Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning

Jingyu Hu[1], Weiru Liu[1], Mengnan Du[2]

[1]University of Bristol [2]New Jersey Institute of Technology

{ym21669,weiru.liu}@bristol.ac.uk, mengnan.du@njit.edu

Abstract

Recent studies highlight the effectiveness of using in-context learning (ICL) to steer large language models (LLMs) in processing tabular data, a challenging task given the structured nature of such data. Despite advancements in performance, the fairness implications of these methods are less understood. This study investigates how varying demonstrations within ICL prompts influence the fairness outcomes of LLMs. Our findings reveal that deliberately including minority group samples in prompts significantly boosts fairness without sacrificing predictive accuracy. Further experiments demonstrate that the proportion of minority to majority samples in demonstrations affects the trade-off between fairness and prediction accuracy. Based on these insights, we introduce a mitigation technique that employs clustering and evolutionary strategies to curate a diverse and representative sample set from the training data. This approach aims to enhance both predictive performance and fairness in ICL applications. Experimental results validate that our proposed method dramatically improves fairness across various metrics, showing its efficacy in real-world scenarios.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023), Claude-3 (AnthropicAI, 2023), and LLaMA-2 (Touvron et al., 2023), have achieved state-of-the-art performance in many natural language processing tasks. These LLMs can adapt to different tasks by adding in-context prompts, without needing to retrain on the entire new dataset. This optimization technique for LLMs is called in-context learning (ICL), which leverages specific input prompts to guide LLMs to generate more accurate outputs. Recent research suggests that incorporating specially selected demonstrations into these prompts can significantly enhance LLM performance (Brown et al., 2020; Schick and Schütze, 2020).

Due to prompt length limitations, traditional LLMs have faced challenges in processing demonstrations from tabular datasets, which have a large number of features. However, with recent LLMs relaxing input length constraints, new avenues for applications in tabular datasets are opening up. (Hegselmann et al., 2023) has confirmed the predictive capabilities of LLMs on datasets from UCI repository. Considering the usages of tabular data in high-stakes domains (Grinsztajn et al., 2022), ensuring fairness alongside prediction performance is crucial for increasing trust in LLMs. (Liu et al., 2023) has highlighted biases in LLM predictions with tabular datasets, but there are limited further investigations on how the fairness of LLMs performance varies with different ICL demonstrations.

To bridge this gap, we aim to answer the following research question: *How do different demonstration strategies impact the fairness performance of LLMs on tabular classification tasks? And is there a demonstration strategy better than other strategies?* To better understand the impact of in-context learning on fairness, our proposed demonstration strategy considers the distribution of both demographic groups and target labels. A dataset can be divided into subgroups by demographic features, labelling the smallest as the minority (underrepresented) and the larger one as the majority. The fairness investigation compares differences between majority and minority groups. Our investigation includes evaluating five advanced LLMs, i.e., Text-davinci-003, GPT-3.5-turbo, GPT-4-turbo [1], Claude-3 Haiku, and Claude-3 Sonnet[2], across two fairness-focused tabular datasets: Credit and Adult. We found that prioritizing underrepresented samples and conscientiously including minority demographic groups and target labels during few-shot learning can significantly improve the fairness performance in LLMs output.

Despite the experimental observations, we are still wondering: *Why does prioritizing minority group demonstrations benefit the fairness performance of LLMs in tabular-based classification tasks?* To further clarify this phenomenon, we perturb prediction labels and sensitive features in selected demonstrations and compare how the prediction outcomes of LLMs would be altered. Through these perturbation experiments, we found that increasing the proportion of underrepresented labels enhances fairness, but can lead to a decline in prediction performance, and vice versa.

Up until now, the above findings and explanations have been based on random demonstration selection. We hypothesize that: *We can deliberately select demonstrations to further improve fairness performance.* Motivated by the fiLter-thEn-Search (LENS) (Li and Qiu, 2023) for textual classification, we adopt a similar process for extracting tabular demonstrations: first refine the training data set into a candidate pool and then sample and evaluate these candidates to identify the most supportive demonstrations. To this end, we introduced the Fairness via Clustering-Genetic (FCG) algorithm to effectively extract representative samples, to further enhance the fairness of LLM. Unlike LENS, which relies on progressive iterations on LLMs for candidate selection, our FCG method utilizes clustering. Clustering does not require access to LLMs and maintains diversity among the selected shots, effectively addressing concerns related to the time required for iterations and the cost of LLM access. Additionally, previous studies often assume the same selection probabilities for candidates across evaluation iterations, requiring enormous iterations to ensure that each sample is equally considered. Inspired by genetic evolution concepts, we adopt dynamic probabilities which give priority to representative samples with higher selection chances. Sample representativeness is measured by the LLM performance score, whose score is updated for each iteration. In this way, FCG can narrow the final sample set more efficiently, and drastically reduce the number of iterations needed. We implement experiments to evaluate the proposed FCG demonstration selection method. The results confirm that FCG algorithm improves LLMs performance in almost all strategies, with prioritizing the minority group still yielding the best results.

To conclude, the main contributions in this paper are as follows:

- We find that prioritizing underrepresented samples and conscientiously including minority demographic groups and target labels during few-shot learning can dramatically improve the fairness performance in LLM output (Section 3).
- We explain why prioritizing minorities leads to a fairer solution, and find the trade-off between LLMs' performance and demographic labels: increasing the ratio of underrepresented labels enhances fairness, but can lead to a decline in prediction performance, and vice versa (Section 4).
- We propose the FCG (Fairness via Clustering-Genetic) algorithm, an efficient approach to retrieve a diverse and representative set of demonstrations from training data. Across almost all strategies, FCG enhances fairness in LLMs under in-context learning (Section 5).

2 Experiment Setup

Our primary goal is to investigate how different few-shot demonstration choices influence the fairness performance of LLMs under the in-context learning (ICL) setting. Detailed related work on this area is discussed in Appendix 6. In this section, we introduce the overall experimental setups.

Notations. Given a dataset $D = (X, Y, Z)_{i=1}^{n}$ where features $X \in \mathcal{R}$, the binary classification labels $Y \in \mathcal{Y} := \{0, 1\}$, and sensitive feature $Z \in \mathcal{Z} := \{0, 1\}$. We set $Z = 0$ to represent the minority group and $Z = 1$ as the majority group. D is split into training dataset D_{tr}, validation dataset D_{dev} and testing dataset D_{test}. For each data point $d \in D := \{x, y, z\}$, a classifier f predicts the label $f(x)$ based on the input features x.

Given a subset $D' \in D_{tr}$, the proportion of samples where $Z = 0$ within D' is denoted as r_z. Specifically, $r_z = 1$ means all samples in D' belong to a minority group, whereas $r_z = 0$ implies that every sample in D' is from the majority group. Similarly, the proportion of samples for which $Y = 0$ within D' is represented by r_y.

$$r_z = \frac{\text{Count}(D'_{Z=0})}{\text{Count}(D')}, r_y = \frac{\text{Count}(D'_{Y=0})}{\text{Count}(D')} \quad (1)$$

Models and Datasets. We use five LLMs as f: Text-davinci-003 (Davinci), GPT-3.5-turbo, GPT-4-turbo, Claude-3 Haiku, and Claude-3 Sonnet. The temperature in the model parameter is set to zero to ensure consistent responses. We select two tabular-based fairness datasets: default of credit card clients dataset (Credit, (Yeh, 2016)) and adult

income (Adult, (Becker and Kohavi, 1996)). The Credit dataset covers information on credit card clients in Taiwan, including demographics, bills, payment history, etc. Its target is to predict whether there will be an overdue payment next month. The Adult dataset is to predict whether an individual's annual income exceeds 50K based on their individual features. Appendix A contains further descriptions of dataset structures.

