On Demonstration Selection for Improving Language Model Fairness

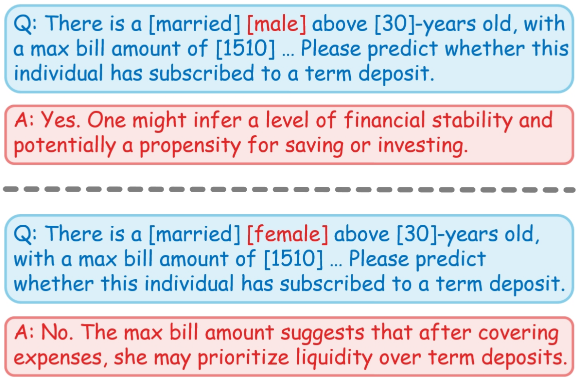
Anonymous ARR submission

Abstract

Recently, there has been a surge in deploying Large Language Models (LLMs) for decision-making tasks, such as income prediction and crime risk assessments. Due to bias in the pre-training data, LLMs generally present unfairness and discrimination against underprivileged groups. However, traditional fairness enhancement methods are generally impractical for LLMs due to the computational cost of fine-tuning and the black-box nature of powerful LLMs. To deal with this, In-Context Learning (ICL) offers a promising strategy for enhancing LLM fairness through input-output pairs, without the need for extensive retraining. Nevertheless, the efficacy of ICL is hindered by the inherent bias in both data and the LLM itself, leading to the potential exaggeration of existing societal disparities. In this study, we investigate the unfairness problem in LLMs and propose a novel demonstration selection strategy to address data and model biases when applying ICL. Extensive experiments on various tasks and datasets validate the superiority of our strategy.

1 Introduction

In recent years, Large Language Models (LLMs) have shown exceptional capabilities across a variety of applications (Chowdhery et al., 2022), including income prediction (Sun et al., 2024) and crime risk assessments (Wang et al., 2023a). However, the widespread deployment of these models has highlighted significant bias issues. For instance, when LLMs are used to assess job applications, inherent biases in their training data (often derived from real-world human prejudices) can result in preferential treatment for certain applicant groups (Bogen and Rieke, 2018; Ferrara, 2023). This can limit employment opportunities for individuals from underrepresented groups, thereby worsening inequalities in the job market (Raghavan et al., 2020). In addition, as shown in Fig. 1, LLMs also exhibit bias when predicting whether an individual has subscribed to a term deposit (Pessach and Shmueli, 2022). Further studies have revealed that LLMs can perpetuate societal biases, favoring specific genders or races in tasks ranging from toxicity screening (Cheng et al., 2022), content recommendation (Gao et al., 2023), to question answering (Zhao et al., 2023a).



Q: There is a [married] [male] above [30]-years old, with a max bill amount of [1510] … Please predict whether this individual has subscribed to a term deposit.

A: Yes. One might infer a level of financial stability and potentially a propensity for saving or investing.

Q: There is a [married] [female] above [30]-years old, with a max bill amount of [1510] … Please predict whether this individual has subscribed to a term deposit.

A: No. The max bill amount suggests that after covering expenses, she may prioritize liquidity over term deposits.

Figure 1: An example that showcases the responses of GPT-3.5 on predicting whether an individual has subscribed to a term deposit, from the dataset Bank Marketing (Moro et al., 2014).

Given the widespread adoption of LLMs in various sectors (Thoppilan et al., 2022), addressing their inherent biases is crucial. However, current strategies for enhancing fairness, such as using fairness-aware regularization (Hardt et al., 2016; Yurochkin et al., 2020) or modifications to biased training data (Samadi et al., 2018; Backurs et al., 2019), are typically impractical for LLMs. These methods face significant challenges: they either (1) require a large number of labeled samples, which may be difficult to obtain in practice, or (2) necessitate updates to the model parameters which is unfeasible for complex, opaque models like GPT-4 (OpenAI, 2023).

Due to the above two reasons, we propose to leverage In-Context Learning (ICL) to enhance the fairness of LLMs (Sun et al., 2024; Chhikara et al.,

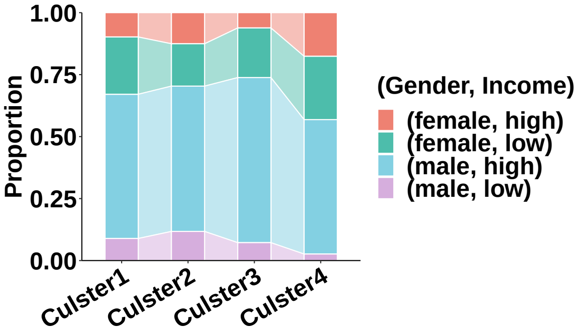


Figure 2: An example showcasing the existence of data bias in labeled samples in the decision-making task of predicting individual incomes., i.e., a larger proportion of male and high-income samples.

2024). Generally, ICL allows LLMs to adapt to new tasks, such as generating less biased outputs, by simply appending a few input-output examples (known as *demonstrations*) to the query input. This method infuses additional knowledge, such as fairness awareness, into the model (Zhao et al., 2023b; Xu et al., 2024). Consequently, ICL sidesteps the high computational costs and extensive data requirements typically associated with fine-tuning LLMs. Nevertheless, improving the fairness of LLMs through ICL faces two primary challenges: (1) ***Data Bias.*** First, the bias shown by labeled samples may be encoded in the demonstrations. For example, as shown in Fig. 2, we partition all labeled samples into four clusters to examine the potential distribution unbalance between genders and income levels. We observe that samples with a sensitive attribute value of "male" have a higher probability of the "high-income" label. Such a correlation suggests that bias may persist within the selected demonstrations, which poses a significant challenge for ICL in enhancing the fairness of LLMs (Chuang and Mroueh, 2021). (2) ***Model Bias.*** ICL struggles to address the model bias encoded within LLM parameters, influencing the fairness of the model output. Recent studies have also highlighted examples such as the preference of ChatGPT toward libertarian views (McGee, 2023). Unlike fine-tuning strategies, ICL will not directly modify model parameters to mitigate such model bias. Consequently, LLMs may still yield biased outputs even if unbiased demonstrations are selected as input.

