Towards Fair In-Context Learning with Tabular Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Transformer-based tabular foundation models have recently demonstrated promising in-context learning (ICL) performance on structured data, emerging as competitive alternatives to gradient-boosted trees. However, the fairness implications of this new paradigm remain largely unexplored. We present the first investigation of fairness in tabular ICL, evaluating three recently proposed foundation models—TabPFNv2, TabICL, and TabDPT—on multiple benchmark datasets. To mitigate biases, we explore three pre-processing fairness-enhancing methods: correlation removal (decorrelating input features from the sensitive attribute), group-balanced sample selection (ensuring equal representation of protected groups in context examples), and uncertainty-based sample selection (prioritizing context examples with high sensitive-attribute prediction uncertainty). Our experiments show that the uncertainty-based strategy consistently improves group fairness metrics (e.g., demographic parity, equalized odds, and equal opportunity) with minimal impact on predictive accuracy. We release our code to facilitate reproducibility (https://anonymous.4open.science/r/Fair-TabICL-Anonymized).

1 Introduction

Tabular data, represented in rows and columns, is a data modality widely used for prediction tasks in domains such as finance and healthcare (Asuncion et al., 2007). Tree-based models such as XGboost (Chen et al., 2015) and Gradient-Boosted Trees (Ke et al., 2017) have shown the strongest generalization performance on tabular data. Recently, with the emergence of foundation models, Deep Learning (DL) based models have challenged the dominance of tree-based models (Hollmann et al., 2025). Foundation models are models pretrained on vast datasets, without a specific task in mind, and they can be adapted across a wide range of downstream tasks. Large language models (LLMs) such as GPT-3 are common examples of foundation models, and they have demonstrated emerging capabilities such as in-context learning (ICL) with few labelled data (Brown et al., 2020). In-context learning (ICL) has primarily been applied to natural language tasks using large language models (LLMs). For example, in text classification, labeled examples are formatted as textual demonstrations and provided as context to a language model, enabling it to predict the label of a new instance without any parameter updates or fine-tuning (Radford et al., 2019; Brown et al., 2020). More recently, efforts have extended ICL to tabular data by serializing table rows into text or sentences (Hegselmann et al., 2023). However, since LLMs are not pretrained to model the complex structural relationships inherent in tabular data—such as interactions between rows and columns—their performance on large-scale tabular tasks still lags behind tree-based methods (Hegselmann et al., 2023).

Alternatively, recent work has proposed foundation models explicitly tailored for tabular data, achieving competitive performance with tree-based models while reducing the need for extensive model selection and hyperparameter tuning. For instance, TabPFN (Hollmann et al., 2022) is a transformer-based model pretrained on synthetic datasets, and its successor, TabPFNv2 (Hollmann et al., 2025), extends support to larger datasets with up to 10k samples. Similarly, TabICL (Qu et al., 2025), also pretrained on synthetic data, scales to datasets with up to 500k samples. These models leverage synthetic pretraining to encode a wide range of statistical priors, enabling effective target inference from in-context examples.

To better reflect the priors found in real-world datasets, other models incorporate real data during pretraining. Tabular Discriminative Pre-trained Transformer (TabDPT) (Ma et al., 2024) is pretrained directly on real-world datasets, while Real-TabPFN (Garg et al., 2025) builds on synthetic pretraining with additional fine-tuning on real data. These approaches generally yield improved performance by aligning the learned representations more closely with real-world data distributions.

Given their strong performance and in-context learning capabilities, transformer-based tabular foundation models will likely see widespread adoption in real-world decision-making tasks. This shift could mark a turning point in how tabular data problems are approached. However, the use of ICL-based models in high-stakes domains—such as healthcare, finance, or criminal justice—raises important ethical concerns. In particular, it is critical to assess their potential to perpetuate or even amplify existing social biases. Traditional machine learning models have already been shown to replicate biases present in the data (Mehrabi et al., 2022), and recent studies indicate that LLM-based ICL can also produce biased predictions (Hu et al., 2024; Bhaila et al., 2024). However, these studies rely on serialized representations of tabular data and therefore inherit the limitations of LLMs in handling tabular structures (Ma et al., 2024).

This paper investigates the fairness of ICL prediction using transformer-based tabular foundation models. First, our study reveals, perhaps unsurprisingly, that while these models focus on improving prediction accuracy, they can also amplify bias. Motivated by recent studies on the sensitivity of ICL performance—in terms of fairness and accuracy— to demonstration selection, we aim to address the following research question: *What in-context selection/transformation method can improve the fairness of ICL predictions?* In the fairness literature, fairness-enhancing methods are generally grouped into three categories: pre-processing, in-processing, and post-processing (Mehrabi et al., 2022). Projecting these categories into the ICL paradigm, pre-processing methods perform demonstration transformation or selection before predicting in context (Hu et al., 2024). In-processing methods would fine-tune or retrain the foundation model with fairness constraints (Robertson et al., 2024). Post-processing methods would alter the ICL predictions to improve a given fairness metric (Hardt et al., 2016). Pre- and Post-processing methods are more computationally friendly since they do not require model updates. This motivates our choice to focus on the pre-processing techniques and leave post-processing interventions for future exploration. More specifically, we propose and investigate three pre-processing fairness interventions: (i) Correlation Remover (Feldman et al., 2015), a method that alters each input feature to reduce their correlation with the sensitive attribute; (ii) group-balanced[1] in-context selection, ensures that the in-context set is group-balanced; (iii) Uncertainty-based in-context selection, estimates the uncertainty of predicting the sensitive attribute of in-context samples and only selects samples with uncertain predictions. We performed intensive experiments on eight fairness benchmark datasets to investigate the effectiveness of each method in terms of fairness and accuracy. Our results reveal that the uncertainty-based method can provide better fairness performance across datasets, fairness metrics, and foundational models, with marginal impact on accuracy. Our contribution can be summarized as follows:

- While most existing studies focus on fair ICL with serialized tabular data, we provide, to our knowledge, the first investigation into preprocessing methods for fair prediction in ICL using transformer-based tabular foundation models.
- We propose and investigate three pre-processing intervention methods to enforce fair ICL predictions. These methods aim to reduce the information about the sensitive attributes of in-context samples. We demonstrate that uncertainty-based in-context sample selection can significantly improve the fairness of ICL predictions with a slight drop in accuracy.
- We perform extensive experiments on a broad range of start-of-the-art fairness benchmarks and provide insights into contexts where a given fairness intervention performs best in terms of fairness accuracy tradeoff.
- We release the code to ease reproduction of the results and help researchers and practitioners integrate the proposed methods.

[1] Underlined represent the method's name throughout the paper and in the results.

2 Related works

Fairness. Numerous methods have been developed to enforce group fairness in classical machine learning models (Mehrabi et al., 2022; Mbiazi et al., 2023; Kenfack et al., 2024a). These methods are often categorized as pre-processing, in-processing, or post-processing approaches. Model-agnostic methods in the pre-processing category typically modify or reweight the input data to reduce information correlated with sensitive attributes (Madras et al., 2018; Creager et al., 2019; Kamiran & Calders, 2012; Celis et al., 2020; Balunović et al., 2021; Feldman et al., 2015). In contrast, post-processing techniques adjust the model's prediction outcomes after training to satisfy fairness constraints (Hardt et al., 2016; Petersen et al., 2021). Finally, in-processing approaches embed fairness constraints directly into the training objective (Agarwal et al., 2018; Zhang et al., 2018; Roh et al., 2020). Unlike prior work, our approach focuses on pre-processing interventions applied in in-context learning (ICL) settings, where downstream predictions are made by foundation models without any model updates. We emphasize that model-agnostic methods—particularly pre- and post-processing—are especially suitable in ICL because they do not rely on access to or retraining of the model. However, their effectiveness in this setting, especially for tabular foundation models, remains largely unexplored and is the focus of our evaluation.

Tabular Fondation Models. In-context learning with tabular foundation models presents a notable advantage over traditional machine learning approaches by enabling models to adapt dynamically to new data without the need for retraining (Hollmann et al., 2022; Qu et al., 2025; Hollmann et al., 2025). Conventional ML methods typically depend on predefined training datasets, meaning that any alteration in the data or task necessitates a time-consuming and resource-intensive retraining process. In contrast, tabular foundation models utilize in-context learning to execute tasks based on the specific context of the data provided at inference time. This allows these models to interpret and process new tabular data with minimal prior preparation, facilitating more flexible and efficient decision-making (Hollmann et al., 2022). The advantages of this approach are particularly apparent in scenarios where data distributions change over time or when models must quickly adjust to various data tasks without undergoing retraining. Thus, as in-context learning emerges as a powerful tool for real-time, adaptive predictions in complex and dynamic environments, assessing and mitigating biases in the prediction can make its use more socially acceptable. Existing models are pre-trained using synthetic (Qu et al., 2025; Hollmann et al., 2022) or real-world data (Ma et al., 2024). Pre-training on real-world data often provides competitive or better performance, and the performance of synthetic pre-trained tabular foundation models can be boosted with continued pre-training on real-world data (Garg et al., 2025). In this work, we consider tabular foundation models pretrained on synthetic (Hollmann et al., 2025; Qu et al., 2025) and real-world data (Ma et al., 2024) and assess the fairness implications of each pretraining strategy.

Fairness in ICL. Fairness in in-context learning has primarily been studied in the context of large language models (LLMs) applied to tabular data serialized as text(Bhaila et al., 2024; Hu et al., 2024; Ma et al., 2023). For instance, Hu et al. (2024) explore group-aware sampling strategies, finding that prioritizing minority group demonstrations improves fairness outcomes. Similarly, Bhaila et al. (2024) propose a data augmentation technique that guides demonstration selection to reduce bias during inference. While related in spirit, our approach differs in two key ways: (1) we focus on numerical tabular foundation models (rather than LLMs), and (2) our uncertainty-based selection method aims to reduce model reliance on sensitive attributes rather than optimize informativeness per se. Prior uncertainty-driven methods(Mavromatis et al., 2023; Kung et al., 2023) focus on selecting informative examples under a labeling budget, whereas we use uncertainty to guide fair demonstration selection. It is also important to highlight that LLMs, though flexible, are not optimized for tabular data and often perform worse than specialized numeric models, such as gradient-boosted trees (Hegselmann et al., 2023). As such, our work fills a key gap by investigating fairness interventions tailored to numeric tabular foundation models. To our knowledge, this is the first study to evaluate pre-processing fairness methods in this emerging model class.

Fairness-Aware Tabular Foundation Models. Recent work by Robertson et al. (2024) introduced FairPFN, a TabPFN-style model trained to suppress the causal effect of sensitive attributes during pretraining. Their approach seeks counterfactual fairness (Kusner et al., 2017), ensuring that model predictions remain invariant when sensitive attributes are counterfactually changed. Our work is conceptually distinct in two respects. First, we do not require model retraining and rely on model-agnostic pre-processing methods, making our approach broadly applicable to any pretrained tabular foundation model. Second, we aim for group fairness, focusing on improving performance disparities between subgroups, rather than enforcing counterfactual invariance at the individual level.

3 Methodology

Problem Setup We consider a classification task with the giving training data $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$ where x_i is an input feature vector, y_i is the corresponding class label, and s_i the corresponding demographic group. The goal is to obtain a classifier f, via ICL, to accurately predict the target y given a sample x while being *fair* w.r.t. demographic information s. Several metrics have been proposed to measure fairness at the group or individual levels (Dwork et al., 2012). In this work, we focus on group fairness notions, measuring the performance disparity across different demographic groups, i.e., demographic parity, equalized odds, and equal opportunity (Hardt et al., 2016). A detailed description of these fairness metrics can be found in Appendix B.1.

This section presents three pre-processing techniques proposed in this work to ensure fairer ICL inference on tabular data. In particular, we consider *correlation remover* (Feldman et al., 2015), group-balanced demonstration selection, and uncertainty-based demonstration selection.

3.1 In-context Samples Transformation

Correlation remover (Feldman et al., 2015; Bird et al., 2020) is a preprocessing method that reduces the correlation between the sensitive and non-sensitive attributes before fitting the model. More specifically, a linear transformation is applied to each non-sensitive feature to reduce its correlation with the sensitive feature. We use the correlation remover as a preprocessing step over the training (in-context example) and testing sets before performing in-context prediction. Ultimately, transforming input features to reduce their linear dependency on the sensitive feature can reduce the reliance on sensitive features in the downstream models. However, as we will see in the results, nonlinear and complex downstream models can still infer the nonlinear dependencies over the sensitive feature and provide unfair results. A more detailed description of correlation remover can be found in Appendix B.2.

3.2 In-context Samples Selection

In this work, we posit that in-context sample selection can have a significant impact on the fairness of ICL prediction. We analyze several demonstration selection methods that can improve the fairness of ICL predictions without any fairness intervention.

Group-balanced demonstration set selection. Representation bias is a common source of bias in machine learning models (Mehrabi et al., 2022). It occurs when the collected training data does not reflect the demographic diversity of the population. As a result, some demographic subgroups are under-represented, if not represented at all. Recent studies have demonstrated the benefits of group-balanced training data on the fairness properties of the downstream model. Several methods have been proposed to mitigate representation bias in the data, including *subsampling the majority group* or *reweighting the training data* based on group proportions (Kamiran & Calders, 2012; Celis et al., 2020). In this paper, we focus on *subsampling* since current tabular foundation models do not handle sample weights (Hollmann et al., 2025; Qu et al., 2025). Specifically, we perform ICL with a group-balanced demonstration set sampling from each group uniformly at random. When the demonstration set size does allow equal group representation, we subsample the majority group at random. A similar strategy is employed by (Hu et al., 2024) to select demonstrations for few-shot ICL prediction with LLMs. In this paper, we evaluate the effectiveness of this fairness intervention with tabular foundation models instead of using LLMs on serialized tabular data.

