Fair In-Context Learning via Latent Concept Variables

Anonymous ACL submission

Abstract

The emerging in-context learning (ICL) ability of large language models (LLMs) has prompted their use for predictive tasks in various domains with different data types, including tabular data, facilitated by serialization methods. However, with increasing applications in high-stakes domains, it has been shown that LLMs can inherit social bias and discrimination from their pre-training data. In this work, we investigate inherent bias in LLMs during in-context learning with tabular data. We focus on an optimal demonstration selection approach that utilizes latent concept variables for resource-efficient task adaptation. We design data augmentation strategies that reduce the correlation between predictive outcomes and sensitive variables, helping promote fairness during latent concept learning. We utilize the learned concept to select demonstrations and obtain fair predictions. The latent concept variables are learned using a smaller internal LLM and generalized to larger external LLMs. We empirically verify that the fair latent variable approach improves fairness results on tabular datasets compared to multiple heuristic demonstration selection methods. Code and data are available at `https://anonymous.4open.science/r/fairicl-AF6D`.

1 Introduction

LLMs have demonstrated immense capabilities in performing various natural language processing (NLP) tasks. A factor contributing to widespread LLM usage is their in-context learning (Brown et al., 2020) ability, which allows adaptation to downstream tasks without costly training or fine-tuning. With a few demonstration examples, ICL equips LLMs with the ability to infer task-specific context and perform inference with impressive utility. Recent research has also explored the applicability of LLMs on tabular data through serialization methods that facilitate ICL by transforming the data into natural language formats (Hegselmann et al., 2023). With the increasing integration of LLM inference in domains such as healthcare (Wu et al., 2023), finance (Li et al., 2023a), and the legal sector (Sun, 2023) with various data formats, it has become crucial to scrutinize their use from a trustworthiness perspective.

LLMs have been shown to exhibit discriminatory behavior in their outputs due to stereotypes and prejudices inherent in pre-training data (Abid et al., 2021; Basta et al., 2019). When used for decisive tasks, LLMs may mirror social inequalities and biases from the real world, leading to harmful consequences. Furthermore, in ICL settings with tabular data classification, recent research has empirically verified the presence of bias in LLM outputs. Liu et al. (2023) investigated unfairness in ICL with tabular data by flipping the labels of in-context demonstration examples and observed bias reduction but with significant trade-offs in model utility. Li et al. (2024) similarly implemented multiple heuristic methods for demonstration selection based on sensitive attributes and label distribution in the demonstrations. Hu et al. (2024) discovered that increasing the representation of minority groups and underrepresented labels in demonstrations helps to improve fairness at some cost to utility. They further developed a strategy that uses clustering to extract representative samples and selects demonstrations from the extracted samples based on their performance on a validation set.

In this work, we similarly explore optimal demonstration selection for ICL to promote fairness in LLM predictions, but utilize the latent concept variable mechanism (Wang et al., 2024) to achieve fair in-context learning. Wang et al. (2024) formulated ICL via a Bayesian perspective and theorized that inference with a finite number of demonstrations selected using latent concept approximates the optimal Bayes predictor. The latent concept is learned from an observed set of task-specific training data with a small LLM and used to obtain demonstrations that can be generalized to larger LLMs for improving performance.

Motivated by the influence of latent concepts on model performance, we formulate a fair demonstration selection approach for in-context learning, dubbed as **FairICL**. As the latent concepts are learned from an observed set of task-specific examples, we conjecture that the training data distribution may affect the quality of the learned latent concepts and ultimately the model predictions from both accuracy and fairness perspectives. Therefore, in FairICL, we incorporate an effective data augmentation technique that promotes decorrelation between the sensitive attributes and the outcome variables by randomizing the relationship between them. This augmentation allows us to obtain a fairer representation of the task-specific data used to learn the fair latent concept variables while preserving relevant information among non-sensitive attributes and the label. We then utilize the learned concepts to select demonstrations from the observed training examples such that the probability of observing the learned latent variable is maximized when conditioned on the corresponding example. The selected demonstrations are used to perform in-context learning with external LLMs larger than the one used for learning. This framework can support private businesses or organizations to obtain fair LLM predictions on their local data without having to train/fine-tune large models with fairness objectives. We empirically validate FairICL on real-world tabular datasets known to represent social biases and demonstrate that FairICL can effectively achieve fairness goals while maintaining predictive utility. We compare the performance of FairICL with multiple heuristic approaches and conduct a comprehensive analysis of the influence of different hyperparameters. Our empirical results show that FairICL can generalize demonstration selection to external LLMs and outperform baseline methods.

2 Preliminaries

2.1 In-Context Learning

The in-context learning (Brown et al., 2020) ability of LLMs has prompted multiple works that investigate how LLMs learn from demonstration examples for certain tasks without being explicitly trained. Let us denote a pre-trained LLM as $\mathcal{M}$ with parameters $\mathbf{W}$. Let $D = \{(x_i, y_i)\}_{i=1}^n$ denote a tabular dataset observed for an arbitrary task where $x_i \in \mathcal{X}$ represents attributes of the i-th instance and $y_i \in \mathcal{Y}$ its corresponding outcome. Assume $a_i \in \mathcal{A}$ denotes its sensitive attribute. For in-context learning, the LLM is provided with k examples from D as demonstrations to guide the model in structuring its response for a test example x. Conditioned on a task description $inst$, a set of sampled demonstrations $\{(x_1, y_1), \cdots, (x_k, y_k)\}$, and a test query x, the prediction output $\hat{y}$ from $\mathcal{M}$ can be formally formulated as

$$\hat{y} \leftarrow \mathcal{M}(inst, \underbrace{g(x_1, y_1), \cdots, g(x_k, y_k)}_{\text{demonstration examples}}, g(x)), \tag{1}$$

where $g(x_k, y_k)$ denotes a prompt function (e.g., a template) that transforms the k-th demonstration into natural language text. To simplify, we omit the task description and prompt function thereafter and represent the output probability as

$$P_{\mathcal{M}}(y|(x_1, y_1), \cdots, (x_k, y_k), x; \mathbf{W}). \tag{2}$$

ICL performance has been found to be significantly influenced by demonstration examples and their ordering (Liu et al., 2021; Rubin et al., 2021; Su et al., 2022; Lu et al., 2021; Ma et al., 2024). Consequently, recent works explore effective demonstration selection based on similarity to query input (Liu et al., 2021; Rubin et al., 2021; Su et al., 2022), entropy of predicted labels (Lu et al., 2021), and low predictive bias (Ma et al., 2024).

