Can Bias in Large Language Models for Tabular Data Classification Be Mitigated?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) perform well in tabular prediction tasks with limited data, using their ability to understand instructions and learn from examples. However, their reliance on training data can perpetuate social biases, leading to unfair outcomes and disproportionately impacting underprivileged groups. Addressing these biases is critical as LLMs see wider adoption in tabular data tasks. Traditional bias mitigation strategies in machine learning, such as balancing datasets or applying fairness constraints, are less effective with LLMs. Our research explores whether bias in LLMs for tabular data classification can be mitigated. Through extensive experiments, we found that using LLMs in a zero-shot setting introduces bias, and in-context learning slightly reduces the bias. Meanwhile, fine-tuning and retrieval augmented generation show limited effectiveness in bias mitigation. We introduced three instruction-based prompting strategies to enhance fairness: *Fair Prompting*, *Generalised Prompting*, and *Descriptive Prompting*. The results show that combining descriptive prompting with in-context learning, particularly the Equal Samples Across Demographics approach, significantly improved fairness metrics such as Statistical Parity Ratio and Equal Opportunity Ratio and yielded accuracy gains ranging from 3.27% to 15.05% across multiple datasets, underscoring its potential as a powerful strategy in the ongoing effort to mitigate bias in LLMs.

1 Introduction

Large Language Models (LLMs) have marked a substantial leap forward in artificial intelligence (Zhao et al., 2023a). A prime example is the Generative Pre-trained Transformer (GPT) model (Achiam et al., 2023), which has demonstrated its robust capabilities across diverse tasks, including machine translation (Xu et al., 2023), text generation (Li et al., 2024), and complex question-answering (Fergus et al., 2023). Recently, LLMs have found applications far beyond their original uses in language processing. Recent studies have uncovered the potential of LLMs for predictive tabular data tasks (Fang et al., 2024; Hegselmann et al., 2023; Slack and Singh, 2023; Yang et al., 2024). In these studies, tabular data is converted into natural language and presented to LLMs with a brief task description to generate predictions. The findings conclude that LLMs for tabular data classification achieve significant performance, demonstrating the method's capacity to leverage their encoded prior knowledge (Slack and Singh, 2023).

However, as LLMs are capable of generating human-like content, they can perpetuate social biases present in the extensive datasets they were trained on, potentially causing significant harm to underprivileged groups (Abid et al., 2021; Basta et al., 2019; Ganguli et al., 2022). Undoubtedly, the issue of LLM-generated unfair responses and biases is a multifaceted problem (Gallegos et al., 2024). With the widespread adoption of LLMs across various industries and the extensive use of tabular data in high-stakes domains (Grinsztajn et al., 2022), it is essential to thoroughly examine the fairness implications and mitigate biases when using LLMs for tabular data classification. Bias in traditional machine learning models was addressed by ensuring datasets were diverse and balanced through careful data collection and preprocessing. Fairness constraints and mitigation algorithms were applied during model training, and outputs were adjusted to promote fairness, with regular monitoring and audits to maintain accountability (Mehrabi et al., 2021). However, these strategies are unsuitable for LLM in tabular data classification tasks because LLMs are pre-trained, and biases may arise from the training dataset. To address this, we pose the research question: *Can bias in LLMs for tabular data classification be mitigated? If yes, to what extent?* The answer to this question has profound implications for LLM applications with the potential to enhance fairness, equity, and trustworthiness across various domains. We aim to investigate the challenge of mitigating biases generated by LLMs in tabular data classification tasks. Extensive experiments using both open-source and proprietary models across tabular datasets demonstrate that:

- **Using LLMs for tabular data classification can introduce bias.** Experiments in a zero-shot setting show that LLM transfer social biases from their pre-training data into tabular tasks (details in Results). This evidence demonstrates that LLMs adopt social biases from their pre-training data and frequently rely on these biases when classifying tabular data, leading to potentially unfair outcomes.

- **In-context learning in LLMs enhances model performance for classification tasks; however, it only slightly reduces bias.** We develop a framework of strategies to mitigate bias in LLMs for tabular data classification, categorising approaches for clarity and practical use. Providing LLMs with few-shot examples using strategies like mitigation through unawareness, counterfactuals, and equal samples across demographics improves accuracy and F1 score. While fairness shows some improvement, the overall impact remains limited, highlighting the need for more robust solutions.

- **The effectiveness of fine-tuned model and Retrieval Augmented Generation (RAG) in LLMs for tabular data classification is limited in terms of bias mitigation.** We also fine-tuned the LLMs using an extensive training dataset and the RAG technique. While this approach contributes to bias reduction, we observed only slight effects, highlighting the need for more effective bias mitigation techniques.

To enhance the fairness of LLMs as tabular data classifiers, we propose three instruction-based prompting approaches: *Fair Prompting*, *Generalised Prompting*, and *Descriptive Prompting*. These strategies guide LLMs toward equitable predictions. Experiments show descriptive prompting with in-context learning improves fairness, achieving Statistical Parity Ratio and Equal Opportunity Ratio values closer to 1 and accuracy gains of 3.27% to 15.05% across datasets. These strategies advance bias mitigation, marking significant progress in promoting fairer AI outcomes.

2 Related Work

2.1 LLM for Tabular Data

LLMs have been trained on vast amounts of data, enabling them to achieve impressive performance across various downstream tasks (Brown et al., 2020). Recent studies have used LLMs for tabular data classification (Zhao et al., 2023b; Wang et al., 2024). For example, the TABLET benchmark reveals improved LLM performance from instructions in tabular data predictions (Slack and Singh, 2023). Hegselmann et al. (2023) explored using LLMs for the classification of tabular data by converting tables to natural language and providing problem descriptions, finding this method outperforms traditional techniques and competes with strong baselines. Yang et al. (2024) enhanced LLMs' ability to handle tabular data for classification, regression, and imputation tasks by training Llama-2 on a comprehensive corpus of annotated tables, demonstrating significant improvements. However, the research concludes that for LLM-based tabular data prediction methods, the fairness metric gap between different subgroups is larger than that observed in traditional machine learning models (Ma et al., 2024). Therefore, while it is established that fairness issues exist in LLMs for tabular data classification tasks, the methods to mitigate this bias remain largely unexplored. To our knowledge, our work represents one of the most comprehensive investigations into mitigating bias when using LLMs to classify tabular data.

2.2 Fairness and Biases in LLMs

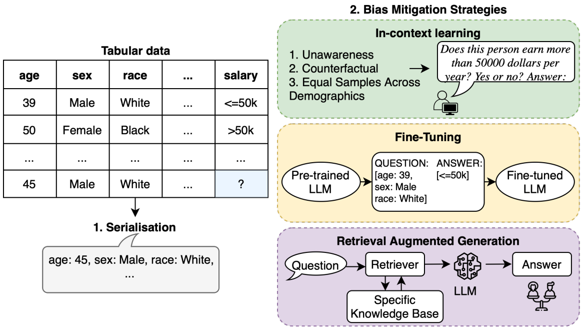
While LLMs are rapidly advancing in capabilities and applications, biased systems can produce discriminatory and stereotypical outcomes, negatively impacting underprivileged or vulnerable groups and causing societal harm (Kumar et al., 2022). LLMs may produce biased or prejudiced responses when the training data contains stereotyped or discriminatory information (Nadeem et al., 2020). Research has shown that these models frequently display biases concerning gender (Cai et al., 2024; Kotek et al., 2023), profession (Nadeem et al., 2020), race (Haim et al., 2024), and religion (Gallegos et al., 2024). Researchers are addressing these issues by developing improved benchmarks, such as CrowS-Pairs (Nangia et al., 2020) and RealToxicityPrompts (Gehman et al., 2020), to assess and mitigate unfairness in LLMs. Additionally, regarding prompt engineering, Chisca et al. (2024) proposed a novel prompt-tuning to reduce these biases in models, effectively mitigating gender bias with minimal impact on performance. Ma et al. (2024) introduced a metric to evaluate bias in prompts and propose a greedy search strategy to identify near-optimal prompts. Although there is research on fairness in LLMs, there remains a significant gap in studies specifically addressing mitigation strategies for tabular data classification.

