Adversarial Group Fairness in Contextual Bandits: When Robust is Not Fair

Ray Telikani[1,*] **Jaber Valizadeh**[2], **Amir H Gandomi**[3,4], **Bao Quoc Vo**[1], **Ming Ding**[5]

[1]Swinburne University of Technology, Australia.
[2] Western Sydney University, Sydney, Australia.
[3]University of Technology Sydney, Australia.
[4] Obuda University, Hungary.
[5]CSIRO's Data61, Australia.

{rtelikani,bvo}@swin.edu.au
j.valizadeh@westernsydney.edu.au, gandomi@uts.edu.au
ming.ding@data61.csiro.au

Abstract

Neural contextual bandits support high-stakes decision systems—from clinical trial allocation to content recommendation. This paper investigates an adversarial threat, *group-targeted suppression attacks* (GTSA), in which an adversary selectively degrades performance for a demographic subgroup while preserving global performance metrics. We first show that under standard aggregate monitoring, GTSA with budget $B = \Omega(\sqrt{T})$ can induce $\Omega(T)$ group regret while maintaining only $o(1)$ deviation in observable global statistics, establishing that any defense that ignores group structure is fundamentally vulnerable. To address this vulnerability, we propose ANCHORFAIR, a robust mitigation framework comprising three components: (i) a trust-weighted policy anchor with provable bounded drift, (ii) representation-stable demographic discovery via online clustering in a slowly evolving embedding space, and (iii) a demographic-aware exploration mechanism that adaptively amplifies learning for underrepresented or low-trust groups. We prove that ANCHORFAIR achieves sublinear group regret $\tilde{O}(\sqrt{T})$ for all groups under GTSA. Source codes are available at https://github.com/RayTelikani/AnchorFair.

1 Introduction

Neural contextual bandit (NCB) algorithms extend multi-armed bandits by incorporating side information (i.e., context vectors) to guide arm selection and by leveraging neural networks to model complex, nonlinear reward functions [Zhou et al., 2020, Ban et al., 2022, Wang et al., 2025]. These algorithms now serve as core decision engines in recommendation systems [Kveton et al., 2015], cloud resource allocation [Dang'ana et al., 2026], clinical trials [Kuleshov and Precup, 2014], dynamic pricing [Tullii et al., 2024], and increasingly in the alignment and fine-tuning of large language models, such as ChatGPT Achiam et al. [2023], Gemini Team et al. [2023], and DeepSeek Guo et al. [2025]. When context vectors encode or correlate with sensitive demographic groups, the bandit's decisions carry fairness implications, and reward distributions may differ across groups, and policies that maximize aggregate reward can disadvantage these subpopulations [Grazzi et al., 2022, Joseph et al., 2016, Baek and Farias, 2026].

In addition, NCBs have shown to be sensitive to adversarial corruptions Qi et al. [2024], Telikani and Gandomi [2026], Ding et al. [2022]. From the perspective of fairness, this vulnerability directly undermines group fairness objectives, such as demographic parity or equalized opportunity across protected attributes embedded in the context vector Grazzi et al. [2022], Chzhen et al. [2023]. It

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

enables covert attacks that selectively violate fairness constraints (e.g., under-serving target demographic groups while preserving overall performance metrics) or trigger complete policy collapse, resulting in linear regret and systemic bias amplification. The adversarial vulnerability of contextual bandits has attracted growing attention. Recent work has demonstrated that subtle perturbations to rewards [Garcelon et al., 2020, Ding et al., 2022], actions [Liu and Lai, 2023], or contexts [Ma et al., 2018, Zeng et al., 2025, Hosseini et al.] can mislead learners into persistently suboptimal decisions.

