Fair-CCD: Mitigating Bias in Large Language Models for Tabular Classification Through Context-Contrastive Decoding

Donghan Liu[1], **Han Sun**[1], **Zhaohui Wang**[1], **Qin Li**[1*], **Min Zhang**[1*]
[1]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
51275902146@stu.ecnu.edu.cn, hsun@stu.ecnu.edu.cn
sternstund22@gmail.com, qli@sei.ecnu.edu.cn, mzhang@sei.ecnu.edu.cn

Abstract

While recent studies show the effectiveness of in-context learning (ICL) for tabular data prediction, they also reveal significant fairness issues in large language models (LLMs). Prior work to mitigate fairness issues often employs interventions relying on subjective demonstration selection. Its effectiveness varies significantly with the specific demonstration content, leading to low controllability. Moreover, the improvement of fairness is highly unstable across different models and tasks. To address the challenges of low controllability and limited stability in fairness interventions, we propose **Fair**ness-Aware **C**ontext-**C**ontrastive **D**ecoding (**Fair-CCD**). Fair-CCD first constructs Structural Bias Templates (SBTs), motivated by behavioral patterns observed in demonstrations, to encode the relationship between sensitive attributes and predicted labels in a structured and controllable form. During inference, Fair-CCD injects multiple SBTs and contrasts the model's responses, generating two differential signals that guide fairness adjustment and preserve task performance. By leveraging attention signals to scale decoding adjustments guided by the difference signals, Fair-CCD achieves stable and adaptive bias mitigation across models and tasks. Extensive experimental results demonstrate that Fair-CCD consistently improves fairness metrics without degrading task accuracy.

1 Introduction

In recent years, large language models (LLMs) have achieved remarkable progress in natural language processing and multimodal learning, and have been increasingly applied to structured tabular prediction tasks (Hegselmann et al., 2023) such as income classification, credit scoring, and recidivism risk assessment (Slack and Singh, 2023). However, these models often inherit social biases embedded in their training data (Abid et al., 2021a; Bender et al., 2021). In particular, when sensitive attributes such as gender or race are involved, they may produce systematically biased predictions against certain demographic groups, raising serious concerns about fairness (Liu et al., 2023; Zhang et al., 2025a; Yang et al., 2024).

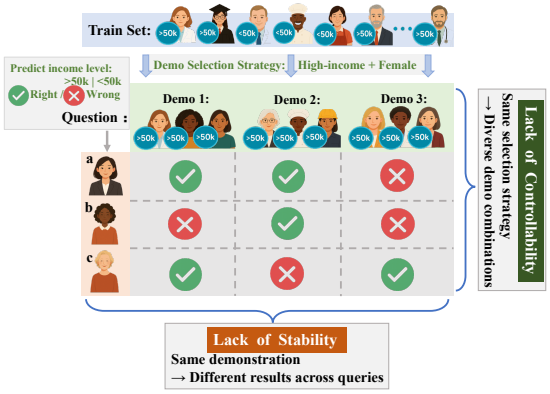

Figure 1: Demonstration-based methods suffer from limited controllability and stability.

To address this issue, recent work has leveraged in-context learning (ICL) by inserting targeted demonstrations into the prompt to steer the model's predictions (Hu and Du, 2024; Halim et al., 2025; Wang et al., 2024). However, such approaches face two fundamental challenges illustrated in Figure 1.

First, these methods lack controllability (Voronov et al., 2024; Dong et al., 2024). Even with fixed demonstration selection rules, different valid demonstrations often lead to inconsistent model behavior. This is because each demonstration contains additional attributes, such as occupation, education, or age, which may be correlated with the target label and influence the model's prediction. These confounding factors cannot be removed or isolated, making it difficult to ensure that the intervention works as intended. As a result, the effectiveness of the intervention is shaped by uncontrolled content within the demonstrations, fundamentally limiting the

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28561–28575

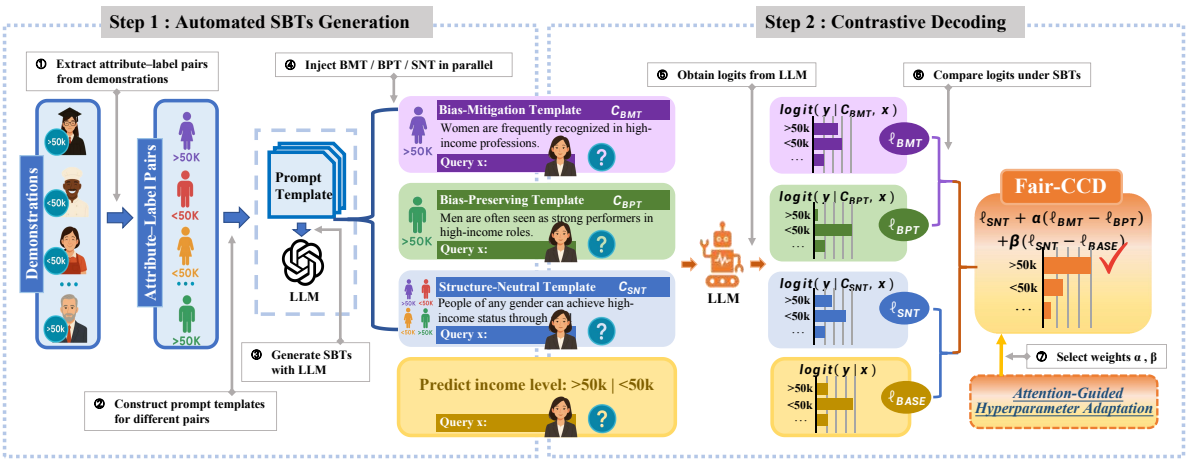

Figure 2: **The Framework of Fair-CCD on the Adult Dataset.** Fair-CCD improves the controllability of bias mitigation by introducing Structural Bias Templates (SBTs) and automating their generation, using real demonstrations sampled from the dataset. It constructs two contrastive signals from different SBTs, with one used for fairness adjustment and the other for accuracy preservation, enabling stable bias mitigation during inference.

controllability of demonstration-based methods.

Second, these methods suffer from limited stability (Zhang et al., 2025b; Cobbina and Zhou, 2025). Even when the same set of demonstrations is used, changing their order in the prompt can lead to different model predictions (Xiang et al., 2024). This is because LLMs tend to assign more weight to demonstrations that appear closer to the query, causing different demonstrations to dominate under different configurations. In addition, the same prompt can produce inconsistent results across models or tasks due to differences in pretraining data, decoding behavior, or domain-specific biases. These factors make the intervention effects difficult to reproduce and generalize, highlighting the inherent instability of demonstration-based approaches.

To address the controllability and stability issues of demonstration-based fairness interventions in ICL, we propose **Fair**ness-Aware **C**ontext-**C**ontrastive **D**ecoding (**Fair-CCD**), as shown in Figure 2, a two-step inference time framework. **(1)** We first introduce Structural Bias Templates (SBTs) to provide a standardized and controllable representation of attribute–label relationships, along with an automated construction mechanism. **(2)** We secondly design a contrastive decoding strategy with attention-guided parameter scaling, which determines the direction of bias adjustment from differential SBT responses and adaptively controls its strength using attention signals, improving the stability of bias mitigation.