2.3 In-context Learning, Fine-tuning and RAG for Tabular Data

In-context learning uses examples to guide LLMs toward desired outputs. Guo et al. (2023) reported a 30.38% accuracy drop when switching from one-shot to zero-shot settings, while Chen (2022) found that increasing shots from one to two improved performance. This approach is crucial for integrating contextual information and fairness guidelines to enhance equitable outcomes. Chhikara et al. (2024) introduced a framework incorporating fairness rules, demonstrating GPT-4's superior accuracy and fairness using in-context learning. Liu et al. (2024) highlighted that in-context learning reduces fairness gaps between subgroups, and Hu and Du (2024) showed that including minority samples in prompts improves fairness without compromising performance. Fine-tuning involves training pre-trained LLMs on specific datasets to improve accuracy. Zhang et al. (2023) fine-tuned Llama-2 for better tabular task performance, and similar methods have been explored (Hegselmann et al., 2023; Jaitly et al., 2023; Liu et al., 2024; Wang et al., 2023). Further, RAG adds domain-specific context to prompts but faces challenges in relevance extraction. Sundar and Heck (2023) addressed this with a dual-encoder Dense Table Retrieval model for better table cell ranking. These techniques enhance LLM performance, driving the need for a fairness framework to mitigate bias and assess their understanding of fairness in classification tasks.

3 Methodology

This section outlines bias mitigation methods for LLMs in tabular data classification, including detection, in-context learning, fine-tuning, and RAG. Figure 1 illustrates the methodology, covering tabular data serialisation and three mitigation strategies.


Tabular data

age	sex	race	...	salary
39	Male	White	...	<=50k
50	Female	Black	...	>50k
...	...	...	...	...
45	Male	White	...	?

1. Serialisation
age: 45, sex: Male, race: White, ...
2. Bias Mitigation Strategies
In-context learning
1. Unawareness
2. Counterfactual
3. Equal Samples Across Demographics
Does this person earn more than 50000 dollars per year? Yes or no? Answer:
Fine-Tuning
Pre-trained LLM
QUESTION: [age: 39, sex: Male race: White]
ANSWER: [<=50k]
Fine-tuned LLM
Retrieval Augmented Generation
Question
Retriever
Specific Knowledge Base
LLM
Answer

Figure 1: **Framework for Bias Mitigation in LLM Tabular Data Classifications.** We serialise the tabular data, then apply three bias mitigation strategies to improve fairness: (a) in-context learning, (b) fine-tuning, and (c) retrieval-augmented generation. These strategies can be used separately or in combination.

3.1 Bias Detection

Bias in LLM tabular classifications refers to the systematic favouritism or discrimination against certain groups based on their demographic characteristics or other attributes (Liu et al., 2024). Such bias can arise from various sources, including the source of bias coming from the pre-training step, model architecture, as well as the societal and historical context in which the models are developed and deployed (Hegselmann et al., 2023). To detect and measure this bias, we conduct evaluations from two perspectives: model utility and fairness.

3.1.1 Model Utility

We evaluate the model using accuracy and F1 score. Accuracy measures overall performance across subgroups, while F1, the harmonic mean of precision and recall, accounts for imbalances in the datasets.

3.1.2 Fairness Definition

We assess fairness using statistical parity and equal opportunity. Statistical Parity Ratio (SPR) ensures that the probability of a positive outcome is similar across different demographic groups (Garg et al., 2020). The formula is as follows:

$$SPR = \frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)}$$

where $\hat{Y}$ is the predicted outcome, and A is the demographic attribute. Statistical parity requires that the probability of a positive outcome is the same across different demographic groups. This metric highlights the relative difference in outcomes, showing how much more likely one group is to receive a positive outcome than the other. When the SPR is less than 1, it suggests potential bias against the demographic $A = 1$. On the other hand, when the SPR is greater than 1, it suggests potential bias against the demographic $A = 0$.

Further, the Equal Opportunity Ratio (EOR) ensures that individuals who qualify for a positive outcome have an equal chance of being correctly identified by the model, regardless of their demographic group (Garg et al., 2020). The formula is as follows:

$$EOR = \frac{P(\hat{Y} = 1 \mid Y = 1, A = 1)}{P(\hat{Y} = 1 \mid Y = 1, A = 0)}$$

where $\hat{Y}$ is the predicted outcome, Y is the actual outcome, and A is the demographic attribute. Equal opportunity requires that individuals from different demographic groups who qualify for a positive outcome (i.e., $Y = 1$) should have an equal probability of being assigned a positive outcome by the model. This metric highlights the relative difference in true positive rates, showing how often the protected group is more likely to have a true positive prediction than the unprotected group. Both SPR and EOR, being close to their ideal values (i.e., 1), are indicators of a fair model with respect to the specified fairness criteria.

3.2 Framework for Bias Mitigation in LLMs

3.2.1 Serialisation, Prompt Templates and In-context Learning

Tabular data is organised into rows and columns, where each column represents a feature and each row corresponds to an instance. Transforming this structured data into a format suitable for LLMs involves either flattening the data into sequences by concatenating all features of each instance (row-wise) or using embeddings for both categorical and continuous features. Building on previous studies on LLMs for tabular data classification (Hegselmann et al., 2023; Slack and Singh, 2023), we format the feature names and values into strings in the format "$f_1 : x_1, \ldots, f_k : x_k$," where f represents the feature names and x represents the corresponding values. After that, we propose three few-shot prompt strategies, each using n examples extracted from the training dataset, where n can be any number depending on the specific case; This study focuses on scenarios with limited or no training data, where LLMs perform well by utilising their knowledge for classification (Slack and Singh, 2023). In this research, we use a ten-shot approach as an example, with three prompt strategies: 1. *Unawareness*: Sensitive or protected attributes are removed from in-context learning. 2. *Counterfactual*: Sensitive attributes are altered to evaluate and adjust model predictions for fairness. 3. *Equal Samples Across Demographics*: An equal number of samples from each demographic group is ensured. The example template is provided in Appendix A.

3.2.2 Fine-Tuning and RAG

We fine-tune on the entire training set to assess its impact on reducing biases in LLMs. For the OpenAI model, we utilise the publicly available API from OpenAI for fine-tuning (OpenAI, 2024a). For open-source models, we employ Low-Rank Adaptation (LoRA) fine-tuning. LoRA assumes weight matrices in the form of low-rank factors, significantly reducing the number of trainable parameters and the computational resources required for fine-tuning (Hu et al., 2021). The data format and parameters for fine-tuning are shown in Appendix B. On the other hand, RAG operates on the principle that while LLMs are designed to be versatile and capable of generating a wide range of responses, they can be significantly enhanced by embedding relevant context within the prompts, thereby directing the LLMs towards domain-specific responses (Fang et al., 2024). This approach improves accuracy and relevance by incorporating external knowledge sources. Specifically, the studies in the Related Work Section on fairness and bias using LLM for tabular data classification can be classified as RAG systems for our research.

