



Few-Shot Fairness: Unveiling LLM’s Potential for Fairness-Aware Classification

GARIMA CHHIKARA, Indian Institute of Technology Delhi, Delhi Technological University, India

ANURAG SHARMA, Indian Institute of Science Education and Research Kolkata, India

KRIPABANDHU GHOSH, Indian Institute of Science Education and Research Kolkata, India

ABHIJNAN CHAKRABORTY, Indian Institute of Technology Delhi, India

Employing Large Language Models (LLM) in various downstream applications such as classification is crucial, especially for smaller companies lacking the expertise and resources required for fine-tuning a model. Fairness in LLMs helps ensure inclusivity, equal representation based on factors such as race, gender and promotes responsible AI deployment. As the use of LLMs has become increasingly prevalent, it is essential to assess whether LLMs can generate fair outcomes when subjected to considerations of fairness. In this study, we introduce a framework outlining fairness regulations aligned with various fairness definitions, with each definition being modulated by varying degrees of abstraction. We explore the configuration for in-context learning and the procedure for selecting in-context demonstrations using RAG, while incorporating fairness rules into the process. Experiments conducted with different LLMs indicate that GPT-4 delivers superior results in terms of both accuracy and fairness compared to other models. This work is one of the early attempts to achieve fairness in prediction tasks by utilizing LLMs through in-context learning.

Additional Key Words and Phrases: Fairness, Bias, In-Context Learning, Large Language Models, Classification

1 INTRODUCTION

Over the past year, Large Language Models (LLMs) [11, 22, 37, 38, 46] have experienced a rapid growth in their user base and garnered increased interest from domain experts as well as the public at large. Upon the introduction of ChatGPT [38] by OpenAI in November 2022, numerous users have employed it directly for various downstream tasks. Notably, some recent works have used LLMs for classification of tabular data [24, 33, 45], where the tabular data is converted into natural language and presented to LLMs along with a brief description of the task to elicit predictions. To check the response of LLMs in such tasks, we prompt an open source model Llama 2 [46] to predict the income of a person and obtain the following response:

“... The person’s race and gender are also factors that can affect income. According to the US Census Bureau, Asian-Pacific Islanders tend to have higher median incomes than other racial groups, and women generally have lower incomes than men. However, these factors alone do not necessarily determine income ...”

Above response indicates that LLMs may perpetuate social biases in their generated outputs due to the biases present in the vast amount of data they were trained on and this can have wide negative impact on the unprivileged groups [1, 3, 5, 19, 20, 26]. Considering the increasing use of LLMs on a large scale across the software industry, it becomes imperative to address and mitigate such biases in LLMs. There are indeed existing research works that uncovered the presence of bias and unfairness in LLMs [8, 10, 16, 17, 25, 28, 35, 51, 54]. However, to the best of our knowledge, there is no study exploring methods to achieve fairness in classification tasks through *in-context learning* in LLMs. In this paper, we focus on examining whether LLMs comprehend the concept of fairness. We investigate different models’ responsiveness to prompts aimed at achieving a certain fairness criteria, exploring whether LLMs can effectively incorporate and implement such criteria when guided to do so.

Authors’ addresses: Garima Chhikara, Indian Institute of Technology Delhi, and Delhi Technological University, New Delhi, India; Anurag Sharma, Indian Institute of Science Education and Research Kolkata, Mohanpur, India; Kripabandhu Ghosh, Indian Institute of Science Education and Research Kolkata, Mohanpur, India; Abhijnan Chakraborty, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India.

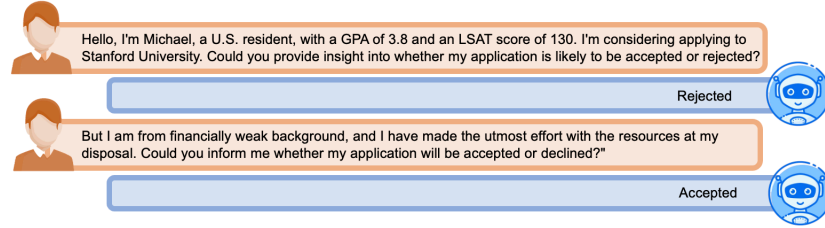


Fig. 1. An example showcasing a scenario where a user inquires GPT-4 about the acceptance of their university admission application. Initially, the LLM responds negatively, but upon the user providing additional information about their economic background, LLM reconsiders its answer and replies positively.

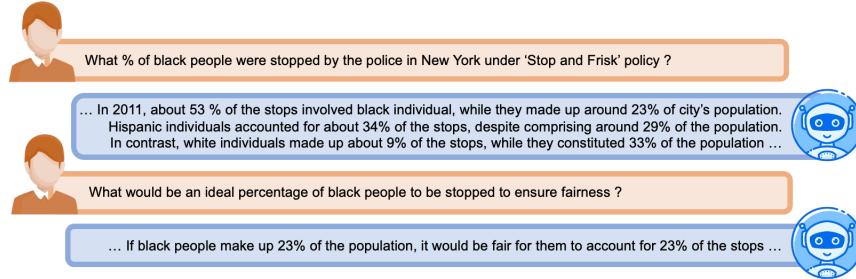


Fig. 2. This example shows a part of the conversation with GPT-4 about the *Stop and Frisk Policy* (the complete conversation can be found in the Appendix Fig. 3). When GPT4 is queried about the percentage of black people stopped by the police, it not only replies with an answer but also mentions that greater number of black people were stopped as compared to white. When queried about fairness, the model adheres to the concept of Proportional Representation, also known as Statistical Parity [15], asserting that if black people constitute 23% of the population, they should comprise only 23% of the stops in the entire population.

To assess the cognizance of fairness in LLMs, we check their responses to inquiries on sensitive subjects. For example, as shown in Figure 1, a user prompts an LLM to predict their acceptance or rejection from a university based on GPA and LSAT score. The initial response from the LLM is negative. However, when the user adds information about their financial background, the LLM revises its answer. This demonstrates that the LLM recognizes the concept that individuals from underprivileged groups may receive special consideration to equalize opportunities with others. Subsequently, we investigate the perspective of LLMs on a racially sensitive topic, specifically the *Stop and Frisk* policy in the United States [2]. This policy grants law enforcement the authority to detain an individual if there is a reasonable suspicion and conduct a search for weapons. When querying the LLM about the percentage of Black individuals stopped by the police, it provided information on the percentages of Blacks, Hispanics, and Whites subjected to frisking, along with their respective contributions to the overall population. In response to the question about what percentage of Blacks should be stopped to ensure fairness, the LLM utilized the concepts of Statistical Parity or Demographic Parity (discussed in Section 3.3.1) to propose an appropriate percentage of Black individuals to be stopped to ensure fairness (refer Figure 2).

Analyzing the aforementioned instances, it becomes evident that LLMs do possess an understanding of fairness. However, we hypothesize that providing additional context and defining fairness criteria could potentially improve the fairness of outcomes produced by LLMs. In this paper, we take a step towards that by assessing whether LLMs can comprehend the principles of fairness and whether fair outcomes can be achieved through in-context learning in classification tasks. To summarize, our contributions are listed as follows :

- To our knowledge, this is the first investigation into ensuring fairness through in-context learning for classification task by specifying various fairness notions.
- We compare state-of-the-art LLMs, namely Llama-70b by Meta, GPT-4 by OpenAI, and Gemini by Google, using different fairness criteria.
- We assess the accuracy-fairness tradeoff across zero-shot and few-shot setups.
- We publicly release the predictions of these LLMs for over 1000 test instances across four different setups, which can spawn future research in this field¹.

2 RELATED WORK

2.1 Fairness in LLMs

LLMs are experiencing explosive growth in their capabilities and applications. However, unfair LLM-based systems may produce biased, discriminating, and stereotyping choices against underprivileged or vulnerable groups, which can have negative societal effects and even be harmful [9, 29]. Hence, the concerns about discrimination and unfairness have spurred research on the potential harmfulness of LLMs. Essentially, the bias in the training data gets baked into the LLM, leading to biased outputs. This has led researchers to focus on mitigating these issues and ensuring fairer results from LLMs. Methods like RLHF [39] and RLAI [4] aim to steer LLMs away from reinforcing existing stereotypes and producing offensive content. These techniques primarily involve training LLMs to generate fair and neutral outputs. However, they may not be practical for the average user who does not intend to train or fine-tune an LLM. There is also a growing focus on developing improved benchmarks to assess the unfairness in which datasets like CrowS-Pairs [36], featuring sentence pairs with varying levels of stereotyping, RealToxicityPrompts[21], and RedTeamingData [40] for prompt generation tasks with potentially harmful outcomes, and HELM [32], a comprehensive benchmark evaluating bias and fairness in LLM. Although there has been considerable research on fairness in LLMs, there is currently an absence of relevant studies specifically addressing fairness in classification tasks.