We evaluate Fair-CCD on two widely used structured fairness benchmarks, Adult (1996) and COMPAS (2016). Compared with five baselines, Fair-CCD consistently achieves stronger bias mitigation while preserving predictive accuracy, demonstrating its stability and effectiveness across models and tasks.

Our main contributions are as follows:

- We introduce SBTs, which explicitly model the attribute–label structure underlying biased predictions. This approach offers a clearer alternative to demonstration selection and improves the controllability of fairness interventions.
- We propose Fair-CCD, a decoding-time method for bias mitigation. It contrasts predictions under multiple SBTs to determine the adjustment direction, and uses attention signals to adaptively control the adjustment strength at the instance level. This design enhances bias mitigation stability through the disentanglement of adjustment direction and magnitude.
- We show that Fair-CCD consistently outperforms five state-of-the-art baselines on Adult and COMPAS, delivering more stable and controllable bias mitigation without sacrificing accuracy.

2 Method

Fair-CCD consists of two key steps that respectively address controllability and stability in prior fairness interventions.

2.1 Structural Bias Templates Generation

Existing bias mitigation methods based on ICL typically adjust model behavior by incorporating demonstrations from specific demographic groups, such as adding positive examples of minority groups to reduce bias. However, these methods often suffer from controllability, as even demonstrations with the same attribute configurations can yield inconsistent model responses due to variations in semantic content. This observation leads us to hypothesize that model behavior may be influenced less by surface-level semantics and more by the attribute–label relationships implicitly encoded in the input. Such structural preferences may stem from pretraining data, where certain attribute–label pairs co-occur more frequently (Kossen et al., 2024). For instance, if "male–high income" appears more often than "female–high income" during pretraining, the model may internalize this imbalance as a default association. Given this hypothesis, explicitly modeling such attribute–label structures could enable more consistent and controllable interventions into model behavior.

To operationalize this idea, we propose to abstract the structural combinations of sensitive attributes and target labels into a controllable form, termed **Structural Bias Templates (SBTs)**. Rather than relying on the specific semantics of individual demonstrations, these templates represent structural attribute–label patterns through consistent textual forms, enabling direct and controllable interventions in model predictions.

We formalize each SBT as a triplet:

$$\tau = \langle a,\ y,\ s \rangle, \quad a \in \mathcal{A},\ y \in \mathcal{Y},\ s = f(a, y) \in \mathcal{S} \tag{1}$$

where $\mathcal{A}$ is the set of sensitive attribute values, $\mathcal{Y}$ the set of target labels, and $\mathcal{S}$ the set of linguistic templates. The function $f : \mathcal{A} \times \mathcal{Y} \to \mathcal{S}$ maps each attribute–label pair into a standardized natural language sentence s, which serves as a controllable input template.

$$\begin{aligned} \text{BMT} &= \{\, \tau \mid a \in \mathcal{A}_{\text{min}},\ y \in \mathcal{Y}_{\text{pos}},\ s = f(a, y) \,\}, \\ \text{BPT} &= \{\, \tau \mid a \in \mathcal{A}_{\text{maj}},\ y \in \mathcal{Y}_{\text{pos}},\ s = f(a, y) \,\}, \\ \text{SNT} &= \{\, \tau \mid a \in \mathcal{A},\ y \in \mathcal{Y},\ s = f(a, y) \,\} \end{aligned} \tag{2}$$

To specify the construction domains of different template types, we define the following subsets:

- $\mathcal{A}_{\text{min}} \subseteq \mathcal{A}$: the set of minority groups (e.g., *female*, *African American*),
- $\mathcal{A}_{\text{maj}} \subseteq \mathcal{A}$: the set of majority groups (e.g., *male*, *White*),
- $\mathcal{Y}_{\text{pos}} \subseteq \mathcal{Y}$: the set of favorable outcome labels (e.g., *high income*, *low risk*).

We categorize SBTs into three types based on the underlying attribute-label configurations. (1) **Bias-Mitigation Templates (BMTs)** are designed to introduce positive attribute-label associations for minority groups, encouraging the model to make fairer predictions. (2) **Bias-Preserving Templates (BPTs)** are constructed to represent attribute-label pairs where majority groups are associated with positive outcomes, thereby reinforcing the model's default bias tendencies. (3) **Structure-Neutral Templates (SNTs)** cover the full range of attribute-label combinations in a balanced and semantically minimal form, serving as neutral prompts to stabilize predictions and reduce structural sensitivity. The formal definitions of the three template sets are given in Equation 2.

All templates are constructed through a unified and automated mechanism, differing only in their structural domains. By abstracting away instance-level semantics, these templates provide enhanced consistency, transferability, and controllability compared to raw demonstrations, serving as a principled interface for fairness-aware interventions (see Appendix A for detailed generation procedures).

2.2 Bias Mitigation with Contrastive Decoding

Different SBTs guide the model in distinct directions: BMT favors fairer outcomes for minority groups, BPT amplifies favorable predictions for majority groups, and SNT acts as a neutral baseline. However, across tasks, model architectures, and data distributions, no single SBT consistently achieves optimal fairness, indicating that stability remains a key challenge despite improved input controllability.

These findings motivate our design in two parts: a contrastive decoding strategy that leverages structural response differences to determine adjustment directions, and an attention-guided scaling mechanism that modulates the influence of these differences at inference time.

2.2.1 Contrastive Decoding Strategy

To overcome the instability of relying on a single SBT, we introduce a contrastive decoding strat-

egy that contrasts model responses under different SBTs and uses the resulting differences to adjust prediction outputs in a fairness-aware manner.

To implement Fair-CCD at inference time, we operate at the token level. At each generation step t, a language model produces a logits vector $\ell_t \in \mathbf{R}^{|V|}$ over the vocabulary V , which is converted into a next-token distribution via softmax. Traditional decoding relies solely on these logits and offers no explicit control over biased behavior.

To introduce structural guidance, Fair-CCD prepends Structural Bias Templates (SBTs) as contextual inputs and performs decoding in parallel under different structural conditions. Specifically, we inject Bias-Mitigation (BMT), Bias-Preserving (BPT), and Structure-Neutral (SNT) templates to obtain three corresponding logits vectors, denoted as $\ell_{\text{BMT}}, \ell_{\text{BPT}}, \ell_{\text{SNT}}$. We additionally retain the base logits ℓ_{BASE} computed without any template as a reference for perturbation regularization.

The final prediction distribution is computed through a contrastive fusion of logits:

$$\hat{p}(y_t) = \text{softmax}\left[\begin{array}{c} \ell_{\text{SNT}} + \alpha \cdot (\ell_{\text{BMT}} - \ell_{\text{BPT}}) \\ +\beta \cdot (\ell_{\text{SNT}} - \ell_{\text{BASE}}) \end{array}\right] \tag{3}$$

where α and β are scaling coefficients. This decoding mechanism consists of three components:

- **Backbone Generation** (ℓ_{SNT}): provides semantically neutral structural guidance to ensure generation stability and fluency;
- **Bias Contrast** ($\ell_{\text{BMT}} - \ell_{\text{BPT}}$): captures disparities between minority- and majority-associated contexts and serves as the primary control signal for generation-time bias adjustment;
- **Perturbation Regularization** ($\ell_{\text{SNT}} - \ell_{\text{BASE}}$): penalizes excessive deviation from the model's original prediction, maintaining output consistency.