2.2 Latent Concept Learning

An essential research question in in-context learning is effective demonstration selection to enable optimal downstream performance. Towards this objective, (Xie et al., 2021) put forth an interpretation of ICL based on latent concept variables, concluding that pre-trained models learn latent concepts during next-token prediction training and infer a shared latent concept among demonstrations used for inference. They show that under assumptions of a hidden Markovian data generation process, discrete latent concept tokens, and an approximately infinite number of demonstrations, in-context learning is an optimal predictor. Similarly, (Wang et al., 2024) studied latent concepts in LLMs but under a more general assumption of continuous latent concepts and that the data generation process is governed by an underlying causal mechanism given as $X \rightarrow Y \leftarrow \theta$ or $Y \rightarrow X \leftarrow \theta$ where X represents the input, Y the output, and θ the latent concept

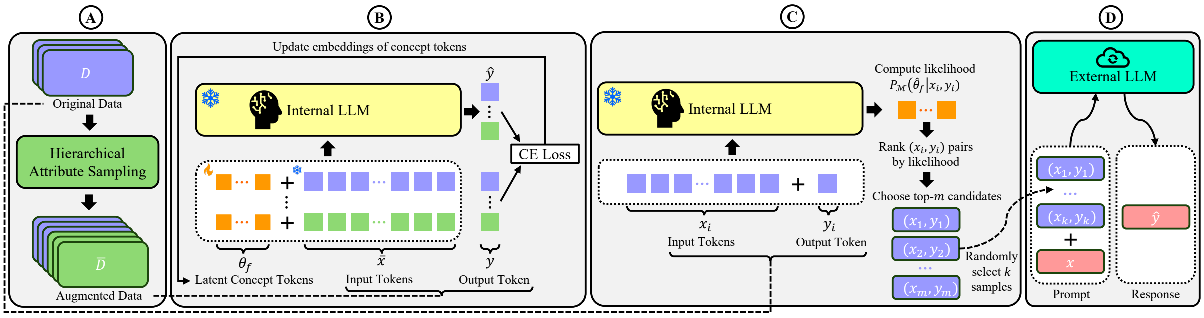


Figure 1: Overview of FairICL including steps from **A** to **D**, **A**: A hierarchical attribute sampling approach is proposed to craft synthetic samples and create augmented training data $\bar{D}$; **B**: Samples in $\bar{D}$ are utilized to learn latent concept tokens with an internal LLM; **C**: A corresponding likelihood score is computed for each sample in D, and all samples are then ranked to choose k demonstrations from top-m candidates; **D**: Selected demonstrations and test input x are used to prompt an external LLM to get prediction $\hat{y}$.

variable. Then, in-context learning can become an optimal predictor with a finite number of demonstrations chosen using the latent concept variable θ. To find the optimal value of θ when considering the $X \rightarrow Y \leftarrow \theta$ direction, (Wang et al., 2024) formulated a latent concept variable framework for learning task-specific concept tokens that capture sufficient information for next-token prediction by minimizing a loss jointly conditioned on X and the learned $\hat{\theta}$ as

$$l(x, y; \hat{\theta}) = -\log P_{\mathcal{M}}(y|\hat{\theta}, x), \tag{3}$$

where $\hat{\theta}$ represents the learned latent concept variable, x denotes an input token sequence and y the discrete target variable. In practice, $\hat{\theta}$ is optimized by adding new tokens to $\mathcal{M}$'s vocabulary with corresponding embedding vectors, which we refer to as $\mathbf{W}_\theta$. During training, $\mathbf{W}_\theta$ is updated using the loss defined above. The learned $\hat{\theta}$ is then used to select k most suitable demonstrations based on the likelihood of observing the concept tokens when conditioned on the demonstration pairs formulated as $P_{\mathcal{M}}(\hat{\theta}|(x_i, y_i), \ldots, (x_k, y_k))$. Assuming independence among the sampled demonstrations, the top-ranked examples are obtained based on latent concept likelihood for individual examples,

$$\underset{(x_i,y_i)\in D}{\text{argmax}}\, P_{\mathcal{M}}(\hat{\theta}|x_i, y_i). \tag{4}$$

The selected demonstrations are used to perform in-context learning and are further generalizable for inference with LLMs larger than the ones used to learn $\hat{\theta}$.

3 Fair Latent Concept Learning

LLMs have been shown to replicate bias and prejudice likely present in their pre-training corpus. Providing LLMs with biased examples as demonstrations during ICL may further corroborate the prediction bias, potentially leading to discriminatory outcomes in classification tasks. However, filtering pre-training data and re-training/fine-tuning LLMs on unbiased data is often practically infeasible due to resource constraints. Moreover, removing discrimination from pre-training data may not entirely address the unfairness resulting from biased demonstrations during inference. Here, we focus on the demonstration selection process, which can guide LLM predictions by providing task-specific contextual information. Researchers have empirically shown that varying demonstrations can affect the bias and fairness outcomes of LLMs (Hu et al., 2024; Ma et al., 2024). Furthermore, the proportion of samples from minority and majority groups in demonstrations affects the trade-off between fairness and performance metrics (Hu et al., 2024).

3.1 Problem Setup

In the latent concept variable model, demonstrations are selected based on the likelihood of observing latent concept variable $\hat{\theta}$ (Wang et al., 2024). Generally, the concept variables capture format and task information and can help improve in-context learning performance. However, the quality of the learned latent concept variables highly depends on the observed data D. We hypothesize that using a biased dataset D to learn the latent concept can lead to selecting demonstrations that favor the majority group. For instance, consider a dataset containing a comparatively higher number of positive/advantaged class instances for the majority group, reflecting real societal bias. The latent concept variables may associate the positive outcome with the majority class, as this biased prediction can lead to better prediction accuracy owing to imbalanced label distributions. Consequently, demonstrations selected using the latent concept variables can reinforce the bias originating from the dataset. In the following, we propose FairICL, a fair latent concept learning framework with data augmentation to mitigate unfairness in ICL predictive outcomes arising from demonstration selection. An overview of the method is presented in Fig. 1.

3.2 Constructing Augmented Training Data

To ensure fair predictive outcomes, we consider the correlation between the sensitive attribute a and the outcome variable y in the dataset D used to learn the latent concept variable θ. We conjecture that learning latent concept variables from an unbiased dataset can prevent $\hat{\theta}$ from incorporating bias into the task-specific contextual information that improves ICL performance. To this end, we design and implement a data pre-processing strategy on D aimed at decorrelating the sensitive attribute and the label. Assuming we obtain a dataset $\tilde{D}$ that preserves task-relevant information from D and not the biased correlation between a and y, we then construct an augmented training dataset $\bar{D}$ from both D and $\tilde{D}$ to promote fairness while learning task-specific contextual information in a fair representation of latent concepts $\hat{\theta}_f$. Note that our focus is on LLM classification with ICL on tabular data, which is the most commonly used data representation in fairness literature.

For hierarchical attribute sampling, we define an order for all non-sensitive attributes and construct a synthetic sample based on the order. First, we randomly sample a label from a uniform distribution and obtain a subset of D conditioned on the sampled label value. We then uniformly sample the first non-sensitive attribute in the ordered list from the values occurring in the subset. We further constrain the subset to include only the sampled value of the first non-sensitive attribute, and sample the second non-sensitive attribute uniformly, and so on. To populate the sensitive attribute value, we randomly sample it from a uniform distribution independent of the label and any non-sensitive attributes. Furthermore, if D contains any proxy-sensitive attributes that may allude to the sensitive attribute, we condition its sampling on the sensitive attribute value to promote complete decorrelation. In this manner, we generate $\tilde{D} = \{(x_i, y_i)\}_{i=1}^{\tilde{n}}$ as an unbiased representation of D.