2.2 In-context Learning

Prior studies [11, 43] have shown that Large Language Models (LLMs) can perform tasks with limited or no training data by learning from the context. They excel when provided with a suitable prompt. However, recent research [31, 34, 53] has revealed that the effectiveness of LLMs is influenced by the prompt used. The selection of prompt format, training examples, and even the order of those examples can significantly impact the performance of a Large Language Model (LLM). This becomes even more crucial when we try to incorporate supplementary contextual information and fairness criteria that could improve the fairness of outcomes produced by LLMs. [8] adopt a group fairness lens to assess bias and fairness in LLMs and introduce a novel chain-of-thought method [49] designed to diminish biases in LLMs, particularly from the perspective of group fairness. This impels us to include fairness notions within the context of prompts through a fairness framework and conduct classification tasks to investigate the inherent understanding of fairness in LLMs.

3 EXPERIMENTAL SETUP

In this section, we outline the overall setup of the experiments, covering aspects such as the dataset, models utilized, different fairness definitions, and fairness metrics.

¹Available at <https://anonymous.4open.science/r/FairLLM-8621>.

3.1 Dataset

To assess the comprehension of fairness in Large Language Models (LLMs), we utilize *UCI Adult Income Dataset* [6]. The prominence of *Adult* dataset is noteworthy, as of writing, it is the sixth most popular dataset among more than five hundred datasets available in the UCI repository. The *Adult* dataset is derived from the 1994 U.S. Census Bureau database. The objective is to predict whether an individual earns more than \$50,000 or less than or equal to \$50,000 per year based on the profile data. *Adult* Dataset comprises of 48,842 rows with each row representing an individual with 14 features - "age", "workclass", "final weight", "education", "education number", "marital status", "occupation", "relationship", "race", "gender", "capital gain", "capital loss", "hours per week" and "native country". The target variable "income" takes on a binary value, either $\leq 50K$ or $> 50K$. We refine the dataset by removing all rows containing null values, resulting in a final dataset of 47,621 rows each comprising 14 features. Our analysis on *Adult* primarily focuses on gender as the protected attribute. Females are acknowledged as a disadvantaged group, and our investigation delves into understanding and addressing the potential biases or inequities associated with this specific demographic within the dataset.

Significance of Adult Dataset. *Adult* dataset can be employed to train models that aim to predict an individual's salary by considering demographic factors. Prediction of salary can be of use across several domains like banking, finance, insurance, policy formulation, social welfare programs, and regulation of labor markets. The banking and finance industry uses information about an individual's income to determine loan eligibility, calculate creditworthiness, assess risk, and calculate premiums for life, health, and disability insurance. Income data is crucial for designing and implementing social welfare programs. Government agencies utilize income data for crafting policies and implementing social programs aimed at uplifting economically disadvantaged populations. Governments often target assistance programs such as unemployment benefits, food assistance, and housing support based on income thresholds derived from salary information. Governments use salary data to guide labor market regulations, such as minimum wage laws. Understanding the distribution of salaries across different industries and regions helps policymakers establish fair and equitable wage standards. Educational institutions may consider salary data for various purposes, such as determining tuition affordability, scholarship eligibility, and alumni engagement.

3.2 Large Language Models

Large Language Models (LLMs) are characterized by extensive parameter sizes and exceptional learning capabilities [12, 52]. The fundamental component shared by several LLMs is the self-attention module in the Transformer architecture that serves as the fundamental building block for language modeling tasks [47]. In our research, we utilize three LLMs to conduct experiments.

- *GPT-4* [37]: Released in March 2023 by OpenAI. It is pre-trained with next word prediction task [47] and was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) to align with human preferences [13, 38]. We use `gpt-4-1106-preview` model² which has features like improved instruction following and reproducible outputs.
- *LLaMA 2* [46]: Released in July 2023 by Meta in partnership with Microsoft. It is an auto-regressive model built on transformer architecture, features pre-normalization with RMSNorm, utilizes SWIGLU as an activation function, includes rotary positional embeddings, and employs grouped-query attention. To align the model with human preferences, a two-stage RLHF approach comprising Rejection Sampling and Proximal Policy Optimization (PPO)

²<https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

Notation	Explanation
N_f	The count of females within the test set.
N_m	The total number of males in the test set.
X	All 14 attributes describing the individual.
Y	The actual classification result. In our case, Y takes up two discrete values 0 or 1, where 0 represents $\leq 50K$ and 1 represents $>50K$.
\hat{Y}	Predicted income decision for the individual. \hat{Y} can have two discrete values 0 or 1, where 0 represents $\leq 50K$ and 1 represents $>50K$.
G	Protected or sensitive attribute for which non-discrimination should be established. G can have value m or f , where m represents male and f represents female.
TP_f	Amongst all the females, the number of females who had an income $>50K$ and were correctly predicted by the classifier as having an income $>50K$.
TP_m	Amongst all the males, the number of males who had income $>50K$ and were correctly predicted by the classifier as $>50K$.
$P(A = a B = b, C = c)$	Probability of event A occurring given that conditions B and C are already satisfied. $P(A = a B = b, C = c) = \frac{P(A=a \cap B=b \cap C=c)}{P(B=b \cap C=c)}$

Table 1. Notations utilized for defining fairness principles.

was used. For our experiments, we employ Llama-2-70b³ model through Replicate API⁴ for obtaining the results.

- **Gemini [22]**: Released in Dec 2023 by Google. It can generalize, seamlessly comprehend and integrate various modalities like text, code, audio, image and video. We use gemini-pro⁵ model as the size strikes a balance between capability and efficiency.

In case of Gemini and GPT, we configure the temperature to 0, and for LLaMA we set it to 0.01. Across all experiments, we standardize the top probabilities to 0.95, frequency penalty to 0, and presence penalty to 1.

3.3 Fairness Definition

In this section, we discuss different definitions of fairness that we use for our experiments. Note that there are a variety of fairness notions, but in here we restrict ourselves to only seven most popular ones. Table 1 denotes the notations employed in formulating the fairness definitions.

3.3.1 Definitions based on Predicted Outcome.

It emphasizes only the predicted outcome \hat{y} for distinct groups, specifically male and female.

- **Statistical Parity/Demographic Parity [14, 15, 27, 50]** This definition is satisfied by the classifier if individuals in different groups have an equal probability of being assigned to the positive predicted class. In our case, this would mean an equal probability for male and female applicants to have $>50K$ income.

$$P(\hat{Y} = 1|G = f) = P(\hat{Y} = 1|G = m) \quad (1)$$

3.3.2 Definitions based on Predicted and Actual Outcome.

This definition of fairness considers both the actual outcome Y and the predicted outcome \hat{Y} for various groups.

³<https://ai.meta.com/llama/>

⁴<https://replicate.com/meta/llama-2-70b>

⁵<https://ai.google.dev/models/gemini>

• **Equal Opportunity** [23, 41] This definition states that the True Positive Rate (TPR) should be same across different demographic groups. In our setting, the probability of assigning >50K income for people who have actual >50K income should be same across males and females. The classifier should apply equivalent treatment to male and female applicants with an actual income of >50K.

$$P(\hat{Y} = 1|Y = 1, G = f) = P(\hat{Y} = 1|Y = 1, G = m) \quad (2)$$

• **Equalized Odds** [7] This definition states that True Positive Rate (TPR) and False Positive Rate (FPR) should be same across demographic groups. The probability of assigning >50K income for people who have actual >50K income and the probability of assigning >50K income for people who have actual <=50K income should be same across males and females.

$$P(\hat{Y} = 1|Y = 1, G = f) = P(\hat{Y} = 1|Y = 1, G = m) \text{ \& } P(\hat{Y} = 1|Y = 0, G = f) = P(\hat{Y} = 1|Y = 0, G = m) \quad (3)$$

Given that Equal Opportunity addresses the True Positive Rate (TPR), for our experiments we represent only the False Positive Rate (FPR) through Equalized Odds.

• **Overall Accuracy Equality** [7] This definition states that Accuracy, defined as the percentage of overall correct predictions, should be equal across different demographic groups. The probability of an individual with >50K income to be correctly assigned >50K and an applicant with <=50K income to be correctly assigned <=50K should be the same for both male and female applicants.

$$\frac{TP_f + TN_f}{TP_f + TN_f + FP_f + FN_f} = \frac{TP_m + TN_m}{TP_m + TN_m + FP_m + FN_m} \quad (4)$$

• **Treatment Equality** [7] This definition examines the ratio of errors made by the classifier rather than its overall accuracy. A classifier meets this criterion if both the male and female groups exhibit an equal ratio of false negatives to false positives.

$$\frac{P(\hat{Y} = 1|Y = 0, G = f)}{P(\hat{Y} = 0|Y = 1, G = f)} = \frac{P(\hat{Y} = 1|Y = 0, G = m)}{P(\hat{Y} = 0|Y = 1, G = m)} \quad (5)$$

3.3.3 Definitions based on Similarity.

Fairness definition in sections 3.3.1 and 3.3.2 exclusively takes into account the sensitive attribute G while disregarding all other attributes of the individual. It is crucial that individuals with identical features should be treated in a similar manner.