However, the existing models focus on attacks that degrade *global* cumulative regret, overlooking a more pernicious class of attacks in which the adversary selectively suppresses performance for a specific demographic group while leaving the aggregate signal intact. In clinical trial allocation, for instance, an attacker could degrade treatment recommendations only for the elderly group while preserving favorable outcomes for others. We term this class of threats *group-targeted suppression attacks* (GTSA), where harmful group-level effects remain hidden behind apparently optimal global performance. We make the following contributions:

1. **Threat model and impossibility result.** We formalize GTSA as a context-perturbation attack targeting a demographic subgroup. We prove that under aggregate-only monitoring, an attacker with budget $B = \Omega(\sqrt{T})$ can inflict $\Omega(T)$ group regret on a target subgroup while keeping global regret sublinear and all observable aggregate statistics within $o(1)$ of their unattacked values, establishing that group-oblivious defenses are fundamentally insufficient.
2. **Robust mitigation framework.** We propose ANCHORFAIR, a defense with three components: (i) a *trust-weighted policy anchor* whose exponential moving averages (EMA) update freezes under suspected attack and recovers exponentially in benign periods; (ii) *demographic discovery* in augmented anchor embedding space, which resists adversarial fragmentation and merging of groups; and (iii) a *demographic-aware exploration bonus* that amplifies learning for low-trust or underrepresented groups and concentrates exploration on arms where anchor–policy disagreement signals manipulation.
3. **Theoretical and experimental guarantees.** We prove that ANCHORFAIR achieves $R_T^{(g)} = \tilde{O}(\sqrt{T})$ group regret for *all* groups under GTSA with budget $B = O(\sqrt{T})$, and $\tilde{O}(\sqrt{G/T})$ regret equity across G groups. These bounds degrade gracefully with the attack budget and the clustering error rate. We conduct experiments on real-world datasets (COMPAS and MovieLens) comparing ANCHORFAIR against five NCB baselines spanning undefended, robust, and fair bandit algorithms.

2 Problem Formulation

2.1 Neural Contextual Bandit Setting

Consider a K-armed contextual bandit problem over a horizon of T rounds. At each round $t \in [T]$, the learner observes a d-dimensional context set $\mathcal{X}_t = \{\mathbf{x}_{t,i} \in \mathbb{R}^d : i \in [K]\}$, selects an arm $a_t \in [K]$, and receives a stochastic reward

$$r_{t,a_t} = h(\mathbf{x}_{t,a_t}) + \xi_t, \tag{1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is an unknown reward function belonging to a reproducing kernel Hilbert space $\mathcal{H}$ with bounded RKHS norm $\mathcal{R} \geq \|h\|_{\mathcal{H}}$ [Valko et al., 2013, Bogunovic et al., 2022], approximated in practice by an over-parameterized neural network operating in the NTK regime [Zhou et al., 2020, Qi et al., 2024], and ξ_t is conditionally zero-mean σ-sub-Gaussian noise. The learner's objective is to minimize cumulative pseudo-regret

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t})\big], \tag{2}$$

where $a_t^* = \arg\max_{i \in [K]} h(\mathbf{x}_{t,i})$ is the optimal arm at round t.

2.2 Demographic Structure

Let $F_{\mathrm{g}} : \mathbb{R}^d \to [G]$ be a function mapping each context to one of G latent demographic groups. At each round t, the learner interacts with context $\mathbf{x}_{t,a_t}$ and maps the context. We denote by

$\tau_g = \{t \in [T] : F_{\mathrm{g}}(\mathbf{x}_{t,a_t}) = g\}$ the set of rounds assigned to group g, and by $n_g = |\tau_g|$ the group's effective sample size. The *group regret* for subpopulation g is

$$R_T^{(g)} = \sum_{t \in \tau_g} \mathbb{E}\big[h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t})\big]. \tag{3}$$

Standard regret decomposes as $R_T = \sum_{g=1}^{G} R_T^{(g)}$. Sublinear global regret $R_T = o(T)$ is compatible with $\Theta(1)$ per-capita regret for a targeted group: when $n_{g^\dagger} = o(T)$, the group's total regret $n_{g^\dagger} \cdot \Delta_{\max}$ is small relative to T and thus invisible in the aggregate, where $\Delta_{\max}$ denotes the worst-case loss per interaction (i.e., the largest gap between the best action and the worst one).