Evaluation Metrics. The predictive performance of LLMs on labels Y is evaluated by metrics $Accuracy$, $Precision$, $Recall$, and $F\text{-}score$ [3]. We introduce Δ_{dp}, R_{dp}, Δ_{eo}, and R_{eo} to evaluate fairness [4]. They refer to the differences and ratios of Demographic Parity (DP) (Dwork et al., 2012) and Equalized Odds (Eodds) (Hardt et al., 2016) between subgroups.

The demographic parity of the two groups partitioned by Z is defined by Equation 2. DP difference Δ_{dp} represents the difference between two, and DP ratio R_{dp} is the ratio of the DP_0 and DP_1.

$$DP_0 = P(f(x) = 1 \mid Z = 0)$$
$$DP_1 = P(f(x) = 1 \mid Z = 1) \quad (2)$$

The True Positive Rate (TPR) and False Positive Rate (FPR) for both subgroups ($Z = 0$ and $Z = 1$) are defined as follows.

$$TPR_0 = P(f(x) = 1 \mid y = 1, Z = 0)$$
$$TPR_1 = P(f(x) = 1 \mid y = 1, Z = 1)$$
$$FPR_0 = P(f(x) = 1 \mid y = 0, Z = 0)$$
$$FPR_1 = P(f(x) = 1 \mid y = 0, Z = 1) \quad (3)$$

Eodds difference Δ_{eo} is defined as the greater metrics of TPR and FPR differences (Equation 4) where $\Delta TPR = (TPR_1 - TPR_0)$ and $\Delta FPR = (FPR_1 - FPR_0)$.

$$\Delta_{eo} = \max(\Delta TPR, \Delta FPR) \quad (4)$$

Eodds ratio R_{eo} is the smaller ratio of TPR and the ratio of FPR between two groups, as shown below.

Here ϵ is used to avoid the setting where the denominator is zero, where we set $\epsilon = 1^{-6}$:

$$R_{eo} = \min \left(\frac{TPR_0}{TPR_1 + \epsilon}, \frac{FPR_0}{FPR_1 + \epsilon} \right) \quad (5)$$

The four fairness metrics range from 0 to 1. Lower Δ_{dp} and Δ_{eo} show smaller performance differences between groups, which points to fairer

[3] https://scikit-learn.org/stable/
[4] https://fairlearn.org/main/user_guide/fairness_in_machine_learning.html

predictions. Higher R_{eo} and R_{dp} reflect more consistent performance across subgroups, suggesting better fairness.

Prompt Template. The output answer from the LLMs is based on the input prompt. As shown in Figure 1, the structure of a prompt can be divided into three parts: task description, in-context demonstrations, and questions. Part ❶ clarifies the task and defines the format of output prediction label options. Part ❷ contains a series of demonstrations as references. Part ❸ is the sample to be predicted.

We consider both zero-shot learning and few-shot learning in our experiments. Zero-shot learning refers to LLMs with a prompt exclude demonstration references (without ❷) and is set as the baseline. Few-shot learning, sometimes also called in-context learning (ICL), consists of all three parts as input prompts. We compare how different demonstrations in part ❷ influence the fairness of LLMs. The prompt example in Figure 1 simplifies the tabular dataset, the detailed template is provided in Appendix B.

3 How Demonstrations Impact Fairness of LLMs for Tabular Inputs?

In this section, we aim to answer: how do few shot demonstrations influence the fairness performance of LLMs for processing tabular inputs under the in-context learning setting?

To investigate this, we examine fairness performance variances across different demonstrations. We propose different combinations of prediction feature distribution r_z and sensitive feature distribution r_y, expecting to explore the potential correlation between these feature distributions and LLM fairness. In the experiment, different demonstrations are based on three distinct sampling strategies denoted as $S1$, $S2$, and $S3$, each with unique distribution combinations of r_z and r_y.

- S1: Balanced Samples with Balanced Labels ($r_z = 0.5, r_y = 0.5$);
- S2: Prioritize Minority Samples with Balanced Labels ($r_z = 1, r_y = 0.5$);
- S3: Prioritize Minority Samples with Unbalanced Labels ($r_z = 1, r_y \neq 0.5$).

Figure 2 displays the performance of different LLMs on the Credit dataset. r_y is set to 1 in S3. The fairness performance improves when prioritizing samples from minority groups ($r_z = 1$) compared to a balanced sample selection ($r_z = 0.5$).
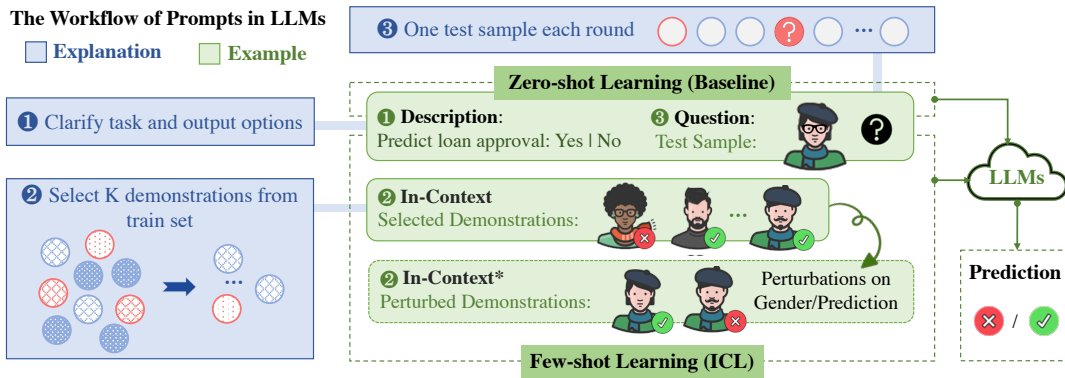
Figure 1: The Prompt Template and Content Example (*Perturbation is optional and is used to test the effectiveness of ICL. We discussed perturbations in Section 4)
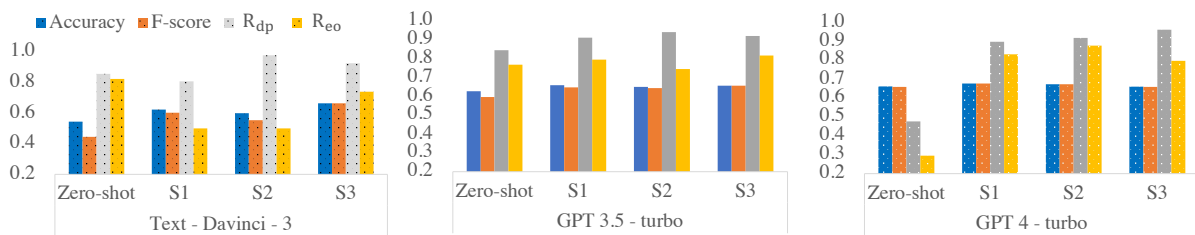

Figure 2: Prediction and fairness performance comparison across different LLMs on Credit dataset. It shows improvements in fairness metrics when samples from minority groups are prioritized.

Similar findings are found in the Adult dataset. Table 1 presents the performance of the GPT-3.5-turbo with zero-shot and different few-shot strategies. To ensure the stability and reliability of the results, we use random seeds set={25, 35, 42, 45, 55} when selecting few-shot samples. The presented table summarizes average values and standard errors for the random seeds set.

Overall, the results show that all few-shot strategies have generally improved fairness compared to zero-shot learning without lowering predictions. Also, prioritizing minorities (S2, S3) is an effective way to improve fairness. In contrast, balanced prompts (S1) show worse fairness performance. To further explain the observed pattern, we implement additional experiments and discussions on GPT-3.5's performance under the Adult dataset in the following sections. Complete results for other LLMs (e.g., Claude), using different seeds, are included in Appendix C.

