Improving LLM Group Fairness on Tabular Data via In-Context Learning

Valeriia Cherepanova[1*], Chia-Jung Lee[1], Nil-Jana Akpinar[1], Riccardo Fogliato[1], Martin Bertran Lopez[1], Michael Kearns[1,2], James Zou[1,3]

[1]Amazon AWS AI
[2]University of Pennsylvania
[3]Stanford University
cherepv@amazon.com, cjlee@amazon.com, nakpinar@amazon.com, fogliato@amazon.com, maberlop@amazon.com, kearmic@amazon.com, jamesz@stanford.edu

Abstract

Large language models (LLMs) have been shown to be effective on tabular prediction tasks in the low-data regime, leveraging their internal knowledge and ability to learn from instructions and examples. However, LLMs can fail to generate predictions that satisfy group fairness, that is, produce equitable outcomes across groups. Critically, conventional debiasing approaches for natural language tasks do not directly translate to mitigating group unfairness in tabular settings. In this work, we systematically investigate four empirical approaches to improve group fairness of LLM predictions on tabular datasets, including fair prompt optimization, soft prompt tuning, strategic selection of few-shot examples, and self-refining predictions via chain-of-thought reasoning. Through experiments on four tabular datasets using both open-source and proprietary LLMs, we show the effectiveness of these methods in enhancing demographic parity while maintaining high overall performance. Our analysis provides actionable insights for practitioners in selecting the most suitable approach based on their specific requirements and constraints.

1 Introduction

In recent years, the scope of large language models (LLMs) has broadened significantly beyond traditional natural language processing tasks, with recent research demonstrating their effectiveness in tackling challenges on tabular data, including predictive tasks (Hegselmann et al. 2023; Yin et al. 2020). Typically, structured data is converted into textual format and provided to the language model along with a concise task description and key features. Notably, it has been shown that language models are particularly beneficial in scenarios with limited training data, as they can utilize internal knowledge about world from pre-training combined with textual instructions and few-shot examples to make predictions (Slack and Singh 2023).

Although considerable research has been devoted to exploring and addressing issues of stereotypical bias and fairness in language models applied to natural language tasks, tabular datasets present distinct challenges, particularly in group fairness. It is important to differentiate group fairness in the context of tabular data from conventional notions of fairness in NLP tasks: group fairness in tabular problems hinges on class labels and the representation of various demographic groups within these labels, while stereotypical fairness in NLP has primarily focused on bias in model representations. Notably, achieving fairness in the typical NLP sense does not automatically ensure group-fair predictions in tabular tasks due to potential disparities in class distributions.

Recent studies have started exploring how language models handle group fairness when applied to tabular data, revealing noticeable fairness discrepancies among different demographic groups. Liu et al. (2024) and Li, Zhang, and Zhang (2024) evaluate a few baseline methods for improving group fairness in tabular tasks, including resampled fine-tuning, and few-shot learning with label flipping and find these methods to have limited effectiveness. A recent survey paper (Fang et al. 2024) recognizes the challenge of mitigating inherent biases in large language models through conventional fine-tuning and few-shot learning and highlights the need for more effective strategies to address group unfairness in tabular tasks.

In this work we examine four approaches for empirically improving demographic parity of LLMs when applied to making predictions on tabular data. These approaches include in-context methods such as prompt optimization, soft prompt tuning, few-shot in-context learning, and self-refining predictions to promote fairness. We empirically evaluate these methods using both open-source and proprietary models across four tabular datasets, demonstrating their effectiveness. Based on our analysis, we provide actionable recommendations to practitioners on the most suitable method for different scenarios, and discuss how these approaches may be adapted to other notions of fairness.

2 Related Work

Large Language Models on Tabular Data

A growing body of work has applied deep learning algorithms to tabular data (Yin et al. 2020; Herzig et al. 2020; Huang et al. 2020; Levin et al. 2022; Zhu et al. 2023). Relevant to our setting, some of these studies have employed LLMs to analyze tabular data that is serialized into formatted text. They show that descriptive feature names, well-defined

*Correspondence: cherepv@amazon.com

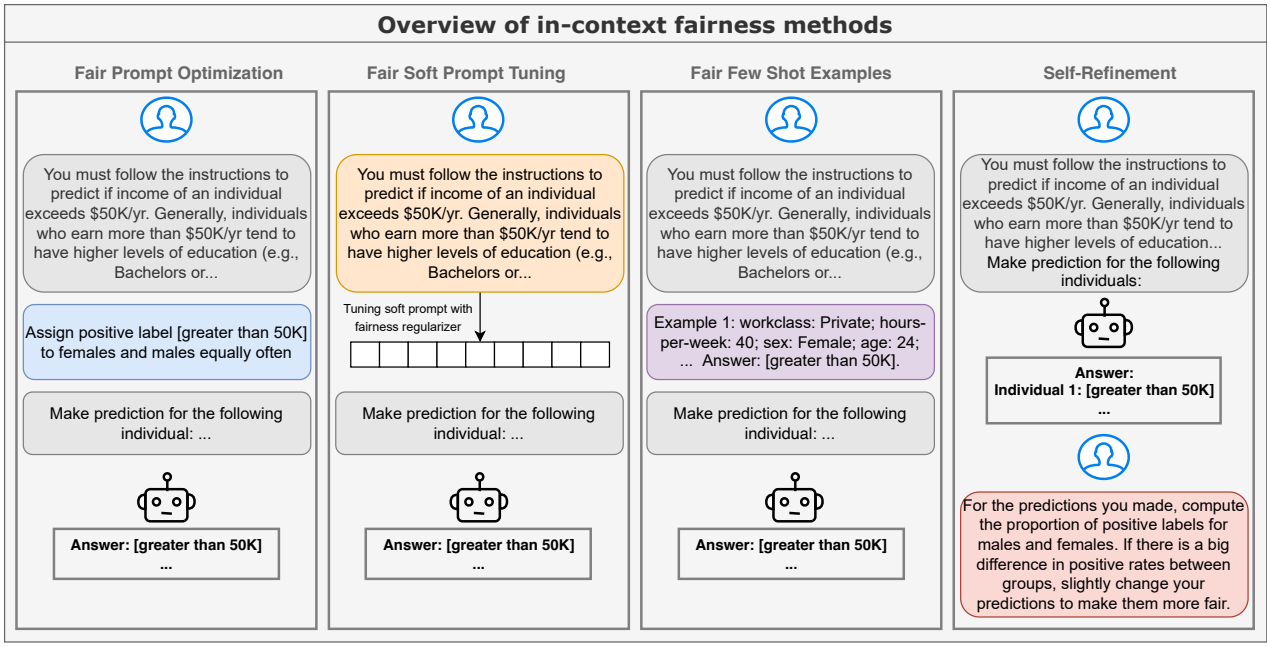

Figure 1: **Overview of fairness methods explored in this work.** We focus on in-context learning approaches, including fair prompt optimization and soft prompt tuning, fair few-shot examples, and self-refinement. For each method, we highlight the specific prompt components optimized in these approaches using different colors, while components of the prompts highlighted in gray do not change across strategies.