We then construct our augmented training dataset, which contains $n + \tilde{n}$ instances, and each augmented instance contains q demonstration examples from D and one query sample from either D or $\tilde{D}$ to facilitate in-context learning. Formally each instance takes the form $\langle (x_1, y_1), \cdots, (x_q, y_q), x, y \rangle$ which we denote as $(\overline{x}, y)$ thereafter. We also denote this formatted dataset containing augmented samples as $\bar{D} = \{\bar{x}_i, y_i\}_{i=1}^{n+\tilde{n}}$. The following discusses how we learn the fair latent concept variables from $\bar{D}$.

3.3 Learning Fair Latent Concept Variable

We learn the latent concept variables by implementing prompt tuning to optimize a set of new token embeddings that is prepended to each training input token sequence (Wang et al., 2024). More importantly, we utilize the augmented dataset $\bar{D}$ to construct input sequences for learning θ_f to promote improvements in fairness and utility simultaneously. Directly optimizing θ_f as a sequence of words is inefficient due to the discrete nature of text space. Typically, large language models (LLMs) process inputs as sequences of tokens, which are subsequently transformed into embeddings. Therefore, we optimize the fair latent concept in the LLM $\mathcal{M}$'s embedding space, where θ_f is represented as a sequence of c learnable tokens, each associated with an embedding vector. We denote the subset of weights in $\mathbf{W}$ corresponding to θ_f as $\mathbf{W}_{\theta_f}$. During training, we prepend $\hat{\theta}_f$ to the input sequences and learn $\mathbf{W}_{\theta_f}$ by minimizing the negative log-likelihood objective as follows

$$\mathcal{L} = -\sum_{i=1}^{n+\tilde{n}} \log P_{\mathcal{M}}(y_i | \hat{\theta}_f, \overline{x}_i). \tag{5}$$

During gradient optimization, parameters $\mathbf{W}_{\theta_f}$ corresponding to $\hat{\theta}_f$ are updated, and all other parameters are frozen. The ultimate goal of learning fair latent concept variables is to derive the optimal $\widehat{\mathbf{W}}_{\theta_f}$ using the task-specific data D to improve performance and the generated data $\tilde{D}$ to promote fairness simultaneously.

3.4 Demonstration Example Selection with θ_f Likelihood

The learned fair latent concept $\hat{\theta}_f$ is then used to select top-ranking examples from D, which will be provided as context to a larger external LLM during inference via ICL. This demonstration selection follows the rationale that training examples that maximize the likelihood of predicting the trained task-specific latent concept variables are optimal demonstrations for the corresponding task objective (Wang et al., 2024). For each training example $(x_i, y_i) \in D$, the likelihood of $\hat{\theta}_f$ is expressed using the probability distribution shown as

$$P_{\mathcal{M}}(\hat{\theta}_f | x_i, y_i). \quad (6)$$

In our implementation, we obtain this likelihood as the probability of observing the trained $\hat{\theta}_f$ when postfixed to a sample (x_i, y_i). Subsequently, training examples are sorted based on their computed likelihood values. We then select the top m examples that maximize the likelihood of $\hat{\theta}_f$ and form the demonstration candidate set. We subsample this candidate set to allow each test query to be paired with varying demonstrations during testing. Finally, we randomly select k demonstration examples from the candidate set for each test instance, combining these to construct the final prompt for inference with an external LLM. The augmented dataset generation, latent concept learning, and demonstration selection procedures are summarized in Algorithm 1 (Appendix B).

4 Experimental Evaluation

4.1 Datasets

We evaluate the effectiveness of fair demonstration selection with FairICL using three benchmark fair machine learning datasets: Adult Income dataset (Becker and Kohavi, 1996), COMPAS (Larson et al., 2016), and LawSchool (Quy et al., 2022). We focus on the Adult dataset for the main experiments and include discussion and results for the other two in Appendix D and E. Following previous work (Liu et al., 2023), we use a subset of 10 attributes from the dataset named in Fig. 5a. We also subsample a training dataset of 30,000 records after preprocessing. We perform serialization on the tabular Adult dataset similar to (Hegselmann et al., 2023; Carey et al., 2024), i.e., we convert each row in the dataset to a natural language format to facilitate LLM prompting. The specific serialization template and in-context learning format are included in Appendix D.

As discussed in Section 3.2, we generate an augmented dataset to enable fair latent concept learning. For Adult, we use *sex* as the sensitive attribute and distinguish *relationship* and *marital status* as the proxy-sensitive attributes, as some instances of the *relationship* attribute contain gender-specific vocabulary and the attribute *marital status* may depend on values of *relationship*. To generate augmented samples, we specify a hierarchical order for the non-sensitive attributes and a separate order for the sensitive and proxy-sensitive attributes based on the analysis in (Quy et al., 2022). Using this attribute sampling technique, we generate $\tilde{n} = n$ number of unique augmented data samples and construct our training dataset $\bar{D}$ by combining D and $\tilde{D}$. For the test dataset, we randomly sample 1000 instances with equal representation for majority and minority groups for each experimental run. Please refer to Appendix D for details regarding the attribute hierarchy and dataset distributions.

4.2 Baselines

We compare FairICL against several baselines that implement different demonstration selection approaches. *Random* refers to standard in-context learning where k training examples are randomly sampled as demonstrations for each test instance (Brown et al., 2020). *Balanced* implements in-context learning with equal representation for each sensitive attribute and class label combination in the demonstrations (Li et al., 2023b). *Instruction* is used to evaluate an LLM for fair and unbiased decisions based on manual prompting-based guidance with a balanced demonstration set (Li et al., 2023b; Atwood et al., 2024). *Removal* omits the sensitive attribute from the demonstrations of Balanced (Li et al., 2023b). As we serialize tabular data, we further replace gendered pronouns with gender-neutral ones in the training data. *Counterfactual* is another heuristic technique and constructs demonstrations using $k/2$ examples from the majority (minority) group and the remaining examples by flipping the sensitive attribute of the previously sampled examples (Li et al., 2023b). *LatentConcept* is the latent concept variable-based approach from (Wang et al., 2024) where the latent concept variables are learned using the training dataset and then used to select top-k demonstrations.