4 Why Prioritizing Minority Group Demonstrations Benefit Fairness?

The above analysis points out a strong correlation between prioritizing minority group demonstrations with improved fairness performance of LLMs.
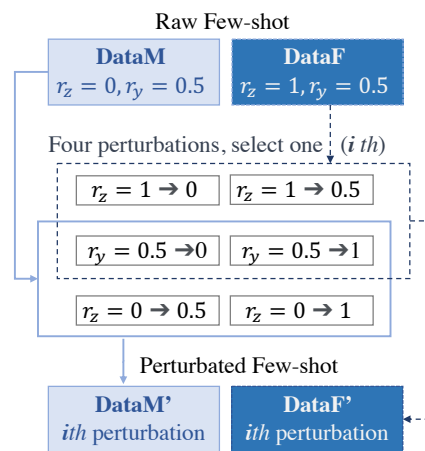
Figure 3: The Workflow of Perturbations

However, it is not yet clear how and why this phenomenon occurs. Thereby our next step is to clarify which part of the demonstrations most influenced the performance of LLMs. Specifically, we perturb the prediction label Y and the sensitive feature Z in selected demonstrations and compare how the prediction outcomes of LLMs would be altered.

The following experiment is performed on the Adult dataset with 'income' as feature Y and 'gender' as feature Z. We set the random seed to 55

7463

Table 1: Performance of GPT3.5-turbo with zero-shot and different few-shot strategies (S1, S2, S3) on Adult Income dataset. It demonstrates that strategic inclusion of demonstrations, particularly those from minority groups, can significantly enhance both predictive performance and fairness outcomes.

Prediction	Zero-shot	$r_z = 0.5, r_y = 0.5$ (S1)	$r_z = 1, r_y = 0.5$ (S2)	$r_z = 1, r_y = 1$ (S3)
$Accuracy \uparrow$	0.6855	0.7312 ± 0.0009	**0.7328 ± 0.0028**	0.7230 ± 0.0014
$Precision \uparrow$	**0.8519**	0.7936 ± 0.0012	0.7841 ± 0.0051	0.7808 ± 0.0038
$Recall \uparrow$	0.4492	0.6250 ± 0.0012	**0.6461 ± 0.0130**	0.6122 ± 0.0036
$F\text{-}score \uparrow$	0.5882	0.6993 ± 0.0011	**0.7060 ± 0.0062**	0.6915 ± 0.0017
Fairness	Zero-shot	$r_z = 0.5, r_y = 0.5$ (S1)	$r_z = 1, r_y = 0.5$ (S2)	$r_z = 1, r_y = 1$ (S3)
$R_{dp} \uparrow$	0.4063	0.6470 ± 0.0019	0.6769 ± 0.0080	**0.6732 ± 0.0095**
$R_{eo} \uparrow$	0.1111	0.3682 ± 0.0044	0.4152 ± 0.0125	**0.4722 ± 0.0187**
$\Delta_{dp} \downarrow$	0.2227	0.1688 ± 0.0009	0.1578 ± 0.0031	**0.1555 ± 0.0046**
$\Delta_{eo} \downarrow$	0.3203	0.1875 ± 0.0019	**0.1859 ± 0.0058**	0.1906 ± 0.0071

to extract two groups with balanced labels $DataF$ and $DataM$ from D_{tr} as raw demonstrations. (1) $DataF$: balanced high-income and low-income females ($r_y = 0.5, r_z = 1$). (2) $DataM$: balanced high-income and low-income males ($r_y = 0.5, r_z = 0$). Figure 3 illustrates the perturbation workflow. We define four perturbations, each consisting of the feature to be perturbed and the new proportions after perturbation. For example, perturbing $r_y = 0.5 \rightarrow 1$ means that the quantity of high-income and low-income samples are balanced in raw demonstrations ($r_y = 0.5$), and the perturbed demonstrations will become all low-income samples ($r_y' = 1$) by flipping high-income labels to low-income.

The next part will discuss how perturbations at different proportions affect the overall prediction and fairness performance of LLM, along with a deeper performance comparison within subgroups.

4.1 Perturbations Impact on Overall Fairness

Table 2 compares the prediction and fairness performance with different perturbations on gender and income.

Perturbing the income labels for $DataF$ and $DataM$ leads to a certain degree of decline in predictive performance (1% to 6%). Min et al., 2022 mentioned a similar phenomenon that replacing gold labels with random labels only presents the marginal impact on performance. Nevertheless, we also found that altering the ground truth labels (income) can greatly affect fairness performance, resulting in a drastic drop in all scenarios. Especially when replacing labels with high income, the R_{dp} in DataF decreased from 71.32% to 50.54%, and in DataM, it decreased from 50.94% to 43.06%.

When we perturbed gender labels, results show

that fairness performance improves with a higher proportion of females. The fairness performance decreases when we perturb from female to male in $DataF$, as R_{dp} decreases from 71.32% to 59.15%. Similar patterns are observed in $DataM$, where fairness gradually increases by 8.1% when modifying from male to female.

In most cases, the perturbation results align with the intuition that distorting real data can degrade its quality, thus potentially leading to negative impacts on LLMs performance. However, we also find that perturbing to a higher ratio of minority labels ($r_z' = 1$) can positively enhance fairness, suggesting a strong connection between fairness performance and sensitive labels. To further validate this finding, Section 4.2 compares performance variations at the subgroup level.

4.2 Perturbations Impact across Subgroups

Table 3 displays the model performance of TPR and FPR on both minority (female) and majority (male) subgroups after income and gender perturbations. Similar to DP and EO, the metrics ΔTPR and ΔFPR assess performance disparities between female and male subgroups. Equal treatment is achieved when these differences approach zero, hence, lower values of ΔTPR and ΔFPR are preferable. We also consider absolute values of FPR and TPR within each subgroup to fully assess fairness changes in perturbations.

In income perturbations, replacing the income labels results in a decrease in TPR and FPR for both female and male groups, with a more significant decline observed in the female group. This reduction is most notable when income labels are changed to high-income. In a few cases, the relative metrics ΔTPR and ΔFPR show improve-

Table 2: Prediction and Fairness Performance on Income and Gender Perturbations

	Different perturbations on income						Different perturbations on gender					
	$r_z=1, r_y=0.5$ (DataF)			$r_z=0, r_y=0.5$ (DataM)			$r_z=1, r_y=0.5$ (DataF)			$r_z=0, r_y=0.5$ (DataM)		
Prediction	Raw	$r_y'=1$	$r_y'=0$	Raw	$r_y'=1$	$r_y'=0$	Raw	$r_z'=0.5$	$r_z'=0$	Raw	$r_z'=0.5$	$r_z'=1$
$Accuracy \uparrow$	**0.7480**	0.7383	0.6992	**0.7148**	0.6914	0.6543	0.7480	0.7422	**0.7617**	0.7148	**0.7168**	0.7090
$Precision \uparrow$	0.7873	0.7933	**0.8643**	0.8438	0.8451	**0.8835**	0.7873	0.7925	**0.7965**	**0.8438**	0.8364	0.8204
$Recall \uparrow$	**0.6797**	0.6445	0.4727	**0.5273**	0.4688	0.3555	0.6797	0.6563	**0.7031**	0.5273	0.5351	**0.5352**
$F\text{-}score \uparrow$	**0.7296**	0.7112	0.6111	**0.6490**	0.6030	0.5070	0.7296	0.7179	**0.7469**	0.6490	**0.6556**	0.6478
Fairness	Raw	$r_y'=1$	$r_y'=0$	Raw	$r_y'=1$	$r_y'=0$	Raw	$r_z'=0.5$	$r_z'=0$	Raw	$r_z'=0.5$	$r_z'=1$
$R_{dp} \uparrow$	**0.7132**	0.6508	0.5054	**0.5094**	0.4639	0.4306	**0.7132**	0.6308	0.5915	0.5094	0.5421	**0.5905**
$R_{eo} \uparrow$	**0.3824**	0.3438	0.0556	0.1364	**0.1579**	0.0000	**0.3824**	0.2571	0.1500	0.1364	0.2273	**0.3043**
$\Delta_{dp} \downarrow$	**0.1445**	0.1719	0.1797	0.2031	0.2031	**0.1602**	**0.1445**	0.1875	0.2266	0.2031	0.1914	**0.1680**
$\Delta_{eo} \downarrow$	**0.1641**	0.1797	0.226	0.2578	0.2813	**0.2266**	**0.1641**	0.2031	0.2656	0.2578	0.2500	**0.2109**

ment compared to the results without perturbations. However, the corresponding absolute metrics TPR and FPR do not show consistent trends and worsen instead. This inconsistency makes it difficult to validate the impact of income on fairness performance. Therefore, we conclude that the ground truth labels in the demonstrations are not the source of benefit for LLMs' fairness.