Unlike traditional demonstration-based or static prompting approaches, Fair-CCD shifts the focus from semantic content to structural contrast. By directly leveraging logit-level differences induced by SBTs, Fair-CCD identifies explicit adjustment directions for bias mitigation, enabling a controllable, interpretable, and training-free decoding-time intervention.

2.2.2 Attention-Guided Hyperparameter Adaptation

Contrastive logit differences provide clear directions for bias mitigation, but using fixed coefficients can lead to either insufficient correction or excessive perturbation (Sun et al., 2026). An instance-wise mechanism is therefore needed to adaptively control the strength of these contrastive terms at inference time, as illustrated in Figure 3.

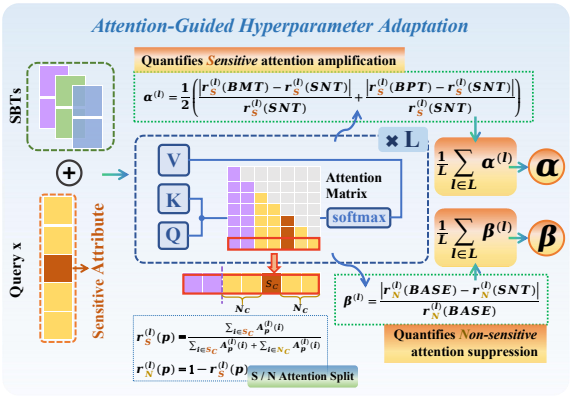

Figure 3: Changes in the final-row attention weights over sensitive token groups scale the Bias Contrast logit term, while changes over non-sensitive token groups scale the Perturbation Regularization logit term.

At each decoding layer l, we take the final-row attention vector $A^{(l)}$ (corresponding to the current query token) and compute attention statistics over two base-content token sets: the sensitive-token group S_C and the non-sensitive group N_C. Here S_C and N_C are defined only on the original input content (the base prompt without SBTs) and are kept fixed across all SBT-injected contexts to ensure aligned and comparable measurements.

These quantities characterize the relative influence of sensitive and non-sensitive content in the original input on the current prediction. Based on this partition, we compute attention-based influence estimates for the sensitive and non-sensitive token groups as

$$\begin{aligned} r_S^{(l)} &= \frac{\sum_{i \in S_C} A^{(l)}(i)}{\sum_{i \in S_C} A^{(l)}(i) + \sum_{i \in N_C} A^{(l)}(i)}, \\ r_N^{(l)} &= 1 - r_S^{(l)}. \end{aligned} \tag{4}$$

These influence estimates are computed separately under each structural context (i.e., BASE, BMT, BPT, and SNT).

Based on these attention-based influence estimates, we compute two layer-wise scaling coefficients $\alpha^{(l)}$ and $\beta^{(l)}$ to control the strength of the

contrastive terms. The coefficient $\alpha^{(l)}$ quantifies how strongly the Bias Contrast $(\ell_{\text{BMT}} - \ell_{\text{BPT}})$ is driven by attention to sensitive attribute–related content, measured as the relative deviation of sensitive-token influence under BMT and BPT from the structure-neutral context (SNT):

$$\alpha^{(l)} = \frac{1}{2}\left(\frac{\left|r_S^{(l)}(\text{BMT}) - r_S^{(l)}(\text{SNT})\right|}{r_S^{(l)}(\text{SNT})} + \frac{\left|r_S^{(l)}(\text{BPT}) - r_S^{(l)}(\text{SNT})\right|}{r_S^{(l)}(\text{SNT})}\right) \tag{5}$$

Similarly, $\beta^{(l)}$ quantifies how strongly the Perturbation Regularization $(\ell_{\text{SNT}} - \ell_{\text{BASE}})$ is driven by changes in non-sensitive-token influence, measured as the relative deviation between the base prediction and the structure-neutral context:

$$\beta^{(l)} = \frac{\left|r_N^{(l)}(\text{BASE}) - r_N^{(l)}(\text{SNT})\right|}{r_N^{(l)}(\text{BASE})} \tag{6}$$

Finally, we compute the inference-time scaling factors by aggregating and normalizing the layer-wise coefficients:

$$\alpha = \frac{\sum_{l\in L}\alpha^{(l)}}{\sum_{l\in L}\alpha^{(l)} + \sum_{l\in L}\beta^{(l)}}, \quad \beta = \frac{\sum_{l\in L}\beta^{(l)}}{\sum_{l\in L}\alpha^{(l)} + \sum_{l\in L}\beta^{(l)}} \tag{7}$$

3 Experimental Setup

Our experiments are designed to address three main research questions: **RQ1:** *Is the structural relationship between sensitive attributes and target labels a key factor in shaping biased model behavior?* **RQ2:** *Can Fair-CCD improve both the controllability and stability of bias mitigation across different tasks and models, compared to demonstration-based interventions?* **RQ3:** *How do the core components of Fair-CCD contribute to its overall effectiveness in controlling biased predictions?*

In this section, we introduce the overall experimental setup.

3.1 Datasets and Metrics

We evaluate fairness using two widely adopted structured datasets: Adult (Becker and Kohavi, 1996) and COMPAS (Angwin et al., 2016), which assess model behavior with respect to gender and race, respectively. The Adult dataset involves predicting whether an individual's annual income exceeds \$50K, where the sensitive attribute is gender (male or female). The COMPAS dataset focuses on predicting the likelihood of recidivism among criminal defendants, using race (Caucasian or African-American) as the sensitive attribute.

	Method	Prediction		Fairness			
		$\text{Acc}_{\to 1}$	$\text{F1}_{\to 1}$	$R_{\text{EO}\to 1}$	$\Delta_{\text{EO}\to 0}$	$R_{\text{SP}\to 1}$	$\Delta_{\text{SP}\to 0}$
Qwen2.5-7B	Zero-shot	0.649	0.633	0.669	0.300	0.592	0.369
	FADS	0.623	0.616	0.713	0.233	0.651	0.280
	FCG	0.646	0.628	0.760	0.211	**0.690**	**0.245**
	JUDGE	0.641	0.618	0.759	0.223	0.677	0.270
	Fair-CCD	**0.652**	**0.637**	**0.771**	**0.196**	0.685	0.253
Qwen2.5-14B	Zero-shot	0.653	0.640	0.716	0.248	0.587	0.331
	FADS	0.634	0.636	0.778	0.187	0.767	0.178
	FCG	0.646	0.646	0.728	0.202	0.620	0.244
	JUDGE	0.650	0.640	0.768	0.193	0.642	0.243
	Fair-CCD	**0.709**	**0.700**	**1.061**	**-0.017**	**0.917**	**0.053**
Mistral-7B	Zero-shot	0.662	0.641	0.772	0.207	0.627	0.315
	FADS	0.463	0.518	1.647	-0.259	1.469	-0.297
	FCG	0.642	0.635	0.750	0.199	0.639	0.257
	JUDGE	0.662	0.651	0.768	0.158	0.668	0.213
	Fair-CCD	**0.705**	**0.691**	**0.869**	**0.105**	**0.772**	**0.169**
LLaMA3-8B	Zero-shot	0.581	0.574	0.180	0.526	0.135	0.479
	FADS	0.595	0.675	0.554	0.287	0.472	0.192
	FCG	0.488	0.412	2.541	-0.106	1.565	-0.038
	JUDGE	0.560	0.523	1.435	-0.238	1.473	-0.389
	Fair-CCD	**0.687**	**0.734**	**0.928**	**0.048**	**0.772**	**0.067**
LLaMA3-11B	Zero-shot	0.610	0.610	0.321	0.536	0.249	0.534
	FADS	0.572	0.511	0.677	0.239	0.670	0.238
	FCG	0.486	0.418	3.020	-0.139	1.672	-0.053
	JUDGE	0.601	0.588	0.757	0.194	0.733	0.211
	Fair-CCD	**0.628**	**0.613**	**0.927**	**0.054**	**0.819**	**0.111**