4.3 Experimental Setup

In the FairICL framework, we use LLaMA-2-7B (Touvron et al., 2023) as the internal LLM for fair latent concept learning and LLaMA-2-13B (Touvron et al., 2023) as the external LLM for inference. We fix the learning rate at 0.0001 for all experiments and optimize the concept token embeddings over 5 epochs. For main experiments on the Adult dataset, we fix the number of added tokens

Table 1: Performance and fairness metrics of FairICL on the Adult dataset compared with baselines on LLaMA-2-7B and LLaMA-2-13B as external LLMs and LLaMA-2-7B as the internal LLM for latent concept learning; bold denotes best performance among fairness-aware methods and underline denotes best performance among all models

External LLM	Method	Acc(%)↑	F1(%)↑	$\lvert\Delta SP\rvert\downarrow$	$\lvert\Delta EO\rvert\downarrow$
LLaMA-2-13B	Random (Brown et al., 2020)	$76.00_{1.19}$	$75.75_{1.44}$	$0.14_{0.04}$	$0.11_{0.08}$
	LatentConcept (Wang et al., 2024)	$\underline{77.48}_{0.70}$	$77.22_{0.74}$	$0.16_{0.02}$	$0.12_{0.01}$
	Balanced (Li et al., 2023b)	$74.58_{1.38}$	$71.92_{2.19}$	$0.13_{0.05}$	$0.10_{0.07}$
	Counterfactual (Li et al., 2023b)	$68.18_{2.05}$	$57.39_{4.55}$	$0.13_{0.06}$	$0.17_{0.08}$
	Removal (Li et al., 2023b)	$75.72_{0.98}$	$76.48_{2.16}$	$0.14_{0.03}$	$0.09_{0.02}$
	Instruction (Li et al., 2023b)	$\mathbf{76.20}_{1.09}$	$77.21_{1.61}$	$0.20_{0.07}$	$0.15_{0.06}$
	FairICL	$75.72_{1.60}$	$\underline{\mathbf{77.61}}_{1.35}$	$\underline{\mathbf{0.08}}_{0.02}$	$\underline{\mathbf{0.03}}_{0.03}$
LLaMA-2-7B	Random (Brown et al., 2020)	$69.92_{0.87}$	$62.80_{1.25}$	$0.08_{0.02}$	$0.08_{0.04}$
	LatentConcept (Wang et al., 2024)	$\underline{70.04}_{1.69}$	$\underline{64.79}_{2.42}$	$0.17_{0.02}$	$0.17_{0.04}$
	Balanced (Li et al., 2023b)	$63.30_{6.33}$	$44.09_{17.37}$	$0.04_{0.03}$	$0.04_{0.03}$
	Counterfactual (Li et al., 2023b)	$59.58_{1.02}$	$34.44_{2.87}$	$0.08_{0.01}$	$0.13_{0.02}$
	Removal (Li et al., 2023b)	$64.60_{6.25}$	$47.92_{17.34}$	$0.09_{0.05}$	$0.11_{0.06}$
	Instruction (Li et al., 2023b)	$63.86_{6.35}$	$46.64_{14.86}$	$0.07_{0.06}$	$0.09_{0.05}$
	FairICL	$\mathbf{68.48}_{0.89}$	$\mathbf{64.42}_{1.01}$	$\underline{\mathbf{0.02}}_{0.03}$	$\underline{\mathbf{0.01}}_{0.04}$

c at 10, and the number of demonstrations during training q at 2. We randomly sample $k = 4$ demonstrations from a top-ranked candidate set of $m = 100$ training examples for each test query. We conduct our experiments on NVIDIA A100 GPUs with 40GB RAM. We report performance as an average of 5 runs with standard deviations for different test splits. For model utility, we report accuracy and F1 scores. To evaluate fairness, we compute commonly used fairness metrics, namely, **Statistical Parity** (ΔSP) (Dwork et al., 2012) and **Equalized Odds** (ΔEO) (Hardt et al., 2016). Please refer to Appendix C for a detailed description and formulation of the fairness metrics.

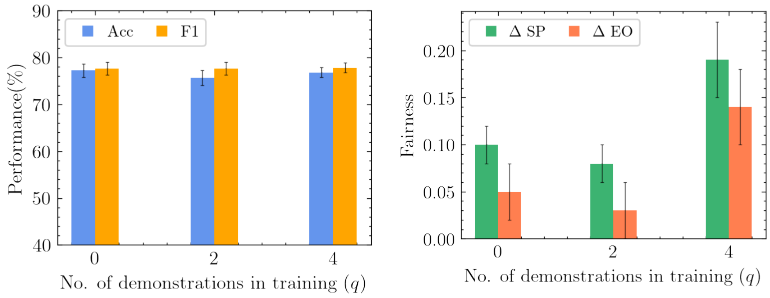
We also evaluate the learned fair latent concepts with LLaMA-2-7B as the external LLM. We investigate the impact of FairICL hyperparameters on fair latent concept learning via its overall performance. To this end, we report results when varying q as $\{0, 2, 4\}$ and evaluate the effect of $\tilde{n}$, i.e., the size of the generated dataset $\tilde{D}$, on FairICL. To analyze the effectiveness of latent concept learning with an augmented dataset, we conduct an ablation study where the augmented samples are created via complete random sampling as opposed to hierarchy-based sampling. We also evaluate the learned fair latent concepts directly by prepending them to test queries during inference. Finally, we vary k among $\{2, 4, 6, 8\}$ to analyze the influence of ICL demonstration size on inference results.

4.4 Results

Model Performance and Comparison with Baselines We report results for the Adult dataset from inference with LLaMA-2-7B and LLaMA-2-13B in Table 1. Firstly, we observe the performance of Random, where LLaMA-2-13B has increased utility compared to LLaMA-2-7B, undoubtedly due to the model's complexity. However, the fairness metrics ΔSP and ΔEO are larger, indicating a significant presence of bias in the outputs generated by LLaMA-2-13B. With the LatentConcept method, which optimizes demonstration selection for utility, performance is improved, but the bias is further amplified for both 7B and 13B models. These results motivate our study of bias in LLMs specifically for tabular classification and methods that can promote fairness in a resource-efficient manner.

In Table 1, we observe that FairICL can noticeably improve SP and EO measures for LLaMA-2-7B compared to the Random and LatentConcept methods while achieving comparable performance. Similarly, FairICL significantly reduces unfairness for LLaMA-2-13B with minimal loss of utility. Note that the latent concept variables are learned using the smaller LLaMA-2-7b as the internal model, and the selected demonstrations are utilized to construct inference prompts for LLaMA-2-13b. This shows that FairICL can generalize the fair demonstration selection process to larger LLMs, thus making the method resource-efficient. Since the external LLMs are used only for few-shot inference, FairICL also enables generalization to black-box LLMs.

