.5766 ± 0.0856	0.6586 ± 0.2337	0.5689 ± 0.2693	0.5078 ± 0.2984	0.4182 ± 0.3396	0.4922 ± 0.2984	0.5818 ± 0.3396
Generic	0.5505 ± 0.0638	0.4790 ± 0.1090	0.6680 ± 0.1819	**0.7840 ± 0.2417**	0.7433 ± 0.3017	0.7177 ± 0.2959	0.2567 ± 0.3017	0.2823 ± 0.2959
Real Data	0.677	0.6698	0.7129	0.3932	0.3623	0.1518	0.6377	0.8482

Table A.8 Subgroup level analysis on XGBoost classifier trained on synthetic data generated using different fairness notion

Fairness Notion	**Acc_AA (↑)**	**Acc_C (↑)**	**Recall_AA (↑)**	**Recall_C (↑)**	**FPR_AA (↓)**	**FPR_C (↓)**	**TNR_AA (↑)**	**TNR_C (↑)**
DP	0.5456 ± 0.0512	0.5669 ± 0.0756	0.5498 ± 0.1981	0.3583 ± 0.0959	0.459 ± 0.1987	0.2986 ± 0.1737	0.541 ± 0.1987	0.7014 ± 0.1737
EOP	0.5652 ± 0.036	0.5833 ± 0.0622	**0.6701 ± 0.0739**	0.4485 ± 0.0622	0.5154 ± 0.0848	0.3064 ± 0.0659	0.4846 ± 0.0848	0.6936 ± 0.0659
EOD	0.5583 ± 0.0461	0.5825 ± 0.0236	0.6258 ± 0.1018	0.4854 ± 0.0653	0.5154 ± 0.1489	0.3549 ± 0.0685	0.4846 ± 0.1489	0.6451 ± 0.0685
Causal	**0.5698 ± 0.0327**	**0.5975 ± 0.0263**	0.5312 ± 0.0785	0.3752 ± 0.1098	**0.3881 ± 0.0654**	**0.2826 ± 0.1591**	**0.6119 ± 0.0654**	**0.7174 ± 0.1591**
Unawareness	0.5583 ± 0.0371	0.5652 ± 0.0490	0.5440 ± 0.0636	0.4617 ± 0.1042	0.4261 ± 0.1111	0.3681 ± 0.1399	0.5739 ± 0.1111	0.6319 ± 0.1399
Generic	0.5203 ± 0.0401	0.5372 ± 0.0302	0.5160 ± 0.0816	**0.5087 ± 0.1341**	0.4751 ± 0.1482	0.4444 ± 0.1336	0.5249 ± 0.1482	0.5556 ± 0.1336
Real Data	0.665	0.6565	0.6695	0.4078	0.3399	0.1831	0.6601	0.8169