instructions, in-context examples, and chain-of-thought reasoning enhances LLM performance (Zhao et al. 2023; Marvin et al. 2023; Chen 2022). Some specifically focus on classification tasks (Hegselmann et al. 2023; Wang et al. 2024; Liu, Yang, and Wu 2022; Fang et al. 2024; Jaitly et al. 2023), which is also the focus of our work. The prior knowledge of LLMs allows them to perform better than traditional algorithms such as XGBoost in low-data regimes (Slack and Singh 2023; Hegselmann et al. 2023). However, LLM predictions can reflect inherent biases, affecting the fairness of their outcomes (Hu and Du 2024; Liu et al. 2023). Liu et al. (2023)'s work is closely related to ours: they analyze the accuracy and fairness of LLM predictions, concluding that traditional ML models exhibit fewer disparities. Although in-context learning and finetuning do not fully close the fairness gap, label-flipping in in-context examples significantly reduces biases, albeit at the cost of prediction performance. Our work contributes to this literature by introducing four in-context learning approaches for mitigating the demographic parity gap in tabular data predictions, demonstrating their effectiveness across widely-used fairness datasets.

Bias and Stereotypes in LLMs

Despite their promising capabilities, language models also exhibit biases and stereotypes (Bolukbasi et al. 2016; Bender et al. 2021; Chu, Wang, and Zhang 2024). These biases mostly originate from the training data, which often contain historical and societal prejudices embedded within the text. Biases have been reported with respect to several demographic groups, e.g., gender, race, ethnicity, and socioeconomic status (Wan et al. 2023; Haim, Salinas, and Nyarko 2024; Santurkar et al. 2023). With the use of these models becoming more widespread, these biases have the risk to substantially reinforce harmful stereotypes and perpetuate existing inequalities, especially when deployed in high-stakes settings (Zou and Schiebinger 2018). Addressing these biases is essential, and several mitigation strategies have been proposed for this purpose, including data augmentation, prompt tuning and few-shot learning (Zhao et al. 2017; Wang et al. 2021; Sun et al. 2019; Zmigrod et al. 2019; Mattern et al. 2022; Fatemi et al. 2021; Aguirre et al. 2023). However, effectively applying these strategies to the large-scale pretraining corpora remains challenging. Finally, biases can be hard to detect and several datasets and methods have been proposed to help identify them (Caliskan, Bryson, and Narayanan 2017; May et al. 2019; Webster et al. 2020; Kurita et al. 2019; Nangia et al. 2020; Blodgett et al. 2021).

Fairness on Tabular Data

Much of the work on classification and algorithmic fairness has focused on tabular datasets (Mehrabi et al. 2021; Caton and Haas 2024; Pessach and Shmueli 2023; Fabris et al. 2022; Barocas, Hardt, and Narayanan 2023; Chouldechova and Roth 2018; Fogliato, Chouldechova, and G'Sell 2020). Consequently, there is a wide range of research describing the properties and trade-offs of predictive algorithms on this type of data (Dutta et al. 2020; Black, Raghavan, and Barocas 2022; Akpinar et al. 2022). Multiple works have pro-

posed fairness-enhancing techniques for traditional ML algorithms (e.g., logistic regression), which generally work by debiasing the data, including a fairness constraint in the optimization problem, or post-processing model predictions (Zafar et al. 2017; Dwork et al. 2012; Berk et al. 2017; Lum and Johndrow 2016; Hardt, Price, and Srebro 2016; Martinez, Bertran, and Sapiro 2020; Akpinar, Lipton, and Chouldechova 2024). Our work employs related techniques, although some of them are not directly applicable. The formalization of fairness definitions has also been extensively discussed (Castelnovo et al. 2022). Fairness metrics evaluated on tabular data typically measure the equality of some target measure across demographic groups, such as accuracy or recall (Chouldechova 2017), which fall under the umbrella of group fairness definitions (as opposed to individual fairness definitions). One such widely-adopted measure, which we also employ in this work, is demographic parity, which ensures that the frequency of positive predictions is approximately equal across different demographic groups.

3 Experimental Details

In our experiments we focus on scenarios with little to no training data available. This is a particularly attractive setting for using language models since they typically outperform classical tabular models in low-data regimes as they can leverage their inherent knowledge for predictions (Hegselmann et al. 2023; Slack and Singh 2023). To make prediction on a single sample, we prompt the model with task-specific instructions, along with the relevant features of the sample of interest; see Appendix A for more details on prompting templates. Optionally, we may also include fairness-specific instructions and few-shot examples, depending on the method used to improve fairness. We then extract the answer either by directly generating a response to the question (for closed-source models) or by calculating the likelihood of tokens corresponding to labels.

In experiments that involve selecting the "best prompt" from several iterations, such as in prompt engineering, we utilize a small validation set of 50 labeled examples to assess model accuracy. We then select (empirically) Pareto-optimal prompts, which represent those where any improvement in either accuracy or fairness would necessitate a compromise in the other metric. Due to its limited size, the validation set is used solely to assess accuracy, while demographic parity is directly evaluated on the test set to identify Pareto-optimal prompts. For the experiments requiring selecting the "best model" from several iterations, we assume the availability of a small validation set comprising 50 labeled examples, used to evaluate accuracy of the models for selecting Pareto-optimal points. We additionally compare our methods against a tabular model, specifically, CatBoost implementation of gradient boosted decision trees trained on 50 examples (Prokhorenkova et al. 2018) and fairness constraint enforced by GridSearch[1] function following the reductions approach by Agarwal et al. (2018).