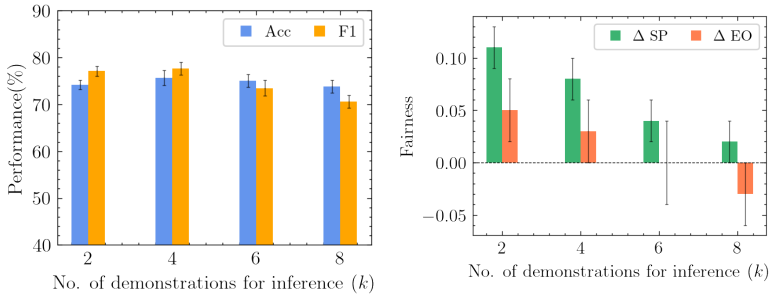
We also evaluate the effectiveness of FairICL compared to multiple fair demonstration selection baselines. As discussed in Section 4.2, these meth-



(a) Performance metrics

(b) Fairness metrics

Figure 2: FairICL performance on LLaMA-2-13B for varying number of demonstrations (q) during learning



(a) Performance metrics

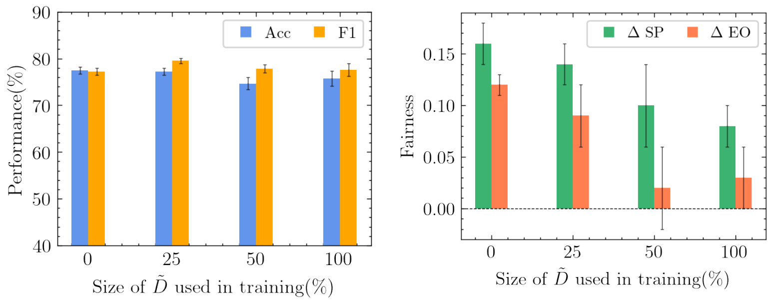
(b) Fairness metrics

Figure 3: FairICL performance with LLaMA-2-13B for varying number of demonstrations (k) during inference

ods address the LLM fairness issue via heuristic approaches. For LLaMA-2-7B, the fair baselines reduce some unfairness compared to LatentConcept but incur a significant loss in performance. Compared to Random, only the Balanced approach shows a notable reduction in SP and EO. FairICL, however, achieves the best fairness results without negatively affecting utility. For LLaMA-2-13B, the baselines mostly maintain performance but do not achieve fair outcomes. In contrast, FairICL shows a large decline in fairness metrics with similar or improved accuracy and F1. Our results for the COMPAS and LawSchool show similar trends of lower fairness metrics without utility losses (Appendix E). For a more granular analysis in the appendix, we also show how FairICL gradually reduces fairness metrics as training progresses while maintaining utility. These results demonstrate that decorrelation of sensitive attributes and outcomes helps learn fair latent concepts, resulting in demonstration selection that promotes fair predictions.

Number of Demonstrations We investigate the influence of the number of demonstrations during latent concept learning, denoted by q, and during inference, denoted by k, on the overall performance of FairICL with the Adult dataset. First, we vary q among $\{0, 2, 4\}$ and report results in Fig. 2 for LLaMA-2-13B while keeping the other parameters fixed at $c = 10$ and $k = 4$. From Fig. 2, we observe that accuracy and F1 remain fairly unchanged across different values of q. However, SP and EO are noticeably higher at $q = 4$, with the best metrics observed at $q = 2$. Since the q demonstrations for training are obtained from the original dataset containing biased examples, training prompts constructed with more biased samples negatively affect fairness during inference. In contrast, fewer demonstrations do not affect model utility as the augmented samples preserve useful correlations from the original dataset.

We then vary k among $\{2, 4, 6, 8\}$ for LLaMA-2-13B with fixed $q = 2$ and $c = 10$, and include the results in Fig. 3. Here, the fairness metrics demonstrate a sharper decline as the number of demonstrations during inference increases. We also observe a slight decrease in utility as the number of demonstrations increases, most likely due to the trade-off between utility and fairness. Since the k demonstrations are obtained from the top-m training examples ranked by the fair latent concept variable, having a larger k allows the inference prompt to guide the LLM towards fairer predictions.



(a) Performance metrics

(b) Fairness metrics

Figure 4: Performance and fairness metrics of FairICL with LLaMA-2-13B for different sizes of $\tilde{D}$

Table 2: Ablation results on LLaMA-2-7B

Method	Acc(%)↑	F1(%)↑	$\lvert\Delta\text{SP}\rvert$ ↓	$\lvert\Delta\text{EO}\rvert$ ↓
FairICL	$68.48_{0.89}$	$64.42_{1.01}$	$0.02_{0.03}$	$0.01_{0.04}$
FairICL-LC	$75.96_{1.20}$	$70.70_{1.67}$	$0.06_{0.01}$	$0.08_{0.02}$
FairICL-R	$58.58_{0.62}$	$31.72_{0.94}$	$0.01_{0.01}$	$0.00_{0.03}$

Ablation Study We investigate the role of data augmentation and latent concept learning by implementing two variations of FairICL. FairICL-LC directly evaluates the learned latent concepts as we prepend them to test prompts containing k randomly sampled demonstrations. FairICL-R adopts a random sampling mechanism for all attributes to create the augmented dataset and follows an inference procedure similar to FairICL. In other words, the generated dataset does not preserve the useful correlation between the non-sensitive attributes and outcomes. We report ablation results in Table 2 for LLaMA-2-7B since FairICL-LC can be evaluated only for the internal LLM whose vocabulary contains additional tokens corresponding to the latent concept variable. FairICL-LC achieves the best accuracy and F1 score, indicating that the latent concept learns information relevant to the task. Also, the low fairness metrics imply that training with the augmented dataset prompts the latent concept to favor fair predictions. FairICL-R achieves almost ideal fairness metrics but does not maintain model accuracy as the randomly generated dataset removes even the useful correlation between non-sensitive attributes and labels. FairICL, however, preserves relevant information in $\tilde{D}$, thus achieving fair and accurate predictive results.

Size of Augmented data In this section, we conduct a sensitivity analysis of $\tilde{n}$, the size of $\tilde{D}$, to evaluate the influence of the augmented dataset on FairICL's performance. We vary $\tilde{n}$ as {0, 25, 50, 100}% of its original size of 30,000 generated samples in the Adult dataset. We fix the other parameters q at 2, c at 10, and k at 10 to perform latent concept learning and obtain results for LLaMA-2-13B shown in Fig. 4. Note that the $\tilde{n} = 0\%$ setting corresponds to the LatentConcept baseline method in Table 1. From the results, we notice that the accuracy and F1-scores are generally unchanged when more augmented examples are included in the training prompt. This indicates that the data augmentation process in FairICL does not negatively affect an LLM's predictive performance. We further observe significant drops in the fairness metrics as the size of $\tilde{D}$ used for latent concept learning is increased. This demonstrates the positive impact of the data augmentation strategy in FairICL.

5 Conclusion

We investigated the issue of fairness in large language models during in-context learning for tabular data classification. We focused on a latent concept learning framework that optimizes the demonstration selection process for improved model utility via latent concept variables learned from a training dataset. We hypothesized that learning latent concepts from a biased dataset can cause the selection of biased demonstrations, resulting in unfair predictions, and empirically verified this phenomenon. To tackle this issue, we explored resource-efficient ways to influence LLM outputs without modifying model parameters and presented a fairness-aware latent concept learning framework, FairICL, that incorporates data augmentation to enable learning concept tokens that promote fairness while preserving task-relevant contextual information. Our experimental analysis showed that FairICL can effectively mitigate unfairness without causing significant tradeoffs in model utility for multiple datasets.

Limitations

We acknowledge certain limitations of the proposed framework. As FairICL utilizes a latent concept learning framework, it requires white-box access to a small LLM and resources to train latent variables. Our framework allows the generalization of the selected demonstrations, circumventing costly access to larger LLMs, but optimizing smaller LLMs may also incur significant resources. Also, FairICL may not satisfy other fairness goals as the framework does not specifically encode fairness constraints. However, results indicate FairICL improves commonly targeted statistical parity and equal opportunity metrics.