In gender perturbations, however, subgroup performance is greatly affected by these gender label changes. For absolute values, flipping female labels to male in $DataF$ leads to a 4.69% increase in FPR and a 5.47% increase in TPR for the male group. Similarly, transforming male labels to female in $DataM$ results in increases in both TPR and FPR for the female group. Similar trends are found in their relative values. Increasing the proportion of male labels leads to higher ΔTPR and ΔFPR, illustrating greater difference in subgroup treatment. Conversely, an increase in the ratio of female labels leads to reductions in both ΔTPR and ΔFPR, suggesting enhanced fairness.

In general, the above results show a trade-off between the LLMs performance and demographic labels. LLMs exhibit improved performance when the proportion of minority groups increases: they become fairer compared to the use of original data when perturbing demographic labels from male to female. Therefore, we conclude that prioritizing demonstrations from minority groups can maximize these advantages and promote fairness in LLMs. In contrast, perturbing labels leads to demonstrations becoming less reliable, as they can lead models to learn noise and perform worse. The perturbation on prediction labels (income) conforms with this pattern.

5 Mitigation Algorithm for Fair Demonstration Selection

The above results confirm that the application of diverse demonstrations, particularly those from the minority, can drastically influence the fairness of LLMs. Experiments on different sets of selected shots under the same strategy also reveal a similar trend, albeit with different absolute performance values. This leads to the question: *how to extract representative demonstrations that yield better performance among these sets?*

Enumerating and evaluating the outcomes of LLMs across all possible sets is impractical due to the sheer number of combinations. Thus, in this section, we propose a fairness via clustering-genetic (FCG) algorithm to efficiently select influential demonstrations, leading LLMs to a better performance without having to explore all possible combinations. The core idea of FCG includes three aspects: (1) Use clustering to shrink the selection sets while maintaining diverse samples. (2) Define a score that considers both prediction and fairness performance, applying a genetic approach to iteratively select and score these samples within the sets. (3) Rank samples from highest to lowest based on their scores to select the most influential ones.

5.1 The Proposed FCG Algorihtm

We introduce details of the proposed FCG mitigation algorithm in this section.

Clustering. Based on the value combinations of the sensitive feature Z and label Y, we divide the training data D_{tr} into four subgroups denoted as SG $= \{g_1(Z = 1, Y = 0), g_2(Z = 1, Y = 1), g_3(Z = 0, Y = 0), g_4(Z = 0, Y = 1)\}$. For each subgroup, we apply k-means clustering to extract a diverse and representative initial population.

Table 3: TPR and FPR Assessment across Subgroups on Income and Gender Perturbations

| | Different perturbations on income | | | | | | Different perturbations on gender | | | | | |
| | $r_z=1, r_y=0.5$ (DataF) | | | $r_z=0, r_y=0.5$ (DataM) | | | $r_z=1, r_y=0.5$ (DataF) | | | $r_z=0, r_y=0.5$ (DataM) | | |
	Raw	$r'_y=1$	$r'_y=0$	Raw	$r'_y=1$	$r'_y=0$	Raw	$r'_z=0.5$	$r'_z=0$	Raw	$r'_z=0.5$	$r'_z=1$
TPR_M	**0.7422**	0.7344	0.5859	**0.6563**	0.6094	0.4688	0.7422	0.7422	**0.7969**	**0.6563**	0.6641	0.6406
TPR_F	**0.6172**	0.5547	0.3594	**0.3984**	0.3281	0.2422	**0.6172**	0.5703	0.6094	0.3984	0.4141	**0.4297**
$\Delta_{TPR}\downarrow$	**0.1250**	0.1797	0.2266	0.2578	0.2813	**0.2266**	**0.1250**	0.1719	0.1875	0.2578	0.2500	**0.2109**
FPR_M	0.2656	0.2500	0.1406	**0.1719**	0.1484	0.0938	0.2656	0.2734	**0.3125**	**0.1719**	0.1719	0.1797
FPR_F	0.1016	0.0859	0.0078	**0.0234**	0.0234	0.0000	**0.1016**	0.0703	0.0469	0.0234	0.0391	**0.0547**
$\Delta_{FPR}\downarrow$	0.1641	0.1641	**0.1328**	0.1484	0.1250	**0.0938**	**0.1641**	0.2031	0.2656	0.1484	0.1328	**0.1250**

Algorithm 1 FCG Algorithm

```
 1: procedure STEP1: DIVERSE CLUSTERING
 2:     all_idx_set =[]
 3:     for each g_i in SG do
 4:         selected_idx_set = []
 5:         X = g_i.get_all_data()
 6:         centroids = Kmeans(clusters = n).fit(X)
 7:         for each center in centroids do
 8:             dis_set = distance(X,center)
 9:             closest = argsort(dis_set)[:m]
10:             selected_idx_set.extend(closest)
11:         end for
12:         all_idx_set.append(selected_idx_set)
13:     end for
14:     return SG'.set(all_idx_set)
15: end procedure
16: procedure STEP2:UPDATE EVOL SCORE
17:     ch = [M_pred, M_fair]
18:     for each g'_i in SG' do
19:         shots = g'_i.genetic_algorithm(k)
20:         evol_score = CAL_SCORE(shots,ch)
21:         shots.update(evol_score)
22:         repeat iters times
23:     end for
24: end procedure
25: function CAL_SCORE(shots,ch)
26:     Y_base= LLM.predict(X_dev)
27:     Base_pred, Base_fair = eval(Y_base,Y_dev, ch)
28:     Y_ICL = LLM.predict(shots, X_dev)
29:     ICL_pred, ICL_fair = eval(Y_ICL, Y_dev, ch)
30:     Δpred = max((ICL_pred − Base_pred),p)
31:     Δfair = max((ICL_fair − Base_fair),p)
32:     return α · Δpred + (1 − α) · Δfair
33: end function
```

Each subgroup in SG is clustered into n clusters, with m neighbors selected around each centre of the cluster. The filtered new subgroups are denoted as SG'=$\{g'_1, g'_2, g'_3, g'_4\}$.

Genetic Evolution for Score Updates. Next, for each subgroup within SG', we select K-demonstrations for $iters$ times using the roulette wheel genetic evolutionary approach and validate their ICL performance on D_{dev}. The evolutionary method means that data with a higher score is more likely to be chosen in each round. The score is first set as the default initialisation score p and then updated as the average of EvolScore computed during the iterations when the sample is selected. EvolScores integrates the performance of both prediction M_{pred} and fairness M_{fair} metrics, with α serving as the trade-off coefficient. The metrics provide options for selecting either $Accuracy$ or F-$score$ as M_{pred}, and either R_{dp} or R_{eo} as M_{fair}. EvolScores in SG' will be updated and then used for subsequent selecting iterations.

In the testing stage on D_{test}, demonstrations in SG' are ranked by their average EvolScores, enabling different ICL strategies to select the top-performing demonstrations from their corresponding subgroups. The detailed steps of FCG pseudocode is given in Algorithm 1. Figure 4 gives an example of the whole process of representative sample selections with FCG on Adult dataset.

5.2 Experimental Results

Table 4 presents the experimental results evaluating the debiasing performance of the proposed FCG algorithm. The experiments are performed on the Adult dataset, setting the number of clusters to $n = 8$ and the number of neighbors to $m = 5$. We start with an initial score of $p = 0.05$ and perform 10 iterations to update EvolScore. F-$score$ and R_{eo} are selected for calculating Evolscore and the balancing parameter α is set to 0.5.