Uncertainty-based demonstration set selection Kenfack et al. (2024b) demonstrated that models trained without fairness constraints can have better fairness properties when the training data consists of samples with uncertain sensitive attributes. We hypothesize that *the uncertainty of the sensitive attribute prediction can be a good measure to select demonstrations that improve the fairness of in-context predictions.* To validate this, we measure the uncertainty of predicting the sensitive attribute in the demonstration set and use samples with high uncertainty for in-context learning. We focus on conformal prediction (Shafer & Vovk, 2008; Vovk et al., 2005) as uncertainty measure since it provides strong theoretical guarantees for the coverage. Instead of returning a single label, a conformal predictor returns a prediction set containing the true label with a probability of at least $1 - \epsilon$, with ϵ being a user-defined coverage parameter of the conformal prediction (Angelopoulos et al., 2023). For example, setting $\epsilon = 0.1$ ensures the prediction set contains the true sensitive attribute value with at least 90% probability. Specifically, samples with prediction sets containing more than one value are uncertain. Intuitively, the coverage parameter ϵ controls the fairness-accuracy tradeoff, with $\epsilon \approx 1$ meaning no fairness intervention where all the datapoints are used and $\epsilon \approx 0$ meaning maximal fairness intervention where only uncertain samples are included in the in-context examples. We show in Section 4.2.2 how the coverage parameter ϵ of the conformal predictor consistently controls the tradeoff between accuracy and fairness across fairness metrics and downstream foundational models.

Since conformal prediction is model agnostic, we considered both classical methods, e.g., Logistic Regression (LR), and foundation models, e.g., TabPFN for training the sensitive attribute classifier to measure the prediction uncertainties. Note that the method could be applied using other uncertainty measures, such as Monte Carlo dropout and confidence interval (Kenfack et al., 2024b). We focus on conformal prediction due to its rigorous theoretical guarantees, and it does not require a hyperparameter to threshold the level from which a prediction is considered uncertain (Angelopoulos et al., 2023). More details about uncertainty measurement with conformal prediction can be found in Appendix B.3.

3.3 In-context Prediction

After performing demonstration selection or transformation using the fairness intervention methods presented previously, we pass them through the tabular foundation model as in-context examples for predicting class labels on the test set. In this paper, we consider three tabular foundation models: TabDPT (Ma et al., 2024), TabICL (Qu et al., 2025), and TabPFNv2 (Hollmann et al., 2025). While all these models use Transformer architectures as their backbones and are pretrained on a large amount of datasets, TabICL and TabPFN use only synthetic data generated from structured causal networks, and TabDPT uses real-world data. The current version of TabPFN can handle a maximum of 10k samples with 500 features, and TabICL can handle up to 500K samples. We randomly subsample in-context examples when the context set size exceeds the maximum number of samples allowed by the foundation model being evaluated.

4 Experiments

We describe the experimental setup and perform intensive experiments to answer the following research questions: Does correlation remover and group-balanced demonstration selection effectively reduce information about the sensitive attribute and improve the fairness of ICL? What fairness intervention provides a better fairness-accuracy tradeoff? And what foundation tabular (TabPFN vs TabICL) model can provide a better fairness-accuracy tradeoff?

4.1 Experimental Setup

Datasets We experiment on tasks from the recently proposed folktables (Ding et al., 2022), which contains data extracted from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2022). More specifically, we experiment with the following ACS PUMS tasks: ACSIncome, ACSMobility, ACSTravelTime, ACSEmployment, ACSPublicCoverage. These tasks reflect a range of real-world predictive challenges with fairness concerns. A limitation of the ACS PUMS datasets is that they are US-centric; we diversify the experimental setup by including other tasks and datasets. Specifically, we also experiment on the other tabular datasets and tasks including: Diabetes (Gardner et al., 2023), German Credit (Frank, 2010), and CelebA (Liu et al., 2018). More details about each dataset, including the sensitive attributes used for fairness evaluation, the number of samples, and the number of features, can be found in the Appendix A.

Metrics In addition to the classic accuracy, we consider three group fairness metrics, i.e., Demographic Parity (ΔDP), Equal Opportunity (ΔEOP), and Equalized Odds (ΔEOD). More details about group fairness metrics can be found in B.1.

Evaluation For evaluation, we split each dataset into 80%-20%, where the 20% is held out and used to train the sensitive attribute classifier for uncertainty quantification. For the remaining 80% we use 5-fold cross-validation with random shuffle across ten independent runs. This ensures our evaluation is robust and reliable since every data point is used as a test or in-context example across k-fold evaluations. We report the average and standard deviation of fairness and accuracy performance across the k-fold test sets and the ten random independent seeds. As aforementioned, we use TabICL, TabDPT and TabPFN as foundation models with their respective default parameters.

Baselines We measure and compare the fairness and accuracy performance of ICL under four different in-context selection methods:

- `Vanilla`: the *vanilla* method performs ICL using randomly selected in-context examples from the training data without any fairness consideration.
- **Balanced**: group-balanced in-context examples selection where context examples are randomly drawn with equal demographic group ratio. The majority group is uniformly downsampled across k-fold evaluations and independent runs.
- `Correlation Remover`: in-context transformation with correlation remover (Feldman et al., 2015) where each non-sensitive feature is transformed to reduce correlation with the sensitive ones. We use the fairlearn toolkit (Bird et al., 2020) implementation and fixed the parameter α controlling the fairness-accuracy tradeoff to one, meaning maximal fairness.
- `Uncertain`: using the uncertainty of the sensitive attribute prediction to select in-context examples. We use the Mapie (Taquet et al., 2022) implementation of conformal prediction (Cordier et al., 2023) to estimate the uncertainty of the training data. We fix the coverage parameter ϵ, of conformal prediction to 0.05 and only select as in-context example samples with prediction sets equal to two, i.e., samples with uncertain sensitive attributes. We consider two variants of the method under different model classes used to train the sensitive attribute classifier for uncertainty estimation: a variant that uses the traditional Logistic Regression model (`Uncertain+LR`) and a variant that uses a foundation model (`Uncertain+TabPFN`).

4.2 Results and Discussion

We evaluate several aspects of fairness in ICL prediction with foundational tabular datasets. First, we compare the different baselines considered in terms of fairness and accuracy; For the methods with controllable tradeoff between fairness and accuracy, we vary the hyperparameter controlling the fairness-accuracy tradeoff and compare their Pareto fronts. We then compare the foundation model under the best-performing fairness intervention method. Finally, we provide an ablation study on the impact of the size of the in-context set on fairness and accuracy and discuss the failure case of the correlation removal method.

4.2.1 Baseline Comparison

Figure 1 summarizes the accuracy and fairness outcomes of various baselines across eight datasets, using TabPFN as the foundation model. The group-balanced method yields only marginal improvements in fairness compared to the `Vanilla` ICL approach, suggesting that representation bias alone does not account for the observed disparities. In other words, even when the in-context sample set is group-balanced, performance disparities across subgroups persist—likely due to other sources of bias such as historical or measurement bias (Mehrabi et al., 2022).

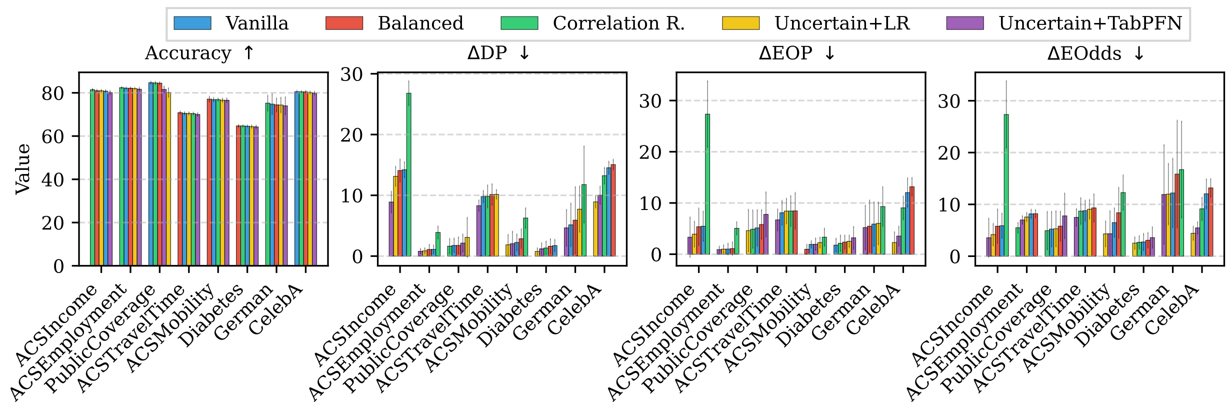


Figure 1: **Fairness-accuracy performance of different ICL methods**. ↑ indicates higher is better (accuracy) and ↓ lower is better (unfairness). To ease the interpretation of the plot, we sorted the bar plots from best performing to worst, i.e., for fairness plots, the rightmost bar represents the best-performing method. Uncertainty-based instance selection tends to provide better fairness while preserving accuracy. Table 5 in the Appendix supplements the figure with the actual values.

More strikingly, the `Correlation Remover` method often exacerbates unfairness. For instance, on the ACSIncome dataset, demographic parity disparity increases from 0.14% to 0.26% following the application of `Correlation Remover`. To investigate this failure mode, we assess whether the foundation model can reconstruct the sensitive attribute after `Correlation Remover`'s preprocessing step. Specifically, we perform ICL where the sensitive attribute is treated as the prediction target, and measure the model's test accuracy in reconstructing it.

We find that TabPFN and TabICL can reconstruct the sensitive attribute with up to 100% accuracy (see Table 4 in Appendix), even after `Correlation Remover`'s intervention. This suggests that these transformer-based models, pretrained on synthetic data with diverse structural priors, are capable of uncovering and leveraging the transformation patterns introduced by `Correlation Remover`. Since `Correlation Remover` modifies each non-sensitive feature based on the sensitive attribute, it unintentionally leaks sensitive information. The foundation model learns to exploit this leakage, ultimately relying on the hidden sensitive signal to predict the target variable—thereby amplifying unfairness.

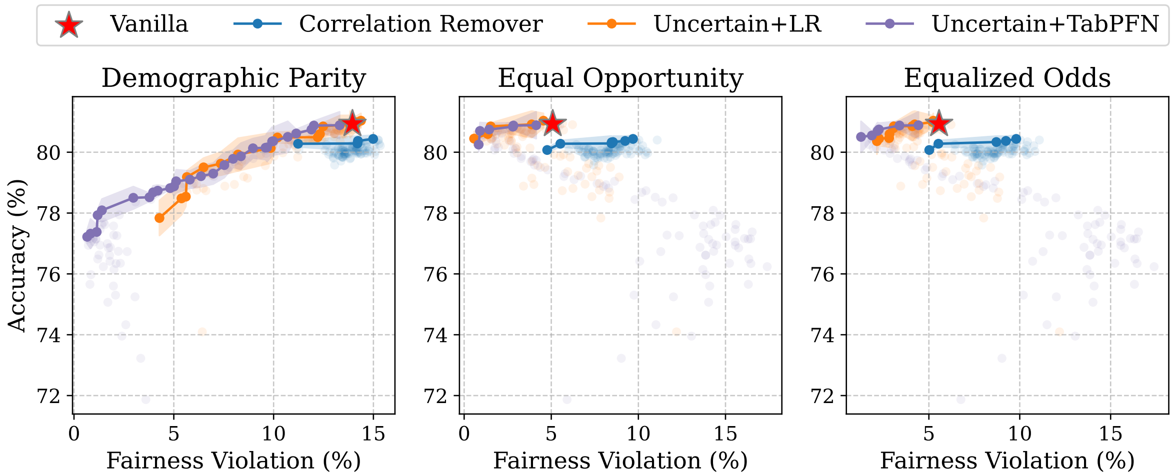
Further results and discussion are presented in Section 4.5, where we show that applying the feature transformation only to the training set—while keeping the test set unchanged—leads to improved fairness outcomes compared to the `Vanilla` baseline.

On the other hand, the results also demonstrate that `Uncertain` methods can significantly improve the fairness of the ICL prediction compared to the different baselines. Table 4 in the Appendix shows that `Uncertain` methods yield the smallest accuracy of ICL predictions of the sensitive attribute across datasets. This indicates that the selected in-context samples do not encode sufficient information about the sensitive attribute that the foundation model can rely upon for inference, thereby reducing unfairness.

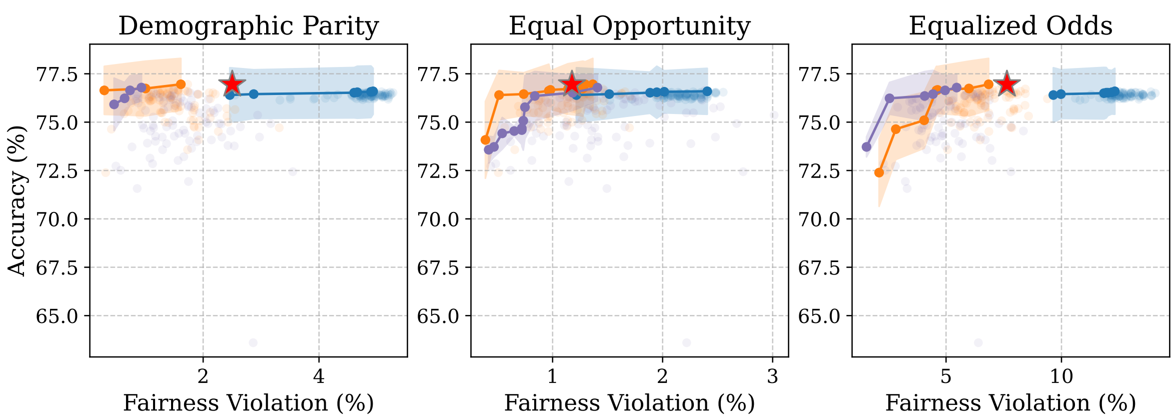
4.2.2 Fairness-Accuracy Tradeoff

In the previous experiment, we fixed the parameters α and ϵ, which control the tradeoff between fairness and accuracy for the `Correlation Remover` and `Uncertain` methods, respectively. To better evaluate this tradeoff, we now vary α and ϵ over the interval $[0, 1]$ and compare the resulting Pareto fronts of each method.