3 Threat Model: Group-Targeted Suppression Attacks

Definition 1 (Group-Targeted Suppression Attack)**.** *A* GTSA *is specified by $(g^\dagger, B, \epsilon)$, where $g^\dagger \in [G]$ is the target group, $B > 0$ is the total perturbation budget, and $\epsilon > 0$ is the per-round magnitude. The attacker is positioned* upstream *of the learner in the data pipeline (e.g., as a data broker, a compromised feature-store vendor, an adversarial content provider, or a tampered edge client [Crain, 2018, Hosseini et al.]) and at each attack round tobserves the true context set $\mathcal{X}_t$ and perturbs each arm's context:*

$$\tilde{\mathbf{x}}_{t,i} = \mathbf{x}_{t,i} + \boldsymbol{\delta}_{t,i}, \quad \textit{where } \|\boldsymbol{\delta}_{t,i}\|_\infty \le \epsilon, \tag{4}$$

subject to $\sum_{t=1}^{T}\sum_{i=1}^{K} \|\boldsymbol{\delta}_{t,i}\|_\infty \le B$. This corruption forces the learner (victim) to select $\hat{a}_t \neq a_t \in [K]$ (arm for corrupted context) and receive a corrupted arm reward $\hat{r}_{t,\hat{a}_t} = h(\tilde{\mathbf{x}}_{t,\hat{a}_t}) + \xi_t$.

Assumption 1 (Attacker Capabilities)**.** *The attacker operates under the following constraints:* ***(1) Black-box access:*** *The attacker observes $(\mathcal{X}_t, a_t)$ but has no access to the learner's internal parameters $\boldsymbol{\theta}_t$, reward observations $r_{t,i}$, or gradient information;* ***(2) Targeted objective:*** *The attacker seeks to maximize $R_T^{(g^\dagger)}$ while keeping $|R_T^{\mathrm{att}} - R_T^{\mathrm{clean}}|$ small, where R_T^{att} and R_T^{clean} are the learner's cumulative global regret under attack and without attack, respectively;* ***(3) Group knowledge:*** *The attacker possesses a group classifier $\hat{g}_{\mathrm{att}} : \mathbb{R}^d \to [G]$ satisfying probability $\boldsymbol{P}[\hat{g}_{\mathrm{att}}(\mathbf{x}) = F_{\mathrm{g}}(\mathbf{x})] \ge 1 - \eta$ for some misclassification rate $\eta \in [0, 1)$. The attacker's knowledge of group structure may arise from (a) a pre-trained classifier (small η) on external data unavailable to the defender (e.g., due to legal or institutional constraints) or (b) online discovery (large η that shrinks) from the context stream Shaham et al. [2025]; and* ***(4) non-adaptive strategy:*** *The attacker's perturbation strategy is fixed w.r.t. the defender's clustering assignments $\hat{g}_t$; i.e., the attacker does not observe or adapt to the defender's internal group discovery.*

4 ANCHORFAIR: Robust Bandit Algorithm

The capabilities of GTSA (Assumption 1) make aggregate-statistic monitoring impossible (proven in Appendix E). ANCHORFAIR (Algorithm 1) counters this through two principles: (i) policy stability via trust-mediated inertia and (ii) group-level visibility via representation-stable clustering. The defender has *no access to group labels* and must discover group structure online from the corrupted context stream. This constraint mirrors real-world regulations (GDPR, Equal Credit Opportunity Act) and fairness-through-unawareness policies that prohibit explicit use of protected attributes [Chen et al., 2019, Höltgen and Oliver, 2025, Zhao et al., 2025, Deldjoo et al., 2024]. The defender cannot explicitly query or condition on protected attributes, and its embedding space $\phi(\cdot)$ encodes group-correlated structure (which is why clustering succeeds). The attacker, conversely, can exploit demographics or external data, inducing $\eta > 0$ in Assumption 1.