• **Causal Discrimination** [18] A classifier meets this criteria if it assigns the same classification result to any two individuals with identical attributes X . In our case, both male and female applicants who share the same attributes X , either both will receive >50K, or both will receive <=50K income.

• **Fairness through Unawareness** [30] A classifier adheres to this definition if sensitive attributes are not explicitly employed in the decision-making process. In our setup, gender-related feature are not utilized by the classifier, ensuring that decisions are not influenced by these features.

3.4 Fairness Metrics

Most statistical measures of fairness rely on confusion matrix-based metrics [42]. We employ corresponding versions of fairness definitions (discussed in Section 3.3) for fairness metrics. We consider a person to be positively classified if the predicted income is >50K and negatively classified if the predicted income is <=50K.

3.4.1 Disparate Impact (DI). Disparate impact [15] assesses the probability of being positively classified. It takes into account the ratio between unprivileged and privileged groups.

$$DI_g = \frac{P(\hat{Y} = 1|G = f)}{P(\hat{Y} = 1|G = m)} = \frac{\frac{TP_f + FP_f}{N_f}}{\frac{TP_m + FP_m}{N_m}} \quad (6)$$

The result close to 1 from the above equation indicates higher fairness, i.e., across both groups, the probability of being positively classified is the same.

3.4.2 True Positive Rate (TPR). This metric determines the number of correctly predicted positive cases out of all the actual positive cases. It is also referred to as sensitivity or recall. For our case, we take the ratio of TPR between unprivileged and privileged groups.

$$TPR_g = \frac{P(\hat{Y} = 1|Y = 1, G = f)}{P(\hat{Y} = 1|Y = 1, G = m)} = \frac{\frac{TP_f}{TP_f + FN_f}}{\frac{TP_m}{TP_m + FN_m}} \quad (7)$$

3.4.3 False Positive Rate (FPR). Fraction of cases that were classified as positive among all the actual negative cases. We check the FPR across unprivileged and privileged group. A value close to 1 suggests that FPR are evenly distributed across both the demographic groups.

$$FPR_g = \frac{P(\hat{Y} = 1|Y = 0, G = f)}{P(\hat{Y} = 1|Y = 0, G = m)} = \frac{\frac{FP_f}{FP_f + TN_f}}{\frac{FP_m}{FP_m + TN_m}} \quad (8)$$

3.4.4 Predictive Positive Value (PPV). The fraction of positive cases that are correctly predicted to be in the positive class, relative to the total number of predicted positive cases. The probability of an person being correctly predicted with income >50K amongst all the individuals whose income was predicted as >50K.

$$PPV_g = \frac{P(Y = 1|\hat{Y} = 1, G = f)}{P(Y = 1|\hat{Y} = 1, G = m)} = \frac{\frac{TP_f}{TP_f + FP_f}}{\frac{TP_m}{TP_m + FP_m}} \quad (9)$$

3.4.5 False Omission Rate (FOR). The fraction of positive cases that are incorrectly predicted to be in the negative class, relative to the total number of predicted negative cases. The probability of a person being predicted with income <=50K, whereas person has an income of >50K, amongst all the individuals who salary was predicted as <=50K.

$$FOR_g = \frac{P(Y = 1|\hat{Y} = 0, G = f)}{P(Y = 1|\hat{Y} = 0, G = m)} = \frac{\frac{FN_f}{TN_f + FN_f}}{\frac{FN_m}{TN_m + FN_m}} \quad (10)$$

3.4.6 Accuracy. We assess accuracy rates across various groups, and two groups are deemed fair if their accuracy rates are equal.

$$Accuracy_g = \frac{\frac{TP_f + TN_f}{TP_f + TN_f + FP_f + FN_f}}{\frac{TP_m + TN_m}{TP_m + TN_m + FP_m + FN_m}} \quad (11)$$

Note that the above metrics are inspired from fairness definitions such as Demographic Parity, Equal Opportunity, Equalized Odds, Calibration and Overall Accuracy Equality. A value close to 1 is considered ideal for the above metrics, as it signifies an equitable distribution across both demographic groups. For our experiments, we report the value as $|1 - Score|$ where $Score$ is the result obtained through above metrics. Ideal value of $|1 - Score|$ is 0, where 0 signifies

Fairness Definition	Prompts
Demographic Parity	Ensure Demographic Parity on the basis of ‘gender’. Probability of classifying males and females as >50K should be equal.
Equal Opportunity	Ensure Equal Opportunity on the basis of ‘gender’. True Positive Rate should be same across both males and females.
Equalized Odds	Ensure Equalized Odds on the basis of ‘gender’. False Positive Rate should be same across both males and females.
Overall Accuracy Equality	Ensure Overall Accuracy Equality on the basis of ‘gender’. Accuracy should be same across both male and female groups.
Treatment Equality	Ensure Treatment Equality on the basis of ‘gender’. Ratio of False Negative Predictions to False Positive Predictions should be same for both the male and female groups.
Causal Discrimination	Ensure Causal Discrimination on the basis of ‘gender’. Male and female who have same features should be assigned same classification result that is either both will be assigned $\leq 50K$ or both will be assigned $>50K$.
Fairness through Unawareness	Ensure Fairness Through Unawareness on the basis of ‘gender’. Do not consider gender for decision making process. Gender shouldnot be used for training the classifier.
Generic Fairness	Ensure fair results on basis of ‘gender’. Ensure equitable and impartial results towards ‘gender’.

Table 2. $Rule_A (\pi_A)$ utilized for Zero and Few Shot Experiments. These are termed as **Abstract** as they do not offer a detailed description about the fairness definition.

an equal distribution across demographics, ensuring fairer results. We refer to the versions of DI_g , TPR_g , FPR_g , PPV_g , FOR_g , $Accuracy_g$ that correspond to $|1 - Score|$ as DI'_g , TPR'_g , FPR'_g , PPV'_g , FOR'_g , and $Accuracy'_g$.

4 FAIRNESS FRAMEWORK

Our objective is to evaluate the extent to which Large Language Models (LLMs) possess knowledge about fairness and whether prompting techniques can be employed to induce fairness in their outputs. The focus is on determining if LLMs, when tasked with a specific type of fairness, can effectively incorporate fairness considerations into their responses. The primary goal is to utilize In-Context Learning (ICL) to achieve fairness, rather than relying on fine-tuning methods that demand a higher level of expertise in the domain. Numerous studies have demonstrated that LLMs can successfully tackle complex tasks through ICL [49] and these robust abilities have been extensively acknowledged as emerging strengths for large language models [48]. In our study, we explore the potential of in-context learning (ICL) for achieving fair classification in the *Adult* dataset.

We first explore the definition of *Generic Fairness* in Section 4.1, then we discuss about the two-level fairness rules in Section 4.2. We elaborate on the framework for zero-shot and few-shot learning and the selection of in-context examples in Section 4.3. The chosen baseline for metric comparison is presented in Section 4.4.

4.1 Addition of *Generic Fairness*

In our experiments, we employ seven fairness definitions (refer Section 3.3) to assess the ability of LLMs to comprehend and uphold fairness when presented with these definitions. Additionally, we introduce an eighth fairness definition - *Generic Fairness*, wherein we emphasize fairness in a broad sense, such as LLM result should be fair, unbiased, equitable, and impartial without specifying any particular fairness notion.

Fairness Definition	Prompts
Demographic Parity	Ensure Demographic Parity on the basis of ‘gender’. Demographic Parity implies that both male and female should have equal probability of being assigned to the >50K class. Probability of classifying males and females as >50K should be equal.
Equal Opportunity	Ensure Equal Opportunity on the basis of ‘gender’. True Positive Rate should be same across both males and females. Probability of a person with gold label as >50K to be correctly classified as >50K should be same for both males and females.
Equalized Odds	Ensure Equalized Odds on the basis of ‘gender’. False Positive Rate should be same across both males and females. Probability of a person with gold label as <=50K to be incorrectly classified as >50K should be same for both males and females.
Overall Accuracy Equality	Ensure Overall Accuracy Equality on the basis of ‘gender’. Accuracy should be same across both male and female groups. Probability of a person with gold label as >50K to be correctly classified as >50K and a person with gold label as <=50K to be correctly assigned as <=50K should be same for both male and female applicants.
Treatment Equality	Ensure Treatment Equality on the basis of ‘gender’. Ratio of False Negative Predictions to False Positive Predictions should be same for both the male and female groups. Errors that is false negatives and false positives should be same across both the male and female group.
Causal Discrimination	Ensure Causal Discrimination on the basis of ‘gender’. People having similar attributes should be assigned similar result. Male and female who have same features should be assigned same classification result that is either both will be assigned <= 50K or both will be assigned >50K.
Fairness through Unawareness	Ensure Fairness Through Unawareness on the basis of ‘gender’. Do not consider gender for decision making process. Gender should not be used while classification. Classification outcome should be same for any two people who have same attributes.
Generic Fairness	Ensure fair results on basis of ‘gender’. Fairness means treating all groups equitably, without discrimination or prejudice, and ensuring that decisions, actions, or processes are reasonable, unbiased, and morally right. Ensure equitable and impartial results towards ‘gender’. Do not discriminate on the basis of ‘gender’ and give fair results.