Using TabPFN as the foundation model, Figures 2a and 2b display the Pareto fronts on the ACSIncome and ACSMobility datasets, respectively. These figures compare the `Vanilla` ICL method, `Correlation Remover`, and the two variants of our uncertainty-based method: `Uncertain+LR` and `Uncertain+TabPFN`. Each scatter point corresponds to a different value of α or ϵ for a given method.



(a) ACSIncome



(b) ACSMobility

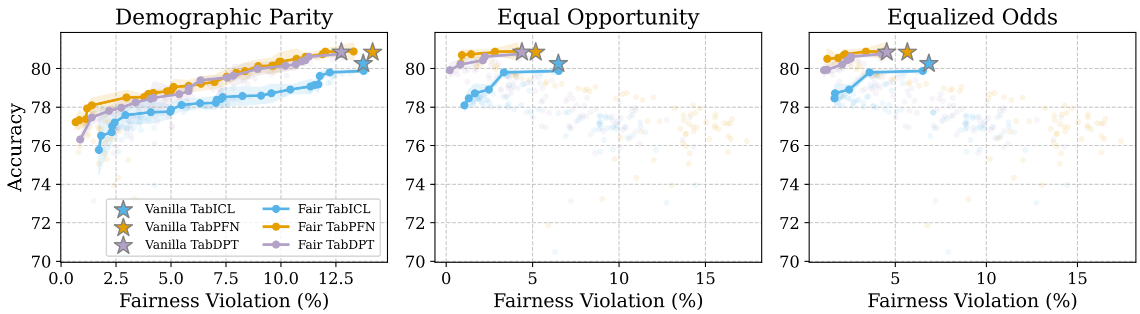
Figure 2: **Fairness-accuracy tradeoffs**. Comparing the fairness-accuracy Pareto-front of different fairness interventions using TabPFN on the ACSIncome and ACSMobility datasets.

Both variants of the `Uncertain` method consistently achieve Pareto-dominant points compared to alternatives. In contrast, `Correlation Remover` can underperform even the `Vanilla` ICL in terms of fairness, suggesting that it may amplify bias when used with foundation models. Notably, our uncertainty-based method provides a more stable control over the fairness–accuracy tradeoff across ϵ values, enabling improved fairness with relatively modest reductions in accuracy.

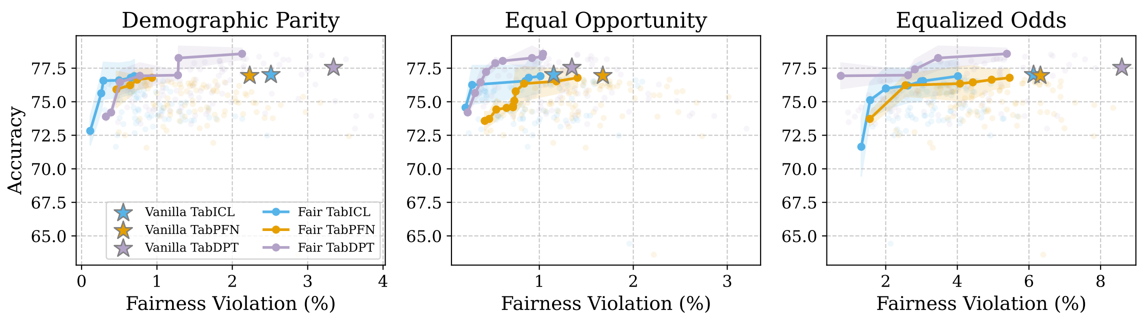
This accuracy drop is primarily attributed to the smaller in-context set size: as ϵ decreases, more samples with high sensitive attribute certainty are excluded from the context, effectively reducing the number of training examples. As discussed further in Section 4.3, the size of the in-context set has a strong influence on model accuracy.

Among the two `Uncertain` variants, `Uncertain+TabPFN`—which uses TabPFN for uncertainty estimation—tends to yield better Pareto fronts. This suggests that TabPFN functions as a stronger conformal predictor than logistic regression, allowing for more accurate estimation of sensitive attribute uncertainty and thus more effective in-context selection for fairness.

We observe similar trends across other datasets. Additional results are provided in the Appendix: Figure 7 for other datasets using TabPFN; Figures 9 and 11 for TabICL; and Figures 10 and 12 for TabDPT.



(a) ACSIncome



(b) ACSMobility

Figure 3: Comparing the fairness-accuracy tradeoffs of tabular foundation models (TabICL, TabDPT, and TabPFN) under uncertainty-based in-context sample selection (`Uncertain+TabPFN`) for different coverage (ϵ controlling the tradeoff). Results with other datasets can be found in the Appendix (Figure 6).

4.2.3 Comparison of Foundation Models: TabICL vs. TabDPT vs. TabPFN

In the previous experiment, we compared the `Correlation Remover` and `Uncertain` methods in terms of their fairness–accuracy tradeoffs by varying their respective control parameters. The results showed that the `Uncertain` methods often achieved better Pareto-dominant points compared to both `Vanilla` ICL and `Correlation Remover`. Notably, using TabPFN for uncertainty estimation (`Uncertain+TabPFN`) outperformed the variant that relies on logistic regression (`Uncertain+LR`).

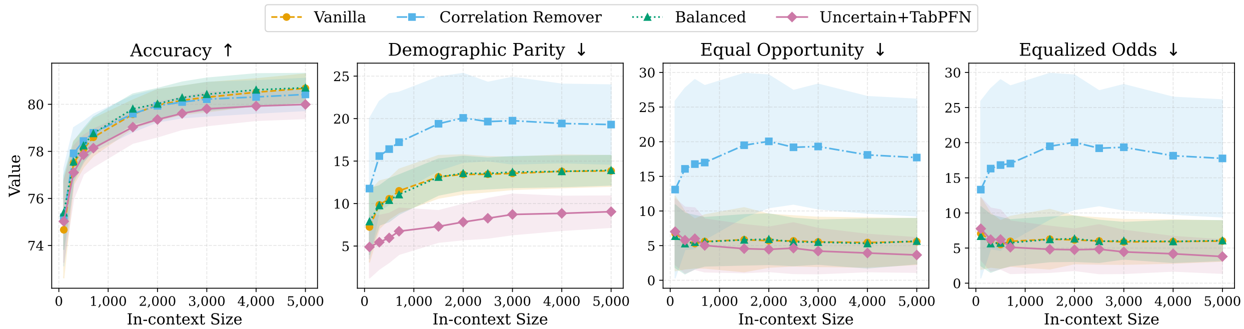
In this experiment, we extend the comparison to different foundation models by applying the same setup using our best-performing fairness intervention: `Uncertain+TabPFN`. Figures 3a and 3b show the fairness–accuracy Pareto fronts on the ACSIncome and ACSMobility datasets, respectively.

Without any fairness intervention (i.e., using the vanilla ICL approach), there is no clear winner among the foundation models in terms of fairness; the results vary across datasets. While one might expect TabDPT—pretrained on real-world data—to encode more real-world biases and thus exhibit poorer fairness performance, this is not consistently observed. In fact, TabDPT often performs comparably to or even better than models pretrained on synthetic data.

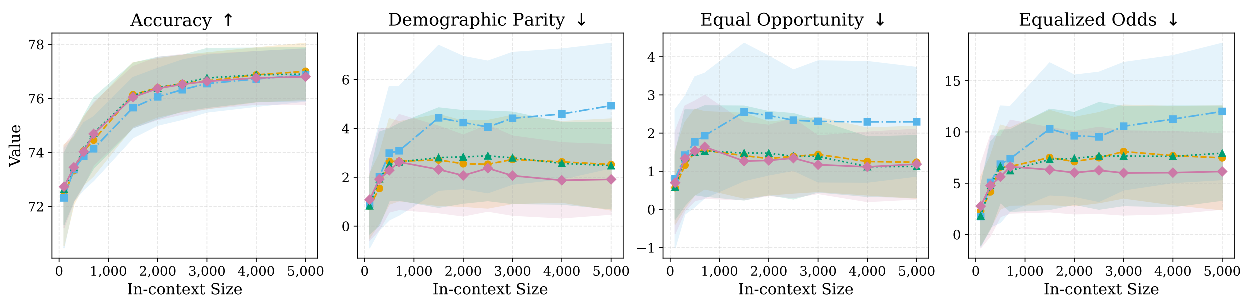
Under fairness intervention with `Uncertain+TabPFN`, all three foundation models exhibit similar fairness peformance, demonstrating that the `Uncertain` method can consistently control the fairness–accuracy tradeoff regardless of the underlying foundation model. However, both TabPFN and TabDPT consistently achieve higher accuracy across datasets. These findings are consistent with previous studies: the authors of TabICL reported strong performance for TabPFN (Qu et al., 2025), while the authors of TabDPT highlighted its competitive performance relative to TabPFN.

We observe similar patterns across additional datasets and results are provided in the Appendix (see Figure 6).

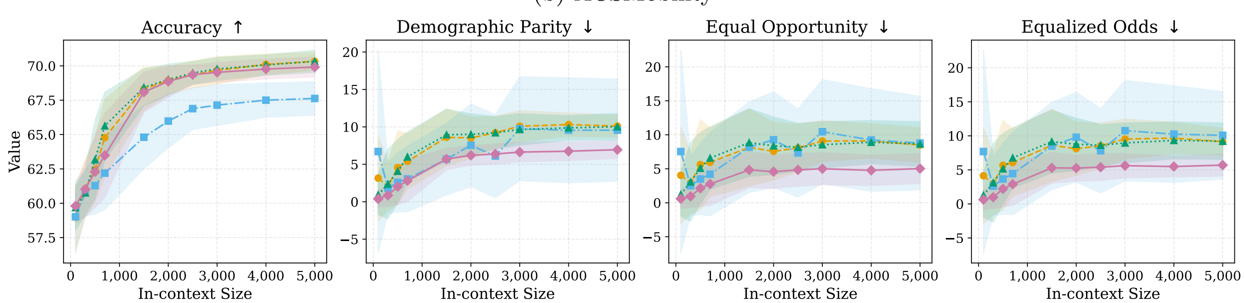
4.3 Ablation on Impact In-context Sample Size



(a) ACSIncome



(b) ACSMobility



(c) ACSTravelTime

Figure 4: **Ablation on the in-context example set size**. Analyzing the impact of the in-context set size on the fairness and accuracy of ICL prediction with TabPFN.

In all previous experiments, we used the full training set as the in-context example set whenever possible. For example, the current version of TabPFN is limited to handling a maximum of 10,000 samples (Hollmann et al., 2025). To understand the impact of in-context set size, we conduct an ablation study by varying the number of in-context examples across the range [100,300,500,700,1500,2000,2500,3000,4000,5000], while keeping the evaluation setup identical to that described in Section 4.2.1.

Figure 4 shows that accuracy increases substantially with larger in-context sets, particularly in the lower range. However, unfairness shows a slight increase initially and then stabilizes once the in-context set exceeds approximately 700 examples. Across all in-context sizes and datasets, the `Correlation Remover` method consistently exhibits the highest level of unfairness, reinforcing earlier observations of its bias amplification. In contrast, the `Uncertain+TabPFN` method consistently achieves the lowest fairness violation, demonstrating robustness to the size of the in-context set. This indicates that even when only a subset of training data can be used, `Uncertain+TabPFN` remains effective at mitigating bias without overly compromising accuracy.

4.4 Comparison with LLM-based ICL Methods

Despite LLMs' in-context learning capabilities, they often lag behind classical machine learning models on large-scale tabular datasets (Hegselmann et al., 2023). To validate the effectiveness of our uncertainty-based demonstration selection method, we compare it against fairness-aware LLM-based fair ICL approaches proposed by Hu et al. (2024) and Bhaila et al. (2024) on the widely studied ADULT dataset (Asuncion et al., 2007). Table 1 presents the results. Vanilla TabPFN achieves an accuracy of 85.72%, significantly outperforming the LLaMA-2-13B baseline, which reaches only 76.00%. This 10-point gap underscores the performance advantage of specialized models for tabular data, consistent with prior findings that LLMs struggle on such inputs when used naively (Hegselmann et al., 2023).

While these comparisons rely on LLMs that are not the most recent generation, and performance may have improved with newer models, the results nonetheless highlight the continued strength of tabular foundation models like TabPFN in this setting. Moreover, when fairness interventions are applied, our uncertainty-based selection strategy consistently preserves high predictive performance while achieving comparable or better fairness outcomes than the LLM-based fair ICL approaches—demonstrating that principled sample selection can improve fairness without sacrificing accuracy.

Method	Foundation model	**Accuracy**	$\mathbf{\Delta}_{\text{DP}}$	$\mathbf{\Delta}_{\text{EOP}}$
`Vanilla`	TabPFN	$\mathbf{85.72}_{\pm\mathbf{0.3}}$	$16.99_{\pm 1.0}$	$\mathbf{8.80}_{\pm\mathbf{5.1}}$
`Vanilla`	LLaMA-2-13B	$76.00_{\pm 1.1}$	$\mathbf{14.00}_{\pm\mathbf{0.4}}$	$11.00_{\pm 0.8}$
Hu et al. (2024)	GPT-3.5-turbo	77.93	6.64	10.9
Bhaila et al. (2024)	LLaMA-2-13B	$75.72_{\pm 1.6}$	$8.00_{\pm 2.0}$	$\mathbf{3.00}_{\pm\mathbf{3.0}}$
`Uncertain+TabPFN`	TabPFN	$\mathbf{82.05}_{\pm\mathbf{1.0}}$	$\mathbf{5.51}_{\pm\mathbf{1.4}}$	$3.86_{\pm 1.4}$

Table 1: Comparison of accuracy and fairness metrics (Δ_{DP}, Δ_{EOP}) on the ADULT dataset.