[1]Fairlearn's implementation is used.

Language Models

We conduct experiments using a variety of widely used language models that vary in size. Due to the computational demands of some methods, we conduct computationally intensive experiments with smaller models and reserve methods that require reasoning for larger language models. Our experiments include Llama 3 8B and 70B (Touvron et al. 2023), Mistral 7B (Jiang et al. 2023), Mixtral 8x7B (Jiang et al. 2024) and Claude Sonnet models (Anthropic 2024).

Datasets

We explore group fairness of LLMs on a set of publicly available datasets widely used in the algorithmic fairness literature. For each of the datasets, we focus on *'gender'* as the protected attribute. We briefly introduce the datasets here and point to Appendix C for additional details.

Adult The Adult Income dataset (Barry Becker 1996) includes 1994 US Census information to predict whether individuals' yearly income exceeds \$50k (*1 = yes, 0 = no*). In accordance with previous work (Liu et al. 2024; Slack and Singh 2023), we retain 10 features for prediction, sample 1000 examples for evaluation and use the remainder of the data for generating task-specific instructions.

German credit The German credit dataset (Hofmann 1994) is used to predict credit default risk (*1 = good, 0 = bad*) based on individual attributes. Following previous work (Liu et al. 2024), we retain 9 features and split the data set into 50% for evaluation and 50% for task instruction generation.

ACS Income & Coverage The American Community Survey (ACS) data (Ding et al. 2021) is sourced from the US Census. For our experiments, we utilize the income (*1 = yearly income* $> \$50k$*, 0 = else*) and coverage (*1 = public health coverage, 0 = else*) prediction tasks for 2018 data from the state of New York. For each classification task, we sample 1,000 examples for evaluation, and use the remaining data to select 10 features with the highest importance.

Serialization and prompts

LLMs require textual input, unlike traditional tabular prediction models. In line with previous work (Slack and Singh 2023; Hegselmann et al. 2023), we serialize data points by (1) mapping categorical values to the respective strings (e.g. *gender = 1* is mapped to *gender = male*), and (2) consolidating column names and entries into one string per row.

Although we assume little to no training data, it is reasonable to expect that practitioners will provide task-specific instructions to the model to facilitate accurate predictions. For this, we construct instructions using prototype clustering on the training folds of the datasets, as suggested by Slack and Singh (2023).

To make instructions more readable, we use GPT-4 to revise prototype information into a single summary paragraph. Please, see Appendix A for details on prompt templates.

Metrics

In this work we focus on optimizing *demographic parity* (DP) which aims to equalize positive label selection rate across groups, i.e.

$$\mathbb{E}[f(X) \mid G = g] = \mathbb{E}[f(X)]$$

for a binary predictor f and $g \in \{\text{male}, \text{female}\}$. Constraint violation is reported as ratio between the smallest and largest group level selection rates $\mathbb{E}[f(X) \mid G = g]$ with values closer to 1 indicating better parity. We use DP primarily because it allows to measure fairness on an unlabeled test set directly and does not require labeled training data. Although our primary focus is on demographic parity, the methods we propose can be adapted to other fairness metrics when labeled training data is available as discussed in Section 7. Additionally, while our main objective is demographic parity, we also evaluate *equalized odds* which aims to balance false positive and false negative rates across groups, i.e.

$$\mathbb{E}[f(x) \mid G = g, Y = y] = \mathbb{E}[f(x) \mid Y = y]$$

for a binary predictor f, $Y \in \{0, 1\}$, and $g \in \{\text{male}, \text{female}\}$, and report equalized odds ratio between groups.

4 Methods

In this work we consider four empirical approaches for improving group fairness of language model predictions on tabular datasets as illustrated in Figure 1. This section provides a detailed overview of each method, with subsequent section delving into experimental results for each approach.

Fair Prompt Optimization. Prompt engineering continues to play an important role in tailoring the capabilities of LLMs to various tasks (Chen et al. 2023; Wang et al. 2023; White et al. 2023). Recently, Tamkin et al. (2023) demonstrated that integrating fairness-specific manually-curated instructions in the prompt, such as "it is illegal to discriminate", can attenuate counterfactual biases in model predictions. Additionally, several works have shown that LLMs can act as prompt engineers producing performant prompts for downstream models (Zhou et al. 2022; Yang et al. 2023). In our work we demonstrate the effectiveness of prompt engineering in achieving group fairness in LLMs and show how prompt optimization can be automated. In particular, we propose to optimize a fairness-specific prompt (highlighted in blue on the left panel on Figure 1), appended to the task-specific instructions. We adopt the prompt optimization approach following Yang et al. (2023) and employ meta-LLM to iteratively refine fairness instructions for the downstream model based on feedback provided from the previous iterations. In particular, we demonstrate the most fair previous instruction and two randomly selected previous instructions, along with their demographic parity scores and selection rates across groups, see example in Figure 2. We refine fairness instructions using the meta-model over 100 iterations. For the meta-LLM we employ the same language model as the one used downstream to make predictions.

Meta-Prompt

You are a great and creative instruction generator that generates instructions for downstream language models. You must generate a one-sentence "fair instruction" to be appended to an existing general task instruction. The fair instruction should improve demographic parity (equality of selection rates for males and females) in the predictions of the downstream model. The following exemplars show how your fair instruction is used by the downstream model:

You must follow the instructions to predict if income of an individual exceeds $50K/yr.

Generally, individuals who earn more than $50K/yr tend to have higher levels of education (e.g., Bachelors or Prof-school), work in managerial or specialty occupations, and are married with a spouse. They also usually work longer hours per week and have higher capital gains. ... [fair instructions]

Input: workclass: Private; hours per week: 17; sex: Female; age: 42; occupation: Exec-managerial; capital loss: 1485; education: Bachelors; capital gain: 0; marital status: Married-civ-spouse; relationship: Wife;
Output: greater than 50K

The previous instructions, along with their corresponding accuracy, demographic parity, and positive label selection rates across genders, are provided below. You should 1. interpret the performance scores, 2. analyze what should be changed in the fair instruction, and 3. provide new short fair instruction to improve demographic parity, which is very different from the previous ones. Wrap the new fair instruction within < > tags.
The previous instructions are:

Assign positive label [greater than 50K] to females and males with equal probability.
Demographic parity ratio is 0.66
Selection rate for Females is 0.23
Selection rate for Males is 0.35
(...more examples..)