Table 1: Results on the **COMPAS** dataset. Bold denotes the best result for each model.

Model performance is assessed using six metrics: two for prediction quality and four for fairness. Accuracy and F1 Score are used to measure overall predictive performance. Fairness is evaluated based on Statistical Parity (SP) and Equality of Opportunity (EO), each measured in terms of both difference and ratio, yielding four fairness metrics:Δ_{SP}, R_{SP}, Δ_{EO}, R_{EO}. Smaller values of Δ_{SP} and Δ_{EO}, as well as ratios R_{SP} and R_{EO} closer to 1, indicate fairer predictions across demographic groups. Formal definitions of all metrics are provided in Appendix C.

3.2 Models and Baselines

To thoroughly evaluate the stability of Fair-CCD, we conduct experiments across five widely used LLMs spanning different families and parameter scales. Specifically, we include Mistral-7B-Instruct-v0.3 (AI, 2023), Qwen2.5-7B-Instruct

and Qwen2.5-14B-Instruct (Team, 2024), as well as LLaMA-3.1-8B-Instruct and LLaMA-3.2-11B-Vision-Instruct (AI, 2024). This selection covers three distinct model series and two model sizes (7B/8B and 11B/14B), aiming to test the stability of our method across a diverse range of architectures. To ensure consistent outputs, all models are evaluated with the temperature set to 0. The detailed prompt templates used in all experiments are provided in Appendix B.

We compare six approaches: Zero-shot, which provides only the task description and test input without any in-context demonstrations; FCG (Hu et al., 2024), which enhances fairness by selecting minority-group demonstrations via semantic clustering and genetic algorithms; FADS (Wang et al., 2024), which improves fairness through structured demonstration selection; Preamble (Oba et al., 2023), which suppresses gender bias by augmenting in-context prompts with counterfactual demonstrations; JUDGE (Halim et al., 2025), which performs fair demonstration selection via incremental greedy evaluation; and our proposed method, Fair-CCD.

4 Results

We organize our results to address the three research questions introduced in the experimental setup. First, we analyze model responses to different types of Structural Bias Templates (SBTs) and their corresponding demonstrations to assess whether the structural relationship between sensitive attributes and target labels plays a key role in shaping biased behavior. Second, we evaluate whether Fair-CCD improves both the controllability and stability of bias mitigation across different models and tasks, compared to demonstration-based baselines. Finally, we conduct component-wise ablations to examine how each part of the Fair-CCD framework contributes to its overall effectiveness in bias mitigation.

4.1 Empirical Validation of SBTs

To evaluate the effectiveness of SBTs in fairness interventions, we compare three context settings: (1) **Zero-shot(Baseline)**, where the model receives only the target query; (2) **Demonstration-based(Demo)**, using five groups of four matched demonstrations per attribute–label pair (a, y), with results averaged across samples; and (3) **SBT-based**, , using BMT, BPT, and SNT templates. All contexts are prepended as input prefixes, with a fixed target query. The example below is based on Qwen2.5-7B on Adult, full results are available in Appendix D.

	Method	Prediction		Fairness			
		$\mathrm{Acc}_{\to 1}$	$\mathrm{F1}_{\to 1}$	$R_{\mathrm{EO}\to 1}$	$\Delta_{\mathrm{EO}\to 0}$	$R_{\mathrm{SP}\to 1}$	$\Delta_{\mathrm{SP}\to 0}$
Qwen2.5-7B	Zero-shot	0.760	0.721	0.911	0.059	1.643	0.126
	FADS	0.743	0.680	0.918	0.052	1.578	0.113
	Preamble	0.751	0.710	0.921	0.051	1.619	0.119
	FCG	0.787	0.723	**0.999**	**0.001**	1.576	0.113
	JUDGE	0.576	0.416	1.737	-0.243	1.344	0.089
	Fair-CCD	**0.813**	**0.742**	1.039	-0.036	**1.042**	**0.018**
Qwen2.5-14B	Zero-shot	0.788	**0.788**	0.897	0.074	1.812	0.164
	FADS	0.733	0.642	0.863	0.089	2.021	0.139
	Preamble	0.772	0.748	0.921	0.061	1.775	0.156
	FCG	**0.797**	0.716	0.722	0.156	2.204	0.146
	JUDGE	0.789	0.735	0.791	0.109	1.737	0.149
	Fair-CCD	0.795	0.752	**1.065**	**-0.036**	**1.378**	**0.091**
Mistral-7B	Zero-shot	0.726	0.725	0.817	0.141	1.401	0.158
	FADS	0.731	0.735	0.882	0.099	1.050	0.022
	Preamble	0.711	0.709	0.784	0.144	1.477	0.153
	FCG	0.698	0.690	0.763	0.145	1.533	0.143
	JUDGE	0.708	0.712	0.937	0.045	1.102	0.064
	Fair-CCD	**0.757**	**0.747**	**1.015**	**-0.011**	**1.032**	**0.017**
LLaMA3-8B	Zero-shot	0.671	0.670	0.653	0.262	1.894	0.279
	FADS	0.684	0.667	0.664	0.260	1.788	0.253
	Preamble	0.642	0.649	0.776	0.151	1.243	0.038
	FCG	0.339	0.415	1.032	-0.008	1.202	0.041
	JUDGE	0.663	0.631	0.728	0.221	1.429	0.155
	Fair-CCD	**0.705**	**0.696**	**0.996**	**0.003**	**1.017**	**0.008**
LLaMA3-11B	Zero-shot	0.623	0.623	0.757	0.172	1.339	0.145
	FADS	0.638	0.657	0.824	0.098	1.173	0.068
	Preamble	0.465	0.438	1.337	-0.132	1.029	0.015
	FCG	0.573	0.550	0.972	0.023	1.044	0.031
	JUDGE	0.616	0.590	1.273	-0.094	1.101	0.056
	Fair-CCD	**0.666**	**0.669**	**0.989**	**0.008**	**1.003**	**0.002**

Table 2: Results on the **Adult** dataset. Bold denotes the best result for each model.