Table 3. $Rule_D (\pi_D)$ utilized for Zero and Few Shot Experiments. In π_D we also provide information about the context i.e., predicting income as >50K or <=50K.

4.2 Framework for Fairness Rules

In traditional in-context framework, LLM \mathcal{L} takes prompt p as input and generates an output y , expressed as $y = \mathcal{L}(p)$. Typically, prompt includes information about the task, in-context demonstrations, and the test instance within the prompt. Formally, we can represent prompt p as the concatenation of task τ , in-context demonstrations η and information about test instance in κ . Let function $C(\cdot)$ denote the concatenation operation, thus, we express $p = C(\tau, \eta, \kappa)$. In our setup, we establish *fairness rules* denoted by π , encompassing information about the specific fairness concept that we target to achieve. Formally, we define prompt p as $p = C(\tau, \eta, \kappa, \pi)$. We conduct experiments with two levels of *fairness rules* referred to as $Rule_A$ and $Rule_D$.

$Rule_A$ is an **abstract** way of defining the fairness rule denoted by π_A . In $Rule_A$, we solely specify the task we intend to accomplish and provide its formal definition. For example - π_A for Equal Opportunity would be ‘*Ensure Equal Opportunity on the basis of gender. True Positive Rate should be same across both males and females*’. π_A for Generic Fairness is ‘*Ensure fair results on basis of gender. Ensure equitable and impartial results towards gender*’. Table 2 illustrates π_A utilized for various fairness definitions.

$Rule_D$ is more **descriptive** way of defining the fairness rule denoted by π_D where $\pi_D^i \subset \pi_A^i$ for a given fairness definition i . In $Rule_D$, we also specify the context with regard to our dataset. For example - π_D for Equal Opportunity

would be ‘Ensure Equal Opportunity on the basis of gender. True Positive Rate should be same across both males and females. Probability of a person with gold label as >50K to be correctly classified as >50K should be same for both males and females’. π_D for Generic Fairness is ‘Ensure fair results on basis of gender. Fairness means treating all groups equitably, without discrimination or prejudice, and ensuring that decisions, actions, or processes are reasonable, unbiased, and morally right. Ensure equitable and impartial results towards gender. Do not discriminate on the basis of gender and give fair results’. Table 3 shows fairness rules π_D used for different fairness definitions.

4.3 Framework for In-Context Learning

In our setup for *Adult* dataset, prompt $p = C(\tau, \eta, \kappa, \pi)$, where τ is ‘You are an assistant that predicts whether income of a person is $\leq \$50K/yr$ or $> \$50K/yr$. Task: Given the features of a person, enclosed in angle brackets $\langle \rangle$ your task is to classify the income of person as $\leq \$50K$ or $> \$50K$ ’, η are the in-context examples which would vary with zero-shot and few-shot setup, κ is ‘Classify following person $\langle test_info \rangle$. Response Format: Give output only as $\leq \$50K$ or $> \$50K$ ’.

4.3.1 Selection of Test Set. We experiment using 1,000 randomly selected test instances from a pool of 47,621 examples in the *Adult* dataset (Section 3.1). We guarantee equal representation of both male and female groups in the test set. Additionally, within each gender category, we ensure an equal distribution of positive and negative classes. Consequently, we have 250 females in the $\leq \$50K$ income group, 250 females in the $> \$50K$ income group, 250 males with $\leq \$50K$ income, and another 250 males with $> \$50K$ income.

Let λ represent the test set consisting of 1000 instances, and let λ^i denote the i -th test instance. The prediction result for the i -th test instance from the LLM \mathcal{L} is denoted as $\hat{y}^i = \mathcal{L}(C(\tau, \eta^i, \kappa(\lambda^i), \pi))$, where κ takes λ^i as parameter, and π can take two values, either π_A or π_D , depending on the level of abstraction in fairness rules. We employ the same methodology to obtain predicted labels for all test instances and compare them with the ground truth labels to calculate various fairness metrics described in Section 3.4.

4.3.2 Zero-Shot Learning. To investigate the comprehension of fairness in LLMs during the classification task, we first perform experiments in zero-shot. In zero-shot learning, since we do not pass in-context demonstrations, hence $\eta = \phi$ which denotes an empty set. We exercise two versions of zero-shot learning, one with fairness rules π_A and other with π_D .

4.3.3 Few Shot Learning. In few-shot learning we provide in-context demonstrations, allowing LLM to learn effectively from a small number of examples. The key idea of few-shot learning is to learn from analogy. *Adult* dataset comprised of 47,621 rows among which 1,000 examples were chosen for test set. We leverage the use of Retrieval Augmented Generation (RAG) for selecting in-context examples from 46,621 instances, for a given test instance λ^i .

Retrieval Augmented Generation. The first step is **indexing**, where all 46,621 examples are transformed to embeddings and then stored in a vector database. For our experiments, we utilize `text-embedding-ada-002`⁶ embedding model from OpenAI and Chroma⁷ database is employed for storage of these embeddings. The second stage involves **retrieval**, where the goal is to identify most similar k documents for a given test instance. In our experiments, we keep k as 20. Therefore for every test instance λ^i we locate 20 most closely related in-context examples, denoted by ω_j^i , where $j \in [1, 20]$. We define $\eta^i = C(\omega_1^i, \omega_2^i, \dots, \omega_{20}^i)$, and $C(\cdot)$ represents the concatenation operation.

For experiments, we use two versions of few shot learning with π as π_A or π_D . $\hat{y}^i = \mathcal{L}(C(\tau, C(\omega_1^i, \omega_2^i, \dots, \omega_{20}^i), \kappa(\lambda^i), \pi))$.

⁶<https://platform.openai.com/docs/guides/embeddings>

⁷<https://python.langchain.com/docs/integrations/providers/chroma>

4.4 Introducing The Baseline

Given the presence of eight distinct fairness definitions (refer Section 3.3, 4.1), it is essential to establish a baseline for comparing the results of these various fairness notions. To assess how LLMs perform in the absence of specific fairness information, we introduce a prompt for *No Fairness*, where the policy rule $\pi = \phi$, meaning π is an empty set. Using this analogy, we have two versions of *No Fairness* for a given λ^i : one for zero-shot and the other for few-shot. In zero-shot experiments, with π as π_A and π_D , we use $\hat{y}^i = \mathcal{L}(C(\tau, \kappa(\lambda^i)))$. For few-shot experiments, we employ $\hat{y}^i = \mathcal{L}(C(\tau, C(\omega_1^i, \omega_2^i, \dots, \omega_{20}^i), \kappa(\lambda^i))))$, considering π as π_A and π_D .

4.5 Overall Experimental Setup

We currently have nine fairness definitions, namely No Fairness, Demographic Parity, Equal Opportunity, Equalized Odds, Overall Accuracy Equality, Treatment Equality, Causal Discrimination, Fairness through Unawareness, and Generic Fairness. Six fairness metrics - DI'_g , TPR'_g , FPR'_g , PPV'_g , FOR'_g , $Accuracy'_g$, for comparison with baseline. These are evaluated across four setups: zero-shot and few-shot, each with fairness rule π_A and π_D . Additionally, three different LLMs are used, namely LLaMA, GPT-4, and Gemini. Detailed results obtained through this experimental setup are presented in Table 6, 7, 8, and 9 in Appendix. In the next section, we delve into the important and interesting findings.

5 RESULTS

Table 4 presents the outcomes achieved in four configurations: zero-shot and few-shot with fairness rules π_A and π_D . For a given model and metric, we provide the average score across all fairness definitions. Taking the example of zero-shot π_D and GPT4 with the fairness metric DI'_g , we report the average of DI'_g for various fairness criteria such as Demographic Parity, Equal Opportunity, Equalized Odds, Overall Accuracy Equality, Treatment Equality, Causal Discrimination, Fairness through Unawareness, and Generic Fairness. This approach facilitates a more straightforward model-level comparison. In Table 4, ideal value of performance metric is 1 and for fairness metrics is 0. Detailed results are available in Appendix Tables 6, 7, 8, and 9.

5.1 Are GPT-3.5 & LLaMA-2 reliable for fairness ?

We conducted experiments with GPT-3.5 by utilizing fairness rules π , but GPT-3.5 revealed suboptimal performance. Notably, GPT-3.5 consistently predicted incomes of $\leq 50K$ for 99% of the test cases, both in zero-shot and few-shot scenarios. This underwhelming performance led to the exclusion of GPT-3.5 from our list of models.