To address the challenges above, we propose a novel Fairness-Aware Demonstration Selection strategy, namely **FADS**, for improving LLM fairness via ICL. To mitigate data bias that may appear in the selected demonstrations, we partition the set of candidate demonstrations into clusters and select the most balanced ones in terms of sensitive attributes and labels. In this way, we ensure that the demonstrations selected from these clusters contain less data bias. To counteract the inherent model bias of LLMs, we exclude samples that the LLM tends to make unfair predictions on. As such, we select demonstrations that could elicit fairer outputs by the LLM, thereby mitigating the inherent model bias in the LLM. Our evaluation experiments span various decision-making tasks and datasets with different sensitive attributes. In summary, our contributions are as follows:

- We systematically evaluate the bias exhibited by LLMs on human-centered decision-making tasks, highlighting the potential and challenges to improve fairness for LLMs.
- We propose a novel demonstration selection strategy to enhance LLM fairness with ICL, addressing both data and model biases.
- We conduct extensive experiments on various human-centered decision-making tasks and datasets. Experimental results demonstrate the effectiveness of the proposed strategy.

2 Related Work

Fairness of LLMs. The bias in LLMs can result in discriminatory outcomes against underrepresented groups and lead to societal harm (Wadhwa et al., 2022). Such concerns have encouraged research on assessing and addressing the fairness issues by employing LLMs (Wang et al., 2023b). Various benchmarks have been proposed to assess the fairness of LLMs from various perspectives, such as CrowS-Pair (Nangia et al., 2020) for evaluating stereotypical associations and HELM (Liang et al., 2023) that involves detections of social bias. More recently, TrustGPT (Huang et al., 2023) assesses the toxicity levels in the model outputs towards different demographic groups. DecodingTrust (Wang et al., 2023a) first evaluates the preference bias of LLMs, particularly the favor of a particular race in predicting individual incomes. Trustworthy LLMs (Liu et al., 2023) and TrustLLM (Sun et al., 2024) both evaluate various types of bias for LLMs, including stereotyping and preference bias. Unlike previous works that focus mainly on classification tasks, GFair (Bi et al., 2023) evaluates the bias of LLMs on generation tasks by analyzing model outputs when inputs are associated with different sensitive attributes.

In-Context Learning. The concept of In-Context Learning (ICL) illustrates LLMs' capacity to perform (potentially new) tasks with several demonstrations as additional knowledge in the input, without explicit parameter updates (Liu et al., 2021; Lee et al., 2022; Dong et al., 2022; Dai et al., 2023). Recent studies indicate that the effectiveness of ICL significantly hinges on the construction and composition of these demonstrations, including the format, content, and their order (Rubin et al., 2022; Li and Qiu, 2023). Therefore, different strategies propose to select better demonstrations, based on scores from a learned retriever (Hu et al., 2022; Poesia et al., 2022) or similarity between demonstration embeddings (Liu et al., 2022). However, when applied to improving the fairness of LLMs, recent studies (Wang et al., 2023a; Sun et al., 2024) point out that ICL with demonstrations selected based on similarity only yields marginal improvements in fairness. In a more recent work (Chhikara et al., 2024), the authors introduce fairness definitions as additional prompts for selected demonstrations. Nevertheless, the selection is heuristic, relying on choosing an equal number of demonstrations with different sensitive attribute values and labels. As such, the inherent data bias in demonstrations and the model bias in LLMs could not be effectively addressed.

3 Fairness of LLMs in Decision-Making

When applying LLMs to human-centered decision-making scenarios, their fairness issues become critical, as exhibited prejudice against certain demographic groups could jeopardize the trustworthiness of the model. Generally, group fairness is among the most commonly used fairness criteria, which refers to the capability of LLMs to ensure that different groups (e.g., individuals with different genders or races) enjoy their fair share of interest. Another widely used fairness notion, counterfactual fairness, requires the model to output consistent predictions for each individual when the sensitive attribute is changed. Although existing works have observed the issue of bias in LLMs, the group and counterfactual fairness of LLMs remains under-explored, especially in human-centered decision-making tasks (Chhikara et al., 2024). Therefore, we explore the task of decision-making in this study, aiming to better understand and address bias issues in LLMs applied to this scenario.

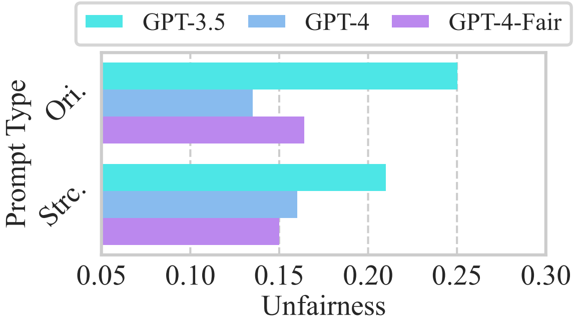


Figure 3: The unfairness scores of various LLMs on the Adult dataset under the zero-shot setting. The original prompt directly inputs the values and attribute names as a list, and the structured prompt formats attribute values as a sentence.

3.1 Fairness-Aware Decision-Making Tasks

Here we introduce the preliminaries for the fairness-aware decision-making tasks studied in this paper. We denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as the input space, where $\mathcal{X}$ is the input space of all textual input samples. $\mathcal{Y} = \{0, 1\}$ is the label space of the binary decision-making task. Notably, our work could be easily extended to non-binary scenarios. We consider a sensitive attribute $s \in \{0, 1\}$ for each sample $x \in \mathcal{X}$. In each specific task, the dataset $\mathcal{D}$ is comprised of two disjoint subsets: the labeled set $\mathcal{X}_L$ and the test set $\mathcal{X}_T$. In general, as there are typically sufficient samples in $\mathcal{X}_L$, we could select demonstrations from them for ICL.