Figure 2: Meta-prompt used to iteratively refine fairness-instructions using a meta-LLM.

Soft Prompt Tuning. In addition to hard prompt optimization, we explore soft prompt tuning, which optimizes the prompt directly in the embedding space instead of discrete token space, see the second-from-left panel in Figure 1. In traditional tabular methods, standard in-processing fairness interventions often involve training machine learning models with a fairness penalty. This encourages the model to equalize selection rates or, depending on the penalty, the error rates across demographic groups (Zafar et al. 2017; Hardt, Price, and Srebro 2016). Drawing inspiration from these techniques and parameter-efficient fine-tuning methods, we propose a similar approach that can be applied to improving group fairness in language models. In particular, rather than optimizing fair prompts in the discrete space of tokens, as done in the previous section, we suggest optimizing a soft prompt by fine-tuning tokens in the embedding space. Continuous optimization in the embedding space allows us to incorporate the fairness penalty into objective directly. Specifically, we fine-tune 50 tokens initial-

ized with task-specific instructions in the embedding space for 20 epochs. This approach applies a penalty designed to equalize the likelihoods of tokens corresponding to positive labels across groups within a batch:

$$|P(Y = 1|A = 0) - P(Y = 1|A = 1)|.$$

To tune the prompt we use 1000 samples with pseudo-labels obtained by the same language model in the zero-shot setup, simulating a scenario without labeled data. To preserve accuracy and ensure predictions remain close to the original model outputs, we include the standard cross-entropy loss for the pseudo-label predictions.

Fair Few-Shot Examples. Prior work (Liu et al. 2024; Hu and Du 2024; Li, Zhang, and Zhang 2024) has leveraged the in-context learning capabilities of language models for this problem space. They hypothesize that, when selected appropriately, few-shot examples can effectively influence the final predictions to more accurately reflect the desired notion of fairness. For instance, it has been demonstrated that flipping the labels of few-shot examples can effectively reduce bias, albeit at the expense of significantly lower classification performance (Liu et al. 2024), while class- and group-balanced selection does not mitigate the bias (Li, Zhang, and Zhang 2024).

With a similar goal, we propose a strategy for constructing fair few-shot examples, which differs from the previous methods in three ways. First, instead of randomly sampling examples from the training data, we apply the nearest neighbor search to select examples that are most similar to a current test instance in the feature space[2]. Also, we always select examples that share sensitive attribute with the test instance. Secondly, as we assume no access to training data labels, we use the language models' default zero-shot predictions as pseudo labels to construct demonstrations (similarly to soft prompt tuning experiments). Finally, we extensively manipulate the distributions of positive and negative in-context examples between groups. In particular, we test varying ratios of positive examples for female test samples, $p_f = [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]$, and for male test samples $p_m = [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]$, resulting in 36 ratio pairs. We hypothesise that increasing the number of positive examples for the minority group increases their selection rate, thereby promoting better parity with the majority group.

Self-Refinement. In addition to in-processing methods, fairness literature also includes a wide array of post-processing techniques (Hardt, Price, and Srebro 2016). These methods work by altering model outputs directly. We propose an LLM-based post-processing method that leverages the reasoning capabilities of language models, along with a chain-of-thought process, to refine their own predictions. The self-refinement approach involves using language models to identify individuals from both minority and majority groups who are near the "decision boundary", and then flipping their labels to achieve the desired demographic parity ratio. Therefore, the prediction process includes two stages. First, the model makes initial predictions on a batch of data samples. After that, the model assesses demographic parity in a batch and adjusts predictions to attain the desired parity, if necessary. An example prompt used to refine predictions is illustrated in Figure 1 most right panel. Given that self-refinement approach relies on the reasoning capabilities of language models to analyze predictions, compute metrics of interest, and adjust individual predictions, we conduct these experiments with larger models, specifically Llama3 70B and Claude Sonnet models.

[2]To compute similarity scores between instances, we use the Jaccard metric as most features are discrete or categorical.

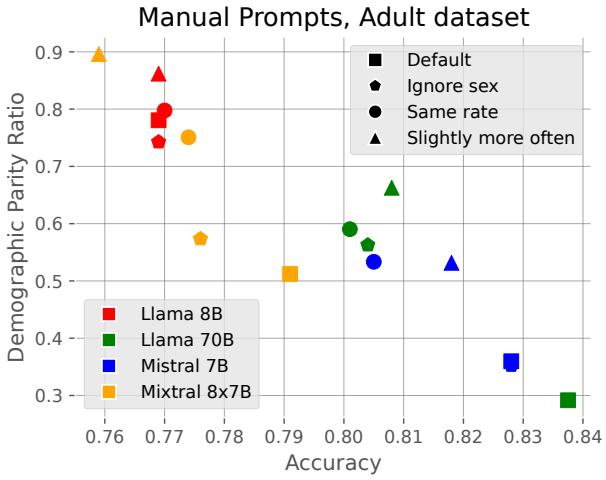

Figure 3: Accuracy and demographic parity for manually constructed fair prompts on Adult dataset, 4 models.

5 Experimental Results

In this section we present performance of the proposed methods. Since certain fairness metrics are not necessarily aligned with accuracy, models producing fairer decisions may suffer from accuracy degradation. Therefore, it is important to identify methods resulting in an optimal fairness-accuracy trade-off.