3.3 Enhancing Fairness with Prompting Strategies

We introduce three instruction-based prompting designed to enhance the fairness of LLMs. These strategies, which can be used alongside in-context learning, aim to improve both fairness and model performance. Below, we detail the three strategies, including exemplar prompts and their integration into the LLMs, as illustrated in Figure 2:

- **Fair Prompting**. We hypothesise that explicitly instructing the LLMs to consider fairness can enhance its ability to make equitable predictions (Gallegos et al., 2024). To achieve this goal, we propose employing instruction-based prompting by adding the phrase *"Ensure fairness for demographic features, such as gender"* to the prompt input. This instruction aims to guide the LLM to incorporate fairness considerations into its decision-making process.

- **Generalised Prompting**. We also hypothesise that fairness can be achieved by providing clear definitions and instructions. To facilitate this, we propose a method called Generalised Prompting. This involves appending specific instructions to the prompt input to guide the model towards fairness. For instance, to achieve both Demographic Parity and Equal Opportunity, we add the instruction: *"Ensure that the probability of predicting an income above $50K is equal for both males and females, and that the probability of correctly predicting an income above $50K is equal for both males and females who actually earn above $50K"*. This instruction is designed to guide the LLM towards making balanced and fair predictions across gender groups.
- **Descriptive Prompting**. We assume that explicitly detailing the fairness criteria can help in achieving fairness. We propose using instruction-based prompting by appending detailed descriptions to the prompt input. For instance, to achieve both Demographic Parity and Equal Opportunity, we add the instruction: *"Demographic Parity ensures that predictions do not favour one group over another based on attributes like gender or race. Ensure that the probability of predicting an income above $50K is equal for males and females. Equal Opportunity ensures equal chances of correct classification for positive outcomes across groups. Ensure that the probability of correctly classifying individuals earning more than $50K is the same for males and females"*. This instruction provides a clear directive to the LLM to maintain fairness in its classifications.

These prompting strategies are appended to the input text before classification. For instance, if the original input for a tabular data prediction task includes demographic and feature values, the corresponding fairness instruction is added as part of the prompt. This ensures that the model processes both the input data and the fairness directive together, influencing the generation of predictions.

4 Experiments

4.1 Dataset

We use four recognised datasets typically employed to assess fairness in traditional machine learning models to explore the fairness of LLMs in classifying tabular data: Adult Income (Adult) (Kohavi et al., 1996), Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Angwin et al., 2022), Diabetes (Strack et al., 2014), and Student Performance (Student) (Cortez and Silva, 2008). The Adult dataset predicts whether an individual's income exceeds $50,000 based on demographic features. The COMPAS dataset assesses recidivism risk, focusing on race as a protected attribute. The Diabetes dataset predicts 30-day hospital readmissions, using gender as the protected attribute. Lastly, the Student Performance dataset predicts final-year grades, with sex as the protected attribute. Due to the time and cost constraints associated with LLMs, we randomly selected 1,000 samples from each dataset as the test set for experiments if 20% of the dataset exceeded 1,000 samples; otherwise, we used 20% of the dataset as the test set. Table 1 provides a summary of these datasets, highlighting their features and classification tasks.

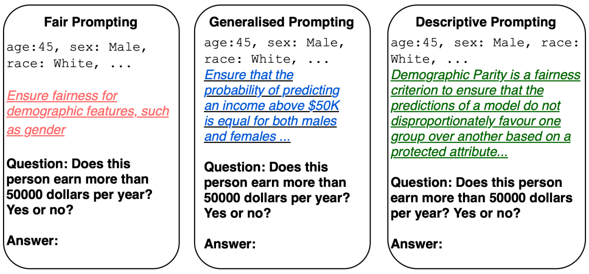

Fair Prompting
age:45, sex: Male, race: White, ...
Ensure fairness for demographic features, such as gender
Question: Does this person earn more than 50000 dollars per year? Yes or no?
Answer:
Generalised Prompting
age:45, sex: Male, race: White, ...
Ensure that the probability of predicting an income above $50K is equal for both males and females ...
Question: Does this person earn more than 50000 dollars per year? Yes or no?
Answer:
Descriptive Prompting
age:45, sex: Male, race: White, ...
Demographic Parity is a fairness criterion to ensure that the predictions of a model do not disproportionately favour one group over another based on a protected attribute...
Question: Does this person earn more than 50000 dollars per year? Yes or no?
Answer:

Figure 2: **Overview of fair prompting, generalised prompting, and descriptive prompting,** we use these to mitigate bias in LLMs for tabular data classification.

Dataset	Features	Label
Adult	Work class, hours per week, sex, age, occupation, capital loss, education, capital gain, marital status, relationship.	Income: $\leq$ 50k or >50k
COMPAS	Sex, race, age, charge degree, priors count, risk.	Two-year recidivism (yes/no)
Diabetes	Excludes weight, payer code, medical specialty features due to missing data.	30-day readmissions (yes/no)
Student Performance	Includes features on demographic, academic, and social factors.	Grade: Low (<10) or High ($\geq$10)

Table 1: Summary of Datasets, Features, and Labels

4.2 Baselines

Establishing a baseline is essential for comparing bias mitigation strategies. Previous studies have shown that users can provide instructions to LLMs to achieve strong performance on tabular datasets without additional data collection (Slack and Singh, 2023). In this study, we use a zero-shot setting as the baseline to evaluate LLM performance without fairness interventions, focusing on their natural strengths and weaknesses in handling tabular data.

4.3 Models and Setting

The selection criteria for the evaluated LLMs include accessibility, a balance between open-source and proprietary solutions, support for tabular data tasks, and suitability for computationally intensive experiments or methods requiring advanced reasoning. There are numerous LLMs available, including both open-source models and proprietary models. Open-source models, like Llama (Meta, 2024) and Gemma (Google, 2024), are publicly available and customisable but may lack production optimisation and long-term maintenance. In contrast, proprietary models, such as OpenAI models (Achiam et al., 2023), are optimised for production but are not publicly accessible, customisable, or free, requiring trust in the model owner for data privacy and responsible AI use. We use GPT-4o as our default LLM and evaluate three other LLMs (GPT-3.5-Turbo, Meta-Llama-3-8B, and Gemma-2) to balance open-source and proprietary models. For all baselines, we set the model temperature to 0. The experiments are conducted on a server equipped with an A40 GPU, boasting 50 GB of memory. Our code and dataset are available at `https://anonymous.4open.science/r/fairllm-AC4D/`.

5 Results

5.1 Bias Introduction in Tabular Data Classification LLMs

To assess LLM fairness in tabular classification, we conducted zero-shot experiments, evaluating fairness metrics without in-context learning or fine-tuning. Each experiment was repeated five times to account for variability, with mean and standard deviation calculated for robustness.

Figure 3 demonstrates the disparities in prediction metrics for various subgroups across different datasets when utilising LLMs for tabular data classification. We use the fifth (last) experiment as an example to plot the figure. Each subplot depicts the performance of the model across different metrics (Accuracy (ACC), Positive Rate (PR), Negative Rate (NR), True Positive Rate (TPR), and True Negative Rate (TNR)) for specified demographic groups. For instance, the Adult dataset compares performance metrics between male and female groups, whereas the COMPAS dataset compares white and black groups. The subplots reveal noticeable discrepancies in model performance between these groups. Specifically, significant variations can be seen in metrics such as PR and TPR, indicating potential bias in the LLM's predictions. Such differences suggest that the model may favour certain subgroups over others, leading to unfair and biased outcomes. This highlights the problem of fairness in LLM-based tabular data classification. The evidence shows that LLMs inherit social biases from their pretraining data and use these biases in classifying tabular data, resulting in potentially unfair outcomes. This highlights the importance of developing strategies to mitigate bias and ensure equitable performance across all demographic groups. Appendix C, Table 6, provides the fairness evaluation of LLMs across four datasets in a zero-shot setting.