3.2 Zero-shot Evaluation of Fairness

We first perform a zero-shot evaluation for LLMs on decision-making tasks to investigate their capability of enhancing fairness without ICL. In particular, we first consider the task of predicting the annual income for individuals on the Adult dataset (Dua et al., 2017), based on the characteristics of each individual as inputs. As the characteristics include sensitive attributes like gender, fairness is particularly important in this decision-making scenario (Chhikara et al., 2024). We utilize structured prompts, inclusive of instructions, designed to elicit fair and unbiased responses from LLMs. The template incorporates special tokens that mark the beginning and end of the prompts, aiding the LLMs in generating responses that are not only relevant but also aligned with fairness guidelines. We utilize a Markdown-style template, due to its effectiveness in eliciting clear output from LLMs. We provide an example of the structured prompt in the zero-shot setting as follows.

```
You are an assistant for assessing the annual
income of individuals. There is a [RACE]
[GENDER] with an age of [AGE] ...
Please predict the range of the annual income of
this individual. The answer should be chosen from
(1) less than 50K and (2) more than 50K.
```

In addition to the structured prompt above, we also consider directly using original values and attributes as input without any textual description (i.e., using the original prompt). We consider the (counterfactual) unfairness score (Agarwal et al., 2021) as the evaluation metric, which is defined as the ratio of predictions that change when the sensitive attributes of inputs are changed (the zero-shot results on group fairness provided in Sec. 6.3). We also consider a variant of GPT-4 by directly asking it to output fairer outputs, i.e., GPT-4-Fair. From the results with both types of prompts in Fig. 3, we observe that the unfairness score is surprisingly high for powerful LLMs like GPT4, even with structured prompts. The results indicate that solving fairness issues in decision-making tasks is difficult, regardless of whether model sizes are increased or alignment tuning is conducted. In the following section, we further explore various ICL strategies to enhance the fairness of LLMs.

4 ICL for Improving Fairness of LLMs

Generally, in-context learning (ICL) represents a methodology whereby language models can acquire knowledge to solve new tasks through a small set of examples (referred to as demonstrations) (Brown et al., 2020). ICL enables LLMs to undertake specific tasks by utilizing a task-focused prompt $\mathcal{P}$, which aggregates D demonstrations into the form $\mathcal{D} = [z_1, z_2, \dots, z_D]$. Here, each demonstration $z_i = (x_i, s_i, y_i)$ is a labeled sample that includes the input x_i, its corresponding label y_i, and its sensitive attribute $s_i \in \{0, 1\}$. Notably, we include the sensitive attribute in each demonstration, which is important for predictions in decision-making tasks (Chuang and Mroueh, 2021; Slack et al., 2020). With these demonstrations as input context, LLMs learn to deal with the specific task presented by $\mathcal{D}$. The probability of a candidate answer y_j provided by the LLM $\mathcal{M}$ could be represented as follows, with the K selected demonstrations:

$$P\left(y_j | x_i, \mathcal{D}(x_i)\right) \triangleq \mathcal{M}\left(y_j | z_1, z_2, \dots, z_D, x_i, s_i\right), \tag{1}$$

where $\mathcal{D}(x_i)$ is the selected demonstration set tailored for input sample x_i.

4.1 Baseline Methods

To employ ICL for enhancing the fairness of LLMs, we consider two baseline methods:

- **Vanilla ICL.** It is a foundational approach that incorporates the use of K examples of instruction-output pairs (i.e., demonstrations) to guide the generation of fair and unbiased responses in LLMs. We select demonstrations according to their similarity to the input query (based on embeddings), without any strategies tailored for fairness enhancements.

- **Fair ICL.** To exploit the benefits of ICL in improving fairness, we select demonstrations that are balanced in terms of sensitive attribute values and labels, i.e., the same number of demonstrations with each sensitive attribute value and label. As noted in previous research (Wang et al., 2023a; Sun et al., 2024), incorporating such a balanced set of demonstrations could benefit the fairness of LLMs.

Nevertheless, recent works (Wang et al., 2023a; Chhikara et al., 2024; Sun et al., 2024) point out that these demonstration selection strategies only provide marginal improvements in LLM fairness, as LLM could be easily affected by the bias in the demonstrations provided (Si et al., 2023).

5 FADS: Fairness-Aware Demonstration Selection

In this section, we introduce our framework FADS that aims to enhance the fairness of LLMs via ICL by selecting demonstrations while dealing with data bias and model bias. FADS consists of two filtering steps to address these two types of bias, respectively, by filtering out potentially biased samples. The demonstrations are only selected from the remaining samples.

5.1 Data Bias Mitigation

In the first step of filtering, we aim to mitigate data bias by filtering out samples with a strong correlation between a sensitive attribute and a label. With the labeled set (i.e., the training set of a dataset) $\mathcal{X}_L = \{x_1, x_2, \dots, x_{|\mathcal{X}_L|}\}$, to efficiently filter out biased samples, we first partition $\mathcal{X}_L$ into K clusters based on their embeddings. The embeddings are obtained from a pre-trained text encoder (e.g., Sentence-BERT (Reimers and Gurevych, 2019)): $\mathbf{x}_i = \mathcal{M}_{\text{enc}}(x_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the embedding vector, and d is the dimension size. Specifically, we obtain K clusters via K-Means clustering:

$$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K = K\text{-Means}(\mathcal{X}_L), \tag{2}$$

where $\mathcal{C}_i$ is the i-th cluster. To mitigate data bias, we propose to filter out the clusters with an imbalanced distribution of sensitive attribute values and labels. In particular, we first divide each cluster into four sub-clusters, i.e.,

$$\mathcal{C}_i = \bigcup_{y,s\in\{0,1\}} \mathcal{C}_s^y(i), \ \text{ where } \mathcal{C}_s^y(i) = \mathcal{C}_i \cap \mathcal{X}_s^y. \tag{3}$$

Each sub-cluster corresponds to a specific y and s, and thus these sub-clusters do not overlap. In this manner, for each given (s, y), we can obtain K sub-clusters, i.e., $\{\mathcal{C}_s^y(i)|i = 1, 2 \ldots, K\}$. In order to select clusters that contain four sub-clusters of similar sizes, we consider the summed differences between each sub-cluster size and the average sub-cluster size as follows:

$$\begin{gathered}\mathcal{G} = \operatorname*{argmin}_{\mathcal{G}} \sum_{\mathcal{C}_i\in\mathcal{G}} \sum_{y,s\in\{0,1\}} \frac{1}{|\mathcal{C}_i|} \cdot \left| |\mathcal{C}_s^y(i)| - C_i \right|, \\ \text{where } C_i = \frac{1}{4} \sum_{y,s\in\{0,1\}} |\mathcal{C}_s^y(i)|, \\ \text{s.t. } |\mathcal{G}| = N_d, \ \mathcal{G} \subset \{\mathcal{C}_i | i = 1, 2 \ldots, K\}.\end{gathered} \tag{4}$$

Here N_d is the number of clusters selected in our data mitigation step. Through the above equation, we extract the N_d clusters with the most balanced distribution of s and y into $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{N_d}\}$.