Fair Prompt Optimization

Prompt engineering continues to play an important role in tailoring the capabilities of LLMs to various tasks (Chen et al. 2023; Wang et al. 2023; White et al. 2023). Recently, Tamkin et al. (2023) demonstrated that integrating fairness-specific manually-curated instructions in the prompt, such as "it is illegal to discriminate", can attenuate counterfactual biases in model predictions. To evaluate the effectiveness of prompt engineering for fairness, we, first, manually construct a few prompts which are supposed to improve group fairness, such as *(1) "Ignore sex completely when making predictions", (2) "Assign positive label to females and males at the same rate.", (3) "Assign positive label to females slightly more often than to males."*, with the last prompt coming from the fact that female group is the one disadvantaged. We include the accuracy and demographic parity for models queried with these prompts

Model	Adult			German Credit			ACS Coverage			ACS Income		
	Acc	DP	EO	Acc	DP	EO	Acc	DP	EO	Acc	DP	EO
Catboost + GS	0.76	0.57	0.67	0.66	0.75	0.44	0.63	0.81	0.65	0.75	0.85	0.88
Llama8B Default	**0.77**	0.78	0.9	0.56	0.8	0.66	0.62	0.73	0.60	0.71	0.88	0.95
Llama8B+FairPrompt	**0.77**	**0.94**	0.79	0.57	0.95	0.81	**0.67**	0.96	0.93	**0.74**	**0.92**	0.82
Llama8B+Few-Shot	0.76	**0.94**	0.77	0.63	0.91	0.85	0.61	0.96	0.9	0.73	0.9	0.91
Llama8B+SoftPrompt	0.73	**0.94**	0.84	**0.66**	**0.97**	0.90	0.59	**0.97**	0.88	0.69	0.89	0.92
Mistral7B Default	**0.83**	0.36	0.43	0.62	0.82	0.73	**0.67**	0.49	0.26	**0.76**	0.86	0.89
Mistral7B+FairPrompt	0.81	0.55	0.77	**0.7**	0.9	0.68	0.66	**0.94**	0.88	0.71	0.92	0.91
Mistral7B+Few-Shot	0.80	**0.93**	0.68	0.57	0.95	0.92	0.66	0.93	0.83	**0.76**	**0.99**	0.59
Mistral7B+SoftPrompt	0.75	0.90	0.62	0.65	**0.97**	0.90	0.56	0.92	0.88	0.75	0.85	0.91
Mixtral8x7B Default	**0.79**	0.51	0.57	0.47	0.72	0.65	**0.65**	0.83	0.75	0.71	0.85	0.96
Mixtral8x7B+FairPrompt	0.78	**0.95**	0.61	**0.58**	0.94	0.83	0.64	**0.99**	0.9	0.72	0.92	0.89
Mixtral8x7B+FewShot	0.76	0.91	0.81	0.46	**0.97**	0.75	0.64	**0.93**	0.86	**0.73**	**0.93**	0.76

Table 1: Performance of optimized fair prompts, tuned soft prompts, and few-shot contexts across 3 models and 4 datasets. We report performance of Pareto-optimal instructions achieving the best validation accuracy and at least 0.9 demographic parity. Bold numbers indicate better accuracy and demographic parity across methods for each model and dataset.

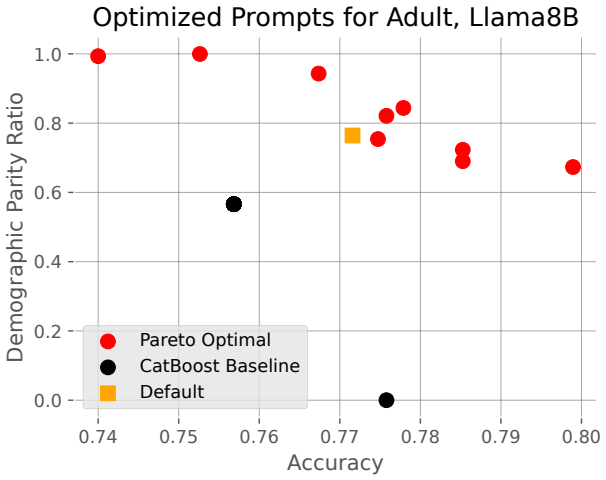

Figure 4: Accuracy and demographic parity for fair prompts optimized via a meta-LLM: red points denote Pareto-optimal fair prompts, the orange square shows the default model's performance, and black points depict a CatBoost model optimized on 50 examples with grid search.

for Adult dataset in Figure 3. We observe that, while these prompts can improve demographic parity in some models, finding a universal "fair instruction" that upholds group fairness consistently across multiple models is challenging.

Next, we experiment with a prompt optimization framework using a meta-LLM, designed to dynamically refine fair instructions based on demonstrated prior instruction candidates with their demographic parity scores and group selection rates. We present the performance of Pareto-optimal fair prompts for the Llama 8B model and the CatBoost model baseline in Figure 4. Additionally, Table 1 lists results for the fair prompts which are Pareto-optimal and achieve at least 0.9 demographic parity ratio. We observe, that including these engineered fair prompts significantly improve fairness of the models, often without sacrificing much accuracy, or even improving it.

Soft Prompt Tuning

Soft prompt tuning enables continuous optimization of a fairness objective by incorporating it directly into the loss function. We tune the soft prompts with a demographic parity fairness regularizer, which aims to equalize the likelihood of positive label predictions across different groups within a batch[3].

To preserve accuracy and ensure predictions remain close to the original model outputs, we include the standard cross-entropy loss for the pseudo-label predictions. The tuning process leverages 1,000 samples with pseudo-labels obtained from the same language model in a zero-shot setup, simulating a scenario without labeled data. Similarly to our fair prompt engineering experiments, we identify Pareto-optimal points among fine-tuning epochs using the validation set and include results for Pareto-optimal soft prompts achieving at least 0.9 test demographic parity in Table 1. We observe that while tuning soft prompts improves demographic parity across all datasets, it results in suboptimal trade-off with accuracy compared to hard prompt optimization approach. This could potentially be attributed to the sensitivity of the tuning procedure to hyperparameters or the reliance on pseudo-labels.

Fair Few-shot Examples

In this section we present results for our fair few-shot example construction strategy, which selects nearest-neighbors to

[3]We use the Prompt Tuning implementation by Hugging Face.

	Adult			German Credit			ACS Coverage			ACS Income		
Model	Acc	DP	EO	Acc	DP	EO	Acc	DP	EO	Acc	DP	EO
Catboost + GS	0.76	0.57	0.67	0.66	0.75	0.44	0.63	0.81	0.65	0.75	0.85	0.88
Llama70B Default	0.78	0.53	0.59	0.58	0.85	0.61	0.61	**0.81**	0.69	0.75	0.84	0.99
Llama70B+Self-Refine	0.71	**0.89**	0.89	0.56	**0.92**	0.78	0.6	0.76	0.64	0.74	**0.87**	0.92
Claude Default	0.79	0.50	0.57	0.66	0.93	0.89	0.63	**0.77**	0.64	0.76	0.82	0.94
Claude+Self-Refine	0.73	**0.98**	0.74	0.63	**0.97**	0.66	0.64	0.72	0.61	0.75	**0.9**	0.88

Table 2: Results for self-refining approach across three models and four datasets. Bold numbers indicate better demographic parity between the original and refined predictions.