4.5 On the Failure of Correlation Remover

In our previous experiments, we observed that applying the correlation removal (CR) to both the training and test data can exacerbate unfairness in ICL predictions. We hypothesized that the foundation model infers the sensitive attribute from the linear transformation applied to each non-sensitive feature, which inadvertently leaks sensitive information and increases unfairness. To validate this hypothesis, we conducted correlation removal prediction of the sensitive attribute after applying fairness interventions. As shown in Table 4, the ICL prediction of the sensitive attribute achieves 100% accuracy following the application of CR. This indicates that the foundation model continues to rely heavily on the sensitive attribute even after correlation removal is performed.

For further verification, we tested a variant of correlation removal where the feature transformation (see Eq. 10 in Appendix) is applied only to the training data, leaving the test data unchanged (referred to as variant S2). Table 2 shows that this variant significantly reduces the ICL prediction accuracy of the sensitive attribute. This demonstrates that the foundation model exploits the transformation applied to the test set in the original correlation removal method (variant S1) as a proxy to fully reconstruct sensitive attributes.

This observation aligns with prior discussions by Aïvodji et al. (2021) regarding scenarios where data transformations reduce correlation with sensitive attributes. Besides the classical correlation removal setting (S1), where transformations are applied to both training and test sets, we evaluated the fairness–accuracy trade-off when applying the transformation only to the training set (S2).

As illustrated in Figure 5, variant S2 of CR substantially improves fairness compared to S1. This confirms the foundation models' capacity to reconstruct sensitive information when the test set is transformed, leading to bias amplification. And the drop in reconstruction accuracy of the sensitive attribute directly supports the observed improvement in fairness when using variant S2

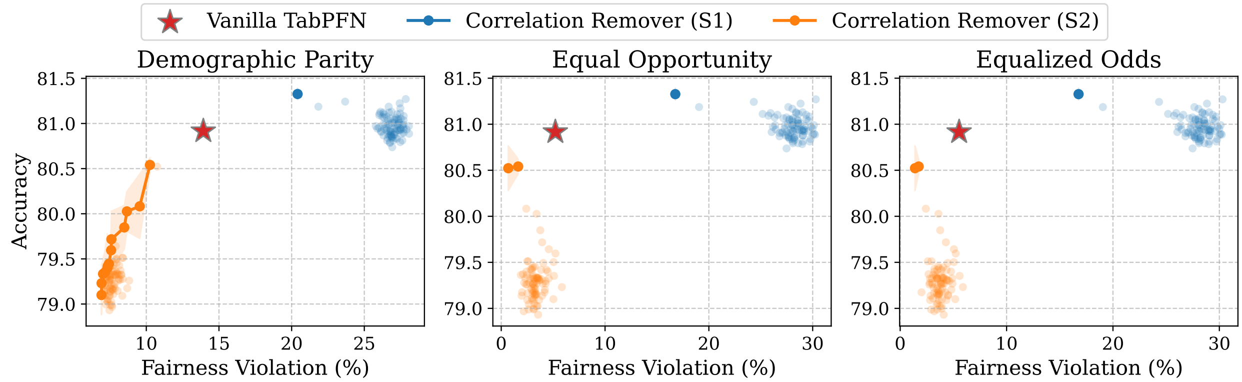
Dataset	Fairness Intervention	TabPFN		TabICL	
		Accuracy ↓	F1 Score ↓	Accuracy ↓	F1 Score ↓
ACSIncome	None	$77.2_{\pm 0.5}$	$78.4_{\pm 0.3}$	$75.0_{\pm 0.5}$	$76.18_{\pm 0.3}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.93_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{53.8}_{\pm 0.4}$	$\mathbf{67.0}_{\pm 0.9}$	$\mathbf{52.9}_{\pm 0.3}$	$\mathbf{68.9}_{\pm 0.3}$
ACSTravelTime	None	$75.9_{\pm 0.4}$	$77.5_{\pm 0.5}$	$72.8_{\pm 0.4}$	$73.75_{\pm 0.5}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{53.6}_{\pm 1.4}$	$\mathbf{54.2}_{\pm 14.5}$	$\mathbf{51.8}_{\pm 1.4}$	$\mathbf{66.4}_{\pm 3.4}$
ACSPublicCoverage	None	$91.4_{\pm 0.2}$	$88.9_{\pm 0.2}$	$91.5_{\pm 0.1}$	$89.23_{\pm 0.2}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{57.8}_{\pm 0.4}$	$\mathbf{0.1}_{\pm 0.1}$	$\mathbf{56.5}_{\pm 1.0}$	$\mathbf{0.1}_{\pm 0.1}$
ACSEmployment	None	$64.0_{\pm 0.4}$	$62.0_{\pm 1.8}$	$65.0_{\pm 0.3}$	$62.23_{\pm 1.4}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{52.6}_{\pm 0.6}$	$\mathbf{27.7}_{\pm 4.4}$	$\mathbf{53.7}_{\pm 5.1}$	$\mathbf{53.0}_{\pm 10.6}$
ACSMobility	None	$68.3_{\pm 0.8}$	$67.8_{\pm 1.0}$	$67.6_{\pm 1.0}$	$67.40_{\pm 1.2}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{49.2}_{\pm 1.5}$	$\mathbf{49.2}_{\pm 13.2}$	$\mathbf{49.2}_{\pm 0.8}$	$\mathbf{40.2}_{\pm 31.2}$
Diabetes	None	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$	$80.4_{\pm 0.1}$	$89.04_{\pm 0.1}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{67.8}_{\pm 13.8}$	$\mathbf{78.9}_{\pm 12.2}$	$\mathbf{79.3}_{\pm 0.9}$	$\mathbf{88.4}_{\pm 0.6}$
German Credit	None	$72.5_{\pm 3.1}$	$70.3_{\pm 3.1}$	$71.8_{\pm 3.1}$	$71.02_{\pm 3.0}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{37.8}_{\pm 6.7}$	$\mathbf{44.0}_{\pm 14.5}$	$\mathbf{46.3}_{\pm 4.2}$	$\mathbf{38.8}_{\pm 14.6}$
CelebA	None	$84.7_{\pm 0.2}$	$83.2_{\pm 0.3}$	$85.0_{\pm 0.2}$	$83.16_{\pm 0.3}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{54.8}_{\pm 0.3}$	$\mathbf{1.0}_{\pm 0.8}$	$\mathbf{45.2}_{\pm 0.2}$	$\mathbf{62.2}_{\pm 0.2}$

Table 2: Accuracy ICL prediction of sensitive attribute after applying `Correlation Remover` on the training and testing datasets (S1) or only to train dataset (S2). Applying the transformation only to the train dataset significantly reduces the accuracy of predicting the sensitive attribute.

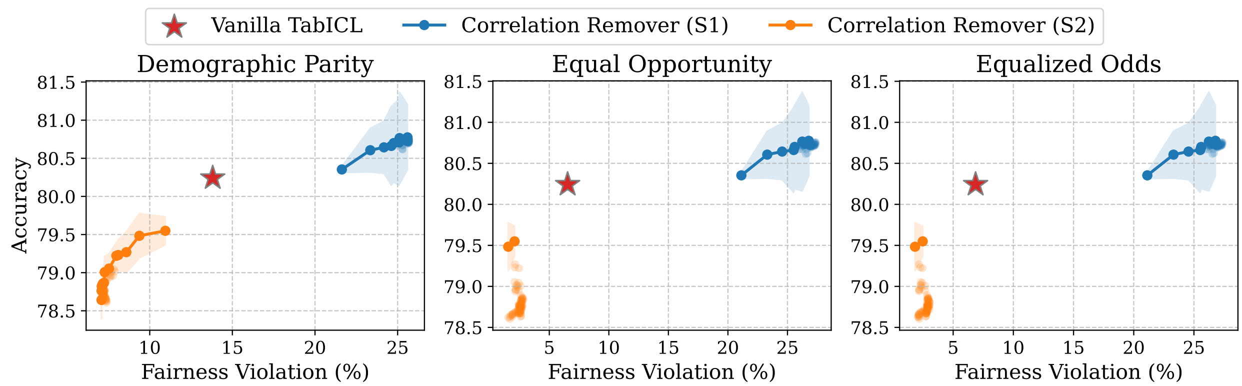
Based on these findings, we recommend practitioners apply the S2 variant of Correlation Remover—transforming only the training data—when using correlation removal in in-context learning frameworks, to avoid sensitive attribute leakage and mitigate bias amplification during testing.

5 Conclusion and future works

In this study, we proposed and analyzed the effectiveness of three preprocessing methods to enhance the fairness of in-context learning (ICL) predictions. Our empirical results, performed on eight fairness benchmarks, posit the uncertainty-based in-context selection method as a strong baseline for improving the fairness of tabular ICL. The key advantages of this method are threefold: (1) it does not require fine-tuning or retraining the foundation model to enforce the desired fairness metrics; (2) it can consistently improve three widely used group fairness metrics; (3) it offers a parameter to control the fairness-accuracy tradeoff. To our knowledge, this is the first work that explores pre-processing fairness intervention on tabular foundation models. We hope this work will trigger more investigations into fair tabular ICL, since in-context learning as a new learning paradigm will be increasingly adopted into decision-making tools. Interesting future research directions include investigating in-processing and post-processing methods and analyzing the effect of distribution shift between in-context and test examples on fairness and accuracy.



(a) TabPFN



(b) TabICL

Figure 5: **Evaluating variants of the correlation remover on the ACSIncome dataset with TabPFN and TabICL**. Applying correlation remover to the training and testing data exacerbates unfairness, while applying the transformation only to the training set improves fairness.

Ethics Statement

This paper explores ways to reduce unfairness in tabular foundation models, emphasizing fair treatment for various groups. We recognize the significance of fairness in machine learning, especially regarding sensitive attributes like race, gender, and socio-economic status. Our research seeks to uncover and tackle potential biases in these models, thereby enhancing transparency, accountability, and inclusivity. While the proposed method uses a sensitive attributes predictor, which could be unlawful in some countries, we emphasized that predicted sensitive values are not used either for training or measuring unfairness. We use the attribute classifier only to quantify uncertainty, and emphasize that this method should not be used for any purpose other than bias measuring or mitigation.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification, July 2018. URL http://arxiv.org/abs/1803.02453. arXiv:1803.02453 [cs].

Ulrich Aïvodji, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu, and Alain Tapp. Local data debiasing for fairness based on generative adversarial training. *Algorithms*, 14(3):87, 2021.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair normalizing flows. *arXiv preprint arXiv:2106.05937*, 2021.

Karuna Bhaila, Minh-Hao Van, Kennedy Edemacu, Chen Zhao, Feng Chen, and Xintao Wu. Fair in-context learning via latent concept variables. *arXiv preprint arXiv:2411.02671*, 2024.

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

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

L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pp. 1349–1359. PMLR, 2020.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library. In *Conformal and Probabilistic Prediction with Applications*, 2023.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly Fair Representation Learning by Disentanglement, June 2019. URL http://arxiv.org/abs/1906.02589. Number: arXiv:1906.02589 arXiv:1906.02589 [cs, stat].

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning, January 2022. URL http://arxiv.org/abs/2108.04884. arXiv:2108.04884 [cs, stat].

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, pp. 214–226, Cambridge, Massachusetts, 2012. ACM Press. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL http://dl.acm.org/citation.cfm?doid=2090236.2090255.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Andrew Frank. Uci machine learning repository. *http://archive. ics. uci. edu/ml*, 2010.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 2023.

Anurag Garg, Muhammad Ali, Noah Hollmann, Lennart Purucker, Samuel Müller, and Frank Hutter. Real-tabpfn: Improving tabular foundation models via continued pre-training with real-world data. *arXiv preprint arXiv:2507.03971*, 2025.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning, October 2016. URL http://arxiv.org/abs/1610.02413. Number: arXiv:1610.02413 arXiv:1610.02413 [cs].

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Jingyu Hu, Weiru Liu, and Mengnan Du. Strategic demonstration selection for improved fairness in llm in-context learning. *arXiv preprint arXiv:2408.09757*, 2024.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-011-0463-8. URL http://link.springer.com/10.1007/s10115-011-0463-8.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=3HE4vPNIfX.

Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. Fairness under demographic scarce regime. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=TB18G0w6Ld.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *arXiv preprint arXiv:2311.00288*, 2023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155, 2023.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L Caterini. Tabdpt: Scaling tabular foundation models. *arXiv preprint arXiv:2410.18164*, 2024.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations, October 2018. URL http://arxiv.org/abs/1802.06309. Number: arXiv:1802.06309 arXiv:1802.06309 [cs, stat].

Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. Which examples to annotate for in-context learning? towards effective and efficient selection. *arXiv preprint arXiv:2310.20046*, 2023.

Dave Mbiazi, Meghana Bhange, Maryam Babaei, Ivaxi Sheth, and Patrik Joslin Kenfack. Survey on ai ethics: A socio-technical perspective. *arXiv preprint arXiv:2311.17228*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning, January 2022. URL http://arxiv.org/abs/1908.09635. Number: arXiv:1908.09635 arXiv:1908.09635 [cs].

Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jake Robertson, Noah Hollmann, Noor Awad, and Frank Hutter. Fairpfn: Transformers can do counterfactual fairness. *arXiv preprint arXiv:2407.05732*, 2024.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. Mapie: an open-source library for distribution-free uncertainty quantification. *arXiv preprint arXiv:2207.12274*, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Appendix

A Datasets

We experiment on tasks from the recently proposed folktables (Ding et al., 2022), which contains data extracted from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2022). More specifically, we experiment with the following ACS PUMS tasks:

- **ACSIncome**: The task involves predicting whether an individual's income exceeds $50,000. The dataset is filtered to include only individuals over the age of 16 who reported working at least 1 hour per week during the past year and earned a minimum of $100.
- **ACSMobility**: This task involves predicting whether an individual had the same residential address one year ago. The dataset is filtered to include individuals aged between 18 and 35. This filtering increases the difficulty of the task, as more than 90% of the general population tends to stay at the same address year-to-year.
- **ACSTravelTime**: This task predicts whether an individual has a commute longer than 20 minutes. The dataset is filtered to include only employed individuals above the age of 16. The 20-minute threshold corresponds to the median commute time in the US, according to the 2018 ACS PUMS data.
- **ACSEmployment**: The objective is to predict whether an individual is employed, using a dataset filtered to include individuals aged between 16 and 90.
- **ACSPublicCoverage**: The goal is to predict whether an individual has public health insurance. The dataset is filtered to include individuals under 65 years of age and those with an income below $30,000, focusing on low-income individuals who are ineligible for Medicare.