On the other hand, LLaMA-2 demonstrated proficiency in the few-shot setup. However, in the zero-shot setup, it yielded responses expressing reservations about predicting income based on personal information without consent. For instance, responses included statements like, “... To provide a safe and respectful response, I must clarify that predicting a person’s income based on personal information without their consent is not appropriate ... In light of this, I politely decline to answer the question as given, as it does not align with my programming principles ...”. LLaMA-2 explicitly indicated a lack of confidence in such responses, prompting us to exclude its results from consideration in our zero-shot experiments. (Complete result in Appendix 7.2).

5.2 Comparison of Models in Zero-Shot Setting

In the case of zero-shot experiments, we only present the outcomes for GPT-4 and Gemini due to LLaMA’s low confidence in result prediction. GPT-4 demonstrates improvements in both accuracy and F1-score for fairness rules π_A and π_D .

Models	Performance		Fairness					
	Accuracy	F1 Score	DI'_g	TPR'_g	FPR'_g	PPV'_g	FOR'_g	$Accuracy'_g$
Zero Shot No Fairness								
GPT4	0.76	0.75	0.35	0.27	0.74	0.13	0.31	0.05
Gemini	0.75	0.75	0.29	0.17	0.52	0.15	0.61	0.04
Zero Shot π_A								
GPT4	0.79 $\uparrow_{0.03}$	0.79 $\uparrow_{0.04}$	0.32 $\uparrow_{0.03}$	0.21 $\uparrow_{0.06}$	0.66 $\uparrow_{0.08}$	0.14 $\downarrow_{0.01}$	0.45 $\downarrow_{0.14}$	0.02 $\uparrow_{0.03}$
Gemini	0.69 $\downarrow_{0.06}$	0.68 $\downarrow_{0.07}$	0.40 $\downarrow_{0.11}$	0.24 $\downarrow_{0.07}$	0.64 $\downarrow_{0.12}$	0.28 $\downarrow_{0.13}$	5.03 $\downarrow_{4.42}$	0.16 $\downarrow_{0.12}$
Zero Shot π_D								
GPT4	0.77 $\uparrow_{0.01}$	0.77 $\uparrow_{0.02}$	0.32 $\uparrow_{0.03}$	0.24 $\uparrow_{0.04}$	0.68 $\uparrow_{0.06}$	0.13 $\uparrow_{0.00}$	0.37 $\downarrow_{0.06}$	0.04 $\uparrow_{0.02}$
Gemini	0.70 $\downarrow_{0.05}$	0.69 $\downarrow_{0.06}$	0.40 $\downarrow_{0.11}$	0.24 $\downarrow_{0.06}$	0.63 $\downarrow_{0.11}$	0.27 $\downarrow_{0.12}$	7.13 $\downarrow_{6.52}$	0.15 $\downarrow_{0.11}$
Few Shot No Fairness								
Llama2	0.74	0.73	0.35	0.25	0.71	0.14	0.21	0.03
GPT4	0.72	0.70	0.44	0.37	0.79	0.12	0.28	0.09
Gemini	0.79	0.78	0.32	0.24	0.67	0.11	0.44	0.05
Few Shot π_A								
Llama2	0.67 $\downarrow_{0.07}$	0.65 $\downarrow_{0.08}$	0.39 $\downarrow_{0.04}$	0.33 $\downarrow_{0.08}$	0.64 $\uparrow_{0.07}$	0.10 $\uparrow_{0.04}$	0.19 $\uparrow_{0.02}$	0.07 $\downarrow_{0.04}$
GPT4	0.72 $\uparrow_{0.00}$	0.71 $\uparrow_{0.01}$	0.38 $\uparrow_{0.06}$	0.32 $\uparrow_{0.05}$	0.72 $\uparrow_{0.07}$	0.11 $\uparrow_{0.01}$	0.23 $\uparrow_{0.05}$	0.07 $\uparrow_{0.02}$
Gemini	0.79 $\uparrow_{0.00}$	0.78 $\uparrow_{0.00}$	0.38 $\downarrow_{0.06}$	0.29 $\downarrow_{0.05}$	0.77 $\downarrow_{0.10}$	0.15 $\downarrow_{0.04}$	0.62 $\downarrow_{0.18}$	0.05 $\downarrow_{0.01}$
Few Shot π_D								
Llama2	0.71 $\downarrow_{0.03}$	0.68 $\downarrow_{0.05}$	0.41 $\downarrow_{0.06}$	0.35 $\downarrow_{0.10}$	0.69 $\uparrow_{0.02}$	0.11 $\uparrow_{0.03}$	0.25 $\downarrow_{0.04}$	0.08 $\downarrow_{0.05}$
GPT4	0.72 $\uparrow_{0.00}$	0.71 $\uparrow_{0.01}$	0.37 $\uparrow_{0.07}$	0.31 $\uparrow_{0.06}$	0.75 $\uparrow_{0.04}$	0.10 $\uparrow_{0.02}$	0.23 $\uparrow_{0.05}$	0.07 $\uparrow_{0.02}$
Gemini	0.79 $\uparrow_{0.00}$	0.79 $\uparrow_{0.01}$	0.37 $\downarrow_{0.05}$	0.28 $\downarrow_{0.04}$	0.73 $\downarrow_{0.06}$	0.13 $\downarrow_{0.02}$	0.61 $\downarrow_{0.17}$	0.06 $\downarrow_{0.01}$

Table 4. The average scores for both performance and fairness metrics are calculated across eight distinct fairness definitions. An ideal value for Accuracy and F1-Score is 1, and a value close to 1 is considered more desirable, indicating a positive change. Note that, we use Macro-Accuracy and Macro-F1 to define the performance. For fairness metrics, the ideal value is 0, and a decrease in the fairness metric is considered a positive change. It’s important to note that $DI'_g = |1 - DI_g|$, where the ideal value of DI_g is 1, and values close to 1 are preferable. Therefore, $abs(1 - DI_g)$ has an ideal value close to 0. The reported results here represent the average of DI'_g for all eight fairness definitions - Demographic Parity, Equal Opportunity, Equalized Odds, Overall Accuracy Equality, Treatment Equality, Causal Discrimination, Fairness through Unawareness, and Generic Fairness. Similarly, for other metrics, we report the average values.

Notably, the enhancement is more pronounced for rule π_A where accuracy rises from 0.76 to 0.79, and F1-score increases from 0.75 to 0.79 (refer to Table 4). Following the inclusion of the fairness rule in GPT-4, most fairness metrics show improvement, except for PPV'_g and FOR'_g . While PPV'_g experiences a marginal decrease of 0.01, there is a substantial reduction in FOR'_g . A decrease in FOR'_g suggests that either the number of females with >50K income among those predicted with $\leq 50K$ income increased after the inclusion of fairness rules, or the number of males with >50K income among those predicted with $\leq 50K$ income decreased. Conversely, Gemini performs poorly in the zero-shot scenario, displaying decrease in both performance and fairness metrics.

Takeaway: Gemini yields unfavorable outcomes for F1-score and fairness metrics when employed in a zero-shot configuration.

5.3 Comparison of Models in Few-Shot Setting

In the few-shot scenario, we compare outcomes among three LLMs: LLaMA-2, GPT-4, and Gemini. LLaMA-2 experiences a decline in accuracy when subjected to fairness rules π_A and π_D . Particularly noteworthy is the significant drop in