Assumption 2 (Regularity conditions)**.** *We assume: (i) the reward function $h \in \mathcal{H}$ with $\|h\|_{\mathcal{H}} \le \mathcal{R}$ (as in Eq.* (1)*), which implies L_h-Lipschitz continuity w.r.t. the ℓ_∞ norm for kernels whose RKHS embeds into the space of Lipschitz functions (e.g., squared-exponential kernel) [Valko et al., 2013]; and (ii) group prevalences $\rho_g = n_g/T$.*

4.1 Trust-Weighted Policy Anchor

The main policy π_t is vulnerable to short-term adversarial corruption (Appendix D). To mitigate this, we introduce a slowly evolving *policy anchor* π_a^t as a lightweight single-layer network:

$$\pi_a^t(a \mid \mathbf{x}) = \frac{\exp\big(\mathbf{w}_a^{t\top}\phi(\mathbf{x})\big)}{\sum_{a' \in [K]} \exp\big(\mathbf{w}_{a'}^{t\top}\phi(\mathbf{x})\big)}, \tag{5}$$

where $\phi(\mathbf{x}) \in \mathbb{R}^{d_a}$ is a low-dimensional embedding ($d_a < d$) and $\{\mathbf{w}_a^t\}_{a\in[K]}$ are trainable per-arm weight vectors. The anchor supports (i) anomaly detection via $D_{\mathrm{KL}}(\pi_t \| \pi_a^t)$ and (ii) robustness via slower adaptation because corrupting the anchor requires sustained manipulation over many rounds.

Trust-modulated EMA update. The anchor must update fast enough to stay relevant in benign conditions, yet slow enough under attack that the adversary cannot corrupt it within budget. We achieve both through a *trust-modulated* EMA. EMA of model weights is widely used to stabilize training [Morales-Brotons et al., 2024], control policy drift in LLM alignment [Noukhovitch et al., 2023], and tame RL instabilities [Zhang et al., 2026]; we extend this principle by making the EMA rate depend on a trust score $\kappa_t \in [0, 1]$:

$$\mathbf{w}_a^{t+1} = (1 - \alpha_t)\, \mathbf{w}_a^t + \alpha_t\, \mathbf{w}_a^{\pi_t}, \qquad \alpha_t = \bar{\alpha} \cdot \kappa_t^{p_t}, \tag{6}$$

where $\mathbf{w}_a^{\pi_t}$ is the main policy's action-head parameters, $\alpha_t = \bar{\alpha}\kappa_t^{p_t}$ is the anchor's learning rate, $\bar{\alpha} \in (0, 1)$ is the maximum learning rate, and $p_t \geq 1$ is a sensitivity exponent. When trust is high ($\kappa_t \to 1$), $\alpha_t \to \bar{\alpha}$ and the anchor tracks the main policy; when trust collapses ($\kappa_t \to 0$), $\alpha_t \to 0$ and the anchor freezes.

Trust score. The trust score κ_t aggregates averaged policy–anchor disagreement over a sliding window of size n_w, providing an online proxy for anchor's stability conditions (formalized in Proposition 1):

$$\kappa_t = \exp\left(-\frac{1}{n_w} \sum_{i=t-n_w+1}^{t} D_{\mathrm{KL}}\big(\pi_i(\cdot|\mathbf{x}_{i,a_i}) \parallel \pi_a^i(\cdot|\mathbf{x}_{i,a_i})\big)\right) \cdot \mathbb{I}\big[\big|\hat{\mu}_w^t - \hat{\mu}_w^{t-n_w}\big| \leq \tau_r\big], \tag{7}$$

where $\hat{\mu}_w^t$ is the windowed mean reward and τ_r a reward-smoothness threshold.