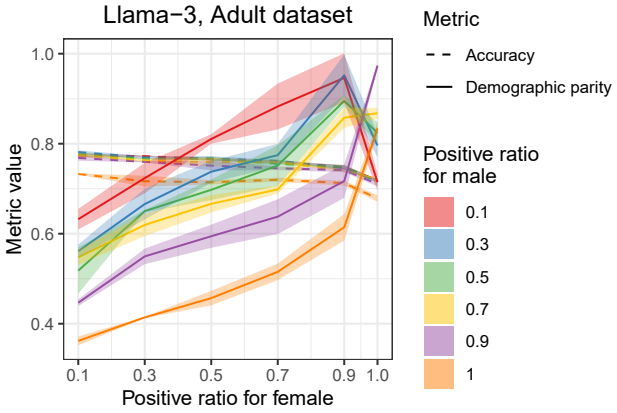

Figure 5: Accuracy and demographic parity ratio metrics for prompts containing few-shot examples with varying number of positive examples across groups, evaluated on Adult dataset using Llama8B.

test instances, uses zero-shot model predictions as pseudo-labels, and adjusts the ratio of positive examples between groups to enhance fairness. Figure 5 illustrates the impact of varying the ratio of positive examples in the prompt. The x-axis represents the ratio of positive examples for female test instances, while the color indicates the ratio of positive examples used for predictions on male samples. The results are averaged across 3 random seeds, with the band indicating the standard deviation across seeds. We observe that increasing the positive ratio for females significantly improves demographic parity, to the extent that the selection rate for females surpasses that for males. These results confirm that adjusting the ratio of positive examples in-context is an effective method for manipulating the prevalence of positive class predictions, and employing different ratios across protected groups can effectively reduce disparities in selection rates.

Additionally, we compare our nearest-neighbor selection strategy with a baseline selecting examples randomly while preserving similar label ratios in-context. Figure 6 in Appendix shows that including random in-context examples results in lower demographic parity with larger variance. Also, unlike the nearest-neighbor approach, there is no apparent trend showing that including more positive samples boosts the selection rate for any demographic group, highlighting the importance to including only relevant examples in-context. Finally, in Table 1 we report demographic parity ratio, equalized odds ratio and accuracy metrics for the Pareto-optimal combination of positive label ratios, which achieves the best validation accuracy and at least 0.9 demographic parity.

Self-Refinement

When making predictions in batches, we can utilize the chain-of-thought and self-refinement capabilities of language models to apply post-hoc corrections to predictions, see the right panel in Figure 1 for an illustration. We make predictions on a batch of 40 samples, and instruct the model to make adjustments only when the difference in positive rates across groups exceeds 15%. We report the results of the self-refinement approach in Table 2. For all models, refined predictions achieve improved demographic parity across all datasets except for ACS coverage, although this sometimes leads to a notable trade-off in accuracy. The effectiveness of this method largely hinges on the batch size used and the chosen disparity threshold. In addition, there is no guarantee for similar individuals to receive similar predictions with this method because of the 'correction step' which is at odds with notions of individual fairness (Dwork et al. 2012).

6 Conclusion

We systematically explore four empirical methods to improve group fairness of language model predictions on tabular datasets, and discuss the key takeaways for each method below.

Fair Prompt Optimization can improve not only fairness but also classification performance, contingent upon the model's "creativity." This method involves an optimization process that requires evaluating the prompt on a dataset for a number of iterations. Although the resulting instructions are interpretable, the reasons why specific instructions yield fairer results are not always clear.

Soft prompt tuning is computationally expensive and sensitive to the choice of hyperparameters. While this method does not yield interpretable instructions, it enables the integration of common fairness regularizers in a differentiable way and may be particularly effective for smaller models.

Fair Few Shot Examples is the most interpretable and predictable method, yielding optimal results across models and datasets when an optimal combination of positive examples ratios is selected. However, it uses a longer context window and may be more computationally expensive for larger datasets because of the number of forward passes needed.
Self-refinement requires a model with strong reasoning capabilities and does not guarantee similar predictions for similar individuals. However, this method offers a computational advantage for larger models, as predictions are made and adjusted in batches, reducing overall processing time.

We recommend fair few-shot examples and fair prompt optimization as universal approaches achieving the optimal accuracy tradeoff. Soft prompt tuning can potentially adapt smaller models, while self-refinement is useful for scenarios with limited budgets and larger language models.

7 Limitations and Potential Risks

Our work has several limitations. Firstly, it exclusively examines in-context approaches and does not address data pre-processing for bias mitigation or post-hoc methods that modify model outputs directly (Hardt, Price, and Srebro 2016). Additionally, we do not consider model training and fine-tuning strategies other than soft prompt tuning. Finally, we focus on a single notion of fairness, that is demographic parity, since it can be applied in little to no training data regime, the most practical scenario for language models on tabular datasets. However, most of the discussed methods can be adapted to optimize for other fairness notions, such as equalized odds, when labeled training data is available. For example, the prompt optimization procedure can incorporate alternative fairness metrics in the feedback component of the meta-prompt. Soft prompt tuning can adopt differentiable proxy regularizers to enforce desired fairness criteria, and the few-shot examples approach can demonstrate more examples with ground-truth labels to underrepresented groups.

While the methods explored in this work show promise for improving demographic parity of large language models on tabular prediction tasks, there are also several potential risks to consider. Enhancing group fairness may come at the expense of overall predictive performance, and focusing solely on demographic parity could neglect other important fairness criteria. Moreover, optimizing for group fairness along one dimension could inadvertently amplify or introduce other biases not directly measured. As such, practitioners should carefully evaluate the appropriateness and potential pitfalls of these methods for their specific use case before deploying language models for high-stakes decision making on tabular data.

A Prompt Details

In this section we provide the prompt templates used in the default setup to obtain predictions from the models. We highlight task instructions in italic. When using an optimized fair prompt, it follows the task instructions. For the fair few-shot examples experiments, we provide selected examples before the test sample in the prompt.