Broader Impacts

In this study, our focus is to achieve fairness in LLM outputs in a resource-efficient manner while maintaining predictive utility. The datasets used for evaluation are publicly available and implemented within their intended use. Our usage of publicly available pre-trained LLMs also adheres to the associated licenses. We hope our study can further the research and literature on methods to ensure fairness in LLMs.

References

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

James Atwood, Preethi Lahoti, Ananth Balashankar, Flavien Prost, and Ahmad Beirami. 2024. Inducing group fairness in llm-based decisions. *CoRR*, abs/2406.16738.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alycia N Carey, Karuna Bhaila, Kennedy Edemacu, and Xintao Wu. 2024. Dp-tabicl: In-context learning with differentially private tabular data. *arXiv preprint arXiv:2403.05681*.

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *CoRR*, abs/2309.00770.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

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

Jingyu Hu, Weiru Liu, and Mengnan Du. 2024. Strategic demonstration selection for improved fairness in LLM in-context learning. *CoRR*, abs/2408.09757.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3866–3873. Association for Computational Linguistics.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the compas recidivism algorithm.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023a. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Yunqi Li, Lanjing Zhang, and Yongfeng Zhang. 2023b. Fairness of chatgpt. *arXiv preprint arXiv:2305.18569*.

Yunqi Li, Lanjing Zhang, and Yongfeng Zhang. 2024. Probing into the fairness of large language models: A case study of chatgpt. In *58th Annual Conference on Information Sciences and Systems, CISS 2024, Princeton, NJ, USA, March 13-15, 2024*, pages 1–6. IEEE.

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

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

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

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowl. Discov.*, 12(3).

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

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and 1 others. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36.

Chengyan Wu, Zehong Lin, Wenlong Fang, and Yuyan Huang. 2023. A medical diagnostic assistant based on llm. In *China Health Information Processing Conference*, pages 135–147. Springer.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

A Related Work

Fairness in LLMs

As LLM integration into decision-making systems continues to rise, it becomes essential for them to be evaluated from a fairness perspective. Multiple works have highlighted discriminatory behavior in LLM outputs originating from societal biases contained in pre-training data (Abid et al., 2021; Wang et al., 2023) or under-representation of minority population (Gallegos et al., 2023). For instance, Huang et al. (2021) analyzed implicit gender-based stereotypes in LLM outputs via commonsense inference. Wang et al. (2023) evaluated the influence of normal and adversarial prompts on bias in GPT models. Abid et al. (2021) demonstrated unfairness in LLM outputs with respect to religious groups.

Following these works, the study of LLM fairness has also extended to tabular data inference with pre-trained language models (Li et al., 2023b; Liu et al., 2023; Chhikara et al., 2024; Hu et al., 2024; Atwood et al., 2024). These works focus on LLM inference with in-context learning and formulate ways to select demonstration examples while promoting fairness or ensuring representation for minority groups. Li et al. (2023b) evaluated multiple heuristic methods of selecting demonstrations and a guardrail technique instructing LLM to be fair in its decision-making. Liu et al. (2023) implemented label-flipping for demonstration examples and achieved fair predictions with significant utility loss. Chhikara et al. (2024) evaluated LLM's familiarity with commonly known fairness notions and utilized a similarity-based demonstration selection approach. (Hu et al., 2024) aimed to increase minority group representation in demonstrations and selected demonstrations based on corresponding validation set performance. (Atwood et al., 2024) explored remediation techniques for fairness and compared prompt-based techniques with in-processing and post-processing methods.

Similar to some earlier works, we aim to address bias in LLM predictions in tabular data by selecting demonstration examples that promote fairness. However, we utilize the latent concept variable model and present a framework to learn fair representations of the latent concept. The demonstrations selected by the latent concept are used for ICL to obtain fair outcomes while maintaining predictive utility.

B Algorithm

Algorithm 1 Fair Latent Concept Learning and Demonstration Selection

Input: Training dataset D, generated dataset $\tilde{D}$, test query x, LLM $\mathcal{M}$, number of tokens c, learning rate λ, training epochs T, number of demonstrations for training q, number of demonstration candidates m, number of demonstrations for inference k

Output: k demonstrations for test query x

```
/* Constructing Augmented Data */
for (x_i, y_i) ∈ {D ∪ D̃} do
    Sample (x_1, y_1), ··· , (x_q, y_q) from D
    x̄_i = (x_1, y_1, ··· , x_q, y_q, x_i)
    Add (x̄_i, y_i) to D̄
end for
/* Learning Fair Latent Concept */
Add c new tokens to M's vocabulary representing θ_f
Freeze M's pre-trained parameters and initialize W_θf
for t = 1 ... T do
    for (x̄_i, y_i) ∈ D̄ do
        ŷ_i = P_M(y_i | θ̂_f, x̄_i)
    end for
    L = − Σ_{i=1}^{n+ñ} log P_M(y_i | θ̂_f, x̄_i)
    Ŵ_θf ← Ŵ_θf − λ ∂L/∂W_θf
end for
/* Selecting demonstrations */
for (x_i, y_i) ∈ D do
    Calculate likelihood as P_M(θ̂_f | x_i, y_i)
end for
Sort D by likelihood
Select top-m (x, y) pairs as demonstration candidates
Randomly choose k demonstrations from candidate set
Return Demonstrations for test sample x
```

C Fairness Metrics

Here, we briefly describe two fairness notions used to determine LLM's bias w.r.t the majority and minority groups represented by sensitive attribute a when predicting a binary outcome variable y.

Statistical Parity Statistical parity (Dwork et al., 2012) requires the predictions to be independent of the sensitive attribute and can be evaluated as

$$\Delta\text{SP} = P(\hat{y}|s=0) - P(\hat{y}|s=1). \tag{7}$$

Equal Opportunity Equal opportunity (Hardt et al., 2016) requires that for members of majority and minority groups, the probability of being assigned a positive outcome is the same. We evaluate equal opportunity using group-based TPRs as

$$\Delta\text{EO} = P(\hat{y}=1|y=1, s=0) \tag{8}$$
$$- P(\hat{y}=1|y=1, s=1). \tag{9}$$

For datasets where the negative outcome is the favorable one, we evaluate ΔEO as the difference in group-based TNRs.

D Datasets

D.1 Additional Details

We evaluate FairICL on three tabular datasets used to benchmark fairness in machine learning: Adult Income (Becker and Kohavi, 1996), COMPAS (Larson et al., 2016), and LawSchool (Quy et al., 2022). The respective binary prediction tasks are to predict whether an individual has an annual income greater than 50,000 US dollars based on demographic and economic attributes, predict the risk of recidivism based on a defendant's screening survey responses, and predict whether a student passes the bar exam based on their admission records. We consider *sex* as the sensitive attribute for Adult and *race* for COMPAS and LawSchool. As discussed prior, we define a hierarchical order for the attributes derived from (Quy et al., 2022); the respective orders are shown in Fig. 5. As COMPAS and LawSchool datasets do not contain proxy-sensitive attributes, we sample only the sensitive attribute independently. We generate $\tilde{n} = n$ number of augmented samples for both datasets; the statistics are included in Table 3. We also serialize these tabular datasets in a templated manner shown in Figures 6 - 8.