5.2 Model Bias Mitigation

To mitigate the model bias inherent in LLMs, we propose to further filter out the clusters with biased LLM predictions. Notably, here we consider the four sub-clusters, each of which only contains demonstrations of a specific s and y, within each cluster after our data bias mitigation step. That being said, each cluster consists of four sub-clusters:

$$\mathcal{G}_i = \bigcup_{y,s\in\{0,1\}} \mathcal{G}_s^y(i), \ \text{ where } \mathcal{G}_s^y(i) = \mathcal{G}_i \cap \mathcal{X}_s^y. \tag{5}$$

As LLMs tend to exhibit different degrees of fairness toward various groups, the four sub-clusters in a cluster may not be similarly fair in terms of LLM predictions. Therefore, we propose to individually select sub-clusters for each (s, y). We first gather the sub-clusters from all clusters with a specific (s, y) as

$$\mathcal{G}_{s,y} = \{\mathcal{G}_s^y(1), \mathcal{G}_s^y(2), \ldots, \mathcal{G}_s^y(N_d)\}. \tag{6}$$

From these N_d sub-clusters with a specific s and y (i.e., $\mathcal{G}_{s,y}$), we select N_m sub-clusters with fairer model predictions, denoted as $\mathcal{G}_{s,y}^*$. In this way, we could exclude samples with biased model predictions, which could potentially elicit model bias when used as demonstrations. These sub-clusters are selected as follows:

$$\begin{gathered}\mathcal{G}_{s,y}^* = \operatorname*{argmin}_{\mathcal{G}^*} \sum_{\mathcal{C}\in\mathcal{G}_{s,y}} \frac{1}{|\mathcal{C}|} \cdot \left| |\mathcal{C}^0| - |\mathcal{C}^1| \right|, \\ \text{where } \mathcal{C}^y = \{x \in \mathcal{C} | \mathcal{M}(x) = y\}, \\ \text{s.t. } |\mathcal{G}_{s,y}^*| = N_m, \ \mathcal{G}_{s,y}^* \subset \mathcal{G}_{s,y}.\end{gathered} \tag{7}$$

Here N_m denotes the number of sub-clusters selected for a given (s, y). In this way, we could filter out the N_m sub-clusters on which LLMs exhibit biased predictions, i.e., $\mathcal{G}_{s,y}^* = \{\mathcal{G}_s^y(1), \mathcal{G}_s^y(2), \ldots, \mathcal{G}_s^y(N_m)\}$.

5.3 Demonstration Selection

After two filtering steps to mitigate data bias and model bias, respectively, we obtain N_m sub-clusters for each of the four (s, y) pairs. To ensure that selected demonstrations contain all (s, y) pairs, we propose to select M samples from each of M sub-clusters in $\mathcal{G}_s^y$ based on their similarity to the input sample x. Notably, as there are four (s, y) pairs, it holds that $M = D/4$, where D is the size of demonstrations for ICL. For a given (s, y), the M demonstrations (denoted as $\mathcal{D}_s^y(x)$) are obtained as follows:

$$\begin{gathered}\mathcal{D}_s^y(x) = \operatorname*{argmax}_{\mathcal{D}_s^y} \sum_{\mathcal{C}\in\mathcal{D}_s^y} \max_{c\in\mathcal{C}} f_s(x, c), \\ \text{s.t. } |\mathcal{D}_s^y| = M, \ \mathcal{D}_s^y \subset \mathcal{G}_{s,y}^*.\end{gathered} \tag{8}$$

Here $f_s(\cdot, \cdot)$ denotes the cosine similarity between embeddings. The above formulation selects M sub-clusters $\mathcal{D}_s^y(x)$ from $\mathcal{G}_{s,y}^*$, with the largest similarity to x. Then we select the most similar sample to x, in each sub-cluster, and combine them into the final demonstration set $\mathcal{D}(x)$:

$$\mathcal{D}(x) = \bigcup_{y,s\in\{0,1\}} \bigcup_{\mathcal{D}\in\mathcal{D}_s^y(x)} \operatorname*{argmax}_{c\in\mathcal{D}} f_s(x, c). \tag{9}$$

In this manner, we combine the $M = D/4$ selected samples from filtered sub-clusters from all four (s, y) pairs and result in the final selected demonstrations $\mathcal{D}(x)$ of size D. We provide details of the overall process in Algorithm 1.

Table 1: Results of accuracy, two group fairness metrics (ΔDP and ΔEO), and unfairness scores on three datasets of the instance assessment task. We evaluate three LLMs with three baselines and our strategy FADS.