Results show that demonstrations selected by FCG perform well, and greatly outperforming random sampling. It is worth noting that using a balanced set from minority samples continues to yield the best performance, proving our finding that prioritizing minority samples ($r_z = 1$) remains an effective strategy in ICL. Besides the minority group, the improvements in accuracy and fairness also happen in the majority group, which affirms the value of considering both factors in FCG selections.

Ablation Study. We implement ablation experiments to verify the utility of the two-step extracting process in FCG mitigation. In the ablation study,
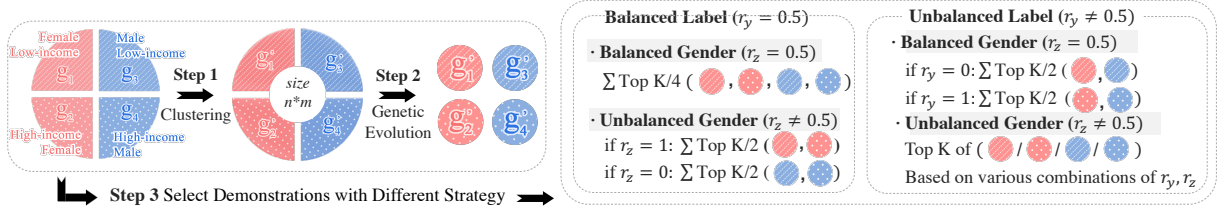
Figure 4: The Workflow of Fairness via Clustering-Genetic (FCG) on the Adult Dataset ($r_y = 1$, all high-income; $r_y = 0$, all low-income; $r_y = 0.5$, balanced samples of high-income and low-income; $r_z = 1$, all females; $r_z = 0$, all males; $r_z = 0.5$, balanced samples of females and males.)

Table 4: The comparative analysis of the predictive and fairness performance achieved by various LLMs with demonstrations selected using the proposed FCG algorithm. The experiments are conducted on the Adult dataset. The table highlights that the FCG algorithm enhances fairness across almost all strategies.

| | Zero-shot | K-Shot (K=8) | | | | | K-Shot (K=4) | | | |
| | Zero-shot | $r_y = 0.5$ (Balanced Labels) | | | $r_y = 1$ | $r_y = 0$ | $r_y = 0.5$ (Balanced Labels) | | | $r_y = 1$ | $r_y = 0$ |
Prediction	Baseline	$r_z = 0.5$	$r_z = 0$	$r_z = 1$	$r_z = 1$	$r_z = 1$	$r_z = 0.5$	$r_z = 0$	$r_z = 1$	$r_z = 1$	$r_z = 1$
$Accuracy \uparrow$	0.6855	0.7344	0.7363	**0.7793**	0.7500	0.7754	0.7656	0.7500	0.7656	0.7539	0.7578
$Precision \uparrow$	**0.8519**	0.7174	0.6997	0.7360	0.7207	0.7640	0.7297	0.7222	0.7194	0.7338	0.7200
$Recall \uparrow$	0.4492	0.7734	0.8281	**0.8711**	0.8164	0.7969	0.8438	0.8125	**0.8711**	0.7969	0.8438
$F\text{-}score \uparrow$	0.5882	0.7444	0.7585	**0.7979**	0.7656	0.7801	0.7826	0.7647	0.7880	0.7640	0.7770
Fairness	Baseline	$r_z = 0.5$	$r_z = 0$	$r_z = 1$	$r_z = 1$	$r_z = 1$	$r_z = 0.5$	$r_z = 0$	$r_z = 1$	$r_z = 1$	$r_z = 1$
$R_{dp} \uparrow$	0.4063	0.7692	0.7719	**0.8938**	0.7576	0.7919	0.7515	0.8000	0.8675	0.7707	0.8072
$R_{eo} \uparrow$	0.1111	0.6250	0.5690	0.7021	0.5577	0.5750	0.5094	0.6000	**0.7059**	0.5417	0.6800
$\Delta_{dp} \downarrow$	0.2227	0.1406	0.1523	**0.0664**	0.1563	0.1211	0.1641	0.1250	0.0859	0.1406	0.1250
$\Delta_{eo} \downarrow$	0.3203	0.1406	0.1953	**0.1094**	0.1797	0.1328	0.2031	0.1563	0.1172	0.1719	0.1250

part of the samples are selected using the same flow of choosing the top K samples based on their EvolScores, while the other part is selected randomly. The results in Table 5 suggest: (1) Even when EvolScores are ignored when selecting partial samples, the results still outperform the raw random selection method (Random ($g_1 + g_2$)), thus proving the effectiveness of the clustering selection in the first stage. (2) Moreover, both ablation test FCG (g_1) and FCG (g_2) performed worse compared to the results using complete FCG ($g_1 + g_2$), further confirming the need for the second stage of genetic selection based on EvolScore scoring.

6 Related Work

6.1 Fairness in Large Language Model

Addressing social biases is crucial for ensuring the trustworthiness of language models (Nangia et al., 2020; Nadeem et al., 2020). LLMs face similar fairness issues: many studies have confirmed that LLMs can capture social biases from unprocessed training data and transmit these biases to downstream tasks (Abid et al., 2021b; Brown et al., 2020; Wang et al., 2023). Abid et al. (2021a) addressed

Table 5: The Ablation Study of FCG under Balanced Labels in Minority Group Strategy (S2) for Selecting K Demonstrations (K=8). S2 strategy is based on minority group ($z = 1$) with two possible labels $y = \{0, 1\}$, The corresponding subgroups are denoted as $g_1(z = 1, y = 0)$ and $g_2(z = 1, y = 1)$.

Shots from g_1	Random	Top K/2	Top K/2	Random K/2
Shots from g_2	Random	Top K/2	Random K/2	Top K/2
Prediction	Random ($g_1 + g_2$)	FCG ($g_1 + g_2$)	FCG (g_1)	FCG (g_2)
$Accuracy \uparrow$	0.7480	**0.7793**	0.7500	0.7559
$Precision \uparrow$	**0.7592**	0.7360	0.7013	0.7079
$Recall \uparrow$	0.7266	**0.8711**	**0.8711**	**0.8711**
$Fscore \uparrow$	0.7425	**0.7979**	0.7770	0.7811
Fairness	Random ($g_1 + g_2$)	FCG ($g_1 + g_2$)	FCG (g_1)	FCG (g_2)
$R_{dp} \uparrow$	0.7254	**0.8938**	0.8276	0.8208
$R_{eo} \uparrow$	0.4390	0.7021	0.6964	0.6140
$\Delta_{dp} \downarrow$	0.1523	**0.0664**	0.1172	0.1211
$\Delta_{eo} \downarrow$	0.1797	**0.1094**	0.1328	0.1719

the issue of GPT-3's output displaying bias to Muslims. Huang et al. (2021) found that bias in LLMs' responses exists even without prompts explicitly asking about it. Liang et al. (2023) tested bias and stereotypes on LLMs using the BBQ dataset for question answering, finding that most models exhibit biases distinct from broader societal trends. Wang et al. (2023) assesses bias by prompting GPT models to express their views on stereotypes.

Most bias studies have focused on representational harms caused by LLM generations, with only limited studies (Hegselmann et al., 2023) addressing fairness concerns classification problems with tabular datasets. Besides investigation on pre-trained LLMs, recent research also focuses on ensuring fairness in other trained machine learning models, such as perturbation-based (Zhang et al., 2022; Wang et al., 2022) and boosting-based (Kim et al., 2019) approaches. Hu and Du (2024) discusses the phenomenon that prioritizing minority demonstrations can enhance fairness in LLMs on tabular classification tasks. This paper builds upon this finding and further validates it across a broader range of LLMs. Furthermore, we delve deeper into the reasons behind this phenomenon and propose a Fairness via Clustering-Genetic (FCG) selection method, which extracts more representative demonstration samples to further improve fairness.