Fast-path detection via Huber warm-up and CUSUM. The windowed KL in Eq. (7) responds only after enough attacked rounds in the window, but the anchor drifts at full rate during the early-onset phase (i.e., early attacked rounds) because the KL signal has not accumulated yet, and the window average is still dominated by benign rounds ($\kappa_t \approx 1$). To address this, we complement the trust score with two methods. First, the anchor's reward predictor $\hat{r}_{\boldsymbol{\theta}_a}(\mathbf{x}) \approx h(\mathbf{x})$ is initialized under Huber loss,

$$\ell_H(\boldsymbol{\theta}_a; \mathbf{x}, r) = \begin{cases} \frac{1}{2}(r - \hat{r}_{\boldsymbol{\theta}_a}(\mathbf{x}))^2 & \text{if } |r - \hat{r}_{\boldsymbol{\theta}_a}(\mathbf{x})| \leq \delta_H, \\ \delta_H |r - \hat{r}_{\boldsymbol{\theta}_a}(\mathbf{x})| - \frac{1}{2}\delta_H^2 & \text{otherwise,} \end{cases} \tag{8}$$

so that the anchor becomes resilient to early-onset contamination (i.e., the transition from benign to adversarial conditions). Second, a per-group CUSUM statistic accumulates anchor prediction errors,

$$S_g^t = \max\big(0,\ S_g^{t-1} + |e_{g,t}| - \mu_g - \delta_{\mathrm{CUSUM}}\big), \tag{9}$$

where $e_{g,t} = r_{t,a_t} - \hat{r}_{\boldsymbol{\theta}_a}$ and μ_g are the anchor's prediction error and the baseline error for group g, respectively. Running G parallel CUSUMs (one per group) sidesteps the aggregate-monitor impossibility (shown in Appendix E): when group g's error rises while others' remain flat, S_g^t crosses the threshold h_{CUSUM} well before the aggregate signal moves, activating trust-mediated slowdown and shortening the pre-detection phase.

Adaptive sensitivity. A fixed exponent p in Eq. (6) forces a one-setting tradeoff: large p is safe under attack but over-conservative in benign conditions; small p is responsive but vulnerable. We resolve this by making p_t *group-conditional*:

$$p_t = p_{\min} + (p_{\max} - p_{\min}) \cdot (1 - \bar{\kappa}_{\hat{g}_t}^t), \tag{10}$$

where $\bar{\kappa}_{\hat{g}_t}^t \in [0, 1]$ is the mean trust for the group assigned to the current context, and $p_{\min}, p_{\max}$ are the baseline and peak sensitivity bounds. Making p_t depend on group trust lets the anchor behave differently per context, in which benign groups contribute to fast tracking at $p_t \approx p_{\min}$ while suspect groups trigger strong suppression at $p_t \approx p_{\max}$.

While the anchor *learns* in parameter space via Eq. (6), the trust mechanism *monitors* in distribution space via $D_{\mathrm{KL}}(\pi_t \| \pi_a^t)$ (full discussion in Appendix F.9). We formalize the resulting stability guarantee:

Proposition 1 (Bounded anchor drift)**.** *Let the trust score satisfy $\kappa_t \leq \bar{\kappa} < 1$ during an attack window $[t_0, t_0 + W]$. Then the total variation distance between the anchor at t_0 and $t_0 + W$ is:*

$$TV\left(\pi_a^{t_0+W}, \pi_a^{t_0}\right) \leq C_\phi K \cdot \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot W \cdot D, \tag{11}$$

where $D = \max_{t \in [t_0, t_0+W]} \max_a \|\mathbf{w}_a^{\pi_t} - \mathbf{w}_a^t\|_\infty$ is the maximum parameter discrepancy and $C_\phi = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|_1/2$ is the embedding norm constant arising from the softmax Jacobian (see Section F.1). The bound uses $p_{\min}$ because $\kappa_t^{p_t} \leq \kappa_t^{p_{\min}}$ for all $\kappa_t \in [0, 1]$ and $p_t \geq p_{\min}$; the actual drift is smaller when $p_t > p_{\min}$.

This formalizes the attacker's *post-detection* dilemma: once $\kappa_t \leq \bar{\kappa}$, shifting the anchor by a constant amount requires sustained perturbations over $W = \Omega(1/(\bar{\alpha}\bar{\kappa}^{p_{\min}}))$ rounds. Before detection, the anchor drifts at the full rate $\bar{\alpha}$; the attacker's total effective window is therefore bounded by the detection delay of Theorem 3(c) plus post-detection budget exhaustion, combined in the group-regret analysis (Appendix F.8, Step 1).