Experimental results support our hypothesis that structural attribute–label relationships determine the direction of bias behavior in in-context learning. As shown in Figure 4, when demonstrations and SBTs encode the same attribute–label configuration, they induce fairness shifts in the same direction relative to the zero-shot baseline, despite differences in semantic content. This directional consistency indicates that the observed bias effects are driven by structural information rather than instance-level semantics.

4.2 Overall Results on Bias Mitigation

Across both COMPAS and Adult, Fair-CCD consistently achieves the strongest bias mitigation while maintaining high accuracy, outperforming prior methods in most settings. Results are presented in Table 1 and Table 2.We attribute this advantage to its dual-level control: the mitigation direction is determined at the logits level, while the intervention strength is adaptively regulated via attention signals, enabling targeted and stable bias correction.

While existing demonstration-based methods also reduce bias, their improvements are generally smaller and less consistent, as the intervention remains driven by the selected demonstrations. Although some works explore instance-level demonstration selection, bias mitigation still depends on the semantic composition of examples. In contrast, Fair-CCD enables stable direction identification and adaptive strength control, resulting in more precise and controllable bias mitigation. Moreover, demonstration-based methods occasionally suffer from severe performance degradation due to overcorrection, such as on COMPAS with FADS on Mistral-7B and with FCG on LLaMA3-8B.

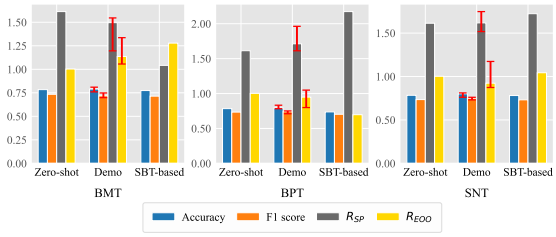

Figure 4: Prediction and fairness performance of Qwen2.5-7B on the *Adult* dataset under different context strategies. Error bars for the Demonstration-based method indicate variation across five sampled demonstrations.

Overall, Fair-CCD delivers stronger and more stable fairness improvements across datasets and model architectures without sacrificing predictive performance, highlighting its superior controllability and stability.

4.3 Ablation Study

4.3.1 Component-wise Ablation

To assess the role of each component within Fair-CCD, we conduct five comparative experiments on the Adult dataset. The evaluated settings include the standard Zero-shot baseline, the complete Fair-CCD method, two ablation variants where the bias contrast term (w/o Bias Contrast) or the perturbation regularization term (w/o Perturb Reg.) is removed, and a simplified version using only the backbone generation module (SNT-only). Figure 5 shows the results on Adult with Mistral-7B, and the full results are given in Appendix E.

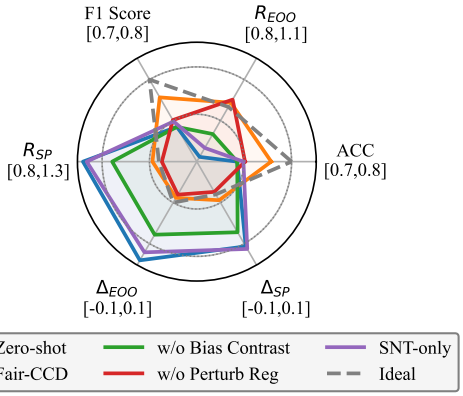

Figure 5: Ablation results of Fair-CCD. The gray dashed contour indicates the ideal target range; closer traces reflect better overall performance.

Figure 5 shows that the full Fair-CCD achieves the strongest and most balanced performance, with competitive accuracy and the largest gains in fairness. Removing the bias contrast term leads to a pronounced degradation in fairness with minimal impact on accuracy, confirming the structural response difference as the primary driver of bias mitigation. In contrast, removing the perturbation regularization term mainly harms accuracy and F1 while leaving fairness relatively unchanged, indicating its role in stabilizing prediction behavior. The SNT-only variant yields limited gains over Zero-shot and fails to produce consistent fairness improvements. Overall, these results demonstrate the necessity and complementarity of all Fair-CCD components.

4.3.2 Effect of Attention Layer Selection

We vary the start attention layer used to compute the averaged attention signal, while keeping all other components fixed. As shown in Figure 6, using attention signals from earlier layers yields limited fairness improvement, which we attribute to these layers primarily capturing low-level lexical and syntactic information. Starting from approximately the first third of layers, fairness mitigation becomes more effective and reaches a peak, suggesting that mid-to-late layers encode more task-relevant and bias-related representations. This

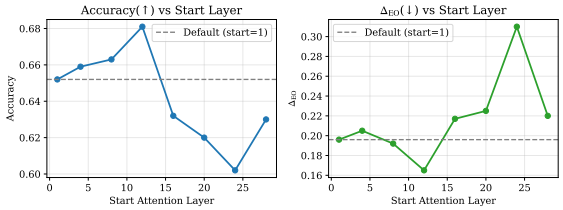

Figure 6: Layer-wise attention ablation on COMPAS with Qwen2.5-7B, with the dashed line indicating the default setting.

trend is consistent with recent findings that generative LLMs exhibit a layer-wise progression from lexical semantics to prediction-oriented representations (Liu et al., 2024). When only very late layers are used, performance becomes less stable, indicating a trade-off between fairness strength and predictive robustness.

4.3.3 Attention-driven vs. Fixed Control

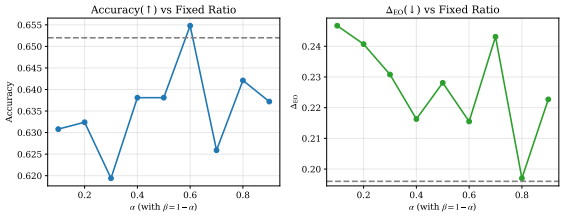

Figure 7: Attention-driven vs. fixed control on COMPAS with Qwen2.5-7B. The dashed line denotes the attention-driven Fair-CCD.

To ensure a fair comparison with attention-driven control, we evaluate fixed control by varying the ratio parameter α while enforcing $\alpha + \beta = 1$, such that both methods operate under the same normalization constraint. As shown in Figure 7, fixed control is highly sensitive to the choice of α: a smaller α leads to insufficient bias mitigation, while a larger α causes over-correction and degrades accuracy. In contrast, attention-driven Fair-CCD adaptively adjusts the mitigation strength at the instance level, achieving consistently stronger fairness improvements without sacrificing predictive performance. This suggests that attention-guided control is more effective than fixed global parameters in balancing accuracy and fairness.

5 Related Work

Among the various lines of research on fairness in LLMs, we focus on two areas most relevant to our work: (1) studies of societal bias in LLMs and their mitigation, and (2) methods for mitigating bias through ICL.