6.2 In Context Learning

In-context learning (ICL), known as few-shot learning, allows LLMs to learn tasks with minimal examples as demonstrations (Dong et al., 2022; Zhao et al., 2021). Positive impacts of ICL on LLMs have been observed in different tasks such as text classification and answering (Gao et al., 2021; Liu et al., 2021), images generations (Bar et al., 2022), speech tasks (Zhang et al., 2023), and multi-modal scenarios (Huang et al., 2023; Wei et al., 2022). Meanwhile, researchers have found that the performance of ICL is highly sensitive to the demonstration prompt (Chen et al., 2023; Lu et al., 2021; Zhao et al., 2021; Shi et al., 2022). Investigations have explored factors that can influence ICL prediction performance, including demonstration retrievals (Tanwar et al., 2023; Sorensen et al., 2022), orderings (Lu et al., 2022), and input-label mapping (Yoo et al., 2022; Work).

Demonstration Retrievals. A common demonstration retrievals approach in ICL involves randomly selecting a subset of examples from the training set (Brown et al., 2020). Given the sensitivity of LLMs to the prompts, there has been investigation into selecting representative samples to enhance outcomes. Selecting the top-K training examples is one mitigation method and has been demonstrated in semantic parsing (Rubin et al., 2021) and semantic classification (Chang and Jia, 2023). LENS (Li and Qiu, 2023) proposed a two-step filter and search algorithm to identify informative support examples. Despite these advances, these retrieval techniques often focus solely on prediction performance and overlook the aspect of fairness. Additionally, most retrieval methods often require extensive experimental iterations, with significant time and resource investment.

7 Conclusions

In this paper, we investigate how the choice of demonstrations impacts the fairness of LLMs on tabular data classification tasks when using in-context learning. Through experiments, we found that prioritizing underrepresented groups and including minority examples in the few-shot demonstrations can significantly enhance fairness performance, without sacrificing predictive accuracy. Further analysis revealed that increasing the proportion of underrepresented labels improves fairness metrics like demographic parity and equal odds. To efficiently retrieve effective demonstrations, we proposed the FCG algorithm that uses clustering and genetic evolution to select a diverse and representative set of examples from the training data. Across multiple strategies and datasets, experimental results indicate that FCG was able to improve fairness compared to random sampling.

8 Limitations

While our study presents significant advancements in understanding and improving fairness in LLMs using in-context learning (ICL), several limitations should be noted. Firstly, we equally weigh fairness and prediction performance in evaluating representative demonstrations using our Fairness via Clustering-Genetic (FCG) algorithm, which might not align with real-world applications that require a dynamic balance between these metrics. Additionally, our focus on binary classification with a single sensitive feature limits the broader applicability of our findings. In future, we plan to explore LLM's intersectional fairness and its performance in multi-classification tasks. Lastly, while we used pre-trained models without fine-tuning, investigating how fine-tuning on curated samples impacts fairness could provide deeper insights.

References

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

Abubakar Abid, Maheen Farooqi, and James Zou. 2021b. Persistent anti-muslim bias in large language models.

AnthropicAI. 2023. Introducing claude.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. 2022. Visual prompting via image inpainting.

Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

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.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning.

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*.

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners.

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

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

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 *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Jingyu Hu and Mengnan Du. 2024. Enhancing fairness in in-context learning: Prioritizing minority samples in demonstrations. In *The Second Tiny Papers Track at ICLR 2024*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. *arXiv preprint arXiv:2109.06437*.

Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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*.

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

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 arXiv:2104.08786*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

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

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

OpenAI. 2023. gpt-4-turbo, gpt-3.5-turbo, text-davinci-003.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment.

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

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. 2022. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388.

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.

What Makes In-Context Learning Work. Rethinking the role of demonstrations: What makes in-context learning work?

I-Cheng Yeh. 2016. default of credit card clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. 2022. Fairness reprogramming. *Advances in Neural Information Processing Systems*, 35:34347–34362.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling.

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

A Dataset

Table 6 describes the data structure in the Default Credit dataset. We calculate the mean values of PAY_AMT_i and BILL_AMT_i, and merge them into Avg_PAY_AMT and Avg_BILL_AMT separately. The raw Adult dataset shown in Table 7 contains 14 features, excluding education-num, fnl-wgt, race, and native-country for this experiment. '> 50K' and '≤ 50K' is mapped to 'greater than 50K' and 'less than or equal to 50K' respectively, for better alignment with the language model. In analysis, we call high income if the person's annual income is over 50K and low income if it is less than 50K. The size ratio of D_{train}: D_{dev}: D_{test} is 9:1:10 in both two datasets. K demonstrations are extracted from D_{train}, 60 samples are extracted from D_{dev}, 512 samples for D_{test}. We consider the balanced group and balanced labels scenario and extract samples with parameter random_seed=42.

B Prompt Architecture

We consider both zero-shot learning and few-shot learning (in-context learning). Zero-shot strategy combines task description and question as its prompt content without providing examples. Few-shot strategy includes three roles, and the in-context content is generated based on selected K-demonstrations using different strategies (Table 10). The default value of K is set to 8, the case of K=4 is disscussed in Section 5. Table 8 and 9 provide templates for few-shot learning in the Adult and Credit datasets respectively.

C More Experimental Results

Tables below present additional detailed results not listed in the main text.

D Discussion between Our FCG and LENs Algorithm

Our proposed FCG shares a similar architecture with LENs, another demonstration selection method. Here, we discuss these two methods and further explore the possibility of combining them. Given the time consumption of LENs, we simplified it by setting batch size to 8, and template index to {0,1}. The training dataset is randomly split ($seed = 42$) into groups of 500 samples for parallel computation, with other settings kept at their defaults. For LENs with FCG, we follow our FCG parameters setting: first extract 160 representative

samples by clustering, then perform LENs to find the final candidates. Figure 5 presents the overall workflow of FCG, LENs, and their possible combination. Both FCG and LENs involve two steps: (1) selecting partial data and (2) searching for the optimal based on the filtered data. There are two key differences in implementations.

Supervised & Unsupervised: LENs algorithm uses LLMs as the classification assistance in both stages. This is a straightforward and effective way. However, since the processing time of language models is related to the amount of information in the input text, the selection time can become very long when the input data is lengthy. This study focuses on tabular datasets, which have longer text when converted into prompts compared to commonly used NLP datasets. Therefore, we consider to optimize the method to reduce processing time and improve efficiency. Our FCG replaces LLMs with simpler unsupervised algorithms in the first stage. On the adult dataset, LENs can take over 50 hours to find supportive demonstrations ($batchsize = 8$), while FCG takes less than 3 hours ($K = 4$). The result in Table 17 validates the effectiveness: even if LLMs are not used initially, using LLMs to search on the validation set in the second stage can still find demonstrations that improve the model's prediction.

Fairness Awareness: Another difference is that LENs use accuracy as the sole evaluation metric when selecting demonstrations. Our FCG takes sensitive features into account and selects demonstrations at the subgroup level. Additionally, FCG considers both accuracy and fairness metrics as constraints when calculating performance scores. Table 17 confirms FCG with minority demonstrations prioritised strategy ($r_z = 1$) shows fairer performance than baseline.

Furthermore, we extend LENs with FCG (as shown on the right side of Figure 5) to make it fairness-aware. Table 17 proves the effectiveness of this combination and also shows the best performance achieved when using more minority demonstrations ($r_z = 1$).