4.2 Prediction-Error-Aware Demographic Discovery

Clustering solely on context embeddings $\mathbf{z}_t = \phi(\mathbf{x}_{t,a_t}) \in \mathbb{R}^{d_a}$ may fail when demographic boundaries are subtle (e.g., activity-based groups). Under GTSA, the anchor's prediction error $e_{g,t}$ is consistently biased for attacked rounds (due to arm-ranking inversion) but unbiased for clean ones. We exploit this by clustering on augmented features $\tilde{\mathbf{z}}_t = [\mathbf{z}_t; \omega_t \cdot \bar{e}_{\hat{g}_t}^t] \in \mathbb{R}^{d_a+1}$, where $\bar{e}_{\hat{g}t}^t$ is the group's mean absolute error, and $\omega_t > 0$ balances the error signal. We maintain G cluster centroids $\{c_g\}_{g=1}^G$ via online k-means with soft assignment:

$$p(g \mid \tilde{\mathbf{z}}_t) = \frac{\exp\left(-\|\tilde{\mathbf{z}}_t - c_g\|^2/(2\sigma_c^2)\right)}{\sum_{g'=1}^G \exp\left(-\|\tilde{\mathbf{z}}_t - c_{g'}\|^2/(2\sigma_c^2)\right)}, \tag{12}$$

where $\sigma_c > 0$ is a temperature parameter. The weight ω_t is updated via $\omega_t = \omega_0 \cdot (1 - \bar{\kappa}^t)$, where $\bar{\kappa}^t$ is the mean trust, ensuring the error signal receives more weight when an attack is suspected.

Assumption 3 (Group separability)**.** *The latent demographic groups exhibit* approximate separability *in the anchor's augmented embedding space $Z = \mathbb{R}^{d_a+1}$. At the beginning of the defense phase (after warm-up), for any two contexts x, x' belonging to the same group $g = F_g(x) = F_g(x')$, $\|\tilde{z}(x) - \tilde{z}(x')\|_{\tilde{\mathcal{Z}}} \leq \frac{\Delta_c}{4} + \nu \cdot \Delta_c$, and for any two contexts belonging to different groups $g \neq g'$, $\|\tilde{z}(x) - \tilde{z}(x')\|_{\tilde{\mathcal{Z}}} \geq \Delta_c - \nu \cdot \Delta_c$, where $\tilde{z}(x) = [\phi(x); \omega \cdot \bar{e}_g]$ is the augmented feature vector, and δ controls the degree of group overlap. When $\nu = 0$, we recover the original strong separation condition. The clustering error rate η_{cluster} degrades gracefully as ν increases.*

Adaptive group count. Rather than fixing G, AnchorFair periodically (every T_{eval} rounds) selects $G \in \{2, \ldots, G_{\max}\}$ via a silhouette criterion and reinitializes centroids $\{c_g\}_{g=1}^G$. Immediately after reinitialization, centroids lack sufficient data to satisfy Assumption 3, leading to transient misclustering and noisier estimates $\hat{\mu}_g^t$ and $\bar{\kappa}_g^t$. Furthermore, the defense mechanism (CUSUM and the trust score) flags such a sudden increase in per-group prediction errors and a sudden drop in policy-anchor agreement. To prevent this self-attack, both CUSUM statistics $\{S_g^t\}$ and trust scores $\{\bar{\kappa}_g^t, \kappa_t\}$ are paused for $n_{\text{stab}} = 2n_w$ rounds after each re-initialization. During this phase, previous trust scores are reused and the exploration bonus reduces to standard UCB ($D_g^t(i) = 0$). Clustering guarantees resume after centroid convergence, requiring $O(\Delta_c^{-2} \log(G/\delta))$ samples per group. In practice, $T_{\text{eval}} \geq 4n_w$ ensures sufficient data between updates.