These tasks were selected to reflect a range of real-world predictive challenges with fairness concerns. We use the data of the year 2018 from the state of Alabama (AL), which is one of the states with the largest fairness violation (Ding et al., 2022)[2].

A limitation of the ACS PUMS datasets is that they are US-centric; we diversify the experimental setup by including other tasks and datasets. Specifically, we also experiment on the following tabular datasets and tasks:

- **Diabetes** (Gardner et al., 2023): The diabetes prediction task uses features related to physical health, lifestyle factors, and chronic conditions, derived from the BRFSS questionnaires. Demographic attributes like race, sex, state, and income are also included. The target is a binary indicator of whether the respondent has ever been diagnosed with diabetes.
- **German Credit** (Frank, 2010): The German Credit dataset contains 20 attributes of 1,000 individuals. We create the task of classifying people according to whether they have a good or bad credit risk using age (over or below 25 years old) as the sensitive attribute.
- **CelebA** (Liu et al., 2018): The dataset contains 202,599 samples described with 40 facial attributes of human annotated images. We create the task of predicting *attractiveness* with facial attributes using gender as the sensitive attribute (Kenfack et al., 2024b). Note that we do not train the model with images and consider this task to diversify the experimental tasks.

[2]We also perform experiments on data from other states, and observed that the results presented in the paper remain consistent.

Table 3: Summary of datasets used in our experiments. For each dataset, we report the number of features (including the sensitive attribute), the number of samples available, and the sensitive attribute used for fairness evaluation.

Dataset	**# Features**	**# Samples**	**Sensitive Feature**	**Prediction Task**
ACSIncome	10	22,268	Gender	Income $\geq$ $50,000
ACSEmployment	16	47,777	Gender	Employment status
ACSTravelTime	16	19,492	Gender	Commute time over 20 minutes
ACSMobility	21	8,625	Gender	Residential mobility
ACSPublicCoverage	19	18,525	Gender	Public health insurance coverage
CelebA	39	202,599	Gender	Attractiveness
Diabetes	183	38,575	Race	Prior diabetes diagnose
German	58	990	Age	Credit risk

B Background

B.1 Fairness Metrics

In this work, we focus on group fairness notions that measure the performance disparity across different demographic groups. More specifically, we consider the following three widely used group fairness metrics:

- **Demographic parity (DP)**: DP enforces equal positive outcome rate for different groups (Dwork et al., 2012) and is defined as follows:

$$P(f(X) = 1|S = s) = P(f(X) = 1) \tag{1}$$

- **Equalized Odds (EOD)**: EOdds is satisfied when the model makes correct and incorrect predictions at the same rate for different demographic groups (Hardt et al., 2016). The metric enforces equal true positive and false positive rates across groups and is measured as follows;

$$P(f(X) = 1|S = 0, Y = y) = P(f(X) = 1|S = 1, Y = y), \ \forall y \in \{0, 1\} \tag{2}$$

- **Equalized Opportunity (EOP)**: In some settings, one can care more about assessing unfairness when the model makes correct predictions. EOP enforces equal true positive rates across groups, i.e., we only consider $y = 1$ in Eq. 2, i.e.,

$$P(f(X) = 1|S = 0, Y = 1) = P(f(X) = 1|S = 1, Y = 1) \tag{3}$$

Empirically, we measure each fairness considered, i.e., Demographic Parity (ΔDP), Equal Opportunity (ΔEOP), and Equalized Odds (ΔEOD) as follows.

$$\Delta\text{DP} = \left| \mathop{\mathbb{E}}_{x|A=0}[\mathbb{I}\{f(x) = 1\}] - \mathop{\mathbb{E}}_{x|A=1}[\mathbb{I}\{f(x) = 1\}] \right| \tag{4}$$

Where $\mathbb{I}(\cdot)$ is the indicator function.

$$\Delta\text{EOD} = \alpha_0 + \alpha_1 \tag{5}$$

$$\Delta\text{EOP} = \alpha_1 \tag{6}$$

Where α_0 and α_1 measure the difference between the false positive and the true positive rates across groups, respectively, and are empirically measured as follows.

Where α_0 and α_1 measure the difference between the false positive and the true positive rates across groups, respectively, and are empirically measured as follows.

$$\alpha_j = \left| \underset{x|A=0,Y=j}{\mathbb{E}}[\mathbb{I}\{f(x)=1\}] - \underset{x|A=1,Y=j}{\mathbb{E}}[\mathbb{I}\{f(x)=1\}] \right| \quad j \in \{0,1\} \tag{7}$$

Since the disparities can be below 0.1 on some datasets, we scaled the fairness values reported throughout the paper by 100 to make it easier to read.

B.2 Correlation Remover

The `Correlation Remover` (Feldman et al., 2015) is a preprocessing technique designed to eliminate linear correlations between sensitive attributes and non-sensitive features in a dataset. This method is particularly useful in mitigating biases that may arise due to such correlations, especially when employing linear models.

Considering a classification task with the given training data $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$ where x_i is an input feature vector, y_i is the corresponding class label, and s_i the corresponding demographic group.

To apply `Correlation Remover`, we assume the training data is formulated as follows:

- $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the training data matrix containing sensitive and non-sensitive features.
- $\mathbf{S} \in \mathbb{R}^{n \times m_s}$ a matrix of the sensitive features. For simplicity, we assumed in this work $m_s = 1$, which corresponds to a single binary sensitive attribute.
- $\mathbf{Z} \in \mathbb{R}^{n \times m_z}$ a matrix of non-senstive features such that $X = [S\ Z]$

The goal of `Correlation Remover` is to transform Z into Z^* such that Z^* is uncorrelated with S, while retaining as much information from the original Z as possible.

For each non-sensitive feature vector $\mathbf{z}^j \in \mathbb{R}^n$ (the j-th column of $\mathbf{Z}$), the algorithm solves the following least squares problem:

$$\min_{\mathbf{w}_j} \left\| \mathbf{z}^j - (\mathbf{S} - \mathbf{1}_n \bar{\mathbf{s}}^\top)\mathbf{w}_j \right\|_2^2 \tag{8}$$

where:

- $\bar{\mathbf{s}} = [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \ldots, \bar{\mathbf{s}}_{m_s}]$ is the mean vector of the sensitive features, i.e., $\bar{\mathbf{s}}_j$ is the mean of j-th the sensitive feature.
- $\mathbf{1}_n$ is an n-dimensional column vector of ones.
- $\mathbf{w}_j \in \mathbb{R}^{m_z}$ is the weight vector that projects the centered sensitive features onto $\mathbf{z}_j$.

After computing the optimal weight vectors $\mathbf{w}_j^*$ for all $j \in \{1, \ldots, m_z\}$, they are assembled into a weight matrix $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_{m_z}^*]$. The transformed non-sensitive features are then obtained by:

$$\mathbf{Z}^* = \mathbf{Z} - (\mathbf{S} - \mathbf{1}_n \bar{\mathbf{s}}^\top)\mathbf{W}^* \tag{9}$$

This operation effectively removes the linear correlations between $\mathbf{S}$ and $\mathbf{Z}$, resulting in $\mathbf{Z}^*$ that is uncorrelated with the sensitive features.

`Correlation Remover` introduces a tunable parameter $\alpha \in [0, 1]$ that controls the extent of correlation removal, i.e., (i) $\alpha = 1$ corresponds to full removal of linear correlations, thus best possible fairness; (ii) $\alpha = 0$ corresponds no transformation, the original data is used; (iii) $0 < \alpha < 1$ corresponds to partial removal, balancing between the original and transformed data, thus controlling the fairness accuracy tradeoff. More specifically, the final transformed dataset $\mathbf{X'}$ is computed as:

$$\mathbf{X'} = \alpha \mathbf{Z}^* + (1 - \alpha)\mathbf{Z} \tag{10}$$

Note that $\mathbf{X'}$ is derived using $\mathbf{Z}^*$, since $\mathbf{S}$ is dropped after transformation. The convex combination 10 allows practitioners to adjust the fairness accuracy tradeoff based on specific requirements of their application.

Equation 8 is optimized on the training dataset, and the optimal weight vectors w_j^* are used to apply the transformation 10 on the test dataset.

B.3 Uncertainty measure with conformal prediction

While any uncertainty measurement method could be used in our method, we employ *conformal prediction* due to its strong coverage guarantees. Specifically, we used Split Conformal Prediction (Angelopoulos et al., 2023), which is a distribution-free and model-agnostic method that provides **prediction sets** for classification tasks, ensuring a user-specified **coverage level** $1 - \epsilon$ with no assumptions beyond data exchangeability.

Given labeled data $\mathcal{D}_{\text{train}} = \{(x_i, s_i)\}_{i=1}^n$ with binary sensitive attribute $s_i \in \{0, 1\}$, we split the data into *proper training* set ($\mathcal{D}_1$) and calibration set ($\mathcal{D}_2$). In the experiments, we did a 50%-50% split of the dataset with sensitive attribute to obtain $\mathcal{D}_1$ and $\mathcal{D}_2$.

We then train a probabilistic classifier $f : \mathcal{X} \to [0, 1]$, on $\mathcal{D}_1$, yielding predictions:

$$\hat{p}_i = f(x_i) = \mathbb{P}(S = 1 \mid X = x_i). \tag{11}$$

In our setup, f could be Logistic Regression (Uncertain+LR) or TabPFN (Uncertain+TabPFN). The starting point for conformal prediction is what is called a *nonconformity measure*, a real-valued function that measures how a prediction is different from any possible class label.

Nonconformity Scores After training f on the proper training set, we use the calibration dataset to compute the nonconformity scores, which measure how far the prediction is from the true label. More specifically, for each calibration point $(x_i, s_i) \in \mathcal{D}_2$, we considered the nonconformity score is defined as:

$$c_i = |s_i - \hat{p}_i|. \tag{12}$$

We then compute the quantile threshold τ based on the user-defined target coverage $\epsilon \in [0, 1]$, e.g., $\epsilon = 0.05$ mean 95% coverage. Specifically, τ is defined based on $1 - \epsilon$ quantile of the nonconformity scores.

$$\tau = \text{Quantile}_{1-\epsilon}\left(\{c_i\}_{i=1}^{n_{\text{cal}}}\right). \tag{13}$$

Prediction Set The quantile threshold is used to build the prediction set of data points from the test set. More specifically, for a test sample x_{test}, we compute the prediction probability $\hat{p}_{\text{test}} = f(x_{\text{test}})$ and derive its prediction set as follows:

$$\Gamma(x_{\text{test}}) = \{s \in \{0, 1\} : |s - \hat{p}_{\text{test}}| \leq \tau\}. \tag{14}$$

When the prediction set only contains $\{0\}$ or $\{1\}$, then the prediction is confident with at least $1-\epsilon$ probability, while when the prediction set contains both labels, i.e., $\{0, 1\}$, the prediction is considered uncertain. Our uncertainty-based demonstration selection method uses only samples whose prediction set contains two values. The coverage guarantee of conformal prediction holds under the assumption that the calibration and test data are *exchangeable*. We use the open-source implementation of split conformal prediction provided by the MAPIE Python package.

Example. Consider a simplified task consisting of two non-sensitive (f^1 and f^2) features one binary target (y) and a binary sensitive attribute (s).

We first train a sensitive attribute classifier, using a fraction of the dataset (20% in our experiment), to estimate the uncertainties. We train the sensitive attribute classifier using f^1 and f^2 to predict s.

A conformal predictor (details provided above) will return for a given (test) sample x a prediction set $\Gamma(x)$ that contains the true sensitive attribute with probability of at least $1 - \epsilon$. For example, setting $\epsilon = 0.05$ means 95% of the data will contain the true label in their prediction sets. The size of the prediction set, therefore, provides information about the prediction uncertainty, i.e., $|\Gamma(x)| = 1$ means the prediction is confident with at least $1 - \epsilon$ probability, and $|\Gamma(x)| = 2$ means the prediction is uncertain. The intuition of our method is that using samples with uncertain predictions makes it challenging for the foundation model to infer and rely on the sensitive attributes to make predictions on the target.

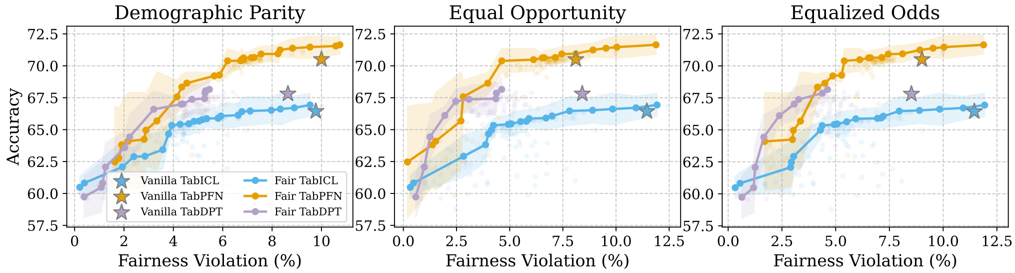
We therefore filter out a demonstration example x if $|\Gamma(x)| = 1$ and select it if $|\Gamma(x)| = 2$. ϵ helps to control the fairness-accuracy tradeoff since smaller ϵ corresponds to dropping more demonstration examples with higher confident sensitive attributes prediction.