5.1 Fairness in Large Language Models

As LLMs are widely adopted, studies show they often encode societal stereotypes during pretraining (Gallegos et al., 2024a; Ganguli et al., 2023), causing systemic bias even in neutral tasks (Abid et al., 2021b; Wang et al., 2023; Huang et al., 2021). Most prior work focuses on generative tasks, using benchmarks like StereoSet and BBQ to evaluate stereotypical outputs (Liang et al., 2022; Venkit et al., 2023; Gallegos et al., 2024b; Xu et al., 2025). In contrast, fairness in structured classification tasks such as tabular prediction has received limited attention. While recent studies show that modifying input demonstrations (e.g., adding minority examples) can improve fairness (Hu et al., 2024), these methods rely on heuristics or sampling strategies without explicitly modeling the structural link between sensitive attributes and labels, limiting their effectiveness in this domain.

5.2 In-Context Learning

In-Context Learning enables LLMs to perform tasks without parameter updates by conditioning on a few demonstrations. It has been widely applied to text classification and answering (Gao et al., 2020; Liu et al., 2021), image generation (Bar et al., 2022), and multimodal reasoning (Huang et al., 2023; Wei et al., 2022). Increasing evidence suggests that ICL performance is strongly influenced by how demonstrations are constructed, with factors such as selection, ordering, and label formatting playing critical roles in shaping model predictions (Tanwar et al., 2023; Sorensen et al., 2022; Lu et al., 2021; Yoo et al., 2022; Oba et al., 2023; Sun et al., 2025).

Recent work on fairness in ICL has focused on selecting demonstrations to guide model predictions. For instance, Hu and Du (2024) found that using minority group examples can improve fairness, while FCG (2024) combines semantic clustering and genetic algorithms to automate such selection. However, these methods largely overlook how LLMs respond to structural differences in context, limiting their explanatory power and controllability. In contrast, our approach models attribute–label structures explicitly through templates and applies contrastive decoding for more stable bias mitigation during inference.

6 Conclusions

This work proposes Fair-CCD, a two-stage inference-time framework for mitigating bias in LLM-based tabular prediction. Fair-CCD encodes attribute–label relationships as structural bias templates and contrasts model responses across templates to adjust predictions. It further employs an attention-guided mechanism to adaptively regulate mitigation strength at the instance level, eliminating manual hyperparameter tuning. Extensive experiments show that Fair-CCD consistently improves fairness without sacrificing accuracy, demonstrating the effectiveness of structure-based, attention-aware bias mitigation.

Limitations

This work investigates bias mitigation for in-context learning through decoding-time interventions. While Fair-CCD demonstrates strong performance across multiple settings, several limitations remain to be addressed. To evaluate Fair-CCD across diverse architectures, we explore several representative open-source LLM families, including Mistral, Qwen, and LLaMA. Specifically, we include Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, LLaMA-3.1-8B-Instruct, and LLaMA-3.2-11B-Vision-Instruct. While these models span multiple model series and two representative scales (7B/8B and 11B/14B), our evaluation does not extend to extremely large models such as LLaMA-3-405B due to hardware constraints. In addition, Fair-CCD requires access to internal decoding processes and attention signals in order to perform contrastive decoding and adaptive control. As a result, our experiments are necessarily limited to open-source LLMs and do not include closed-source models that are accessible only via APIs, such as GPT-4o. Nonetheless, we believe that the diverse and representative set of high-performing open-source LLMs evaluated in this work allows our study to remain comprehensive and informative. Furthermore, our study is limited to binary classification tasks under in-context learning, as well as binary sensitive group settings. We plan to extend our analysis to broader classification scenarios in future work. Finally, in line with prior studies, our evaluation focuses on widely used fairness benchmarks that are tabular in nature and serialized into natural language prompts for LLMs. Exploring other data modalities in the context of fairness in large language models remains an important direction for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62272165.

References

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

Abubakar Abid, Maheen Farooqi, and James Zou. 2021b. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Meta AI. 2024. LLaMA 3. https://llama.meta.com/llama3/. Accessed: April 19, 2024.

Mistral AI. 2023. Introducing mistral 7b and mixtral. urlhttps://mistral.ai/news/mistral-7b/.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias url https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing. *Published on May*, 23:2016.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017.

Barry Becker and Ronny Kohavi. 1996. Adult. uci machine learning repository. *DOI: https://doi.org/10.24432/C5XW20, Accessed*, 15.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Kwesi Adu Cobbina and Tianyi Zhou. 2025. Where to show demos in your prompt: A positional bias of in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29548–29581.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024a. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024b. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, and 1 others. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Sadaf Md Halim, Chen Zhao, Xintao Wu, Latifur Khan, Christan Grant, Fariha Ishrat Rahman, and Feng Chen. 2025. Let the jury decide: Fair demonstration selection for in-context learning through incremental greedy evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18914–18931, Vienna, Austria. Association for Computational Linguistics.

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

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

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

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

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

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

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

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. *arXiv preprint arXiv:2403.01509*.

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*.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2023. In-contextual gender bias suppression for large language models. *arXiv preprint arXiv:2309.07251*.

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

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. 2026. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*.

Zhen Sun, Zongmin Zhang, Deqi Liang, Han Sun, Yule Liu, Yun Shen, Xiangshan Gao, Yilong Yang, Shuai Liu, Yutao Yue, and 1 others. 2025. " to survive, i must defect": Jailbreaking llms via the game-theory scenarios. *arXiv preprint arXiv:2511.16278*.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao'Kenneth' Huang, and Shomir Wilson. 2023. Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Song Wang, Peng Wang, Yushun Dong, Tong Zhou, Lu Cheng, Yangfeng Ji, and Jundong Li. 2024. On demonstration selection for improving fairness in language models. In *Workshop on Socially Responsible Language Modelling Research*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. Addressing order sensitivity of in-context demonstration examples in causal language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6467–6481, Bangkok, Thailand. Association for Computational Linguistics.

Zhenjie Xu, Wenqing Chen, Yi Tang, Xuanying Li, Cheng Hu, Zhixuan Chu, Kui Ren, Zibin Zheng, and Zhichao Lu. 2025. Mitigating social bias in large language models: A multi-objective approach within a multi-agent framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25579–25587.

Jingran Yang, Lingfeng Zhang, and Min Zhang. 2024. Making fair classification via correlation alignment. In *ECAI 2024*, pages 842–849. IOS Press.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

Lingfeng Zhang, Zhaohui Wang, Yueling Zhang, Min Zhang, and Jiangtao Wang. 2025a. Hifi: explaining and mitigating algorithmic bias through the lens of game-theoretic interactions. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, pages 756–768. IEEE.

Xingxuan Zhang, Haoran Wang, Jiansheng Li, Yuan Xue, Shikai Guan, Renzhe Xu, Hao Zou, Han Yu, and Peng Cui. 2025b. Understanding the generalization of in-context learning in transformers: An empirical study. In *The Thirteenth International Conference on Learning Representations*.

A Automated Prompt-Based Generation of SBTs

To generate the natural language sentence $s = f(a, y)$ for each attribute–label pair, we design a structured prompting schema that guides the language model to produce explicit and consistent structural descriptions.