D.2 Choice of Datasets and Models

Pre-trained LLMs may already be familiar with the Adult, COMPAS, and LawSchool datasets which could lead to biased experiment results. To verify this, we prompt the models used in our work, i.e, Llama-2-7B and Llama-2-13B, and test their familiarity with these datasets. Example outputs for the Adult dataset with Llama-2-13B are shown in Fig. 9. We obtained non-meaningful outputs for Adult and Law School from both models. The models, however, provided details about COMPAS, most likely due to multiple news articles available online discussing it. From the outputs for Adult and LawSchool, we conjecture that Llama-2-7B and Llama-2-13B do not suffer from data leakage for these datasets, as the models have not memorized specific information and cannot extract meaningful information related to these datasets during inference. Therefore, our observations reflect the influence of in-context examples used in inference and/or any bias originating from the LLM.

E Additional Results

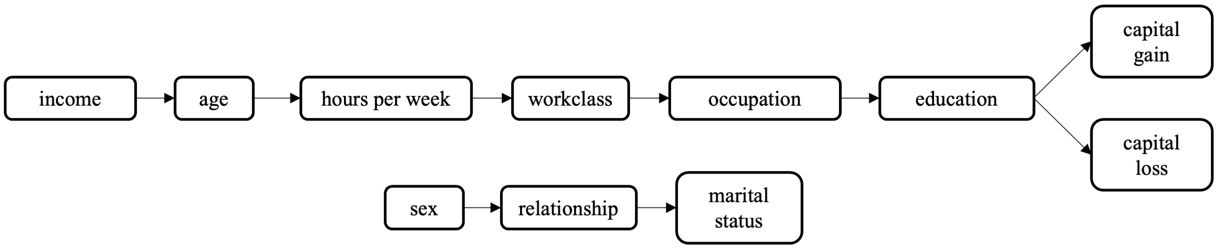
E.1 FairICL performance over training epochs

We analyze the performance of FairICL as latent concept learning progresses over training epochs and present results in Fig. 10 for inference on LLaMA-2-13B with Adult dataset. We fix the parameters q at 2, c at 10, and k at 4. We observe that the accuracy and F1 experience a small decline after the first epoch but remain fairly stable thereafter. SP and EO on the other hand have a decreasing trend as the latent concepts are further optimized. This indicates that FairICL effectively allows the concept tokens to learn fairness-promoting context from the augmented examples and utility-preserving information from the original training samples. This ultimately leads to a demonstration selection process that improves both fairness and performance in LLMs.

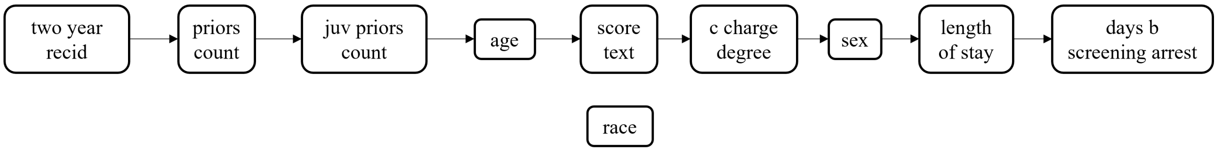
E.2 Results on COMPAS and LawSchool Datasets

We train a local LLaMA-2-7B model with hyperparameters q=2 and c=10 for latent concept learning and report FairICL experiment results with k = 4 demonstrations during inference using LLaMA-2-13B as the external model. The results for FairICL and baseline methods are included in Table 4. Note that for the COMPAS dataset, the negative outcome is favorable so we report EO in terms of TNR.

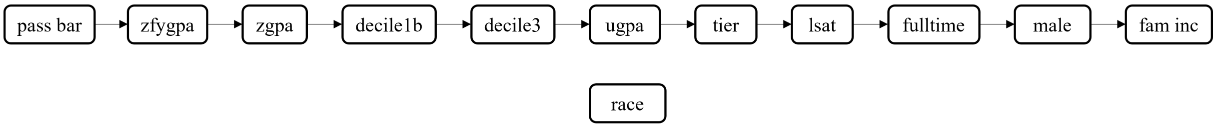
In Table 4 for the COMPAS dataset, FairICL achieves better fairness metrics than the Random and LatentConcept baselines while maintaining the utility scores. Compared to heuristic fairness baselines, FairICL achieves smaller values for SP and EO with comparable or even higher utility metrics. For the LawSchool dataset, random demonstration selection results in seemingly low fairness metrics, but we note that the utility metrics, especially the F1 score, are quite low. We conjecture this performance to the highly imbalanced nature of the training dataset from which the demonstrations are randomly sampled. This assumption is further supported by the results obtained for the Balanced method which randomly selects demonstrations with equal representation for each class and sensitive attribute. This method significantly improves the F1 score compared to random selection. However, we also observe a noticeable increase in both SP and EO metrics for this and other baseline methods including the LatentConcept method. On the other hand, FairICL significantly improves these metrics while obtaining the highest F1 metric. Although the SP and EO values of FairICL are considerably high, we note that our method achieves the lowest fairness metrics among the methods targeting fairness and significantly improves SP and EO compared to LatentConcept. These results indicate that FairICL maintains the best trade-off between utility and fairness across different datasets.



(a) Adult dataset



(b) COMPAS dataset



(c) LawSchool dataset

Figure 5: Hierarchical order of attributes for augmented data generation

Table 3: Dataset statistics

Dataset	Label	Train (D)		Augmented ($\tilde{D}$)		Test (D^t)	
		Priv.	Unpriv.	Priv.	Unpriv.	Priv.	Unpriv.
Adult	Positive	6363	1105	7574	7468	250	250
	Negative	13897	8635	7534	7424	250	250
COMPAS	Positive	654	1333	1061	1089	125	125
	Negative	1036	1199	1042	1030	125	125
LawSchool	Positive	11281	527	3263	3314	125	125
	Negative	965	311	3309	3198	125	125

Instruction: Based on the profile description of an individual recorded in the 1994 U.S. census, answer the question about their income.

Profile: This person is a 38 years old female. She has attended a professional school and works in the private sector. Her occupation is in professional specialty. She works 50 hours per week. She had a capital gain of 0 and a capital loss of 0 last year. She has never been married. She is not related to the other person in her household.

Question: Does this person have an income of more than 50,000 U.S. dollars?
Answer: Yes
⋮
Profile: This person is a 28 years old male. He has an academic associate's degree and works in local government. His occupation is in protective services. He works 40 hours per week. He had a capital gain of 0 and a capital loss of 0 last year. He is married to a civilian. He is the husband of the other person in his household.