Prompt Template for Adult dataset

User: You must follow the instructions to predict if the income of an individual exceeds \$50K / year. *Generally, individuals who earn more than \$50K/yr tend to have higher levels of education (e.g., Bachelors or Prof-school), work in managerial or specialty occupations, and are married with a spouse. They also usually work longer hours per week and have higher capital gains. On the other hand, those earning less than or equal to \$50K/yr typically have lower levels of education (e.g., HS-grad), work in clerical or craft repair occupations, and may be single or married without a spouse. They also usually work fewer hours per week and have lower capital gains.* You must provide an explanation and include answer enclosed in square brackets at the beginning of your response.
workclass: Private; hours-per-week: 40; sex: Male; age: 38; occupation: Craft-repair; capital-loss: 0; education: HS-grad; capital-gain: 0; marital-status: Divorced; relationship: Not-in-family.
You must predict if the individual's income exceeds \$50K/yr. Answer with one of the following: [less than or equal to 50K] or [greater than 50K].
Assistant: Based on the information provided the answer is

Prompt Template for German Credit dataset

User: You must follow the instructions to predict the credit risk of a German bank customer based on their profile data. *Higher credit amounts and longer durations often indicate a 'bad' classification, particularly for loans intended for new cars or high-value purchases. A negative or low checking status (either '< 0' or '$0 <= X < 200$') is also associated with higher risk. Conversely, 'good' classifications tend to involve moderate or lower credit amounts, shorter loan durations, and are frequently for consumer goods like radios, TVs, or used cars. Surprisingly, having 'no checking' status tends to lean towards a 'good' classification. More people tend to have good credit risk.* You must provide an explanation and include answer enclosed in square brackets at the beginning of your response.
age: 36.0; sex: male; job: skilled; housing: own; savings status: no known savings; checking status: $0 <= X < 200$; credit amount: 2181.0; duration: 30.0; purpose: new car.
You must predict the credit risk of a German bank customer. Answer with one of the following: [good] or [bad].
Assistant: Based on the information provided the answer is

Prompt Template for ACS Income dataset

User: You must follow the instructions to predict if income of an individual exceeds \$50K/yr. *Individuals who are most likely have income less than or equal to 50K are typically younger, unmarried, with lower levels of education (high school diploma or some college), and engaged in lower-skilled occupations such as cashiers, janitors, or administrative assistants. Conversely, individuals who are older, married, with higher levels of education (bachelor's or master's degree), and are employed in professions such as teachers, managers, physicians, or registered nurses are more likely to have income greater than 50K.* You must provide

Dataset	Selected Features	Prediction Target
Adult (Barry Becker 1996) (CC BY 4.0)	workclass, hours-per-week, gender, age, occupation, capital loss, education, capital gain, marital status, relationship	Annual income $\geq$ \$50K
German Credit (Hofmann 1994) (CC BY 4.0)	age, gender, job, housing, savings status, checking status, credit amount, duration, purpose	Good / bad credit
ACS Income (Ding et al. 2021) (US Census TOS)	gender, race, educational attainment, relationship, age, usual hours worked per week (past 12 months), marital status, place of birth, class of worker, occupation	Annual income $\geq$ \$50K
ACS Coverage (Ding et al. 2021) (US Census TOS)	sex, race, educational attainment, military service, disability status, total personal income, marital status, employment status, parents' employment status, childbirth (past 12 months)	Has public health insurance

Table 3: Public datasets and the main feature subsets used in our experiments.

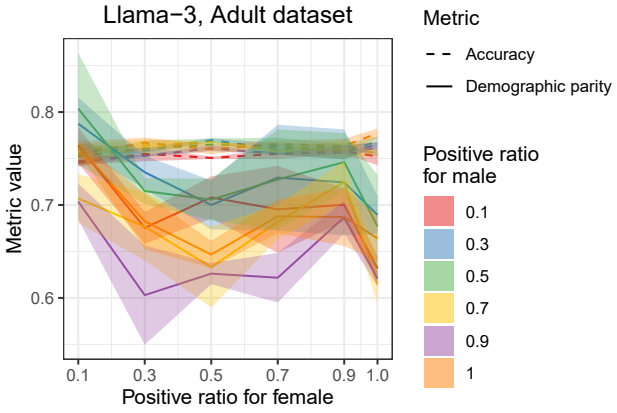

Figure 6: Accuracy and demographic parity ratio metrics for randomly chosen in-context examples with varying number of positive examples across groups, evaluated on Adult dataset using Llama8B.

an explanation and include answer enclosed in square brackets at the beginning of your response.
Sex: Female; Race: White alone; Educational attainment: Bachelor's degree; Relationship: Roomer or boarder; Age: 35.0; Usual hours worked per week past 12 months: 32.0; Marital status: Never married; Place of birth (Recode): Florida/FL; Class of worker: Employee of a private not-for-profit, tax-exempt, or charitable organization; Occupation: EDU-Elementary And Middle School Teachers.
You must predict if the individual's income exceeds \$50K/yr. Answer with one of the following: [less than or equal to 50K] or [greater than 50K].
Assistant: Based on the information provided the answer is

Prompt Template for ACS Coverage dataset

User: You must follow the instructions to predict whether an individual is covered by public health insurance. *Individuals covered by public health insurance tend to have a regular high school diploma, have never served in the military, and generally have lower income. In contrast, features such as being employed, having educational attainment, higher income (above \$20,000) and being married correlate with not being covered by public health insurance. In addition, people with disabilities are more likely to be covered by public health insurance.* You must provide an explanation and include answer enclosed in square brackets at the beginning of your response.
Sex: Female; Race: White alone; Educational attainment: Associate's degree; Military service: Never served in the military; Disability recode: Without a disability; Total person's income: 0.0; Marital status: Never married; Employment status recode: Not in Labor Force; Employment status of parents: N/A (not own child of householder, and not child in subfamily); Gave birth to child within the past 12 months: No.
You must predict if the individual is covered by public health insurance. Answer with one of the following: [covered] or [not covered].
Assistant: Based on the information provided the answer is

B Additional Experimental Details and Hyperparameters

Hyperparameters for Soft Prompt Tuning . In the soft prompt tuning experiments, we fine-tune 50 tokens initialized with the task instructions for 20 epochs. We employ a learning rate of $1e-4$ for Llama 8B models and $1e-5$ for Mistral models, allowing the first three epochs for a warm-up with a linear scheduler. During fine-tuning, we use 1000 train samples with pseudo-labels obtained by using

the language model in a zero-shot setup, we apply demographic parity regularization with a penalty weight of 0.5. We employ a class-balanced sampler and set the batch size to 60 samples for Mistral and 50 samples for Llama models, which were the largest sizes we could use given the computational constraints.