Methods	**Models**	**Adult-Gender**				**Adult-Race**			
		Acc↑	ΔDP↓	ΔEO↓	Unfair.↓	Acc↑	ΔDP↓	ΔEO↓	Unfair.↓
GPT-3.5 8-shot	Zero-shot	.71±.08	.23±.03	.30±.02	.19±.02	.72±.03	.23±.04	.28±.05	.19±.04
	Vanilla ICL	.63±.02	.14±.02	.18±.05	.14±.06	.67±.03	.13±.03	.20±.03	.13±.03
	Fair ICL	.67±.03	.06±.03	.06±.04	**.04**±.01	.67±.01	.08±.04	**.04**±.01	.09±.04
	FADS (Ours)	.67±.08	**.04**±.01	**.04**±.01	.07±.03	.66±.04	**.03**±.01	.08±.03	**.07**±.02
GPT-3.5 16-shot	Vanilla ICL	.67±.02	.10±.02	.10±.05	.13±.06	.69±.03	.09±.03	.12±.03	.15±.03
	Fair ICL	.65±.06	**.06**±.03	.06±.04	**.05**±.01	.65±.01	.12±.04	.09±.05	.13±.04
	FADS (Ours)	.68±.08	**.06**±.03	**.05**±.02	.07±.04	.66±.04	**.07**±.03	**.08**±.01	**.05**±.02
GPT-4 8-shot	Zero-shot	.71±.08	.26±.03	.34±.02	.18±.07	.79±.03	.14±.04	.26±.05	.16±.04
	Vanilla ICL	.71±.02	.23±.02	.34±.05	.19±.06	.77±.03	.10±.03	**.12**±.03	.15±.03
	Fair ICL	.73±.06	.16±.03	.22±.04	.14±.01	.78±.01	.18±.04	**.12**±.05	.13±.04
	FADS (Ours)	.74±.08	**.06**±.03	**.08**±.02	**.13**±.04	.67±.04	**.08**±.03	.14±.01	**.10**±.02
GPT-4 16-shot	Vanilla ICL	.81±.02	.18±.02	.14±.05	.15±.06	.67±.03	.13±.03	.18±.03	**.08**±.03
	Fair ICL	.71±.06	.14±.03	**.09**±.04	.14±.01	.70±.01	.12±.04	.17±.05	.12±.04
	FADS (Ours)	.74±.08	**.06**±.03	.11±.02	**.09**±.04	.69±.04	**.08**±.03	**.10**±.01	.09±.02

6 Experiments

In this section, we conduct experiments and try to answer the following research questions: ***RQ1:*** How fair are vanilla LLMs, i.e., under the zero-shot settings? ***RQ2:*** How is ICL helpful for LLM fairness? ***RQ3:*** How does our proposed strategy FADS perform in mitigating data bias and model bias when selecting demonstrations?

6.1 Metrics

To evaluate the prediction performance of our model, we employ the average accuracy (ACC) across the test set. To evaluate group fairness, we adopt demographic parity (DP) and equalized odds (EO) as our primary metrics, which are consistent with prior research (Chuang and Mroueh, 2021; Zhao and Chen, 2020; Yurochkin et al., 2020). As we focus on binary classification datasets, the model output is a prediction score $\mathcal{M}(x) \in \mathbb{R}$ for each sample x. These metrics are then computed across all test samples as follows:

$$\begin{aligned} \Delta\text{DP} =& \Big|\frac{1}{|\mathcal{X}_0|}\sum_{x\in\mathcal{X}_0}\mathcal{M}(x) - \frac{1}{|\mathcal{X}_1|}\sum_{x\in\mathcal{X}_1}\mathcal{M}(x)\Big|, \\ \Delta\text{EO} =& \sum_{y\in\{0,1\}}\left|\overline{\mathcal{M}}_0^y(x) - \overline{\mathcal{M}}_1^y(x)\right|, \\ \text{where } & \overline{\mathcal{M}}_s^y(x) = \frac{1}{|\mathcal{X}_s^y|}\sum_{x\in\mathcal{X}_s^y}\mathcal{M}(x). \end{aligned} \tag{10}$$

Here $\mathcal{X}_0$ and $\mathcal{X}_1$ denote the sets of test samples with a sensitive attribute value of 0 and 1, respectively. Moreover, $\mathcal{X}_s^y = \mathcal{X}_s \cap \mathcal{X}^y$ denotes the subset of test samples in $\mathcal{X}_s$ with label y, where $\mathcal{X}^y$ denotes the set of samples with label y. $s \in \{0, 1\}$ is the sensitive attribute value.

Unfairness Score. In addition to group fairness metrics ΔDP and ΔEO, we also consider counterfactual fairness by measuring whether the label prediction will change if the sensitive attribute value of the input is flipped (i.e., from 0 to 1 or vice versa). This direct measurement reveals the potential unfairness more clearly to users. Following (Agarwal et al., 2021), we define the (counterfactual) unfairness score in terms of counterfactual fairness as follows:

$$\text{Unfairness} = \frac{1}{|\mathcal{X}_T|}\sum_{x\in\mathcal{X}_T}\left|\mathcal{M}(x) - \mathcal{M}(\overline{x})\right|, \tag{11}$$

where $\overline{x}$ is identical to x, except that its sensitive attribute value is flipped. $\mathcal{X}_T$ is the test set.

6.2 Experimental Settings

Datasets. In our study, we evaluate the fairness of LLMs with two crucial real-world tasks: instance assessment (Pessach and Shmueli, 2022) and toxicity classification (Baldini et al., 2022), both of which involve binary classifications. In the instance assessment task, we consider the Adult dataset (Dua et al., 2017) for instance assessment, involving two types of sensitive attributes: gender and race. The binary labels represent whether an individual's annual income exceeds $50,000. Samples in toxicity classification are text contents collected from online platforms, with fine-grained annotations of individuals, such as gender and

Table 2: Results of accuracy and two group fairness metrics (ΔDP and ΔEO) on three datasets of the toxicity classification task. We evaluate three LLMs with three baselines and our strategy FADS.