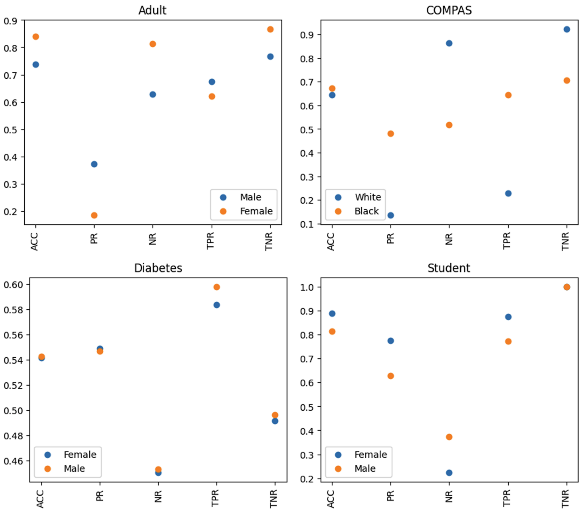


Figure 3: **Disparities in Prediction Metrics Across Demographics using LLMs for Tabular Data.** Subplots show metrics (Accuracy (ACC), Positive Rate (PR), Negative Rate (NR), True Positive Rate (TPR), True Negative Rate (TNR)) for demographic groups in the datasets (Adult, COMPAS, Diabetes, Student). Discrepancies highlight bias and fairness issues. Overlapping dots in the Student dataset indicate identical True Negative Rate (TNR) for male and female students.

5.2 In-context Learning

To evaluate the effectiveness of in-context learning for improving fairness, we provide LLMs with few-shot examples, followed by a test example and task description to guide predictions, a method we refer to as the *foundation* approach. As detailed in the Serialisation, Prompt Templates and In-context Learning, the few-shot examples are positioned before the test example in the prompt. We use ten randomly selected in-context examples from each dataset's training set. Next, we examine the impact of in-context learning on fairness metrics. Table 2 demonstrate that incorporating few-shot examples improves accuracy and F1 scores across the four datasets using GPT-4o, indicating that LLMs can effectively learn input-label mappings within the provided context. More details for other LLMs are included in Appendix C, Tables 7, 8, 9, and 10. Results show that combining few-shot examples with bias mitigation strategies, such as the "Equal Samples Across Demographics" approach, reduces fairness metric gaps between subgroups, particularly in the Adult and COMPAS datasets. However, these improvements are limited for the Diabetes and Student datasets, where metrics such as SPR and EOR remain far from 1. Additionally, standard deviations in these experiments are higher than the baseline, indicating variability in outcomes. While in-context learning paired with targeted bias mitigation strategies demonstrates some potential for enhancing fairness, finding a universal strategy that consistently balances accuracy and fairness across datasets remains a challenge. Further exploration is necessary to refine these methods for broader applicability.

Dataset	Type	Acc	F1	SPR	EOR
Adult	F	0.8185	0.6592	1.6572	**1.0690**
		(0.0154)	(0.0122)	(0.1322)	(0.0163)
	U	0.8631	0.7157	1.6047	1.0829
		(0.0040)	(0.0370)	(0.1151)	(0.1014)
	E	**0.8769**	**0.7290**	**1.5944**	1.0773
		(0.0120)	(0.0444)	(0.1467)	(0.1009)
	C	0.8442	0.6916	1.6250	1.0934
		(0.0193)	(0.0269)	(0.1075)	(0.0928)
COMPAS	F	0.6973	0.6911	1.5110	1.2437
		(0.0069)	(0.0045)	(0.0605)	(0.0580)
	U	0.6925	0.6787	1.5542	1.4056
		(0.0098)	(0.0094)	(0.0541)	(0.1060)
	E	**0.7091**	**0.7041**	**1.4672**	**1.2012**
		(0.0078)	(0.0043)	(0.0864)	(0.0720)
	C	0.6615	0.5498	2.5961	2.0500
		(0.0020)	(0.0037)	(0.0368)	(0.0233)
Diabetes	E	0.6822	0.6809	**1.4195**	**1.1988**
		(0.0384)	(0.0262)	(0.2408)	(0.1359)
	U	0.6849	0.6849	1.4496	1.2351
		(0.0390)	(0.0266)	(0.2456)	(0.1360)
	E	**0.6917**	**0.6967**	1.5123	1.3111
		(0.0406)	(0.0231)	(0.2404)	(0.1428)
	C	0.6887	0.6904	1.4804	1.2686
		(0.0393)	(0.0246)	(0.2498)	(0.1429)
Student	B	0.8414	0.8947	**0.8688**	0.8153
		(0.0246)	(0.0170)	(0.2245)	(0.1920)
	U	0.8463	0.8988	0.8412	0.7793
		(0.0228)	(0.0194)	(0.2331)	(0.1916)
	E	**0.8492**	**0.9039**	0.8568	**0.8202**
		(0.0231)	(0.0177)	(0.2267)	(0.1878)
	C	0.8407	0.8969	0.8058	0.7352
		(0.0234)	(0.0188)	(0.2538)	(0.1954)

Table 2: Fairness evaluation of in-context learning for GPT-4o across datasets. Metrics include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). F: Foundation, U: Unawareness, E: Equal Samples, C: Counterfactual. Best performances are highlighted in bold, with standard deviations presented below the mean.

5.3 Fine-tune and RAG

We present the results of fine-tuning different models in Appendix C, Table 11. Due to GPT-4 lack of support for fine-tuning (OpenAI, 2024a), we focused on fine-tuning the GPT-3.5, Llama3, and Gemma-2 models. Our results indicate that fine-tuning these models leads to improvements in both accuracy and F1 scores. However, the improvement in fairness metrics remains limited. In addition, we utilised RAG with a dataset comprising articles from Related Work (i.e., there are 29 articles). Specifically, we employed OpenAI's embedding methods (OpenAI, 2024b) for generating high-quality vector representations of the texts. The results, detailed in Appendix C, Table 12, show that while RAG led to slight improvements in accuracy and F1 scores, the enhancement in fairness metrics was moderate compared to the in-context learning strategy employing the "Equal Samples Across Demographics" approach. This indicates that while RAG can enhance performance, its impact on fairness is less pronounced than targeted in-context learning methods.

5.4 Enhancing Fairness with Prompting Strategies

The evaluation of fairness metrics for four prompting strategies, including Fair, Generalised, Descriptive, and Descriptive combined with in-context learning through Equal Samples Across Demographics (ESAD), applied to four datasets using GPT-4o is presented in Table 3. Detailed results for other LLMs are provided in Appendix C in Tables 13, 14, 15, and 16.

From the results, we observe that the Descriptive + ESAD (i.e. D + E in Table 3) strategy consistently improves fairness across datasets, demonstrating a clear advantage over standalone Descriptive prompting. Specifically, for the Adult and COMPAS datasets, employing D + E strategy leads to significant improvements in metrics such as SPR and EOR, with values closer to the fairness ideal of 1. These results substantiate the hypothesis that combining descriptive fairness instructions with in-context learning can guide the LLM to make more balanced predictions. While the impact of ESAD is more pronounced in Adult and COMPAS, the Diabetes dataset shows minimal variation in metric values across strategies, likely due to inherent dataset characteristics that make it less sensitive to fairness interventions. In the Student dataset, metrics like SPR and EOR exhibit the least discrepancy, with stable values across strategies.

These findings underline the effectiveness of descriptive prompting enhanced with in-context learning in addressing fairness concerns in LLMs. The D + E strategy not only brings fairness metrics like SPR and EOR closer to the ideal value of 1 but also achieves accuracy improvements ranging from 3.27% to 15.05% across datasets. Importantly, this approach highlights the adaptability of LLMs when fairness definitions are explicitly integrated into input prompts.