Each prompt includes four key components:

- **Task description**: instructs the model to construct an SBT for a given attribute–label pair within a specific classification task, e.g., construct an SBT for the pair (a, y) in the income classification task;
- **Structural target**: specifies the intended attribute–label pair (a, y), e.g., "female" as the sensitive attribute and "high income" as the predicted label;
- **Stylistic prompting**: provides stylistic cues, such as "Women are often perceived to have higher earning potential";
- **Linguistic constraints**: enforces generalizable, group-level expressions and discourages individual-specific or overly verbose descriptions.

We apply the template function f(a, y) across the structural domains defined for BMT, BPT, and SNT to generate corresponding natural language prompts. These are used to construct structured contexts in an automated and consistent way, enhancing control over input design while reducing reliance on manual demonstration crafting.

A.1 Examples of Automatically Generated SBTs

To illustrate how Structural Bias Templates (SBTs) are generated via the prompt-based function $f(a, y)$, we present representative examples from both the *Adult* and *COMPAS* datasets. Each template is constructed using a structured prompting schema that includes a task description, structural target, stylistic reference, and linguistic constraints. The goal is to produce natural language descriptions that encode attribute–label relationships in a consistent and controllable form.

A.1.1 Adult Dataset (Income Classification)

(1) BMT – Bias-Mitigation Template
Prompt Input:

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the high-income classification task.
> **Structural Target:** (Female, $> 50k$)
> **Stylistic Prompting:** (Male, $> 50k$) $\rightarrow$ "Men are often seen as strong performers in high-income roles."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Output:

> *Women are frequently recognized in high-earning professions.*

(2) BPT – Bias-Preserving Template
Prompt Input

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the high-income classification task.
> **Structural Target:** (Male, $> 50k$)
> **Stylistic Prompting:** (Female, $> 50k$) $\rightarrow$ "Women are frequently recognized in high-earning professions."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Template

> *High-income positions are frequently seen as aligned with male performance.*

(3) SNT – Structure-Neutral Template
Prompt Input

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the high-income classification task.
> **Structural Target:** (Male, $> 50k$), (Male, $\leq 50k$), (Female, $> 50k$), (Female, $\leq 50k$)
> **Stylistic Prompting:** (Female, $> 50k$), (Female, $\leq 50k$) $\rightarrow$ "Women may have high or low incomes depending on their circumstances."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Template

> *Men and women are both represented across income levels.*

A.1.2 COMPAS Dataset (Recidivism Classification)

(1) BMT – Bias-Mitigation Template
Prompt Input:

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the recidivism classification task.
> **Structural Target:** (Black, Low Risk)
> **Stylistic Prompting:** (White, Low Risk) $\rightarrow$ "White individuals are often viewed as less likely to reoffend."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Output:

> *Black individuals are frequently recognized as capable of rehabilitation and low recidivism risk.*

(2) BPT – Bias-Preserving Template
Prompt Input:

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the recidivism classification task.
> **Structural Target:** (White, Low Risk)
> **Stylistic Prompting:** (Black, Low Risk) $\rightarrow$ "Black individuals are frequently recognized as capable of rehabilitation and low recidivism risk."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Template:

> *White individuals are generally considered less likely to commit another offense.*

Component	Zero-shot	SBT-based	Demonstration-based
Task Description	Predict whether the individual's income exceeds \$50K per year. Answer with one of the following: [greater than 50K] | [less than or equal to 50K]		
Contextual Content	(none)	Women are frequently recognized in high-earning professions.	**Example 1:** Age is 45, education is Bachelors, marital-status is Married-civ-spouse, occupation is Exec-managerial, sex is Male, income is greater than 50K; **Example 2:** Age is 30, education is HS-grad, marital-status is Never-married, occupation is Sales, sex is Female, income is greater than 50K;
Input Instance	Age is 37, education is Some-college, marital-status is Divorced, occupation is Tech-support, sex is Female, capital-gain is 0, hours-per-week is 40, **please answer the income:**		

Table 3: Prompting Templates for Adult Income Prediction under Different Context Settings

(3) SNT – Structure-Neutral Template
Prompt Input:

> **Task Description:** Construct an SBT for the attribute–label pair (a, y) in the recidivism classification task.
> **Structural Target:** (Black, High Risk), (Black, Low Risk), (White, High Risk), (White, Low Risk)
> **Stylistic Prompting:** (Black, Low Risk), (Black, High Risk) $\rightarrow$ "Black individuals may have varying risk levels depending on their background and circumstances."
> **Linguistic Constraints:** Use general, group-level expressions; avoid individual-specific or overly detailed language.

Generated Template:

> *Both Black and White individuals are found across different levels of recidivism risk.*

B Prompt Template

The output of LLMs is highly sensitive to prompt structure, which typically comprises three components: (1) a task description, (2) contextual content such as SBTs or demonstrations, and (3) the input instance to be predicted. The task description defines the objective and specifies the output format or label space. The contextual content provides prior information to guide the model's reasoning, while the final part presents the actual prediction query. The prompt example in Figure 2 simplifies the tabular dataset, the detailed template is described later in this section.

We consider both zero-shot and few-shot settings. In zero-shot prompts, no contextual content is provided, and the model relies solely on the task description and the input instance, serving as a baseline without fairness intervention. In few-shot prompts, contextual content is included between the task description and the input instance, allowing the model to perform ICL. Our study focuses on the construction of this contextual content, comparing SBTs with conventional demonstrations and analyzing their respective impacts on fairness. These insights form the basis for the proposed bias control mechanism and its empirical evaluation.

B.1 Example Prompt Architecture

To complement the prompt illustration in Figure 2 of the main text, we provide complete prompt template examples illustrating the structural composition of zero-shot and few-shot settings. Each prompt consists of three key components: a task description, contextual content, and an input instance. The contextual content is the primary variable across settings, taking one of three forms: (1) no context (zero-shot), (2) Structural Bias Templates (SBTs), or (3) in-context demonstrations.

In what follows, we present concrete examples

from the *Adult* dataset, demonstrating how each type of contextual content is incorporated into the prompt format. The input instance is kept fixed across all three cases for comparison. Table 3 provides a structured summary of the prompt components and their instantiations under each context strategy.

C Evaluation Metric Definitions

This section provides the formal definitions of the fairness evaluation metrics used in our experiments. We focus on Statistical Parity (SP) and Equality of Opportunity (EO), each quantified using both difference-based and ratio-based measures. These metrics are computed based on group-conditional prediction outcomes and are widely adopted in prior fairness studies. Below, we present the precise mathematical formulations of Δ_{SP}, R_{SP}, Δ_{EO}, and R_{EO}.

For each group a_i defined by the sensitive attribute A, we define statistical parity (SP) as the probability of being predicted as positive:

$$\text{SP}_{a_i} = P(\hat{Y} = 1 \mid A = a_i) \tag{8}$$

The group-wise difference and ratio of statistical parity are given by:

$$\Delta_{\text{SP}} = \text{SP}a_0 - \text{SP}a_1, \quad R_{\text{SP}} = \frac{\text{SP}a_0}{\text{SP}a_1 + \epsilon} \tag{9}$$

Likewise, for individuals whose true label is positive (Y = 1), we define the true positive rate (TPR) as:

$$\text{TPR}_{a_i} = P(\hat{Y} = 1 \mid Y = 1, A = a_i) \tag{10}$$

From this, the corresponding equality of opportunity metrics are defined as:

$$\Delta_{\text{EO}} = \text{TPR}a_0 - \text{TPR}a_1, \quad R_{\text{EO}} = \frac{\text{TPR}a_0}{\text{TPR}a_1 + \epsilon} \tag{11}$$

Here, ϵ is a small constant added to prevent division by zero. Smaller values of Δ_{SP} and Δ_{EO}, as well as ratios R_{SP} and R_{EO} closer to 1, indicate fairer predictions across demographic groups.