C Datasets

We include details on the datasets and features used in our experiments in the Table 3.

D Additional Results

Figure 6 illustrates the trend in demographic parity for prompts including random examples instead of nearest neighbors. In contrast to our strategy, including random examples does not significantly influence the models' selection rates.

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 60–69. PMLR.

Aguirre, C.; Sasse, K.; Cachola, I.; and Dredze, M. 2023. Selecting Shots for Demographic Fairness in Few-Shot Learning with Large Language Models. *arXiv preprint arXiv:2311.08472*.

Akpinar, N.-J.; Lipton, Z.; and Chouldechova, A. 2024. The Impact of Differential Feature Under-reporting on Algorithmic Fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1355–1382. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Akpinar, N.-J.; Nagireddy, M.; Stapleton, L.; Cheng, H.-F.; Zhu, H.; Wu, S.; and Heidari, H. 2022. A Sandbox Tool to Bias(Stress)-Test Fairness Algorithms. arXiv:2204.10233.

Anthropic, A. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Barry Becker, R. K. 1996. Adult.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–863.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Castelnovo, A.; Crupi, R.; Greco, G.; Regoli, D.; Penco, I. G.; and Cosentini, A. C. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1): 4209.

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

Chen, B.; Zhang, Z.; Langrené, N.; and Zhu, S. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

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

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Chouldechova, A.; and Roth, A. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Chu, Z.; Wang, Z.; and Zhang, W. 2024. Fairness in Large Language Models: A Taxonomic Survey. *arXiv preprint arXiv:2404.01349*.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 6478–6490. Curran Associates, Inc.

Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.-Y.; Liu, S.; and Varshney, K. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, 2803–2813. PMLR.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152.

Fang, X.; Xu, W.; Tan, F. A.; Zhang, J.; Hu, Z.; Qi, Y.; Nickleach, S.; Socolinsky, D.; Sengamedu, S.; and Faloutsos, C. 2024. Large Language Models (LLMs) on Tabular Data: Predic-tion, Generation, and Understanding-A Survey. *arXiv preprint arXiv:2402.17944*.

Fatemi, Z.; Xing, C.; Liu, W.; and Xiong, C. 2021. Improving gender fairness of pre-trained language models without catastrophic forgetting. *arXiv preprint arXiv:2110.05367*.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness evaluation in presence of biased noisy labels. In *International conference on artificial intelligence and statistics*, 2325–2336. PMLR.

Haim, A.; Salinas, A.; and Nyarko, J. 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias. *arXiv preprint arXiv:2402.14875*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Herzig, J.; Nowak, P. K.; Müller, T.; Piccinno, F.; and Eisenschlos, J. M. 2020. TaPas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

Hofmann, H. 1994. Statlog (German Credit Data).

Hu, J.; and Du, M. 2024. Enhancing Fairness in In-Context Learning: Prioritizing Minority Samples in Demonstrations. In *The Second Tiny Papers Track at ICLR 2024*.

Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

Jaitly, S.; Shah, T.; Shugani, A.; and Grewal, R. S. 2023. Towards Better Serialization of Tabular Data for Few-shot Classification. *arXiv preprint arXiv:2312.12464*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Levin, R.; Cherepanova, V.; Schwarzschild, A.; Bansal, A.; Bruss, C. B.; Goldstein, T.; Wilson, A. G.; and Goldblum, M. 2022. Transfer learning with deep tabular models. *arXiv preprint arXiv:2206.15306*.

Li, Y.; Zhang, L.; and Zhang, Y. 2024. Fairness of ChatGPT. arXiv:2305.18569.

Liu, G.; Yang, J.; and Wu, L. 2022. Ptab: Using the pre-trained language model for modeling tabular data. *arXiv preprint arXiv:2209.08060*.

Liu, Y.; Gautam, S.; Ma, J.; and Lakkaraju, H. 2023. Investigating the Fairness of Large Language Models for Predictions on Tabular Data. *arXiv preprint arXiv:2310.14607*.

Liu, Y.; Gautam, S.; Ma, J.; and Lakkaraju, H. 2024. Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classifications. arXiv:2310.14607.

Lum, K.; and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, 6755–6764. PMLR.

Marvin, G.; Hellen, N.; Jjingo, D.; and Nakatumba-Nabende, J. 2023. Prompt Engineering in Large Language Models. In *International Conference on Data Intelligence and Cognitive Informatics*, 387–402. Springer.

Mattern, J.; Jin, Z.; Sachan, M.; Mihalcea, R.; and Schölkopf, B. 2022. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *arXiv preprint arXiv:2212.10678*.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Pessach, D.; and Shmueli, E. 2023. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, 867–886. Springer.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Slack, D.; and Singh, S. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Tamkin, A.; Askell, A.; Lovitt, L.; Durmus, E.; Joseph, N.; Kravec, S.; Nguyen, K.; Kaplan, J.; and Ganguli, D. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. " kelly is a warm person, joseph is a role

model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

Wang, T.; Sridhar, R.; Yang, D.; and Wang, X. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Wang, Z.; Zhang, H.; Li, C.-L.; Eisenschlos, J. M.; Perot, V.; Wang, Z.; Miculicich, L.; Fujii, Y.; Shang, J.; Lee, C.-Y.; et al. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Yin, P.; Neubig, G.; Yih, W.-t.; and Riedel, S. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, 962–970. PMLR.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhao, Y.; Zhang, H.; Si, S.; Nan, L.; Tang, X.; and Cohan, A. 2023. 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*, 160–175.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Zhu, B.; Shi, X.; Erickson, N.; Li, M.; Karypis, G.; and Shoaran, M. 2023. Xtab: Cross-table pretraining for tabular transformers. *arXiv preprint arXiv:2305.06090*.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

Zou, J.; and Schiebinger, L. 2018. AI can be sexist and racist—it's time to make it fair.