Overall, while these results are promising, it is important to acknowledge that bias in LLMs cannot be fully mitigated for tabular data classification task. However, through thoughtful in-context learning and designed prompting strategies, bias can be significantly reduced, and fairness can be improved. The findings from these experiments highlight the potential of combining descriptive prompts with in-context learning as a powerful tool in the ongoing effort to create more equitable AI systems. This approach not only advances fairness in model predictions but also contributes to the broader goal of mitigating bias in AI, thereby fostering more responsible and ethical AI development.

Dataset	Type	Acc	F1	SPR	EOR
Adult	F	0.8209	0.6681	1.5836	1.2442
		(0.0087)	(0.0076)	(0.0367)	(0.0200)
	G	0.8296	0.6634	1.5778	1.1719
		(0.0156)	(0.0072)	(0.0337)	(0.0329)
	D	0.8202	0.6729	1.5820	1.1816
		(0.0082)	(0.0122)	(0.0182)	(0.0467)
	D + E	**0.9124**	**0.8206**	**1.3792**	**1.1444**
		(0.0103)	(0.0084)	(0.0362)	(0.0408)
COMPAS	F	0.6980	0.6858	1.2810	1.0606
		(0.0070)	(0.0091)	(0.0767)	(0.0872)
	G	0.7091	0.7227	1.2093	1.0603
		(0.0073)	(0.0082)	(0.1012)	(0.0797)
	D	0.7052	0.7245	1.1828	1.0078
		(0.0028)	(0.0060)	(0.0645)	(0.0628)
	D + E	**0.7108**	**0.7336**	**1.1773**	**0.9427**
		(0.0076)	(0.0032)	(0.0505)	(0.0939)
Diabetes	F	0.6074	0.6339	1.2404	1.0804
		(0.0051)	(0.0065)	(0.0896)	(0.0527)
	G	0.6146	0.6374	1.1964	1.1189
		(0.0054)	(0.0055)	(0.0726)	(0.0730)
	D	0.6169	0.6490	1.2018	1.0965
		(0.0039)	(0.0063)	(0.0971)	(0.0607)
	D + E	**0.6239**	**0.6621**	**1.1479**	**1.0995**
		(0.0082)	(0.0095)	(0.1208)	(0.0663)
Student	F	0.8078	0.8841	0.8714	1.2834
		(0.0063)	(0.0042)	(0.0044)	(0.0528)
	G	0.8113	0.8882	0.8750	1.3292
		(0.0063)	(0.0052)	(0.0049)	(0.0595)
	D	0.8163	0.8932	0.8800	1.3824
		(0.0079)	(0.0065)	(0.0066)	(0.0763)
	D + E	**0.8191**	**0.8982**	**0.8847**	**1.4222**
		(0.0104)	(0.0085)	(0.0068)	(0.0954)

Table 3: Fairness evaluation for GPT-4o across datasets and different prompt strategies. Metrics include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). F: Fair Prompting, G: Generalised Prompting, D: Descriptive Prompting, D + E: Descriptive Prompting + Equal Samples Across Demographics. Best performance is in bold, with standard deviations shown below the mean.

6 Conclusion

This study evaluates several methods to improve fairness in LLM predictions for tabular datasets. **In-context learning** offers a simple approach to incorporate fairness without retraining, though its effectiveness depends on selecting suitable examples, making it less reliable for complex fairness challenges. **Fine-tuning** allows direct adjustment of model parameters to enhance fairness and performance but requires substantial computational resources and careful hyperparameter tuning. **RAG** integrates external knowledge to guide predictions, providing flexibility in addressing fairness issues. However, its success hinges on the quality of the retrieval process and additional infrastructure needs. In this study, we propose **instruction-based prompting**, which proves to be most effective when combined with in-context learning strategies such as ESAD. This approach improves fairness metrics such as SPR and EOR while achieving accuracy gains across datasets. While these methods show promise, fully addressing biases in LLMs remains a challenge. Future research should explore centralized training mechanisms or counting procedures to ensure that improved individual predictions translate into group-level fairness outcomes in alignment with defined metrics.

Limitations

This study highlights effective ways to reduce biases in LLM predictions for tabular data, but several challenges remain. Fully eliminating bias is difficult because LLMs are pre-trained on large datasets that often contain underlying inequalities. While the proposed prompting and in-context learning strategies improve fairness measures like SPR and EOR, they may not work equally well for all datasets, especially those with complex fairness issues. Additionally, the assumption that better individual predictions automatically lead to fairness at a group level may not always hold true, as there is no centralised process to ensure fairness across groups. These methods also depend on the quality of the examples and fairness definitions provided, which can limit their effectiveness in real-world applications. Future research should address these limitations by identifying specific cases where these methods fall short and exploring additional solutions.

Ethical Considerations

This research uses demographic attributes like gender and race solely to assess and improve fairness in LLMs, aiming to identify and mitigate biases that could lead to discriminatory outcomes. These attributes are used only to evaluate fairness metrics such as Statistical Parity and Equal Opportunity, ensuring responsible handling of sensitive information. Publicly available datasets were used, adhering to ethical guidelines, with the goal of promoting equitable AI systems.

References

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

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

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

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

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.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. *arXiv preprint arXiv:2403.14409*.

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

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*.

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

Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. *Psychology*.

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

Suzanne Fergus, Michelle Botha, and Mehrnoosh Ostovar. 2023. Evaluating academic answers generated using chatgpt. *Journal of Chemical Education*, 100(4):1672–1675.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

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

Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE.

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

Google. 2024. Gemma open models: Advancing ai accessibility and performance. https://ai.google.dev/gemma. Accessed: December 10, 2024.

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.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*.

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

Ron Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

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

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

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2024. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3603–3620.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2024. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36.

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

Meta. 2024. Meta llama: Next-generation language models by meta ai. https://llama.meta.com. Accessed: December 10, 2024.

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. *arXiv preprint arXiv:2010.00133*.

OpenAI. 2024a. Fine-tuning guide: Optimizing models for custom applications. https://platform.openai.com/docs/guides/fine-tuning. Accessed: December 10, 2024.

OpenAI. 2024b. Understanding embeddings: Enhancing contextual understanding in models. https://platform.openai.com/docs/guides/embeddings/what-are-embeddings. Accessed: December 10, 2024.

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

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670.

Anirudh S Sundar and Larry Heck. 2023. ctbls: Augmenting large language models with conversational tables. *arXiv preprint arXiv:2303.12024*.

Ruiyu Wang, Zifeng Wang, and Jimeng Sun. 2023. Unipredict: Large language models are universal tabular predictors. *arXiv preprint arXiv:2310.03266*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. 2024. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

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

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175.

Appendix

A Prompt Templates for In-context Learning

Unawareness Example

Given the following individual's profile details, predict if their salary is over $50,000.

ONLY respond with 'yes' or 'no'.

Example: Age: 52, Education: Doctorate, Occupation: Prof-specialty, Capital Gain: 15000, Capital Loss: 0, Hours per Week: 50. The answer is yes.

Counterfactual Example

Given the following individual's profile details, predict if their salary is over $50,000.

ONLY respond with 'yes' or 'no'.

Example: Age: 40, Education: Masters, Occupation: Tech-support, Hours per Week: 45, Sex: Male. The answer is no.

Equal Samples Across Demographics Example

Given the following individuals' profile details, predict if their salary is over $50,000. ONLY respond with 'yes' or 'no'.

Example 1: Age: 29 Education: Bachelors Occupation: Admin-clerical Hours per Week: 40 Sex: Male. The answer is no.

Example 2: Age: 29 Education: Bachelors Occupation: Admin-clerical Hours per Week: 40 Sex: Female. The answer is no.