D Additional Experiments on the Effectiveness of SBTs

This section presents supplementary figures that provide additional empirical evidence on the effectiveness of Structural Bias Templates (SBTs). These results complement the main analysis by showcasing the impact of different SBT configurations across models. They further support our claim that structural attribute–label relationships play a critical role in shaping model behavior and can be leveraged for controllable and stable fairness interventions. Representative results are provided in Figures 8–11.

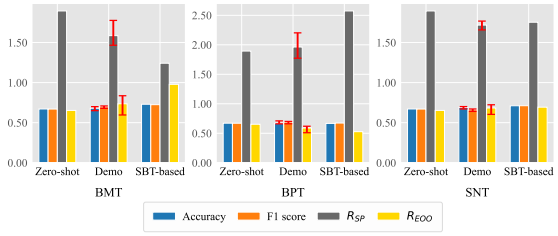

Figure 8: Prediction and fairness performance of LLaMA3-8B on the *Adult* dataset under different context strategies. Error bars for the Demonstration-based method indicate variation across five sampled demonstrations.

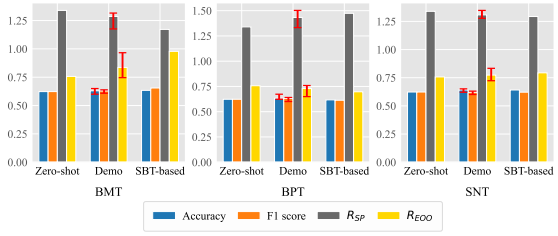

Figure 9: Prediction and fairness performance of LLaMA3-11B on the *Adult* dataset under different context strategies. Error bars for the Demonstration-based method indicate variation across five sampled demonstrations.

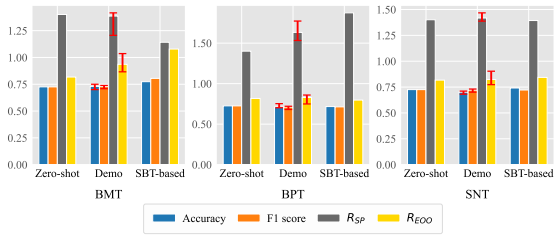

Figure 10: Prediction and fairness performance of Mistral-7B on the *Adult* dataset under different context strategies. Error bars for the Demonstration-based method indicate variation across five sampled demonstrations.

E Supplementary Ablation Results

This section provides supplementary figures reporting detailed results from the ablation study of

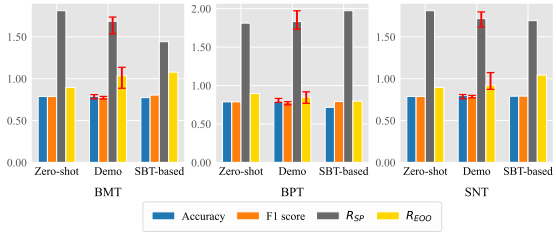

Figure 11: Prediction and fairness performance of Qwen2.5-14B on the *Adult* dataset under different context strategies. Error bars for the Demonstration-based method indicate variation across five sampled demonstrations.

Fair-CCD. These results complement the main text by illustrating additional experimental settings and component-wise analyses. They help further validate the effectiveness and stability of each module, especially under different model configurations. Representative results are provided in Figures 12–15.

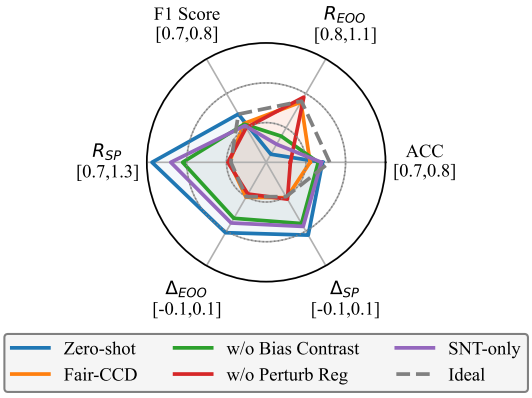

Figure 12: Ablation Results of Fair-CCD with LLaMA3-8B on the *Adult* dataset. The gray dashed contour indicates the ideal target range; closer traces reflect better overall performance.

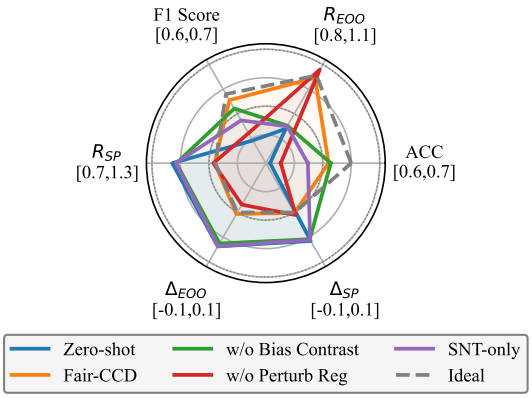

Figure 13: Ablation Results of Fair-CCD with LLaMA3-11B on the *Adult* dataset.

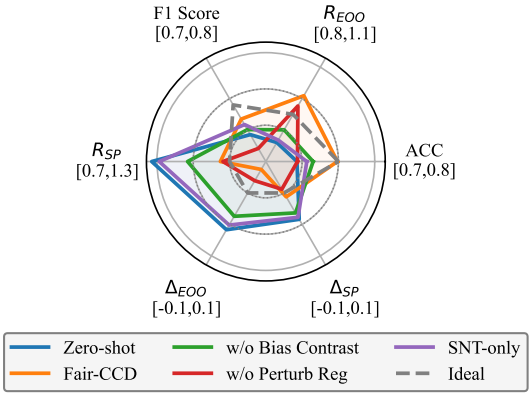

Figure 14: Ablation Results of Fair-CCD with Qwen2.5-7B on the *Adult* dataset.

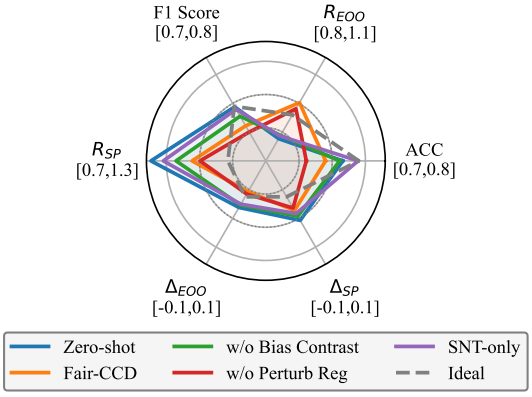

Figure 15: Ablation Results of Fair-CCD with Qwen2.5-14B on the *Adult* dataset.