Table 6: Default Credit Dataset Description

Feature	Default Credit Dataset - Description	
LIMIT_BAL	Amount of given credit	Continuous; NT dollars
SEX	Gender	2 categories, male / female
EDUCATION	Highest education	6 categories; graduate / high school / university / etc
MARRIAGE	Marital status	6 categories of ; married / single / others
AGE	Age in years	Continuous
PAY_i	i∈{1,2,3,4,5,6} , represents the month from April (6) to September (1) in 2005.	10 categories of repayment status for each month; ; pay duly / delay for one month / delay for two months / etc
BILL_AMT_i		Amount of bill statement for each month; Continuous; NT dollar
PAY_AMT_i		Amount of previous payment for each month; Continuous; NT dollar
default_payment_ next_month	If default payment next month	Yes, overdue / no, on-time

Table 7: Adult Income Dataset Description

Feature	Adult Income Dataset - Description
Age	Age in years; Continuous
Workclass	8 general types of employment; private / self-employed / government / etc
Education	16 categories of highest level of education; college / bachelors / masters / etc.
Marital-Status	7 categories; married / divorced / separated / single / etc.
Occupation	15 categories; prof-specialty / craft-repair / Sales / etc
Relationship	6 categories; not-in-family / husband / wife / etc
Sex	2 categories; the biological sex; male / female
Capital-Gain	Person's capital gains; Continuous
Capital-Loss	Person's capital losses; Continuous
Hours-Per-Week	Hours worked per week; Continuous
Salary	2 categories of whether annual income exceeds $50K; >50K / <=50K

Table 8: Few-shot Learning Templates for Adult Dataset

Roles	Prompting Templates for Adult Income Dataset
Description	<Task> Predict if income exceeds $50K per year. Answer with one of the following: [greater than 50K]	[less than or equal to 50K] </Task>
In-context	</Example>Example 1:age is 40, and workclass is Private, and education is HS-grad, and marital-status is Married-civ-spouse, and occupation is Sales, and relationship is Husband, and sex is Male, and capital-gain is 0, and capital-loss is 0, and hours-per-week is 60, and income is <=50K; </Example> <Example>Example 2 ......</Example> </Example>
Question	<Input> Age is 19, and workclass is Private, and education is Some-college, and marital-status is Never-married, and occupation is Other-service, and relationship is Own-child, and sex is Female, and capital-gain is 0, and capital-loss is 0, and hours-per-week is 15, please answer the income: </Input>

Table 9: Few-shot Learning Templates for Credit Dataset

Roles	Prompting Templates for Default Credit Dataset
Description	<Task> Predict if the following data will default payment next month. Answer with one of the following: [No] | [Yes] </Task>
In-context	<Example> Example 1: Amount of given credit is 490000, and SEX is male, and EDUCATION is graduate school, and MARRIAGE is married, and AGE is 45, and PAY_0 is pay duly,......, and default payment next month is No, on-time;</Example> <Example> Example 2: ......</Example>
Question	<Input> Amount of given credit is 90000, and SEX is female, and EDUCATION is university, and MARRIAGE is married, and AGE is 49, and PAY_0 is delay for one month,......, and predict whether default payment: </Input>

Table 10: Demonstrations Selection Strategies

Annotation	Strategy	Adult Income Dataset	Default Credit Dataset
$r_z = 0.5, r_y = 0.5$	S1	A balanced 1/4 ratio of female-low-income : female-high-income : male-low-income : male-high-income.	A balanced 1/4 ratio of female-overdue : female-on-time : male-overdue : male-on-time
$r_z = 1, r_y = 0.5$	S2	Prioritize minority with balanced 1/2 ratio of female-low-income : female-high-income	Prioritize minority with balanced 1/2 ratio of female-overdue: female-on-time
$r_z = 1, r_y = 0$	S3	Prioritize minority with imbalanced targets. All female-low-income samples	Prioritize minority with imbalanced targets. All female-on-time samples
$r_z = 1, r_y = 1$	S3	Prioritize minority with imbalanced targets. All female-high-income samples	Prioritize minority with imbalanced targets. All female-overdue samples

Table 11: Different LLMs performance on Default Credit Dataset

	Text - Davinci - 3				GPT 3.5 - turbo				GPT 4 - turbo			
	Zero-shot	S1	S2	S3	Zero-shot	S1	S2	S3	Zero-shot	S1	S2	S3
Accuracy	0.5449	0.6230	0.5996	0.6641	0.6250	0.6562	0.6484	0.6543	0.6602	0.6758	0.6719	0.6582
F-score	0.4453	0.6030	0.5545	0.6641	0.5947	0.6453	0.6413	0.6543	0.6579	0.6758	0.6716	0.6578
Δdp	0.0117	0.0586	0.0039	0.0391	0.0391	0.0313	0.0234	0.0430	0.2969	0.0547	0.0469	0.0195
R_{dp}	0.8571	0.8077	0.9787	0.9254	0.8413	0.9080	0.9368	0.9160	0.4759	0.8955	0.9155	0.9590
Δeo	0.0234	0.1016	0.0547	0.1016	0.0938	0.1016	0.1094	0.1328	0.3125	0.1250	0.0938	0.1094
R_{eo}	0.8235	0.5000	0.5000	0.7400	0.7647	0.7917	0.7419	0.8132	0.2941	0.8298	0.8750	0.7955

Table 12: Performance of Claude-3-haiku and Claude-3-sonnet with Zero-shot and Different Few-shot Strategies on Adult Dataset ($r_y = 0.5$, K=4)

	Claude-3-haiku				Claude-3-sonnet			
	Zero-shot	$r_z = 0$	$r_z = 0.5$	$r_z = 1$	Zero-shot	$r_z = 0$	$r_z = 0.5$	$r_z = 1$
Accuracy	0.7285	0.7070	0.7012	0.7031	0.6641	0.7266	0.7383	0.7520
Precision	0.7489	0.8118	0.7210	0.7047	0.6556	0.7302	0.7364	0.7699
Recall	0.6875	0.5391	0.6563	0.6992	0.6914	0.7188	0.7422	0.7188
F-score	0.7169	0.6479	0.6871	0.7020	0.6730	0.7244	0.7393	0.7434
	Zero-shot	$r_z = 0$	$r_z = 0.5$	$r_z = 1$	Zero-shot	$r_z = 0$	$r_z = 0.5$	$r_z = 1$
Δdp	0.2305	0.1797	0.1836	0.1563	0.0625	0.1484	0.1094	0.1445
R_{dp}	0.5986	0.5741	0.6643	0.7279	0.8881	0.7379	0.8042	0.7319
Δeo	0.2344	0.2188	0.2031	0.1641	0.0703	0.1563	0.1094	0.1719
R_{eo}	0.3409	0.2800	0.5116	0.5625	0.8235	0.5455	0.6585	**0.5714**

Table 13: Performance of GPT3.5-turbo on the Adult Dataset through Few-shot Strategies (S1) with 5 random seeds

	Zero-shot	S1 ($r_z = 0.5, r_y = 0.5$) Few-shot with Different Random Seed						Few-shot Summary		
	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Accuracy	0.6855	0.7285	0.7363	0.7363	0.7266	0.7285	0.7313	0.0009	0.7363	0.7266
Precision	0.8519	0.7940	0.7980	0.8010	0.7871	0.7882	0.7937	0.0012	0.8010	0.7871
Recall	0.4492	0.6172	0.6328	0.6289	0.6211	0.6250	0.6250	0.0012	0.6328	0.6172
F-score	0.5882	0.6945	0.7059	0.7046	0.6943	0.6972	0.6993	0.0011	0.7059	0.6943
Fairness	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Δdp	0.2227	0.1758	0.1680	0.1680	0.1641	0.1680	0.1688	0.0009	0.1758	0.1641
R_{dp}	0.4063	0.6311	0.6504	0.6475	0.6557	0.6504	0.6470	0.0019	0.6557	0.6311
Δeo	0.3203	0.2031	0.1875	0.1797	0.1797	0.1875	0.1875	0.0019	0.2031	0.1797
R_{eo}	0.1111	0.3667	0.3667	0.3333	0.3871	0.3871	0.3682	0.0044	0.3871	0.3333

Table 14: Performance of GPT3.5-turbo on the Adult Dataset through Few-shot Strategies (S2) with 5 random seeds