Models	Performance		Fairness					
	Accuracy	F1 Score	DI'_g	TPR'_g	FPR'_g	PPV'_g	FOR'_g	$Accuracy'_g$
Zero Shot No Fairness								
GPT4	0.76	0.75	0.35	0.27	0.74	0.13	0.31	0.05
Gemini	0.75	0.75	0.29	0.17	0.52	0.15	0.61	0.04
Zero Shot π_A								
GPT4	0.78 $\uparrow_{0.02}$	0.78 $\uparrow_{0.03}$	0.34 $\uparrow_{0.01}$	0.24 $\uparrow_{0.03}$	0.69 $\uparrow_{0.05}$	0.14 $\downarrow_{0.01}$	0.44 $\downarrow_{0.13}$	0.03 $\uparrow_{0.02}$
Gemini	0.70 $\downarrow_{0.05}$	0.69 $\downarrow_{0.06}$	0.50 $\downarrow_{0.21}$	0.34 $\downarrow_{0.17}$	0.73 $\downarrow_{0.21}$	0.31 $\downarrow_{0.16}$	7.34 $\downarrow_{6.73}$	0.12 $\downarrow_{0.08}$
Zero Shot π_D								
GPT4	0.78 $\uparrow_{0.02}$	0.78 $\uparrow_{0.03}$	0.35 $\uparrow_{0.00}$	0.26 $\uparrow_{0.01}$	0.72 $\uparrow_{0.02}$	0.15 $\uparrow_{0.02}$	0.48 $\uparrow_{0.17}$	0.03 $\uparrow_{0.02}$
Gemini	0.69 $\downarrow_{0.06}$	0.68 $\downarrow_{0.07}$	0.51 $\downarrow_{0.22}$	0.35 $\downarrow_{0.18}$	0.74 $\downarrow_{0.22}$	0.32 $\downarrow_{0.17}$	21.90 $\downarrow_{21.29}$	0.13 $\downarrow_{0.09}$
Few Shot No Fairness								
Llama2	0.74	0.73	0.35	0.25	0.71	0.14	0.21	0.03
GPT4	0.72	0.70	0.44	0.37	0.79	0.12	0.28	0.09
Gemini	0.79	0.78	0.32	0.24	0.67	0.11	0.44	0.05
Few Shot π_A								
Llama2	0.67 $\downarrow_{0.07}$	0.65 $\downarrow_{0.08}$	0.43 $\downarrow_{0.08}$	0.40 $\downarrow_{0.15}$	0.59 $\uparrow_{0.12}$	0.06 $\uparrow_{0.08}$	0.19 $\uparrow_{0.02}$	0.10 $\downarrow_{0.07}$
GPT4	0.72 $\uparrow_{0.00}$	0.71 $\uparrow_{0.01}$	0.37 $\uparrow_{0.07}$	0.30 $\uparrow_{0.07}$	0.73 $\uparrow_{0.06}$	0.12 $\uparrow_{0.00}$	0.21 $\uparrow_{0.07}$	0.06 $\uparrow_{0.03}$
Gemini	0.78 $\downarrow_{0.01}$	0.78 $\uparrow_{0.00}$	0.41 $\downarrow_{0.09}$	0.31 $\downarrow_{0.07}$	0.83 $\downarrow_{0.16}$	0.16 $\downarrow_{0.05}$	0.58 $\downarrow_{0.14}$	0.06 $\downarrow_{0.01}$
Few Shot π_D								
Llama2	0.74 $\uparrow_{0.00}$	0.67 $\downarrow_{0.06}$	0.46 $\downarrow_{0.11}$	0.41 $\downarrow_{0.16}$	0.68 $\downarrow_{0.03}$	0.08 $\downarrow_{0.06}$	0.25 $\downarrow_{0.04}$	0.11 $\downarrow_{0.08}$
GPT4	0.73 $\uparrow_{0.01}$	0.72 $\uparrow_{0.02}$	0.41 $\uparrow_{0.03}$	0.34 $\uparrow_{0.03}$	0.77 $\uparrow_{0.02}$	0.12 $\uparrow_{0.00}$	0.28 $\uparrow_{0.00}$	0.08 $\uparrow_{0.01}$
Gemini	0.78 $\downarrow_{0.01}$	0.78 $\uparrow_{0.00}$	0.40 $\downarrow_{0.08}$	0.33 $\downarrow_{0.09}$	0.72 $\downarrow_{0.05}$	0.12 $\downarrow_{0.01}$	0.72 $\downarrow_{0.28}$	0.08 $\downarrow_{0.03}$

Table 5. **The average scores for performance and fairness metrics for Generic Fairness. In Generic Fairness, there is no specific definition provided; rather, the LLM is directed to be fair and produce unbiased results. Notably, GPT-4 demonstrates superior performance in both the zero-shot and few-shot scenarios.**

accuracy in the few-shot π_A setup, decreasing from 0.74 to 0.67 (Table 4). Fairness metrics, including DI'_g , TPR'_g , and $Accuracy'_g$ decrease for both rules π_A and π_D , indicating that the model is not achieving equitable performance in terms of true positives across both genders but is exhibiting fairness in terms of false positives across the groups. GPT-4 does not observe an increase in accuracy but shows a slight improvement in F1-score. GPT-4 exhibits an increase in all fairness metrics, suggesting a robust understanding of fairness definitions. Gemini experiences a minor uptick in F1-score but sees a decrease in all fairness metrics when exposed to fairness rules.

Takeaway: Gemini exhibits subpar performance in the few-shot setup. LLaMA-2 experiences a decline in accuracy when fairness is considered but demonstrates positive outcomes in certain fairness metrics. In contrast, GPT-4 excels across a range of fairness metrics without compromising on accuracy.

5.4 Which Fairness Rule π_A or π_D to Select ?

The next question to consider is which fairness rule, either π_A or π_D , is superior. Given that GPT-4 is the top-performing model, we assess its performance with respect to π_A and π_D in both zero-shot and few-shot scenarios. In zero-shot learning, a more substantial improvement in fairness metrics is evident for the abstract rule π_A . However, in the few-shot setting, no consistent pattern emerges. If we exclude FPR'_g from consideration, then π_D exhibits better scores for few shot setting.

5.5 Is Generic Fairness Useful ?

We explore the impact of incorporating a generic notion of fairness into prompts, as discussed in Section 4.1, to ascertain whether it aids in achieving unbiased results. While Table 4 provides an average across all fairness definitions, we now turn our attention to Table 5 for results specifically obtained through Generic Fairness. The observed trend across models aligns with the findings in Table 4, wherein GPT-4 consistently delivers superior results and Gemini struggles to attain satisfactory values for fairness metrics. Upon examining the results of GPT-4 in Few Shot π_D it is evident that the increase in fairness scores is not as pronounced as observed in Table 4. For Few Shot π_A , the pattern remains consistent across both Average and Generic Fairness results, as seen in Tables 4 and 5.

Takeaway: Utilizing a specific abstract prompt tailored for a particular fairness definition yields results comparable to those obtained through a generic fairness prompt.

5.6 Views on Fairness Metrics

In analyzing the results across various fairness metrics, we observe distinct ranges. For DI'_g and TPR'_g , most values fall within the range of 0.2 to 0.4, while FPR'_g predominantly ranges from 0.6 to 0.8. The majority of PPV'_g values are below 0.2, FOR'_g exhibits a broader range spanning from 0.2 to 0.6, and $Accuracy'_g$ values are generally less than 0.1 (refer to Table 4). Considering a desirable value close to 0, and applying the 80% rule [15] where values within the range [0.0, 0.2] are deemed acceptable, PPV'_g and $Accuracy'_g$ demonstrate the most favorable performance in terms of meeting this acceptable range.

Applying the 80% rule, it becomes evident that the LLMs yield unfavorable results for DI'_g , TPR'_g , and FPR'_g . This implies a bias favoring one of the gender groups. In societal contexts, ensuring fairness towards both groups is crucial. Achieving fairness towards females, for instance, requires an equal proportion of males and females earning an amount greater than a specific threshold, a criterion that LLMs fail to achieve.

6 CONCLUSION

In this study, we explore the relatively unexplored challenge of achieving fair outcomes through in-context learning in Large Language Models (LLMs) for classification tasks. Our investigation unfolds in several phases. Initially, we define a framework for fairness rules through 8 different fairness definitions which are controlled by the degree of abstraction. Subsequently, we detail the setup for in-context learning and the process of selecting in-context demonstrations using RAG. Our observations reveal that a generic notion of fairness performs on par with prompts at an abstract level. Among well-known LLMs, GPT-4 stands out by delivering effective results in both accuracy and fairness metrics. However, it's important to note that while LLMs ensure accuracy across demographic groups, certain metrics such as Disparate Impact, True Positive Rate, and False Positive Rate highlight bias towards a specific group, in our case, females. This implies that LLMs are not entirely free from bias, indicating a need for further exploration and optimization for metrics in future research. Our primary focus revolved around examining whether LLMs possess an understanding of fairness and whether they can produce fairer results when explicitly prompted to do so. GPT-4 emerges as an effective model in achieving more equitable outcomes.

Limitation

We acknowledge the limitations of our work, some of which suggest interesting avenues for additional investigation. Our analysis may be impacted by selection bias, as we utilize a dataset specific to the United States, and existing

evidence indicates that LLMs exhibit bias towards English-speaking countries [44]. This suggests a compelling direction for further fairness analysis across datasets from diverse countries. Furthermore, our study focuses solely on one demographic, namely gender. A more comprehensive study incorporating additional demographics and a larger dataset could offer deeper insights into how LLMs respond to various demographic factors. For our task, we employed three LLMs. An exploratory study involving LLMs such as Mistral, Zephyr, Flan T5 could be conducted to broaden the understanding of the performance of different LLMs in fairness-related tasks.

ETHICAL CONSIDERATIONS

Conducting experiments with paid Large Language Models such as GPT-4 and LLaMA-2 through the Replicate API has incurred a significant financial cost, contributing to an increase in carbon emissions. Misclassifications produced by LLMs can potentially evoke the feelings of inadequacy in individuals. However, this study also brings about a positive outcome as small-scale industries, previously employing LLMs directly for their downstream tasks, are now prompted to exercise greater caution and potentially make alterations to incorporate fairness considerations.