B Fine-Tuning

We used 1,000 samples as the test set when 20% of the dataset exceeded 1,000 samples; otherwise, 20% was used. For fine-tuning, the training set was further split 80:20 for training and evaluation. An example data format is shown below:

ID	Question	Answer
1	Age: 31 Workclass:Federal-gov ...	0
2	Age: 31 Workclass:Private ...	1
3	Age: 45 Workclass:Private ...	0
4	Age: 31 Workclass:Private ...	0

Table 4: The format of data for fine-tuning

Category	**Details**
General	**Warmup Steps:** 0 **Batch Size per Device:** 2 **Gradient Accumulation Steps:** 1 **Gradient Checkpointing:** Enabled **Maximum Steps:** 500 **Learning Rate:** 1e-5 (suitable for fine-tuning) **BF16:** Enabled **Optimiser:** paged_adamw_8bit
Logging	**Logging Steps:** 25 (reporting loss interval) **Logging Directory:** ./logs **Save Strategy:** Save checkpoints at logging steps **Save Steps:** 100 **Evaluation Strategy:** Evaluate at logging steps **Evaluation Steps:** 100 **Do Evaluation:** Perform at the end of training **Report to:** tensorboard (optional tracking) **Run Name:** Combines run name and current date/time
Data Collator	**Tokeniser:** No masked language modelling (mlm=False)
Configuration	**Use Cache:** Disabled to silence warnings

Table 5: Training and evaluation configurations for fine-tuning. The table outlines general settings, logging and saving configurations, data collator settings, and model configurations.

C Fairness Evaluation

Data	Metric	GPT-4o	GPT-3.5	Llama-3	Gemma-2
Adult	Acc	0.7931 (0.0131)	0.7180 (0.0200)	0.6600 (0.0196)	0.7063 (0.0181)
	F1	0.6275 (0.0216)	0.5801 (0.0238)	0.5478 (0.0174)	0.5767 (0.0067)
	SPR	1.7617 (0.1620)	2.3131 (0.4047)	1.2330 (0.0542)	1.8147 (0.1410)
	EOR	1.3900 (0.1678)	1.3458 (0.2540)	1.1106 (0.0170)	1.1382 (0.0844)
COMPAS	Acc	0.6623 (0.0072)	0.6101 (0.0050)	0.5978 (0.0112)	0.6042 (0.0198)
	F1	0.5601 (0.0088)	0.6470 (0.0096)	0.6614 (0.0130)	0.5328 (0.0249)
	SPR	2.8726 (0.3612)	2.6732 (0.1517)	1.5263 (0.0361)	1.9018 (0.0977)
	EOR	2.4246 (0.2581)	1.9533 (0.1380)	1.1503 (0.0178)	1.5084 (0.1211)
Diabetes	Acc	0.5707 (0.0220)	0.5046 (0.0045)	0.4866 (0.0126)	0.4892 (0.0087)
	F1	0.5896 (0.0239)	0.0118 (0.0041)	0.6002 (0.0070)	0.6350 (0.0187)
	SPR	0.9925 (0.0153)	0.9401 (0.0183)	0.9235 (0.0227)	0.9469 (0.0274)
	EOR	0.9678 (0.0129)	0.9565 (0.3385)	0.9667 (0.0279)	0.9601 (0.0155)
Student	Acc	0.8358 (0.0259)	0.7908 (0.0220)	0.8600 (0.0436)	0.8588 (0.0376)
	F1	0.8927 (0.0186)	0.8577 (0.0150)	0.9235 (0.0261)	0.9206 (0.0218)
	SPR	0.8972 (0.2329)	0.9590 (0.1506)	0.9836 (0.0247)	0.9996 (0.0575)
	EOR	0.8506 (0.1834)	0.9965 (0.0831)	1.0060 (0.0575)	1.0240 (0.0576)

Table 6: Fairness evaluation across datasets. Metrics include accuracy (Acc), F1 score, statistical parity ratio (SPR), and equality of opportunity ratio (EOR).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.7697 (0.0079)	0.5874 (0.0075)	1.9836 (0.1263)	1.3037 (0.0837)
	U	0.7879 (0.0053)	0.6032 (0.0069)	0.4290 (0.0727)	0.6877 (0.0640)
	E	0.7893 (0.0054)	0.6265 (0.0057)	2.0226 (0.1218)	1.2567 (0.0475)
	C	0.7567 (0.0042)	0.6283 (0.0063)	2.5463 (0.0669)	1.2766 (0.0498)
Llama-3	F	0.6170 (0.0065)	0.5623 (0.0084)	1.8415 (0.0744)	1.1953 (0.0753)
	U	0.6449 (0.0094)	0.5876 (0.0108)	1.9572 (0.0682)	1.2930 (0.0853)
	E	0.6513 (0.0119)	0.5452 (0.0108)	1.2890 (0.0678)	1.2428 (0.0911)
	C	0.6466 (0.0205)	0.5487 (0.0080)	1.4667 (0.1338)	1.2633 (0.1132)
Gemma-2	F	0.7171 (0.0151)	0.5874 (0.0082)	1.4658 (0.0958)	1.2779 (0.0945)
	U	0.6606 (0.0118)	0.5750 (0.0057)	1.5492 (0.0705)	1.4366 (0.0699)
	E	0.6569 (0.0135)	0.5687 (0.0138)	1.3985 (0.0366)	1.4435 (0.0742)
	C	0.6712 (0.0097)	0.5777 (0.0107)	1.7938 (0.1073)	1.2818 (0.0631)

Table 7: Fairness evaluation for in-context learning on the Adult dataset. Metrics evaluated are accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are displayed below the mean. Types: F (Foundation), U (Unawareness), E (Equal Samples Across Demographics), C (Counterfactual).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.6399	0.6239	1.9073	1.7210
		(0.0055)	(0.0096)	(0.0854)	(0.0887)
	U	0.6316	0.6364	1.8991	1.6056
		(0.0072)	(0.0091)	(0.0715)	(0.0506)
	E	0.6511	0.5400	1.6130	1.3242
		(0.0069)	(0.0072)	(0.0917)	(0.1252)
	C	0.6338	0.5386	1.6240	1.3122
		(0.0074)	(0.0065)	(0.0581)	(0.0691)
Llama-3	F	0.6253	0.6854	1.5334	1.6107
		(0.0120)	(0.0073)	(0.0656)	(0.0816)
	U	0.6266	0.6869	1.6377	1.5910
		(0.0082)	(0.0070)	(0.1313)	(0.0976)
	E	0.6322	0.6865	1.5586	1.6391
		(0.0075)	(0.0054)	(0.0713)	(0.1215)
	C	0.6317	0.6938	1.6348	1.5942
		(0.0080)	(0.0065)	(0.1236)	(0.0989)
Gemma-2	F	0.6097	0.5543	1.8808	1.5162
		(0.0039)	(0.0096)	(0.0960)	(0.1026)
	U	0.6157	0.5535	1.8876	1.5591
		(0.0044)	(0.0079)	(0.0958)	(0.1319)
	E	0.6240	0.5510	1.8152	1.4852
		(0.0098)	(0.0053)	(0.0443)	(0.1055)
	C	0.6382	0.6098	1.9162	1.6313
		(0.0124)	(0.0347)	(0.1106)	(0.1412)