C Supplementary resutls

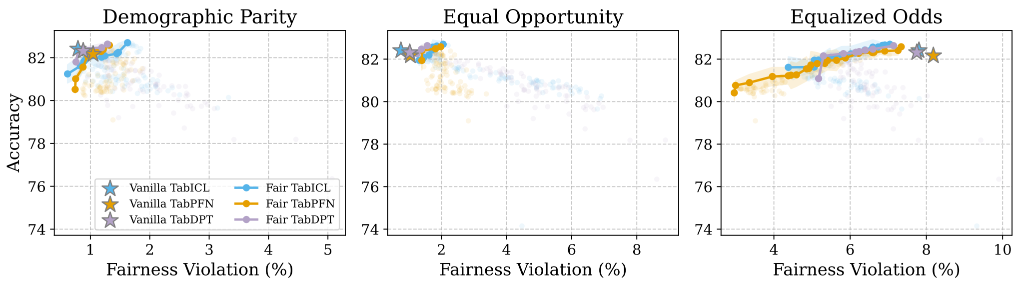
The results in Figure 1 and Table 5 show that improving fairness in the ACSPublicCoverage and Diabetes datasets is particularly challenging. We observe that unfairness is already pretty small, and the fairness methods alter the training data, either by transforming the features (correlation removal) or reducing its size (group-balanced and Uncertain methods), which reduces accuracy, without necessarily improving on fairness. For Table 1, we used a fixed value for the hyperparameter controlling the fairness accuracy tradeoff, $\epsilon = 0.05$ for the uncertain method and $\alpha = 1$ for correlation removal. Figure 7c and 7d show that for different values of ϵ, there are Pareto-optimality points of the uncertain method that significantly reduce unfairness compared to random demonstration selection baselines. These points correspond to values of $\epsilon < 0.05$. In sum, for some tasks, in a low unfairness regime such as PublicCoverage and Diabetes, smaller values of ϵ are needed to reduce unfairness, which means more data points with a confident prediction of the sensitive attribute must be removed from the demonstration set. The value ϵ used in practice can be defined by stakeholders, decision-makers or business needs, considering that smaller ϵ means increased fairness and a drop in accuracy, and higher ϵ means a decrease in fairness and higher accuracy.

Dataset	ICL Method	TabPFN		TabICL	
		Accuracy ↓	F1 Score ↓	Accuracy ↓	F1 Score ↓
ACSIncome	`Vanilla`	$77.2_{\pm 0.5}$	$78.4_{\pm 0.3}$	$75.0_{\pm 0.5}$	$76.2_{\pm 0.3}$
	`Balanced`	$77.1_{\pm 0.5}$	$77.8_{\pm 0.2}$	$75.0_{\pm 0.4}$	$75.5_{\pm 0.1}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.9_{\pm 0.0}$
	`Uncertain+LR`	$74.7_{\pm 1.3}$	$76.7_{\pm 0.6}$	$74.9_{\pm 0.5}$	$76.0_{\pm 0.4}$
	`Uncertain+TabPFN`	$\mathbf{51.3}_{\pm 2.3}$	$\mathbf{66.1}_{\pm 4.4}$	$\mathbf{71.7}_{\pm 0.4}$	$\mathbf{73.6}_{\pm 0.6}$
ACSTravelTime	`Vanilla`	$75.9_{\pm 0.4}$	$77.5_{\pm 0.5}$	$72.8_{\pm 0.4}$	$73.8_{\pm 0.5}$
	`Balanced`	$75.9_{\pm 0.5}$	$77.0_{\pm 0.5}$	$72.5_{\pm 0.6}$	$72.6_{\pm 0.6}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$75.9_{\pm 0.6}$	$77.8_{\pm 0.5}$	$72.8_{\pm 0.5}$	$74.3_{\pm 0.5}$
	`Uncertain+TabPFN`	$\mathbf{74.6}_{\pm 1.1}$	$\mathbf{75.7}_{\pm 1.7}$	$\mathbf{67.4}_{\pm 1.6}$	$\mathbf{66.2}_{\pm 2.4}$
ACSPublicCoverage	`Vanilla`	$91.4_{\pm 0.2}$	$88.9_{\pm 0.2}$	$91.5_{\pm 0.1}$	$89.2_{\pm 0.2}$
	`Balanced`	$91.1_{\pm 0.4}$	$89.0_{\pm 0.3}$	$90.9_{\pm 0.3}$	$88.9_{\pm 0.4}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$56.4_{\pm 10.0}$	$\mathbf{55.3}_{\pm 4.0}$	$57.7_{\pm 11.9}$	$60.0_{\pm 3.6}$
	`Uncertain+TabPFN`	$\mathbf{42.5}_{\pm 0.9}$	$58.7_{\pm 0.6}$	$\mathbf{42.7}_{\pm 1.1}$	$\mathbf{58.6}_{\pm 0.6}$
ACSEmployment	`Vanilla`	$64.0_{\pm 0.4}$	$62.0_{\pm 1.8}$	$65.0_{\pm 0.3}$	$62.2_{\pm 1.4}$
	`Balanced`	$64.0_{\pm 0.5}$	$65.0_{\pm 1.0}$	$64.8_{\pm 0.4}$	$65.3_{\pm 1.1}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$64.3_{\pm 0.4}$	$62.4_{\pm 2.0}$	$64.9_{\pm 0.3}$	$62.8_{\pm 0.9}$
	`Uncertain+TabPFN`	$\mathbf{57.5}_{\pm 3.3}$	$\mathbf{47.0}_{\pm 8.6}$	$\mathbf{61.1}_{\pm 3.0}$	$\mathbf{53.9}_{\pm 6.1}$
ACSMobility	`Vanilla`	$68.3_{\pm 0.8}$	$67.8_{\pm 1.0}$	$67.6_{\pm 1.0}$	$67.4_{\pm 1.2}$
	`Balanced`	$68.1_{\pm 0.7}$	$67.9_{\pm 1.4}$	$67.6_{\pm 1.1}$	$67.6_{\pm 1.5}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$68.1_{\pm 0.9}$	$67.8_{\pm 0.9}$	$67.4_{\pm 0.8}$	$66.9_{\pm 0.8}$
	`Uncertain+TabPFN`	$\mathbf{68.1}_{\pm 0.8}$	$\mathbf{67.7}_{\pm 1.0}$	$\mathbf{67.2}_{\pm 0.8}$	$\mathbf{66.8}_{\pm 0.9}$
German Credit	`Vanilla`	$72.5_{\pm 3.1}$	$70.3_{\pm 3.1}$	$71.8_{\pm 3.1}$	$71.0_{\pm 3.0}$
	`Balanced`	$72.7_{\pm 2.3}$	$70.7_{\pm 2.3}$	$71.9_{\pm 3.0}$	$71.2_{\pm 2.6}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$64.6_{\pm 5.1}$	$62.3_{\pm 8.5}$	$68.4_{\pm 4.8}$	$67.8_{\pm 5.4}$
	`Uncertain+TabPFN`	$\mathbf{60.4}_{\pm 5.5}$	$\mathbf{51.3}_{\pm 26.5}$	$\mathbf{63.8}_{\pm 3.9}$	$\mathbf{60.8}_{\pm 10.9}$
Diabetes	`Vanilla`	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$	$80.4_{\pm 0.1}$	$89.0_{\pm 0.1}$
	`Balanced`	$\mathbf{66.2}_{\pm 0.9}$	$\mathbf{75.9}_{\pm 0.8}$	$\mathbf{65.1}_{\pm 0.4}$	$\mathbf{74.6}_{\pm 0.4}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$70.4_{\pm 20.7}$	$77.1_{\pm 26.8}$	$80.3_{\pm 0.1}$	$89.0_{\pm 0.1}$
	`Uncertain+TabPFN`	$74.9_{\pm 10.0}$	$84.8_{\pm 8.0}$	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$
CelebA	`Vanilla`	$84.7_{\pm 0.2}$	$83.2_{\pm 0.3}$	$85.0_{\pm 0.2}$	$83.2_{\pm 0.3}$
	`Balanced`	$84.6_{\pm 0.2}$	$83.2_{\pm 0.3}$	$84.9_{\pm 0.2}$	$83.3_{\pm 0.3}$
	`Correlation R.`	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	`Uncertain+LR`	$\mathbf{72.4}_{\pm 11.7}$	$\mathbf{61.2}_{\pm 21.7}$	$84.9_{\pm 0.2}$	$83.1_{\pm 0.3}$
	`Uncertain+TabPFN`	$74.5_{\pm 8.4}$	$70.8_{\pm 10.1}$	$\mathbf{81.2}_{\pm 7.2}$	$\mathbf{77.2}_{\pm 11.9}$

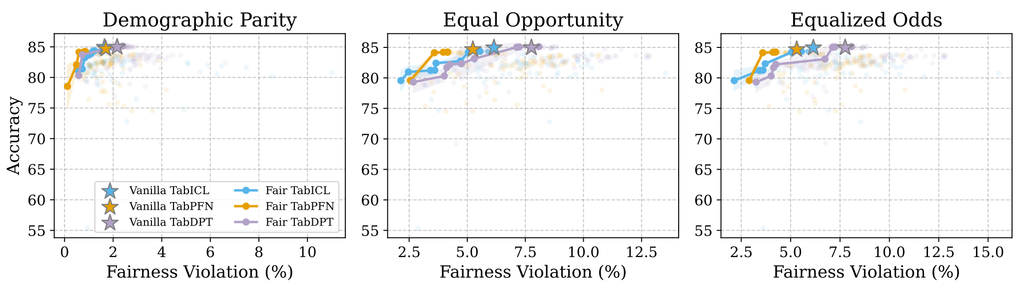
Table 4: **ICL prediction performance of sensitive attributes after applying different fairness interventions**. Smaller accuracy is better since it indicates how well the foundation model can reconstruct the sensitive attribute after the pre-processing fairness interventions. `Uncertain` methods yield the smallest accuracy, which justifies the improved fairness performance.



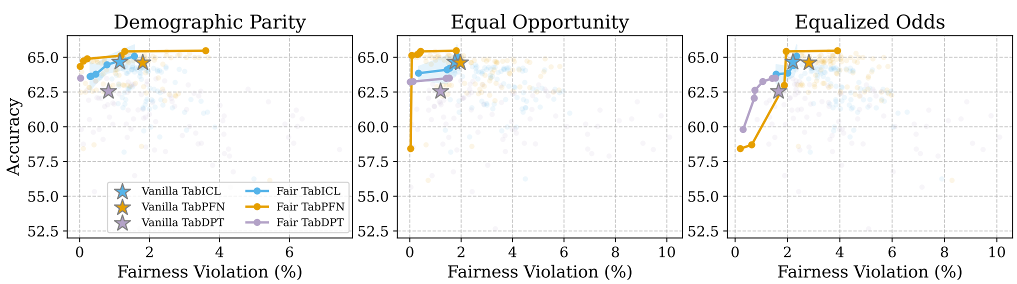
(a) ACSTravelTime



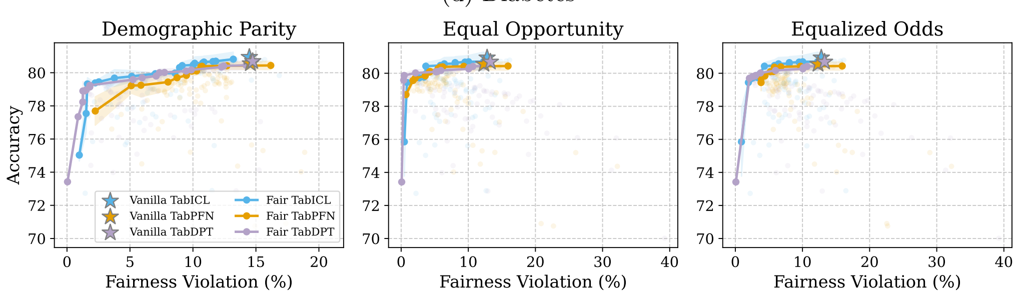
(b) ACSEmployment



(c) ACSPublicCoverage

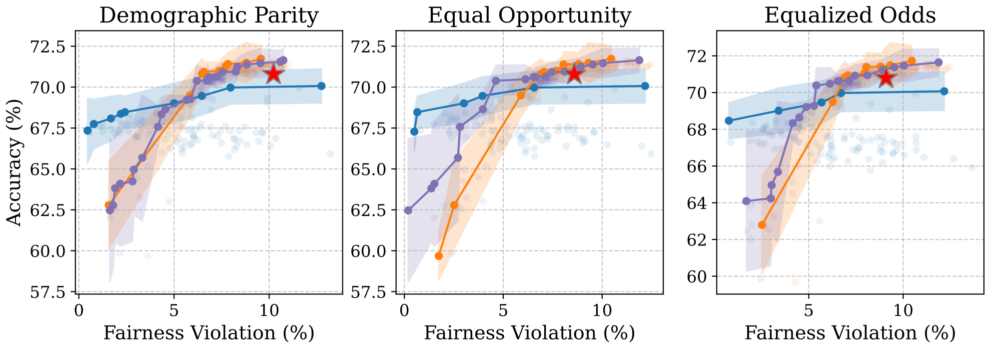


(d) Diabetes

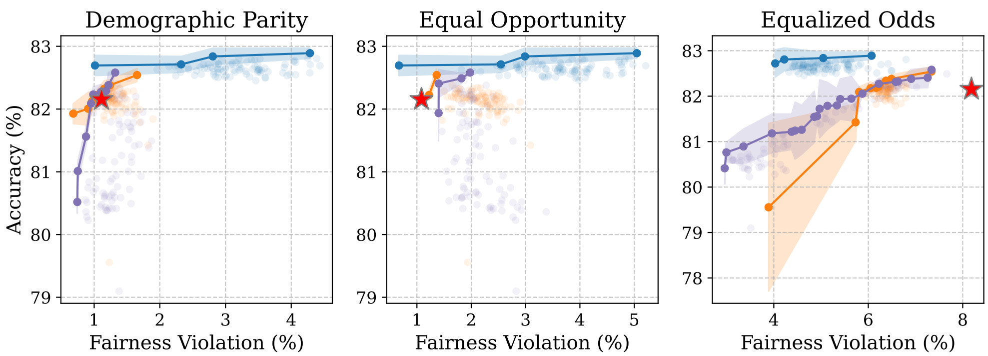


(e) CelebA

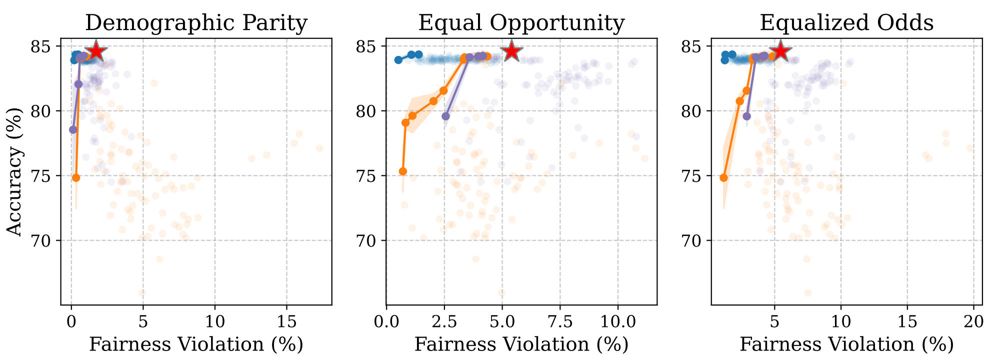
Figure 6: **TabPFN vs. TabICL**. Comparing the fairness-accuracy tradeoffs of tabular foundation models under different fairness interventions. TabPFN generally provides better fairness accuracy tradeoffs.