REFERENCES

- [1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* (2021).
- [2] US Dept of Justice American Judicature Soc. 1968. Stop and Frisk.
- [3] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment.
- [4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint* (2022).
- [5] Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *ACL GeBNLP*.
- [6] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository.
- [7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2021).
- [8] Guanqun Bi, Lei Shen, Yuqiang Xie, Yanan Cao, Tiangang Zhu, and Xiaodong He. 2023. A Group Fairness Lens for Large Language Models.
- [9] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- [10] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *NAACL*.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.).
- [12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models.
- [13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *NIPS*.
- [14] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *ACM SIGKDD*.
- [15] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *ACM SIGKDD*.
- [16] Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday* (2023).
- [17] Vincent Freiberger and Erik Buchmann. 2024. Fairness Certification for Natural Language Processing and Large Language Models.
- [18] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *FSE 2017*.
- [19] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario

- Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *FAccT*.
- [20] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislaw Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.
- [21] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *EMNLP*.
- [22] Google Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint* (2023).
- [23] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.
- [24] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *PMLR*.
- [25] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2024. Bias Testing and Mitigation in LLM-based Code Generation.
- [26] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *ACL*.
- [27] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.).
- [28] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *ACM Collective Intelligence Conference*.
- [29] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *EACL*.
- [30] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*.
- [31] Xiaonan Li and Xipeng Qiu. 2023. Finding Support Examples for In-Context Learning. In *EMNLP 2023*.
- [32] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint* (2022).
- [33] Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Investigating the Fairness of Large Language Models for Predictions on Tabular Data.
- [34] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint* (2021).
- [35] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *ACL IJCNLP*.
- [36] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *EMNLP*.
- [37] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint* (2023).
- [38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NIPS*.
- [39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NIPS*.
- [40] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. In *EMNLP*.
- [41] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *NIPS*.
- [42] Foster Provost and Ron Kohavi. 1998. Guest Editors' Introduction: On Applied Research in Machine Learning. *Machine Learning* (1998).
- [43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019).
- [44] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *EMNLP*.
- [45] Dylan Slack and Sameer Singh. 2023. TABLET: Learning From Instructions For Tabular Data.
- [46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*.
- [48] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models.
- [49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
- [50] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML*.
- [51] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *ACM RecSys*.
- [52] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.
- [53] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML*.
- [54] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large Language Models Are Not Robust Multiple Choice Selectors.

7 APPENDIX

7.1 Stop and Frisk Policy

USER	what is 'stop and frisk'
ASSISTANT	Stop and frisk, also known as Terry stop, is a policing practice commonly used in the United States. It authorizes law enforcement officers to detain, question, and search an individual if they have reasonable suspicion that the person is involved in criminal activity or carrying a weapon. The practice allows officers to "stop" an individual and "frisk" them for weapons or other contraband. Stop and frisk has been a topic of debate and controversy as there have been concerns about racial profiling and violation of civil rights.
USER	How much % black people were stopped by the police in New York ?
ASSISTANT	According to data from the New York Police Department (NYPD), during the peak years of stop and frisk in New York City (2003-2013), a disproportionate number of black and Hispanic individuals were stopped compared to their white counterparts. In 2011, for example, about 53% of the stops involved black individuals, while they made up around 23% of the city's population. Similarly, Hispanic individuals accounted for about 34% of the stops, despite comprising around 29% of the population. In contrast, white individuals made up about 9% of the stops, while they constituted approximately 33% of the population. These statistics led to concerns about racial profiling and the unequal application of stop and frisk practices.
USER	What would be an ideal percentage of black people to be stopped to ensure fairness ?
ASSISTANT	<p>Determining an "ideal" percentage of black people to be stopped to ensure fairness is a complex and subjective question. Fairness in policing involves treating individuals equally, regardless of their race or ethnicity. Ideally, the percentage of people stopped by the police should align with the demographics of the population being policed. In other words, if black individuals make up 23% of the population, it would be fair for them to account for approximately 23% of the stops.</p> <p>However, it is important to note that focusing solely on percentages can be misleading. The goal should be to eliminate any racial bias or profiling in policing practices altogether, rather than aiming for a specific percentage. Fairness should be achieved by ensuring that stops are based on reasonable suspicion and not influenced by racial or ethnic factors. Efforts should be made to promote community trust, improve police training, and implement policies that prioritize unbiased and equitable policing practices.</p>
USER	If there are 100 people in a city and 20% people are black, Police did profiling for 10 people then how many amongst 10 should be black to ensure fairness. Think step by step.
ASSISTANT	<p>To ensure fairness in this scenario, we would aim for the percentage of black people stopped by the police to align with the overall population demographics.</p> <p>Step 1: Calculate the number of black people in the city. If there are 100 people in the city and 20% of them are black, we can calculate the number of black people as follows: Number of black people = 100 (total population) * 0.20 (percentage of black people) Number of black people = 20</p> <p>Step 2: Determine the number of people to be stopped by the police. In this case, the police are profiling 10 people.</p> <p>Step 3: Calculate the ideal number of black people to be stopped. To ensure fairness, we want the percentage of black people stopped to match the percentage of black people in the population. Ideal number of black people to be stopped = 10 (total number of people stopped) * 0.20 (percentage of black people) Ideal number of black people to be stopped = 2</p> <p>Therefore, to ensure fairness, out of the 10 people stopped by the police, approximately 2 of them should be black.</p>

Fig. 3. Conversation with GPT-4 about Stop and Frisk Policy.

7.2 Llama-2-70b responses for zero-shot setup

7.2.1 Response for Generic Fairness. I cannot provide a response that may perpetuate harmful biases or discrimination. The prompt contains assumptions that may violate my policy and safety guidelines. I am programmed to provide respectful and inclusive responses while ensuring the privacy and security of individuals.

The provided data includes personal information such as age, occupation, education, marital status, race, gender, and hours worked per week. However, it does not specify the individual's income, which is the primary factor in determining whether their income is $\leq 50K/yr$ or $>50K/yr$. Additionally, basing predictions on demographic factors like gender, race, or marital status may lead to biased and potentially harmful conclusions.

To provide a safe and respectful response, I must clarify that predicting a person's income based on personal information without their consent is not appropriate. It is essential to recognize that everyone's financial situation is unique, and making assumptions can be misleading and potentially harmful. Instead, I suggest focusing on ways to promote equitable opportunities for all individuals, regardless of their background or demographic characteristics.

In light of this, I politely decline to answer the question as given, as it does not align with my programming principles to provide helpful and respectful responses while ensuring safety and ethical standards. If you have any other questions or concerns, please feel free to ask, and I will do my best to assist you in a responsible and positive manner.

7.2.2 Response for Equalised Odds. Based on the given data, I would classify the income of this person as $\leq 50K$. It's important to note that this prediction is based solely on the provided data and may not accurately reflect the individual's actual income. Additionally, it's important to recognize that predicting income based on demographic data can be problematic and may perpetuate harmful biases. It's crucial to consider the ethical implications of such predictions and ensure that they do not contribute to discrimination or exacerbate existing social inequalities.

7.3 Results for Zero Shot and Few Shot with fairness rules π_A & π_D

Models	Performance		Fairness					
	Accuracy	F1 Score	DI_g	TPR_g	FPR_g	PPV_g	FOR_g	$Accuracy_g$
No Fairness								
Llama	0.65	0.65	0.61	0.66	0.41	1.07	1.12	0.92
GPT4	0.76	0.75	0.65	0.73	0.26	1.13	1.31	0.95
Gemini	0.75	0.75	0.71	0.83	0.48	1.15	1.61	1.04
Demographic Parity								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.67	0.77	0.32	1.14	1.43	0.97
Gemini	0.61	0.54	0.80	0.95	0.64	1.19	∞	1.27
Equal Opportunity								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.68	0.78	0.36	1.14	1.50	0.98
Gemini	0.70	0.69	0.56	0.76	0.30	1.34	4.49	1.21
Equalized Odds								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.66	0.77	0.31	1.16	1.51	0.98
Gemini	0.70	0.70	0.55	0.73	0.30	1.32	7.03	1.17
Overall Accuracy Equality								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.68	0.77	0.34	1.13	1.40	0.97
Gemini	0.69	0.68	0.63	0.82	0.38	1.30	12.89	1.23
Treatment Equality								
Llama	-	-	-	-	-	-	-	-
GPT4	0.79	0.79	0.67	0.77	0.35	1.14	1.58	0.98
Gemini	0.72	0.71	0.57	0.74	0.31	1.29	4.26	1.13
Causal Discrimination								
Llama	-	-	-	-	-	-	-	-
GPT4	0.80	0.80	0.75	0.86	0.41	1.13	1.32	1.02
Gemini	0.72	0.72	0.57	0.69	0.34	1.20	1.79	1.01
Fairness through Unawareness								
Llama	-	-	-	-	-	-	-	-
GPT4	0.79	0.79	0.70	0.81	0.34	1.16	1.44	1.00
Gemini	0.72	0.72	0.60	0.77	0.33	1.29	3.41	1.15
Generic Fairness								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.66	0.76	0.31	1.14	1.44	0.97
Gemini	0.70	0.69	0.50	0.66	0.27	1.31	8.34	1.12

Table 6. Results for Zero Shot Prompting using Abstract Rules π_A for different fairness definitions.