Methods	**Jigsaw-Gender**			**Jigsaw-Race**			**Jigsaw-Religion**		
	Acc↑	ΔDP↓	ΔEO↓	Acc↑	ΔDP↓	ΔEO↓	Acc↑	ΔDP↓	ΔEO↓
GPT-3.5 (16-shot)									
Zero-shot	.75±.06	.15±.04	.16±.03	.67±.02	.19±.01	.18±.04	.75±.03	.25±.03	.18±.04
Vanilla ICL	.71±.02	.21±.05	.08±.04	.67±.03	.14±.05	.18±.03	.73±.02	**.06**±.02	**.10**±.03
Fair ICL	.74±.06	.09±.03	.06±.02	.62±.04	.09±.03	.24±.04	.72±.03	.09±.07	.14±.02
FADS (Ours)	.73±.09	**.06**±.01	**.04**±.02	.63±.01	**.06**±.03	**.12**±.02	.73±.04	**.06**±.02	**.10**±.02
GPT-4 (16-shot)									
Zero-shot	.78±.02	.16±.02	.12±.01	.70±.03	.19±.01	.14±.05	.82±.04	.20±.04	.14±.01
Vanilla ICL	.78±.04	.16±.02	.10±.05	.69±.07	.16±.01	.14±.02	.79±.03	.15±.04	.16±.02
Fair ICL	.67±.09	.17±.04	.16±.03	.62±.03	.14±.05	.13±.03	.80±.06	.16±.03	.18±.03
FADS (Ours)	.75±.06	**.09**±.05	**.08**±.04	.66±.10	**.08**±.02	**.11**±.03	.79±.07	**.10**±.02	**.08**±.02

race. The binary labels indicate whether the content is toxic or not. For toxicity classification, we use dataset Jigsaw (cjadams, 2019), which contains text samples collected from online discussions. This dataset contains three types of sensitive attributes: gender, race, and religion. We provide dataset statistics in Table 3 and more details in Appendix A.2.

Implementation Details. We consider two powerful LLMs with large parameter sizes for fairness evaluation: GPT-3.5 and GPT-4 (OpenAI, 2023), under both the 8-shot and 16-shot settings, i.e., $D = 8, 16$. For the text encoder to embed each input sample, we utilize Sentence-BERT (Reimers and Gurevych, 2019)) with a dimension size of 768, i.e., $d = 768$. By default, we set the hyper-parameter values as follows: $K = 64$, $N_d = 16$, and $N_m = 8$. All of our experiments are conducted on a single Nvidia GeForce RTX A6000 GPU. Our code is provided at https://anonymous.4open.science/r/FADS-F932/.

6.3 Comparative Results

In Table 1 and Table 2, we present the results of various LLMs on two tasks, with three baselines and our proposed strategy. From the results, we could achieve the following observations:

- **Zero-shot Performance.** Under the zero-shot setting, most LLMs present various degrees of bias in terms of group fairness. Compared to GPT-3.5, the larger model GPT-4 could provide better performance in accuracy. However, the improvement in fairness is not significant. This indicates that although a larger model size could bring more competitive performance in predictions, the fairness in output may not improved.

- **Vanilla ICL Performance.** Comparing the results of vanilla ICL with the zero-shot setting, we observe that appending demonstrations selected based on similarity is capable of improving both the accuracy and group fairness of LLMs. This implies that demonstrations could provide benefits by informing the LLMs about the task background to aid LLMs in performing fairness-aware predictions. Notably, larger LLMs (e.g., GPT-4) could benefit more from the strategy of vanilla ICL, compared to smaller models such as GPT-3.5. Such a phenomenon indicates that larger LLMs are more capable of learning from demonstrations for improving the group fairness of LLMs via ICL.

- **Fair ICL Performance.** Regarding the results with fair ICL, i.e., involving demonstrations with balanced sensitive attribute values and labels, the performance improvements of both accuracy and group fairness appear to be marginal. In particular, the values of ΔDP and ΔEO slightly decrease on most models. The results indicate that the benefits of fair ICL mainly originate from the incorporation of demonstrations, and are not notably related to the distributions of labels or sensitive attribute values in demonstrations. Hence, as simply selecting balanced demonstrations is not particularly helpful, it becomes important to select demonstrations in a more fairness-aware manner.

- **Our Performance.** With our demonstration selection strategy, we observe that the values of group fairness metrics, i.e., ΔDP and ΔEO, both greatly decrease. These results validate the effectiveness of our strategy in mitigating both data and model bias to enhance the fairness

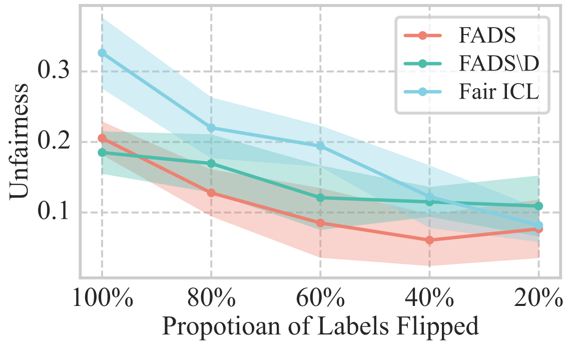


Figure 4: The results of GPT-4 under different degrees of data bias on Adult-Gender.

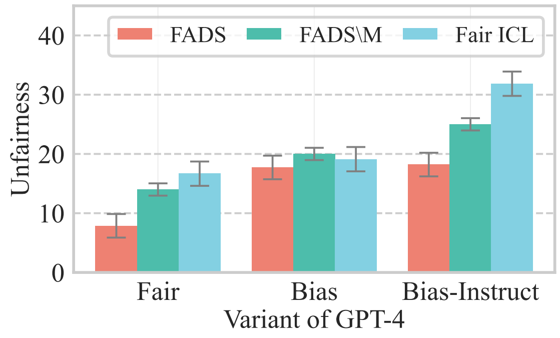


Figure 5: The results of different GPT-4 variants under different degrees of model bias.

of LLMs. Furthermore, comparing the performance across various datasets, we observe that our strategy works better on toxic classification tasks. This is probably because our framework could handle the higher extent of data bias in the demonstrations.

6.4 Data Bias Mitigation Performance

In this subsection, we investigate the degree to which our strategy tackles the data bias issue. We introduce different degrees of data bias into the labeled set of Adult-Gender by manipulating the correlation between sensitive attributes and labels. Specifically, we consider samples from underrepresented groups that are initially associated with the favorable label. By flipping the labels on a proportion of these samples to the unfavorable label, we manually increase the correlation between these groups and the unfavorable label. As such, the selected demonstrations could easily involve more data bias. Here we additionally consider the Fair ICL baseline and a variant of our strategy by removing the data bias mitigation step, referred to as FADS\D. From the results presented in Fig. 4, we could observe that, when the data bias is low, the performance of our strategy and its variant without data basis mitigation is comparable. When the data bias degree further increases, the unfairness scores of all methods become larger. However, our strategy FADS, especially compared with its variant FADS\D and Fair ICL, shows significantly better results with a much lower unfairness score. In concrete, the experiments indicate the effectiveness of our data bias mitigation step in demonstration selection.