Proposition 2 (Representation stability)**.** *Under the conditions of Proposition 1, for any two contexts $\mathbf{x}$ and $\mathbf{x}'$ belonging to the same demographic group at t:*

$$\|\phi^{t+W}(\mathbf{x}) - \phi^{t+W}(\mathbf{x}')\| \leq \|\phi^t(\mathbf{x}) - \phi^t(\mathbf{x}')\| + 2L_\phi \cdot \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot W \cdot D, \tag{13}$$

where L_ϕ is the Lipschitz constant of the embedding w.r.t. its parameters. The augmented clustering inherits this stability: the error component $\bar{e}_g^t$ evolves smoothly via running averages, so the augmented distance $\|\tilde{\mathbf{z}}_t - c_g\|$ is controlled by the same bound up to the ω_t-weighted error term.

Algorithm 1 ANCHORFAIR: Robust Group-Fair Neural Contextual Bandit

Require: Horizon T, arms K, anchor parameters $(\bar{\alpha}, p_{\min}, p_{\max})$, exploration parameters $(\beta, \lambda, \mu, \gamma_0)$, clustering parameters $(\sigma_c, \omega_0, G_{\max})$, CUSUM parameters $(\delta_{\text{CUSUM}}, h_{\text{CUSUM}})$

Initialize main policy π_0, anchor policy π_a^0 (via Eq. (8)), cluster centroids $\{c_g^0\}$, per-group per-arm statistics $\{(A_{g,a}, b_{g,a})\}$, group statistics $\{n_g, \hat{\mu}_g, \hat{\sigma}_g^2, \bar{\kappa}_g\}$, CUSUM states $\{S_g^0 = 0\}$
for $t = 1, 2, \ldots, T$ **do**
  Observe context set $\mathcal{X}_t = \{\mathbf{x}_{t,i}\}_{i \in [K]}$ ▷ *(potentially perturbed)*
  Compute anchor embedding and augmented feature: $\mathbf{z}_t \leftarrow \phi(\mathbf{x}_{t,\cdot})$; $\tilde{\mathbf{z}}_t \leftarrow [\mathbf{z}_t;\, \omega_t \cdot \bar{e}_{\hat{g}_{t-1}}^{t-1}]$
  Assign group: $\hat{g}_t \leftarrow \arg\max_g \, p(g \mid \tilde{\mathbf{z}}_t)$ ▷ *Eq.* (12)
  Compute adaptive exponent: $p_t \leftarrow p_{\min} + (p_{\max} - p_{\min})(1 - \bar{\kappa}_{\hat{g}_t}^{t-1})$ ▷ *Eq.* (10)
  Compute arm-specific exploration bonus: $\mathcal{B}_{\hat{g}_t}^t(i)$ for each arm i ▷ *Eq.* (14)
  Select arm: $a_t \leftarrow \arg\max_{i \in [K]} \big[\pi_t(i \mid \mathcal{X}_t) + \mathcal{B}_{\hat{g}_t}^t(i)\big]$
  Observe reward r_{t,a_t}
  Update main policy π_{t+1} using $(\tilde{\mathbf{x}}_{t,a_t}, r_{t,a_t})$
  Compute trust score κ_t ▷ *Eq.* (7)
  Update anchor: $\alpha_t \leftarrow \bar{\alpha} \cdot \kappa_t^{p_t}$; $\mathbf{w}_a^{t+1} \leftarrow (1 - \alpha_t)\mathbf{w}_a^t + \alpha_t \mathbf{w}_a^{\pi_t}$ ▷ *Eq.* (6)
  Update CUSUM: $S_{\hat{g}_t}^t \leftarrow \max(0,\, S_{\hat{g}_t}^{t-1} + |e_{\hat{g}_t,t}| - \mu_{\hat{g}_t} - \delta_{\text{CUSUM}})$ ▷ *Eq.* (9)
  Update cluster centroids $\{c_g^t\}$, per-group per-arm statistics $\{(A_{g,a}, b_{g,a})\}$, and group statistics
  if periodic evaluation round **then**
    Re-estimate G via silhouette criterion; adjust clusters and σ_c
  end if
end for

For each group $g \in [G]$ and arm $a \in [K]$, the system maintains sufficient statistics $(A_{g,a}, b_{g,a})$, where $A_{g,a}$ is the regularized covariance matrix and $b_{g,a}$ is the reward-weighted context vector for samples assigned to g when pulling a. This separation ensures that corruption in group $g^\dagger$ for arm a does not affect other groups. Additionally, using hard assignments $\hat{g}_t = \arg\max_{g'} p(g' \mid \tilde{\mathbf{z}}_t)$, the system tracks per-group streaming statistics: observation count n_g^t, mean reward $\hat{\mu}_g^t$, reward variance $\hat{\sigma}_g^{2,t}$, and mean trust score $\bar{\kappa}_g^t$ (see Section G).