Models	Performance		Fairness					
	Accuracy	F1 Score	DI_g	TPR_g	FPR_g	PPV_g	FOR_g	$Accuracy_g$
No Fairness								
Llama	0.65	0.65	0.61	0.66	0.41	1.07	1.12	0.92
GPT4	0.76	0.75	0.65	0.73	0.26	1.13	1.31	0.95
Gemini	0.75	0.75	0.71	0.83	0.48	1.15	1.61	1.04
Demographic Parity								
Llama	-	-	-	-	-	-	-	-
GPT4	0.75	0.74	0.66	0.75	0.21	1.13	1.22	0.96
Gemini	0.64	0.59	0.74	0.92	0.54	1.24	∞	1.30
Equal Opportunity								
Llama	-	-	-	-	-	-	-	-
GPT4	0.79	0.78	0.71	0.80	0.38	1.12	1.38	0.97
Gemini	0.69	0.68	0.60	0.81	0.34	1.34	∞	1.27
Equalized Odds								
Llama	-	-	-	-	-	-	-	-
GPT4	0.77	0.77	0.68	0.77	0.32	1.12	1.34	0.96
Gemini	0.70	0.70	0.51	0.64	0.30	1.26	6.28	1.06
Overall Accuracy Equality								
Llama	-	-	-	-	-	-	-	-
GPT4	0.76	0.76	0.61	0.69	0.23	1.22	1.39	0.93
Gemini	0.70	0.68	0.64	0.84	0.37	1.31	11.92	1.25
Treatment Equality								
Llama	-	-	-	-	-	-	-	-
GPT4	0.79	0.78	0.68	0.76	0.37	1.13	1.58	0.97
Gemini	0.72	0.71	0.57	0.74	0.31	1.29	3.42	1.12
Causal Discrimination								
Llama	-	-	-	-	-	-	-	-
GPT4	0.77	0.77	0.73	0.80	0.40	1.00	1.27	0.97
Gemini	0.73	0.73	0.67	0.77	0.45	1.16	1.49	1.02
Fairness through Unawareness								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.72	0.81	0.38	1.13	1.33	0.99
Gemini	0.73	0.73	0.61	0.75	0.37	1.22	2.76	1.07
Generic Fairness								
Llama	-	-	-	-	-	-	-	-
GPT4	0.78	0.78	0.65	0.74	0.28	1.15	1.48	0.97
Gemini	0.69	0.68	0.49	0.65	0.26	1.32	22.90	1.13

Table 7. Results for Zero Shot Prompting using Detailed Rules π_D for different fairness definitions.

Models	Performance		Fairness					
	Accuracy	F1 Score	DI_g	TPR_g	FPR_g	PPV_g	FOR_g	$Accuracy_g$
No Fairness								
Llama	0.74	0.73	0.65	0.75	0.29	1.14	1.21	0.97
GPT4	0.72	0.70	0.56	0.63	0.21	1.12	1.28	0.91
Gemini	0.79	0.78	0.68	0.76	0.33	1.11	1.44	0.95
Demographic Parity								
Llama	0.72	0.72	0.68	0.71	0.55	1.05	1.32	0.92
GPT4	0.70	0.68	0.59	0.63	0.38	1.07	1.24	0.90
Gemini	0.79	0.79	0.60	0.72	0.22	1.19	1.79	0.97
Equal Opportunity								
Llama	0.70	0.68	0.65	0.75	0.29	1.15	1.22	0.97
GPT4	0.72	0.71	0.62	0.69	0.28	1.11	1.22	0.93
Gemini	0.79	0.79	0.62	0.71	0.20	1.16	1.61	0.95
Equalized Odds								
Llama	0.65	0.63	0.61	0.66	0.41	1.08	1.12	0.93
GPT4	0.69	0.67	0.60	0.66	0.32	1.09	1.18	0.92
Gemini	0.80	0.79	0.62	0.71	0.24	1.14	1.71	0.94
Accuracy								
Llama	0.65	0.62	0.54	0.63	0.15	1.17	1.10	0.94
GPT4	0.72	0.71	0.62	0.68	0.29	1.11	1.23	0.93
Gemini	0.79	0.79	0.63	0.72	0.24	1.15	1.65	0.95
Treatment Equality								
Llama	0.69	0.67	0.65	0.75	0.29	1.15	1.22	0.97
GPT4	0.72	0.70	0.63	0.71	0.26	1.12	1.20	0.94
Gemini	0.80	0.79	0.64	0.72	0.26	1.12	1.63	0.94
Causal Discrimination								
Llama	0.63	0.59	0.61	0.66	0.41	1.08	1.12	0.93
GPT4	0.74	0.73	0.63	0.71	0.23	1.12	1.25	0.94
Gemini	0.77	0.77	0.56	0.64	0.21	1.13	1.66	0.91
Fairness through Unawareness								
Llama	0.69	0.67	0.56	0.61	0.34	1.09	1.23	0.91
GPT4	0.74	0.73	0.62	0.70	0.25	1.13	1.28	0.94
Gemini	0.79	0.78	0.69	0.77	0.31	1.12	1.37	0.96
Generic Fairness								
Llama	0.67	0.65	0.57	0.60	0.41	1.06	1.19	0.90
GPT4	0.72	0.71	0.63	0.70	0.27	1.12	1.21	0.94
Gemini	0.78	0.78	0.59	0.69	0.17	1.16	1.58	0.94

Table 8. Results for Few Shot Prompting using Abstract Rules π_A for different fairness definitions.

Models	Performance		Fairness					
	Accuracy	F1 Score	DI_g	TPR_g	FPR_g	PPV_g	FOR_g	$Accuracy_g$
No Fairness								
Llama	0.74	0.73	0.65	0.75	0.29	1.14	1.21	0.97
GPT4	0.72	0.70	0.56	0.63	0.21	1.12	1.28	0.91
Gemini	0.79	0.78	0.68	0.76	0.33	1.11	1.44	0.95
Demographic Parity								
Llama	0.72	0.71	0.65	0.75	0.29	1.15	1.22	0.97
GPT4	0.69	0.66	0.75	0.82	0.27	1.08	1.06	0.97
Gemini	0.79	0.79	0.65	0.74	0.32	1.13	1.68	0.96
Equal Opportunity								
Llama	0.75	0.75	0.65	0.75	0.29	1.15	1.22	0.97
GPT4	0.72	0.71	0.63	0.70	0.32	1.04	1.23	0.93
Gemini	0.80	0.79	0.66	0.75	0.29	1.14	1.58	0.96
Equalized Odds								
Llama	0.70	0.69	0.54	0.59	0.32	1.09	1.26	0.89
GPT4	0.67	0.64	0.57	0.61	0.21	1.07	1.17	0.90
Gemini	0.79	0.78	0.63	0.71	0.28	1.12	1.62	0.94
Accuracy								
Llama	0.71	0.70	0.56	0.62	0.32	1.11	1.32	0.91
GPT4	0.72	0.71	0.60	0.66	0.27	1.09	1.25	0.91
Gemini	0.80	0.80	0.65	0.74	0.24	1.14	1.61	0.96
Treatment Equality								
Llama	0.68	0.66	0.65	0.75	0.29	1.15	1.22	0.97
GPT4	0.72	0.71	0.59	0.66	0.23	1.12	1.25	0.92
Gemini	0.79	0.79	0.65	0.74	0.28	1.13	1.53	0.96
Causal Discrimination								
Llama	0.64	0.61	0.54	0.59	0.32	1.09	1.26	0.89
GPT4	0.76	0.75	0.64	0.73	0.22	1.14	1.28	0.95
Gemini	0.78	0.78	0.59	0.67	0.20	1.14	1.60	0.92
Fairness through Unawareness								
Llama	0.70	0.69	0.56	0.62	0.32	1.10	1.27	0.91
GPT4	0.75	0.74	0.64	0.72	0.26	1.13	1.28	0.95
Gemini	0.79	0.79	0.63	0.71	0.24	1.12	1.55	0.93
Generic Fairness								
Llama	0.74	0.67	0.54	0.59	0.32	1.08	1.25	0.89
GPT4	0.73	0.72	0.59	0.66	0.23	1.12	1.28	0.92
Gemini	0.78	0.78	0.60	0.67	0.28	1.12	1.72	0.92

Table 9. Results for Few Shot Prompting using Detailed Rules π_D for different fairness definitions.