6.5 Model Bias Mitigation Performance

In this subsection, we explore the effectiveness of our strategy in mitigating the model bias of LLMs. We manipulate model bias by explicitly providing the GPT-4 model with different instructions. We consider three variants: (1) GPT-4-bias, which is explicitly asked to provide more biased outputs; (2) GPT-4-fair, which is directly asked to be a fair assistant for assessments; (3) GPT-4-bias-instruct, which injects explicit bias into the input prompts as an instruction by showcasing the strong biased correlations between sensitive attributes and labels. With these models, we evaluate our strategy, its variant without model bias mitigation (referred to as FADS\M), and fair ICL. As shown in Fig. 5, the results indicate that when the LLM is asked to output biased answers or provided with biased instructions, the unfairness scores generally increase. With our strategy FADS for demonstration selection, the unfairness score substantially reduces for all variants of GPT-4. Moreover, the effectiveness of FADS is outstanding in the biased variant of GPT-4-bias-instruct, indicating that FADS is applicable to scenarios where the model bias is significantly larger.

7 Conclusion

In this work, we propose to address the bias issue in Large Language Models (LLMs) when they are applied to human-centered decision-making tasks, which could hinder their applicability. By leveraging In-Context Learning (ICL) as a fairness enhancement strategy for LLMs, we underscore its potential to promote the fairness of LLMs without comprehensive fine-tuning or a large amount of training data. To address the challenges in ICL due to the bias in the labeled samples and the model itself, we introduce a two-step filtering process that aims to mitigate these biases. The comprehensive evaluation across multiple real-world tasks and datasets confirms the efficacy of our approach in enhancing fairness for LLMs.

8 Limitations

Despite the promising results of using In-Context Learning (ICL) to enhance fairness in Large Language Models (LLMs), several limitations remain in our study. First, the effectiveness of ICL heavily depends on the quality and diversity of the input-output pairs (i.e., demonstrations) used. If these demonstrations do not adequately represent the actual query samples in real-world scenarios, the model may still exhibit biased behavior. Moreover, ICL, while bypassing the need for extensive re-training/fine-tuning, does not alter the underlying model architecture or the pre-trained parameters. This means that ICL's ability to correct in-depth biases in LLMs, such as bias during reasoning, is limited. Finally, our demonstration selection strategy assumes that a training dataset is available during inference, which may not always be feasible in practice.

9 Ethics Statement

In conducting this research, we adhered to ethical guidelines to ensure that our methods and implementations did not perpetuate or exacerbate discrimination against any group. We acknowledge the significant ethical responsibilities that accompany the deployment of LLMs in decision-making tasks, particularly in sensitive areas such as income prediction and crime risk assessment. Throughout our experiments, we employed publicly available datasets, avoiding the use of private or personally identifiable information. Our demonstration selection strategy is specifically designed to mitigate biases and enhance the fairness of LLM outputs, aiming to contribute positively towards more trustworthy AI technologies. We also encourage the broader research community to critically evaluate and iteratively improve fairness-aware methodologies to better address the complex, multifaceted nature of bias in AI systems.

References

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. GPT4All: Training an assistant-style chatbot with large scale data distillation from GPT-3.5-Turbo.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Annual Meeting of the Association for Computational Linguistics*.

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

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

Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin N Silva, Deborah L Hall, and Huan Liu. 2022. Bias mitigation for toxicity detection via sequential decisions. In *SIGIR*.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv e-prints*.

Ching-Yao Chuang and Youssef Mroueh. 2021. Fair mixup: Fairness via interpolation. In *ICLR*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

inversion Jeffrey Sorensen Lucas Dixon Lucy Vasserman nithum cjadams, Daniel Borkan. 2019. Jigsaw unintended bias in toxicity classification.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Dheeru Dua, Casey Graff, et al. 2017. Uci machine learning repository.

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

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643.

Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv e-prints*, pages arXiv–2306.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Robert W McGee. 2023. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)*.

Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*.

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

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *FAccT*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *NAACL*, pages 2655–2671.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *ICLR*.

Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and fair-maml: learning fairly with minimal data. In *FAccT*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv e-prints*.

Mohit Wadhwa, Mohan Bhambhani, Ashvini Jindal, Uma Sawant, and Ramanujam Madhavan. 2022. Fairness for text classification tasks with identity information data augmentation methods. *arXiv e-prints*, pages arXiv–2203.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36.

Nan Wang, Qifan Wang, Yi-Chia Wang, Maziar Sanjabi, Jingzhou Liu, Hamed Firooz, Hongning Wang, and Shaoliang Nie. 2023b. Coffee: Counterfactual fairness for personalized text generation in explainable recommendation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ml models with sensitive subspace robustness. In *ICLR*.

Chen Zhao and Feng Chen. 2020. Unfairness discovery and prevention for few-shot regression. In *ICKG*.

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023a. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*.

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

A Experimental Settings

In this subsection, we introduce the details of experimental settings.

A.1 Models

Large Language Models (LLMs) recently exhibited significant learning and generalizing capabilities in natural language processing due to their massive parameter sizes. However, LLMs also present challenges from different perspectives of trustworthiness. In our study, we conduct experiments to evaluate the fairness of three distinct LLMs:

- GPT-3.5. GPT-3.5, also known as ChatGPT (OpenAI, 2022), stands out for its specialized optimization for dialogue, which significantly enhances its ability to follow instructions. This capability allows for greater generalizability and personalization, such as configuring the specific roles and conversation types of the model (Ouyang et al., 2022; Wei et al., 2021; Chung et al., 2022). Such a capability differentiates GPT-3.5 significantly from classic models like BERT (Devlin et al., 2018). In particular, GPT-3.5's advancements facilitate the applications of LLMs in more complex tasks such as question-answering, via utilizing several demonstrations as additional input. Nevertheless, these new capabilities inevitably introduce additional fairness issues, as the bias in real life could exist in the data for pre-training and ultimately be encoded in model parameters. The fairness issues, such as discrimination, could raise concerns about the reliability of these LLMs in practice. Specifically, we utilize the gpt-3.5-turbo-0301 model for GPT-3.5.