Table 8: Fairness evaluation for in-context learning on the COMPAS dataset. Metrics evaluated include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are shown below the mean. Types: F (Foundation), U (Unawareness), E (Equal Samples Across Demographics), and C (Counterfactual).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.5114	0.0160	0.9818	0.9758
		(0.0063)	(0.0044)	(0.0313)	(0.3371)
	U	0.5191	0.0221	1.0253	1.0064
		(0.0096)	(0.0055)	(0.0245)	(0.3208)
	E	0.5235	0.0264	1.0513	1.0508
		(0.0110)	(0.0058)	(0.0263)	(0.3055)
	C	0.5290	0.0300	1.0846	1.1009
		(0.0108)	(0.0082)	(0.0417)	(0.3111)
Llama-3	F	0.5097	0.0163	0.9743	1.0001
		(0.0064)	(0.0073)	(0.0208)	(0.3513)
	U	0.5142	0.0205	1.0146	1.0216
		(0.0075)	(0.0062)	(0.0390)	(0.3696)
	E	0.5219	0.0293	1.0872	1.0789
		(0.0093)	(0.0077)	(0.0668)	(0.3574)
	C	0.5187	0.0269	1.0485	1.0464
		(0.0078)	(0.0073)	(0.0494)	(0.3611)
Gemma-2	F	0.5097	0.0163	0.9743	1.0001
		(0.0064)	(0.0073)	(0.0208)	(0.3513)
	U	0.5127	0.0204	1.0186	1.0421
		(0.0065)	(0.0082)	(0.0212)	(0.3432)
	E	0.5221	0.0298	1.0935	1.1074
		(0.0070)	(0.0099)	(0.0188)	(0.3343)
	C	0.5162	0.0254	1.0661	1.0741
		(0.0076)	(0.0091)	(0.0207)	(0.3298)

Table 9: Fairness evaluation for in-context learning on the Diabetes dataset. Metrics evaluated include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are shown below the mean. Types: F (Foundation), U (Unawareness), E (Equal Samples Across Demographics), and C (Counterfactual).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.7943	0.8625	0.9413	0.9729
		(0.0213)	(0.0148)	(0.1427)	(0.0873)
	U	0.8006	0.8662	0.9115	0.9458
		(0.0219)	(0.0156)	(0.1432)	(0.0807)
	E	0.8050	0.8692	0.8803	0.9013
		(0.0240)	(0.0156)	(0.1535)	(0.0760)
	C	0.8050	0.8692	0.8803	0.9013
		(0.0240)	(0.0156)	(0.1535)	(0.0760)
Llama-3	F	0.8631	0.9288	0.9698	0.9637
		(0.0438)	(0.0255)	(0.0234)	(0.0728)
	U	0.8661	0.9317	0.9412	0.9321
		(0.0415)	(0.0243)	(0.0309)	(0.0673)
	E	0.8727	0.9374	0.9173	0.9106
		(0.0414)	(0.0240)	(0.0436)	(0.0573)
	C	0.8652	0.9317	0.8576	0.8596
		(0.0416)	(0.0238)	(0.0630)	(0.0728)
Gemma-2	F	0.8621	0.9254	0.9678	0.9812
		(0.0386)	(0.0226)	(0.0649)	(0.0454)
	U	0.8667	0.9313	0.8689	0.8703
		(0.0395)	(0.0210)	(0.0796)	(0.0702)
	E	0.8761	0.9391	0.9251	0.9378
		(0.0394)	(0.0245)	(0.0690)	(0.0566)
	C	0.8703	0.9364	0.9023	0.9147
		(0.0391)	(0.0228)	(0.0785)	(0.0609)

Table 10: Fairness evaluation for in-context learning on the Student dataset. Metrics evaluated include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are shown below the mean. Types: F (Foundation), U (Unawareness), E (Equal Samples Across Demographics), and C (Counterfactual).

Dataset	Model	Acc	F1	SPR	EOR
Adult	GPT-3.5	0.7180	0.5801	2.3131	1.3458
		(0.0200)	(0.0238)	(0.4047)	(0.2540)
	Llama-3	0.6600	0.5478	1.2330	1.1106
		(0.0196)	(0.0174)	(0.0542)	(0.0170)
	Gemma-2	0.7063	0.5767	1.8147	1.1382
		(0.0181)	(0.0067)	(0.1410)	(0.0844)
COMPAS	GPT-3.5	0.6116	0.6458	2.7982	2.1265
		(0.0029)	(0.0021)	(0.0195)	(0.0084)
	Llama-3	0.6115	0.6449	2.8073	2.1417
		(0.0021)	(0.0022)	(0.0344)	(0.0231)
	Gemma-2	0.6103	0.6479	2.7963	2.1365
		(0.0032)	(0.0007)	(0.0182)	(0.0054)
Diabetes	GPT-3.5	0.5086	0.0167	0.9953	1.5987
		(0.0032)	(0.0023)	(0.0189)	(0.0313)
	Llama-3	0.5025	0.6458	1.0050	0.9901
		(0.0044)	(0.0025)	(0.0253)	(0.0310)
	Gemma-2	0.5013	0.6497	1.0042	0.9860
		(0.0040)	(0.0021)	(0.0274)	(0.0182)
Student	GPT-3.5	0.7906	0.8759	0.8548	0.9194
		(0.0058)	(0.0058)	(0.0371)	(0.0564)
	Llama-3	0.8200	0.9114	1.0360	1.4869
		(0.0066)	(0.0082)	(0.0233)	(0.0658)
	Gemma-2	0.8825	0.9299	1.1000	1.0924
		(0.0045)	(0.0067)	(0.0450)	(0.0280)

Table 11: Fairness evaluation for fine-tuning across datasets. Metrics include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). The standard deviation is shown below the mean value.

Dataset	Model	Acc	F1	SPR	EOR
Adult	GPT-4o	0.8252	0.6682	1.5004	1.1650
		(0.0115)	(0.0130)	(0.0833)	(0.0886)
	GPT-3.5	0.7188	0.6293	1.8903	1.1745
		(0.0048)	(0.0059)	(0.0841)	(0.0356)
	Llama-3	0.7140	0.5950	1.1435	1.0003
		(0.0092)	(0.0104)	(0.0558)	(0.0762)
	Gemma-2	0.6945	0.5915	1.0749	1.0506
		(0.0046)	(0.0062)	(0.0727)	(0.0237)
COMPAS	GPT-4o	0.6139	0.6458	2.8172	2.1308
		(0.0019)	(0.0026)	(0.0275)	(0.0127)
	GPT-3.5	0.8252	0.6682	1.5004	1.1650
		(0.0115)	(0.0130)	(0.0833)	(0.0886)
	Llama-3	0.6135	0.6443	2.8225	2.1465
		(0.0021)	(0.0017)	(0.0217)	(0.0285)
	Gemma-2	0.6110	0.6475	2.7975	2.1155
		(0.0026)	(0.0014)	(0.0178)	(0.0161)
Diabetes	GPT-4o	0.5069	0.0183	0.9885	1.6049
		(0.0052)	(0.0014)	(0.0203)	(0.0368)
	GPT-3.5	0.6059	0.6204	1.0307	1.0037
		(0.0028)	(0.0037)	(0.0167)	(0.0255)
	Llama-3	0.5008	0.6469	0.9983	1.0142
		(0.0030)	(0.0030)	(0.0304)	(0.0250)
	Gemma-2	0.4982	0.6463	1.0050	1.0106
		(0.0025)	(0.0034)	(0.0210)	(0.0181)
Student	GPT-4o	0.8358	0.8927	0.8972	0.8506
		(0.0259)	(0.0186)	(0.2329)	(0.1834)
	GPT-3.5	0.7908	0.8577	0.9590	0.9965
		(0.0220)	(0.0150)	(0.1506)	(0.0831)
	Llama-3	0.8600	0.9235	0.9836	1.0060
		(0.0436)	(0.0261)	(0.0247)	(0.0575)
	Gemma-2	0.8588	0.9206	0.9996	1.0240
		(0.0376)	(0.0218)	(0.0575)	(0.0576)

Table 12: Fairness evaluation for RAG across datasets. Metrics include accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are shown below the mean values.