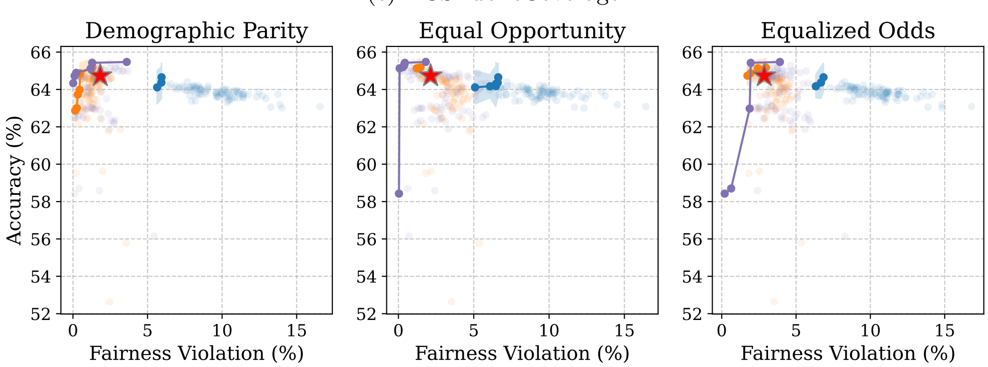
(a) ACSTravelTime



(b) ACSEmployment



(c) ACSPublicCoverage



(d) Diabetes

Figure 7: **Fairness-accuracy tradeoffs**. Comparing the fairness-accuracy Pareto-front of different fairness interventions with TabPFN. Figure 8 shows the results on the CelebA dataset.

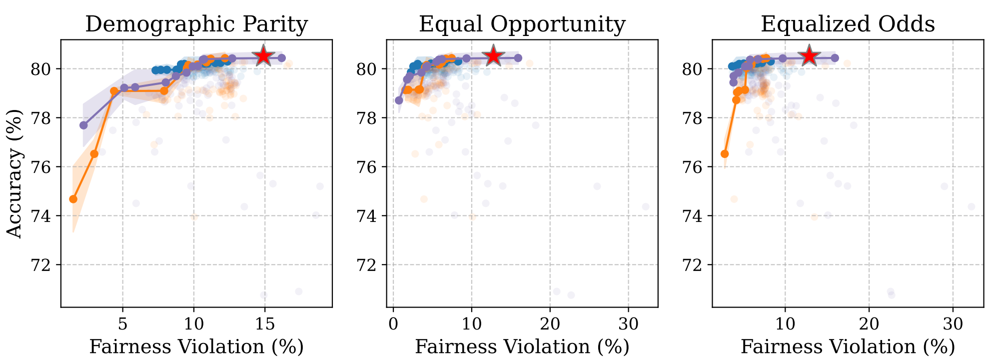
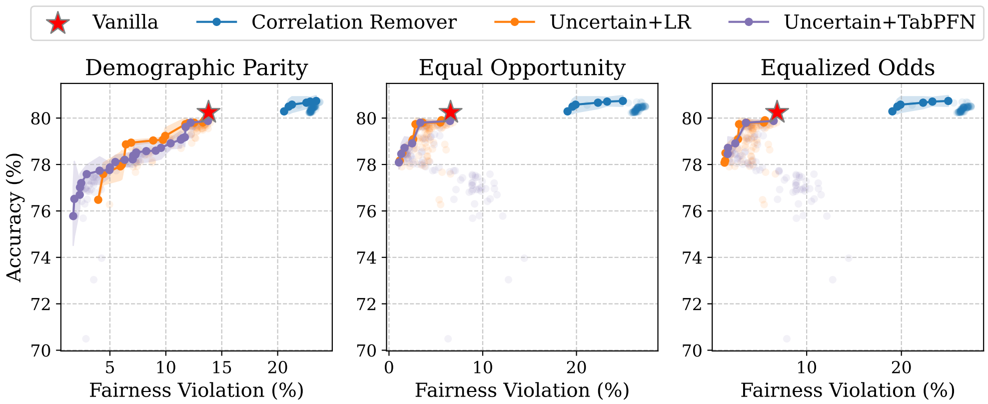
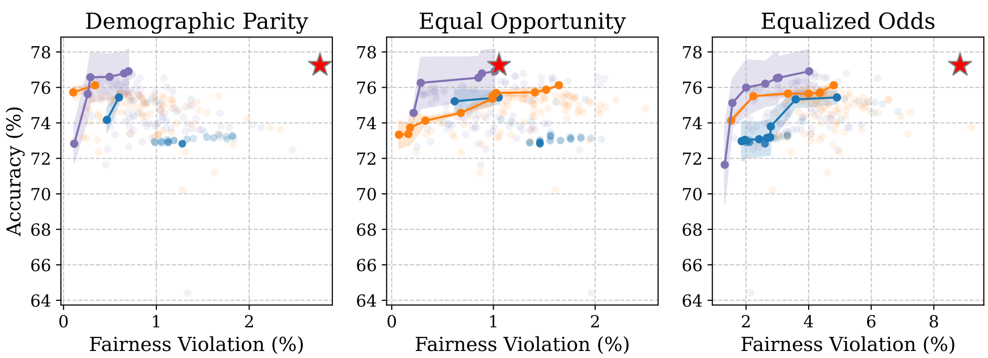


Figure 8: Fairness-accuracy tradeoffs on CelebA with TabPFN

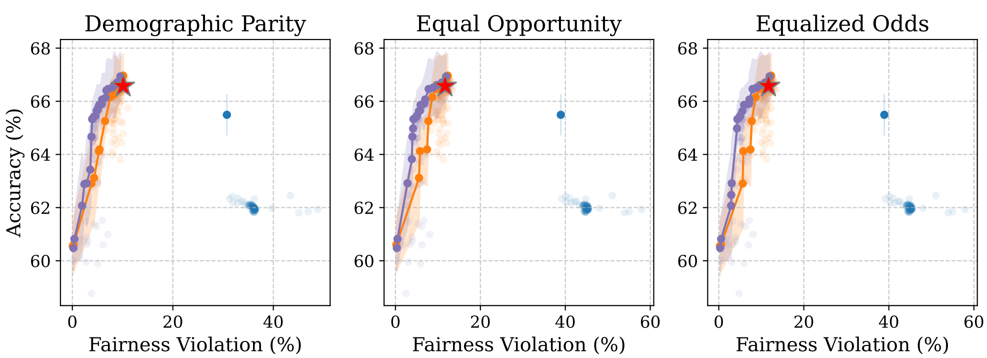
.



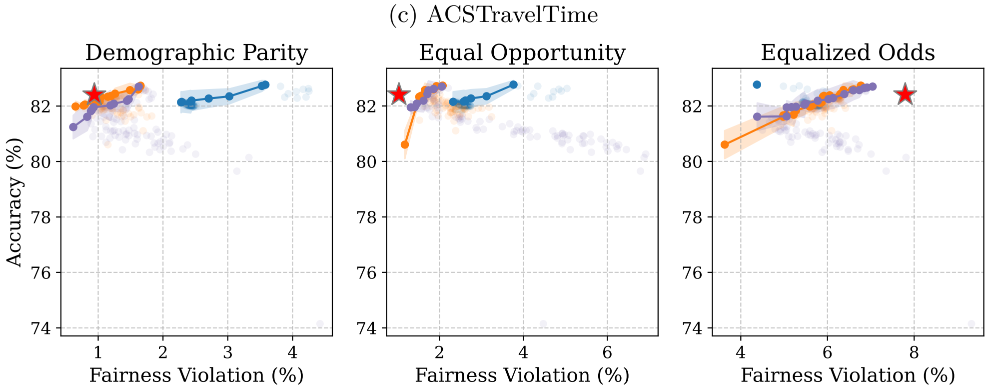
(a) ACSIncome



(b) ACSMobility

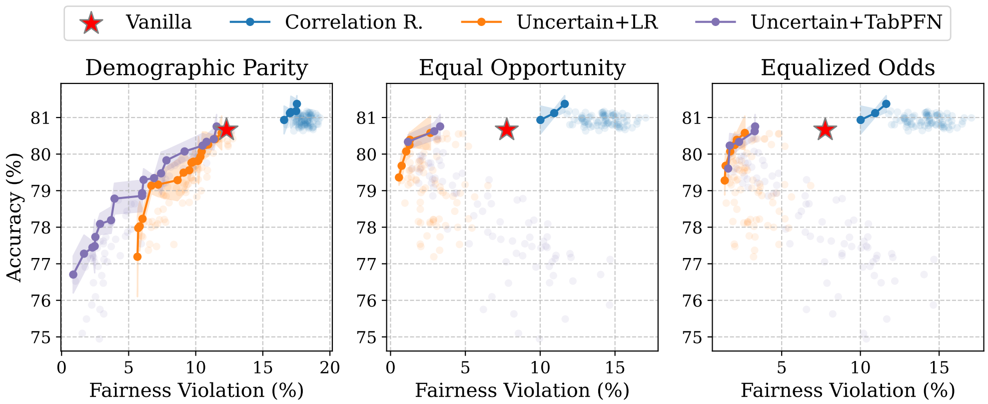


(c) ACSTravelTime

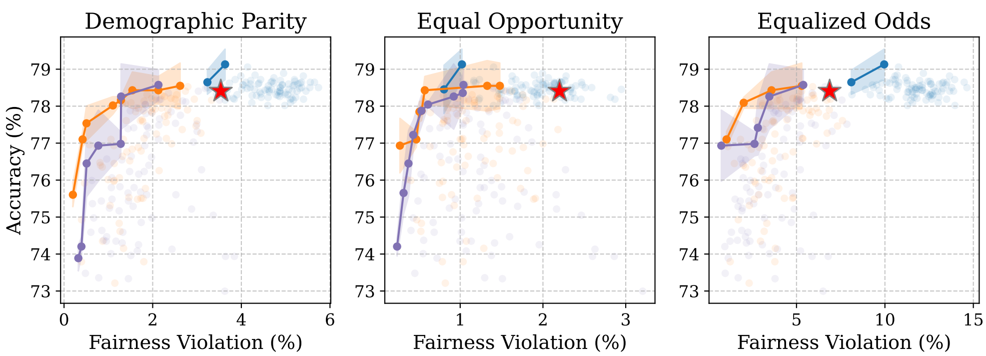


(d) ACSEmployment

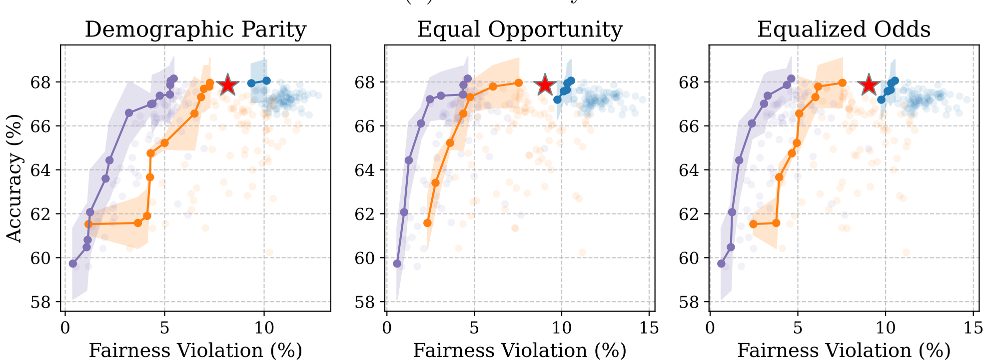
Figure 9: Fairness-accuracy tradeoffs on the ACS datasets with TabICL as foundation model
.