	Zero-shot	S2 ($r_z = 1, r_y = 0.5$) Few-shot with Different Random Seed						Few-shot Summary		
	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Accuracy	0.6855	0.7207	0.7207	0.7480	0.7266	0.7480	0.7328	0.0028	0.7480	0.7207
Precision	0.8519	0.7958	0.8192	0.7592	0.7589	0.7873	0.7841	0.0051	0.8192	0.7589
Recall	0.4492	0.5938	0.5664	0.7266	0.6641	0.6797	0.6461	0.0130	0.7266	0.5664
F-score	0.5882	0.6801	0.6697	0.7425	0.7083	0.7296	0.7060	0.0062	0.7425	0.6697
Fairness	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Δdp	0.2227	0.1680	0.1445	0.1523	0.1797	0.1445	0.1578	0.0031	0.1797	0.1445
R_{dp}	0.4063	0.6325	0.6542	0.7254	0.6593	0.7132	0.6769	0.0080	0.7254	0.6325
Δeo	0.3203	0.2344	0.1641	0.1797	0.1875	0.1641	0.1859	0.0058	0.2344	0.1641
R_{eo}	0.1111	0.5000	0.3333	0.4390	0.4211	0.3824	0.4152	0.0125	0.5000	0.3333

Table 15: Performance of GPT3.5-turbo on the Adult Dataset through Few-shot Strategies (S3) with 5 random seeds

	Zero-shot	S3 ($r_z = 1, r_y = 1$) Few-shot with Different Random Seed						Few-shot Summary		
	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Accuracy	0.6855	0.7168	0.7168	0.7324	0.7285	0.7207	0.7230	0.0014	0.7324	0.7168
Precision	0.8519	0.7707	0.7681	0.7847	0.8128	0.7678	0.7808	0.0038	0.8128	0.7678
Recall	0.4492	0.6172	0.6211	0.6406	0.5938	0.6328	0.6211	0.0036	0.6406	0.5938
F-score	0.5882	0.6855	0.6868	0.7054	0.6862	0.6938	0.6915	0.0017	0.7054	0.6855
Fairness	Baseline	Seed 25	Seed 35	Seed 42	Seed 45	Seed 55	AVG.	SE	Highest	Lowest
Δdp	0.2227	0.1445	0.1445	0.1289	0.1758	0.1836	0.1555	0.0046	0.1836	0.1289
R_{dp}	0.4063	0.6942	0.6967	0.7273	0.6121	0.6357	0.6732	0.0095	0.7273	0.6121
Δeo	0.3203	0.1875	0.1563	0.1563	0.2188	0.2344	0.1906	0.0071	0.2344	0.1563
R_{eo}	0.1111	0.5667	0.4118	0.5517	0.3462	0.4848	0.4722	0.0187	0.5667	0.3462

Table 16: The Ablation Study with Balanced Labels in Minority Group (S2) under FCG Selections on GPT-3.5-turbo. The corresponding subgroups are denoted as $g_1(z = 1, y = 0)$ and $g_2(z = 1, y = 1)$.

K-shots	K=8			K=4		
g_2	Top K/2	Top K/2	Random K/2	Top K/2	Top K/2	Random K/2
g_1	Top K/2	Random K/2	Top K/2	Top K/2	Random K/2	Top K/2
	FCG ($g1 + g2$)	$FCG(g_2)$	FCG (g_1)	FCG ($g1 + g2$)	FCG (g_2)	FCG (g_1)
Accuracy	0.7793	0.7500	0.7559	0.7656	0.7441	0.7539
Precision	0.7360	0.7013	0.7079	0.7194	0.7036	0.7031
Recall	0.8711	0.8711	0.8711	0.8711	0.8438	0.8789
Fscore	0.7979	0.7770	0.7811	0.7880	0.7673	0.7813
	$FCG(g_1 + g_2)$	$FCG(g_2)$	$FCG(g_1)$	$FCG(g_1 + g_2)$	$FCG(g_2)$	$FCG(g_1)$
Δ_{dp}	0.0664	0.1172	0.1211	0.0859	0.1133	0.1016
R_{dp}	0.8938	0.8276	0.8208	0.8675	0.8274	0.8497
Δ_{eo}	0.1094	0.1328	0.1719	0.1172	0.1172	0.1328
R_{eo}	0.7021	0.6964	0.6140	0.7059	0.7170	0.6964

Table 17: ICL Performance of LLaMa-3-8b on the Adult Dataset ($r_y = 0.5$) using Different Demonstration Retrieval Methods (LENs, FCG, and Combined).

LLaMa-3	LENs (K = 4)	FCG (Ours, K = 4)			LENs with FCG (K = 4)		
	Baseline	$r_z = 0$	$r_z = 0.5$	$r_z = 1$	$r_z = 0$	$r_z = 0.5$	$r_z = 1$
Accuracy	0.6270	0.6406	0.6680	**0.7148**	0.5957	**0.6543**	0.6504
Precision	0.7211	**0.8462**	0.7389	0.7778	0.7426	**0.7687**	0.6103
Recall	0.4141	0.3438	0.5195	**0.6016**	0.2930	0.4414	**0.8320**
F − score	0.5261	0.4889	0.6101	**0.6784**	0.4202	0.5608	**0.7041**
	Baseline	$r_z = 0$	$r_z = 0.5$	$r_z = 1$	$r_z = 0$	$r_z = 0.5$	$r_z = 1$
Δdp	0.1523	0.1250	0.1094	**0.0938**	0.1602	0.1445	**0.0039**
R_{dp}	0.5806	0.5294	0.7308	**0.7838**	0.4225	0.5978	**0.9943**
Δeo	0.2344	0.1719	0.1172	**0.0938**	0.2109	0.1953	**0.0469**
R_{eo}	0.5588	0.2308	0.5161	**0.5714**	0.3000	0.4783	**0.9155**

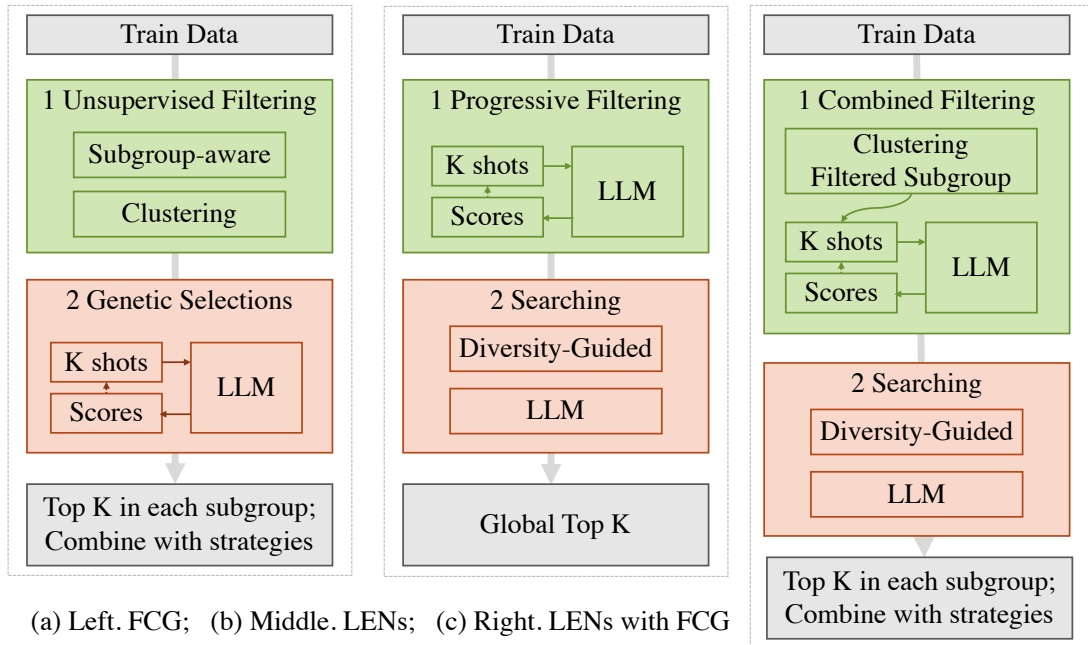(a) Left. FCG;   (b) Middle. LENs;   (c) Right. LENs with FCG

Figure 5: The workflow comparison of demonstration selection algorithms: FCG (Ours, proposed in Section 5), LENs, and LENs with FCG.