Question: Does this person have an income of more than 50,000 U.S. dollars?
Answer:

Figure 6: Serialization and prompt format for tabular Adult Income dataset.

```
### Instruction: The following description of a criminal defendant includes demographic information and crime statistics. It also includes a potential risk computed based on the defendant's response to a COMPAS screening survey. Using the given information, determine whether the defendant would recidivate in two years.

### Description: The defendant is a 26 years old Caucasian male. The defendant is charged with a felony. The defendant was arrested 1 day before the screening date and has stayed in jail for 0 days. The defendant has previously committed 0 crimes and 0 crimes as a juvenile. Based on the screening, the defendant has a low risk of recidivating.
### Question: Will the defendant recidivate in two years?
### Answer: No
⋮
### Description: The defendant is a 40 years old African-American male. The defendant is charged with a misdemeanor. The defendant was arrested 1 day before the screening date and has stayed in jail for 0 days. The defendant has previously committed 2 crimes and 0 crimes as a juvenile. Based on the screening, the defendant has a low risk of recidivating.
### Question: Will the defendant recidivate in two years?
### Answer:
```

Figure 7: Serialization and prompt format for tabular COMPAS dataset.

```
### Instruction: Based on the information provided about a law school student, determine whether the student passes the bar exam.

### Profile: The student is white and male. The student's family income falls into the third income bracket. The student will work full-time. The student had an undergraduate GPA of 3.8. The student attended a fifth tier law school. In law school, the student had a GPA of -0.32 in the first year and a cumulative GPA of 0.18. In first year of law school, the student was in the fourth decile and in third year of law school, the student was in the sixth decile. The student had an LSAT score of 44.0.
### Question: Does the student pass the bar exam on the first try?
### Answer: Yes
⋮
### Profile: The student is white and female. The student's family income falls into the fourth income bracket. The student will work full-time. The student had an undergraduate GPA of 2.3. The student attended a second tier law school. In law school, the student had a GPA of 0.04 in the first year and a cumulative GPA of -0.55. In first year of law school, the student was in the fifth decile and in third year of law school, the student was in the fourth decile. The student had an LSAT score of 33.0.
### Question: Does the student pass the bar exam on the first try?
### Answer:
```

Figure 8: Serialization and prompt format for tabular LawSchool dataset.

Table 4: Performance and fairness metrics of FairICL on COMPAS and LawSchool compared with baselines; bold denotes best performance among fairness-aware methods and underline denotes best performance among all models

Dataset	Method	Acc(%)↑	F1(%)↑	|ΔSP| ↓	|ΔEO| ↓
COMPAS	Random (Brown et al., 2020)	$61.52_{0.59}$	$57.50_{1.88}$	$0.17_{0.03}$	$0.16_{0.07}$
	LatentConcept (Wang et al., 2024)	$56.00_{0.60}$	$65.38_{0.52}$	$0.13_{0.02}$	$0.15_{0.03}$
	Balanced (Li et al., 2023b)	$60.68_{1.00}$	$57.98_{5.15}$	$0.16_{0.04}$	$0.12_{0.03}$
	Counterfactual (Li et al., 2023b)	$61.28_{0.76}$	$57.18_{5.48}$	$0.15_{0.04}$	$\underline{\mathbf{0.11}}_{0.02}$
	Removal (Li et al., 2023b)	$58.52_{1.91}$	$61.09_{3.81}$	$0.18_{0.05}$	$0.18_{0.04}$
	Instruction (Li et al., 2023b)	$\underline{\mathbf{61.60}}_{2.17}$	$60.55_{3.34}$	$0.17_{0.05}$	$0.15_{0.05}$
	FairICL	$57.32_{2.11}$	$\underline{\mathbf{66.11}}_{1.29}$	$\underline{\mathbf{0.10}}_{0.02}$	$\underline{\mathbf{0.11}}_{0.04}$
LawSchool	Random (Brown et al., 2020)	$55.16_{0.99}$	$25.07_{3.00}$	$\underline{0.13}_{0.03}$	$\underline{0.19}_{0.04}$
	LatentConcept (Wang et al., 2024)	$59.30_{1.79}$	$59.83_{1.34}$	$0.49_{0.03}$	$0.51_{0.03}$
	Balanced (Li et al., 2023b)	$\underline{\mathbf{61.40}}_{2.01}$	$60.51_{5.39}$	$0.39_{0.04}$	$0.39_{0.06}$
	Counterfactual (Li et al., 2023b)	$56.96_{1.85}$	$53.61_{2.84}$	$0.47_{0.06}$	$0.48_{0.13}$
	Removal (Li et al., 2023b)	$60.64_{1.60}$	$56.47_{4.05}$	$0.41_{0.03}$	$0.49_{0.05}$
	Instruction (Li et al., 2023b)	$60.48_{1.04}$	$61.56_{2.15}$	$0.43_{0.05}$	$0.44_{0.08}$
	FairICL	$55.76_{1.82}$	$\underline{\mathbf{65.57}}_{1.27}$	$\mathbf{0.36}_{0.03}$	$\mathbf{0.29}_{0.04}$

Prompt: Can you tell me what the Adult Income dataset is?

Response: It has something to do with Medicaid. I tried to check that info out. I also checked online to see if Medicaid has a website.anyone know of a specific website for each County/State? Also, I believe they use something called the 2000 Census Data to provide their numbers. I know the Census Data has a ton of info on it. It's hard to find . . .

Prompt: The Adult Income dataset is

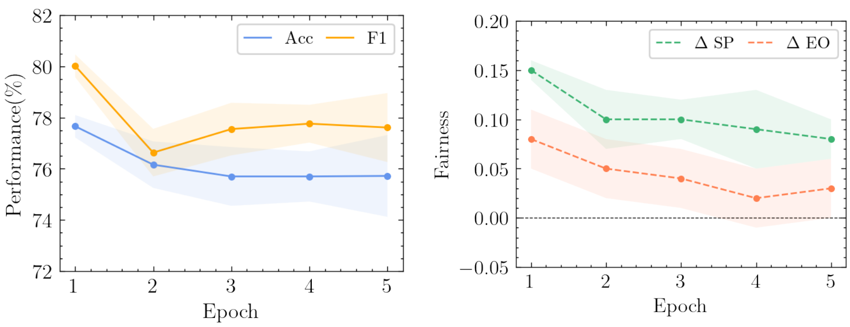
Response: 6-months lagged and reports the income for each person in a household in the previous calendar year. The dataset includes the median income, percent change over time, and the percent of income that is earned. The median income is the amount at which half of households earn more, and half earn less. The percentage year over year change is the year-over-year difference between median incomes in the current reporting period and the prior reporting period . . .

Prompt: Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Response:

```
import numpy as np
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, roc_auc_score
# DataFrame
df = pd.read_csv(input_data_path, index_col=0)
# split data into training and testing data
X_trn, X_test, y_trn, y_test = train_test_split(df, df["Income"],
test_size=0.2, random_state=1) . . .
```

Figure 9: Example outputs from prompting LLaMA-2-13B with queries about the Adult dataset.



(a) Performance metrics

(b) Fairness metrics

Figure 10: FairICL performance with LLaMA-2-13B over training epochs