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.8346	0.5666	2.1110	1.5923
		(0.0121)	(0.0110)	(0.0904)	(0.0824)
	G	0.7966	0.6331	2.0430	2.4656
		(0.0042)	(0.0095)	(0.0739)	(0.0622)
	D	0.8256	0.5819	2.0433	1.5528
		(0.0044)	(0.0499)	(0.0558)	(0.0529)
	D + E	0.8711	0.6852	1.9605	1.4166
		(0.0042)	(0.0058)	(0.0359)	(0.0415)
Llama-3	F	0.7032	0.5876	1.4483	1.2389
		(0.0079)	(0.0030)	(0.0918)	(0.0912)
	G	0.7072	0.5951	1.4260	1.2288
		(0.0095)	(0.0048)	(0.0816)	(0.0974)
	D	0.7014	0.5989	1.4140	1.2034
		(0.0082)	(0.0049)	(0.0860)	(0.0867)
	D + E	0.7175	0.6064	1.4054	1.2597
		(0.0125)	(0.0069)	(0.0887)	(0.0993)
Gemma-2	F	0.6911	0.5762	1.4483	1.2389
		(0.0073)	(0.0119)	(0.0534)	(0.0648)
	G	0.6951	0.5790	1.4483	1.2389
		(0.0094)	(0.0101)	(0.0511)	(0.0515)
	D	0.6987	0.5835	1.4483	1.2389
		(0.0098)	(0.0090)	(0.0536)	(0.0570)
	D + E	0.7025	0.5898	1.4483	1.2389
		(0.0124)	(0.0094)	(0.0635)	(0.0650)

Table 13: Fairness evaluation for different prompt strategies on the Adult dataset. Metrics evaluated are accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are displayed below the mean. Types: F (Fair Prompting), G (Generalised Prompting), D (Descriptive Prompting), D + E (Descriptive Prompting + Equal Samples Across Demographics).

Model	Type	Acc	AUROC	SPR	EOR
GPT-3.5	F	0.6062	0.6563	2.7335	1.9775
		(0.0038)	(0.0051)	(0.0522)	(0.0641)
	G	0.6108	0.6615	2.6718	1.9219
		(0.0044)	(0.0049)	(0.0760)	(0.0604)
	D	0.6158	0.6659	2.5993	1.8840
		(0.0069)	(0.0050)	(0.0781)	(0.0465)
	D + E	0.6203	0.6715	2.5668	1.8204
		(0.0074)	(0.0069)	(0.0749)	(0.0503)
Llama-3	F	0.6216	0.6862	1.5505	1.2590
		(0.0070)	(0.0064)	(0.0522)	(0.0677)
	G	0.6111	0.6655	1.5076	1.2346
		(0.0057)	(0.0081)	(0.0747)	(0.0619)
	D	0.6111	0.6655	1.4451	1.1794
		(0.0057)	(0.0081)	(0.0950)	(0.0599)
	D + E	0.6111	0.6655	1.4153	1.1258
		(0.0057)	(0.0081)	(0.1031)	(0.0366)
Gemma-2	F	0.6270	0.6629	1.4483	1.2389
		(0.0100)	(0.0044)	(0.0667)	(0.0979)
	G	0.6322	0.6674	1.4413	1.2311
		(0.0110)	(0.0051)	(0.0678)	(0.0968)
	D	0.6368	0.6718	1.4023	1.2189
		(0.0099)	(0.0072)	(0.0711)	(0.0972)
	D + E	0.6411	0.6780	1.4001	1.2081
		(0.0116)	(0.0085)	(0.0703)	(0.0992)

Table 14: Fairness evaluation for different prompt strategies on the COMPAS dataset. Metrics evaluated are accuracy (Acc), AUROC, statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are displayed below the mean. Types: F (Fair Prompting), G (Generalised Prompting), D (Descriptive Prompting), D + E (Descriptive Prompting + Equal Samples Across Demographics).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.5215	0.0566	1.1513	1.1185
		(0.0089)	(0.0079)	(0.0809)	(0.0720)
	G	0.5261	0.0546	0.9882	1.1068
		(0.0055)	(0.0039)	(0.0449)	(0.0739)
	D	0.5246	0.0647	0.9514	1.0485
		(0.0063)	(0.0062)	(0.0760)	(0.0730)
	D + E	0.5274	0.0705	0.8196	0.9887
		(0.0080)	(0.0084)	(0.0431)	(0.0614)
Llama-3	F	0.5446	0.0777	1.2166	1.1180
		(0.0070)	(0.0089)	(0.0644)	(0.0251)
	G	0.5557	0.0858	1.2292	1.1484
		(0.0069)	(0.0095)	(0.0493)	(0.0430)
	D	0.5588	0.0904	1.2661	1.1627
		(0.0048)	(0.0070)	(0.0765)	(0.0676)
	D + E	0.5677	0.0960	1.2925	1.2083
		(0.0082)	(0.0067)	(0.0944)	(0.0845)
Gemma-2	F	0.5184	0.0437	1.4483	1.2389
		(0.0048)	(0.0053)	(0.0209)	(0.0864)
	G	0.5198	0.0529	1.4413	1.2312
		(0.0060)	(0.0065)	(0.0883)	(0.0822)
	D	0.5397	0.0750	1.4343	1.2307
		(0.0034)	(0.0031)	(0.0516)	(0.0607)
	D + E	0.5446	0.0789	1.4321	1.2102
		(0.0055)	(0.0037)	(0.0875)	(0.0798)

Table 15: Fairness evaluation for different prompt strategies on the Diabetes dataset. Metrics evaluated are accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are displayed below the mean. Types: F (Fair Prompting), G (Generalised Prompting), D (Descriptive Prompting), D + E (Descriptive Prompting + Equal Samples Across Demographics).

Model	Type	Acc	F1	SPR	EOR
GPT-3.5	F	0.7921	0.9644	0.8672	1.2946
		(0.0052)	(0.0061)	(0.0039)	(0.0494)
	G	0.7956	0.9707	0.8697	1.3369
		(0.0061)	(0.0054)	(0.0026)	(0.0694)
	D	0.8018	0.9752	0.8747	1.3745
		(0.0061)	(0.0082)	(0.0023)	(0.0612)
	D + E	0.8047	0.9773	0.8793	1.4144
		(0.0058)	(0.0090)	(0.0030)	(0.0799)
Llama-3	F	0.8082	0.5355	0.9071	0.8459
		(0.0038)	(0.0034)	(0.0046)	(0.0372)
	G	0.8130	0.5386	0.9112	0.8831
		(0.0045)	(0.0032)	(0.0030)	(0.0598)
	D	0.8144	0.8890	0.8730	1.3084
		(0.0071)	(0.0060)	(0.0050)	(0.0410)
	D + E	0.8144	0.8890	0.8730	1.3084
		(0.0071)	(0.0060)	(0.0050)	(0.0410)
Gemma-2	F	0.8385	0.6295	0.9063	1.2301
		(0.0056)	(0.0048)	(0.0048)	(0.0300)
	G	0.8432	0.6347	0.9090	1.2772
		(0.0057)	(0.0049)	(0.0065)	(0.0228)
	D	0.8144	0.8890	0.8730	1.3084
		(0.0071)	(0.0060)	(0.0050)	(0.0410)
	D + E	0.8174	0.8899	0.8799	1.3088
		(0.0078)	(0.0081)	(0.0059)	(0.0423)

Table 16: Fairness evaluation for different prompt strategies on the Student dataset. Metrics evaluated are accuracy (Acc), F1 score (F1), statistical parity ratio (SPR), and equality of opportunity ratio (EOR). Standard deviations are displayed below the mean. Types: F (Fair Prompting), G (Generalised Prompting), D (Descriptive Prompting), D + E (Descriptive Prompting + Equal Samples Across Demographics).