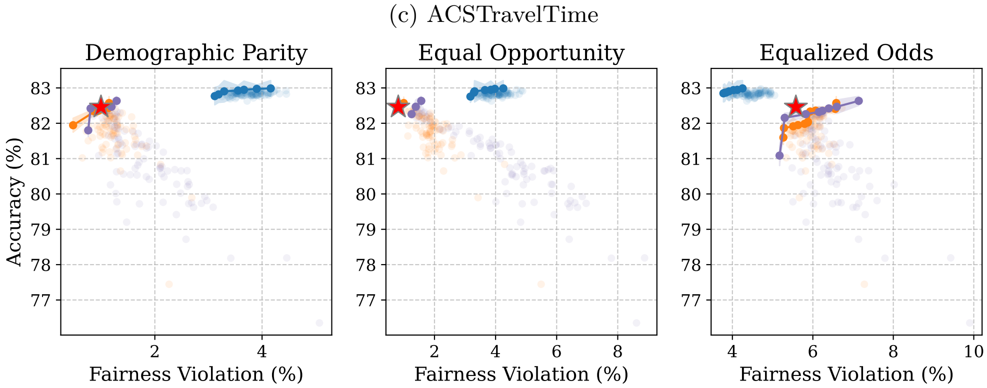
(a) ACSIncome



(b) ACSMobility

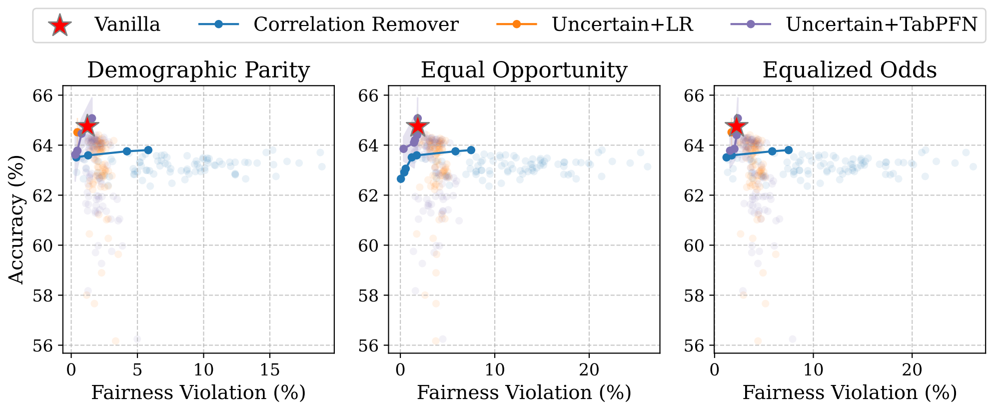


(c) ACSTravelTime

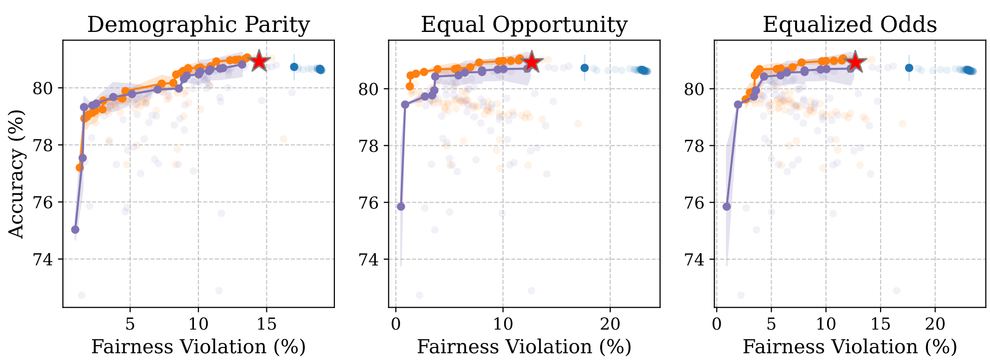


(d) ACSEmployment

Figure 10: Fairness-accuracy tradeoffs on the ACS datasets with TabDPT as foundation model
.



(a) Diabetes



(b) CelebA

Figure 11: Fairness-accuracy tradeoffs on the Diabetes and CelebA using TabICL as foundation model
.

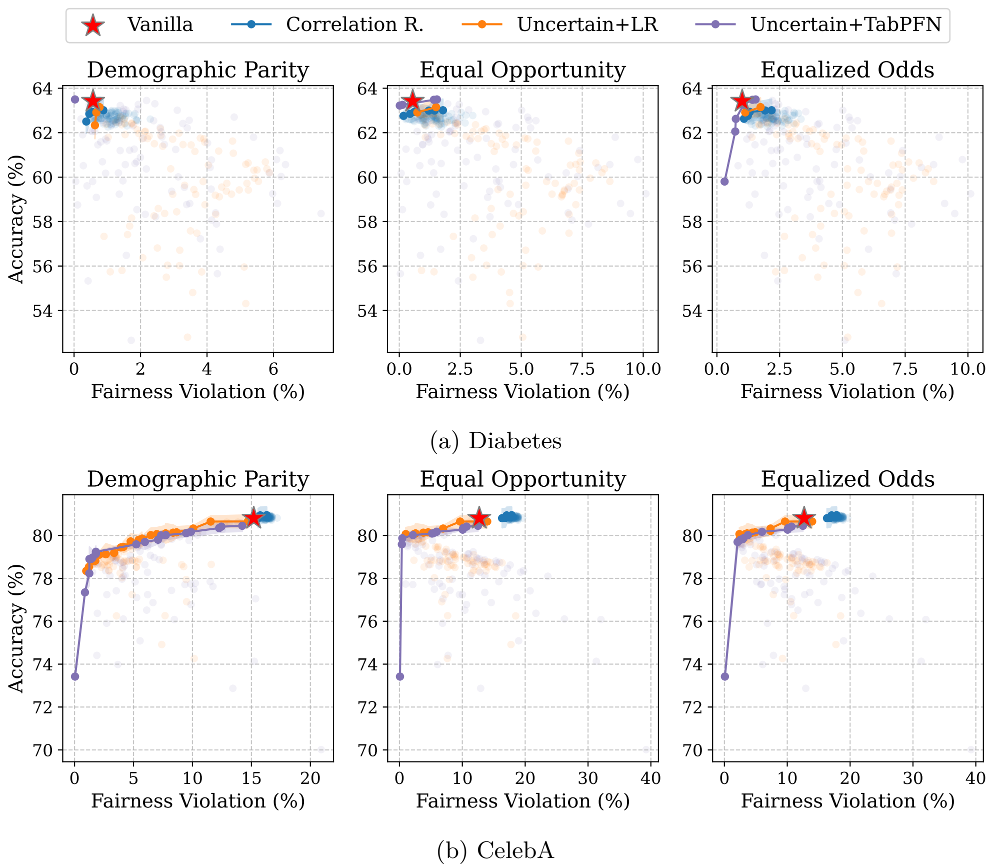


Figure 12: Fairness-accuracy tradeoffs on the Diabetes and CelebA using TabDPT as foundation model
.

Dataset	ICL Method	Accuracy ↑	Δ DP ↓	Δ EOP ↓	Δ EOD ↓
ACSIncome	`Vanilla`	$80.76_{\pm 0.5}$	$14.21_{\pm 1.3}$	$5.46_{\pm 2.9}$	$5.89_{\pm 2.4}$
	`Balanced`	$80.99_{\pm 0.4}$	$14.08_{\pm 1.8}$	$5.36_{\pm 3.6}$	$5.79_{\pm 3.3}$
	`Correlation R.`	$81.44_{\pm 0.5}$	$26.81_{\pm 2.0}$	$27.35_{\pm 6.4}$	$27.35_{\pm 6.4}$
	`Uncertain+LR`	$80.94_{\pm 0.4}$	$13.13_{\pm 1.6}$	$3.91_{\pm 2.4}$	$4.17_{\pm 2.1}$
	`Uncertain+TabPFN`	$80.13_{\pm 0.8}$	$8.90_{\pm 1.7}$	$3.33_{\pm 3.9}$	$3.56_{\pm 3.7}$
ACSEmployment	`Vanilla`	$82.18_{\pm 0.3}$	$1.11_{\pm 0.7}$	$1.01_{\pm 1.1}$	$8.17_{\pm 0.8}$
	`Balanced`	$82.13_{\pm 0.4}$	$1.09_{\pm 0.8}$	$1.10_{\pm 1.2}$	$8.20_{\pm 0.7}$
	`Correlation R.`	$82.41_{\pm 0.4}$	$3.90_{\pm 1.0}$	$5.05_{\pm 1.2}$	$5.49_{\pm 1.0}$
	`Uncertain+LR`	$81.98_{\pm 0.4}$	$0.86_{\pm 0.4}$	$0.99_{\pm 0.8}$	$7.56_{\pm 0.8}$
	`Uncertain+TabPFN`	$81.69_{\pm 0.7}$	$0.80_{\pm 0.4}$	$0.91_{\pm 0.6}$	$6.99_{\pm 0.8}$
ACSPublicCoverage	`Vanilla`	$84.71_{\pm 0.5}$	$1.75_{\pm 1.2}$	$5.13_{\pm 3.5}$	$5.23_{\pm 3.4}$
	`Balanced`	$84.49_{\pm 0.5}$	$1.75_{\pm 1.4}$	$5.80_{\pm 2.8}$	$5.80_{\pm 2.8}$
	`Correlation R.`	$84.57_{\pm 0.6}$	$1.64_{\pm 1.2}$	$4.89_{\pm 3.7}$	$4.94_{\pm 3.6}$
	`Uncertain+LR`	$80.15_{\pm 2.0}$	$3.08_{\pm 3.2}$	$4.61_{\pm 4.1}$	$5.28_{\pm 3.4}$
	`Uncertain+TabPFN`	$81.58_{\pm 1.4}$	$2.14_{\pm 1.4}$	$7.78_{\pm 4.3}$	$7.78_{\pm 4.3}$
ACSTravelTime	`Vanilla`	$70.52_{\pm 0.6}$	$9.79_{\pm 0.9}$	$8.09_{\pm 2.4}$	$8.79_{\pm 1.9}$
	`Balanced`	$70.85_{\pm 0.6}$	$10.14_{\pm 1.7}$	$8.46_{\pm 3.5}$	$9.32_{\pm 2.6}$
	`Correlation R.`	$70.42_{\pm 0.5}$	$9.82_{\pm 1.8}$	$8.43_{\pm 3.0}$	$8.70_{\pm 2.5}$
	`Uncertain+LR`	$70.48_{\pm 0.5}$	$10.16_{\pm 0.7}$	$8.42_{\pm 2.4}$	$9.07_{\pm 1.8}$
	`Uncertain+TabPFN`	$70.00_{\pm 0.6}$	$8.30_{\pm 0.9}$	$6.69_{\pm 2.1}$	$7.49_{\pm 1.6}$
ACSMobility	`Vanilla`	$76.86_{\pm 0.8}$	$2.25_{\pm 1.3}$	$1.91_{\pm 0.6}$	$6.49_{\pm 2.8}$
	`Balanced`	$77.11_{\pm 1.1}$	$2.86_{\pm 1.6}$	$0.97_{\pm 0.9}$	$8.38_{\pm 4.9}$
	`Correlation R.`	$76.86_{\pm 0.6}$	$6.27_{\pm 1.6}$	$3.33_{\pm 1.7}$	$12.28_{\pm 3.3}$
	`Uncertain+LR`	$76.59_{\pm 0.8}$	$1.86_{\pm 1.7}$	$2.27_{\pm 1.0}$	$4.31_{\pm 2.4}$
	`Uncertain+TabPFN`	$76.58_{\pm 0.9}$	$2.05_{\pm 2.0}$	$1.95_{\pm 1.2}$	$4.34_{\pm 4.4}$
Diabetes	`Vanilla`	$64.59_{\pm 0.3}$	$1.74_{\pm 1.0}$	$1.78_{\pm 1.3}$	$2.74_{\pm 1.6}$
	`Balanced`	$64.70_{\pm 0.5}$	$1.62_{\pm 1.2}$	$2.37_{\pm 1.3}$	$3.08_{\pm 1.4}$
	`Correlation R.`	$64.69_{\pm 0.3}$	$1.33_{\pm 1.1}$	$2.13_{\pm 1.5}$	$2.68_{\pm 1.2}$
	`Uncertain+LR`	$64.39_{\pm 0.6}$	$0.77_{\pm 0.5}$	$2.52_{\pm 1.1}$	$2.52_{\pm 1.1}$
	`Uncertain+TabPFN`	$64.24_{\pm 0.6}$	$1.20_{\pm 0.9}$	$3.21_{\pm 2.1}$	$3.59_{\pm 2.0}$
German Credit	`Vanilla`	$74.80_{\pm 4.6}$	$5.14_{\pm 3.5}$	$5.88_{\pm 4.3}$	$12.20_{\pm 6.6}$
	`Balanced`	$74.43_{\pm 3.0}$	$5.92_{\pm 5.5}$	$5.47_{\pm 5.0}$	$15.85_{\pm 10.3}$
	`Correlation R.`	$75.25_{\pm 3.5}$	$11.78_{\pm 6.3}$	$9.30_{\pm 3.8}$	$16.72_{\pm 9.3}$
	`Uncertain+LR`	$74.36_{\pm 3.4}$	$7.72_{\pm 3.8}$	$6.01_{\pm 4.1}$	$11.98_{\pm 5.9}$
	`Uncertain+TabPFN`	$73.98_{\pm 4.0}$	$4.65_{\pm 3.0}$	$5.21_{\pm 4.3}$	$11.91_{\pm 9.6}$
CelebA	`Vanilla`	$80.55_{\pm 0.4}$	$14.54_{\pm 1.0}$	$12.03_{\pm 2.8}$	$12.03_{\pm 2.8}$
	`Balanced`	$80.45_{\pm 0.5}$	$15.06_{\pm 0.8}$	$13.18_{\pm 1.7}$	$13.18_{\pm 1.7}$
	`Correlation R.`	$80.47_{\pm 0.4}$	$13.23_{\pm 1.3}$	$9.04_{\pm 2.2}$	$9.14_{\pm 2.2}$
	`Uncertain+LR`	$80.16_{\pm 0.5}$	$8.92_{\pm 0.9}$	$2.28_{\pm 2.0}$	$4.42_{\pm 1.3}$
	`Uncertain+TabPFN`	$79.86_{\pm 0.7}$	$10.01_{\pm 1.4}$	$3.54_{\pm 1.9}$	$5.46_{\pm 1.1}$

Table 5: This table supplements Figure 1 in main paper. It shows the accuracy and fairness performance of ICL predictions with TabPFN as foundation model under different preprocessing methods. The color range highlights the best ( ) to the worst-performing method ( ) for fairness accuracy. ↑ indicates higher is better (accuracy) and ↓ lower is better (unfairness).