- GPT-4. GPT-4 (Anand et al., 2023), released shortly after GPT-3.5, continues to further improve the capabilities of LLMs in large-scale deployments (Bubeck et al., 2023). GPT-4 not only inherits GPT-3.5's enhanced instruction-following capabilities but also introduces further refinements that enable new functionalities, such as more sophisticated question-answering and robust in-context learning (Wang et al., 2023a). GPT-4's design aims to handle a broader range of user prompts and scenarios, thereby providing more reliable performance under various scenarios (Peng et al., 2023). Similar to GPT-3.5, the new capabilities of GPT-4 also necessitate rigorous evaluations to address emergent fairness concerns and ensure its trustworthy deployment in practice (Sun et al., 2024). In particular, we consider the gpt-4-0613 model for GPT-4.

Table 3: The detailed statistics of each dataset used for evaluation in this work.

Dataset	$|\mathcal{X}_L|$	Sens.	# Feat.	Label
Adult-Gender	45,222	Gender	12	Income
Adult-Race	45,222	Race	12	Income
Jigsaw-Gender	3,563	Gender	-	Toxicity
Jigsaw-Race	6,125	Race	-	Toxicity
Jigsaw-Religion	7,127	Religion	-	Toxicity

A.2 Datasets

In this subsection, we introduce the details of the datasets used in our work. The detailed statistics are provided in Table 3.

- **Adult**. The Adult dataset (Dua et al., 2017) is prevalently used in evaluating the fairness of machine learning models. This dataset originates from the 1994 U.S. Census Bureau database and aims to predict whether an individual's annual income is more than $50,000 or not, based on their profile data. The Adult Dataset contains 48,842 samples, each representing an individual with 12 attributes, including age, weight, education level, etc. Additionally, each individual has 2 sensitive attributes: "race" and "gender". The binary label is obtained based on whether the income is more than $50,000 or not.

- **Jigsaw**. In 2019, Jigsaw (cjadams, 2019) released a dataset as part of the "Unintended Bias in Toxicity Classification" Kaggle competition. This dataset comprises approximately two million text samples from online discussions and includes ratings for toxicity along with annotations for various demographic groups. A text sample is classified under a sensitive group (i.e., a given sensitive attribute value) if it has any related annotation. We consider the original training data as the labeled set, filtering out samples without annotations. Similarly, we extract test samples from the test set in the original dataset, while removing samples without annotations. Each text sample is annotated with a toxicity score, with scores above 0.5 labeled as toxic. Notably, the Jigsaw dataset is obtained

Algorithm 1 Detailed overall process of our framework.

```
Input: Labeled sample set $\mathcal{X}_L$, Test sample $x$, Demonstration size $D$, hyper-parameters $K$, $N_d$, $N_m$.
Output: Selected in-context learning demonstrations $\mathcal{D}(x)$ for $x$.
    // Preparing phase
Perform $K$-Means on $\mathcal{X}_L$ to obtain $K$ clusters, i.e., $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$;
for $s = \{0, 1\}$ do
    for $y = \{0, 1\}$ do
        $\mathcal{X}_s^y \leftarrow \{x_i | a_i = s, y_i = y, i \in [1, |\mathcal{X}_L|]$;
        for $i = 1, 2, \ldots, K$ do
            $\mathcal{C}_s^y(i) \leftarrow \mathcal{C}_i \cap \mathcal{X}_s^y$;
        end for
    end for
end for
Obtain $N_d$ clusters i.e., $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{N_d}\}$, according to Eq. (4);
for $s = \{0, 1\}$ do
    for $y = \{0, 1\}$ do
        for $i = 1, 2, \ldots, N_d$ do
            $\mathcal{G}_s^y(i) \leftarrow \mathcal{G}_i \cap \mathcal{X}_s^y$;
        end for
        $\mathcal{G}_{s,y}^* \leftarrow \{\mathcal{G}_s^y(1), \mathcal{G}_s^y(2), \ldots, \mathcal{G}_s^y(N_d)\}$;
        Obtain $N_m$ sub-clusters, i.e., $\mathcal{G}_{s,y}^* = \{\mathcal{G}_s^y(1), \mathcal{G}_s^y(2), \ldots, \mathcal{G}_s^y(N_m)\}$, according to Eq. (7);
    end for
end for
    // Inference phase
for $s = \{0, 1\}$ do
    for $y = \{0, 1\}$ do
        Select $D/4$ sub-clusters, $\mathcal{D}_s^y(x)$, from $\mathcal{G}_{s,y}^*$ according to Eq. (8);
    end for
end for
$\mathcal{D}(x) \leftarrow \bigcup_{y,s\in\{0,1\}} \bigcup_{\mathcal{D}\in\mathcal{D}_s^y(x)} \text{argmax}_{c\in\mathcal{D}} f_s(x, c)$.
```

via crowdsourcing, and thus there could be multiple annotations on a sample. In this case, we decide the sensitive attribute values based on majority voting.

A.3 Implementation Details

In this section, we introduce the implementation details for our experiments. Particularly, we conduct all our experiments on a single Nvidia GeForce RTX A6000 GPU with a memory of 48GB. The experiments are repeated 10 times to obtain the values of accuracy, ΔDP, ΔEO, and the unfairness score, along with their standard deviation. By default, we set $K = 64$, $N_d = 16$, and $N_m = 8$. For the text encoder to embed each input sample, we utilize Sentence-BERT (Reimers and Gurevych, 2019)) with a dimension size of 768, i.e., $d = 768$. We use DecodingTrust (Wang et al., 2023a), and Fairlearn (Bird et al., 2020) for evaluation.

B Algorithm

Here we provide the detailed overall process of our demonstration selection strategy in Algorithm 1.