4.3 Arm-Specific Demographic-Aware Exploration

To mitigate the suppressive effect of GTSA, ANCHORFAIR augments the base bandit's exploration strategy with an *arm-specific* bonus. Our exploration bonus concentrates exploration on arms where the anchor and main policy *disagree*. The exploration bonus for group g and arm i is

$$\mathcal{B}_g^t(a) = \underbrace{\frac{\beta}{\sqrt{n_{g,a}^t}} \cdot \big(1 + \lambda(1 - \bar{\kappa}_g^t)\big)}_{\text{trust-amplified UCB}} \cdot \underbrace{\big(1 + \mu \cdot D_g^t(a)\big)}_{\text{disagreement factor}} + \underbrace{\gamma_t \cdot \big[\hat{\Delta}_g^t\big]_+}_{\text{regret-gap correction}}, \tag{14}$$

where $\beta > 0$ controls the base exploration intensity, $\lambda \geq 0$ amplifies exploration when group trust is low, and $n_{g,a}^t$ is the per-arm, per-group observation count from the separate $(A_{g,a}, b_{g,a})$ statistics. Here $\hat{\Delta}_g^t = \hat{\mu}_g^t - \bar{\mu}^t$ estimates the per-round reward gap between group g and the global average. The regret-gap correction $\gamma_t \cdot [\hat{\Delta}_g^t]_+$ uses a weight $\gamma_t = \gamma_0$. The *anchor-disagreement factor* is

$$D_g^t(i) = \mathbb{I}\big[\bar{\kappa}_g^t < \bar{\kappa}\big] \cdot \big|\pi_a^t(i \mid \mathbf{x}) - \pi_t(i \mid \mathbf{x})\big|, \tag{15}$$

where $\mu > 0$ controls the disagreement amplification. When group trust is low (attack suspected) and the anchor assigns high probability to arm i but the main policy does not, $D_g^t(i)$ is large, concentrating the bonus on those arms whose ranking has been inverted by the attack. When trust is high, $D_g^t(i) = 0$ and the bonus reduces to the standard UCB form.

Remark 1 (Inter-component error propagation)**.** *The three components of* ANCHORFAIR *have a* one-directional *dependence structure: clustering error biases the regret-gap estimator $\hat{\Delta}_g^t$, but does* not *feed back into the trust score κ_t (which depends only on $D_{\text{KL}}(\pi_t \| \pi_a^t)$ and reward statistics) or*

the anchor update (which depends only on κ_t). This acyclic structure prevents error amplification loops between components.

5 Theoretical Analysis of ANCHORFAIR

We theoretically prove that ANCHORFAIR achieves $\tilde{O}(\sqrt{T})$ group regret for every group under GTSA in terms of three perspectives: (i) the trust score reliability for perturbation detection, (ii) clustering correctness despite adversarial drift, and (iii) the demographic-aware exploration bonus.

5.1 Trust score detection

Theorem 3 (Trust score detection guarantee)**.** *Under Assumption 2 with n_w and threshold τ_r:*

(a) False positive rate: $\boldsymbol{P}[\kappa_t < \bar{\kappa} \mid \textit{no attack}] \leq 1 - \delta_{\mathrm{det}}$*, where δ_{det} is false positive rate.*

(b) Detection power: *With a detection threshold $\rho_w^* = \frac{2\Delta_{\max}}{\Delta_{\min}}\sqrt{\frac{2\log(2/\delta_{\mathrm{det}})}{n_w}}$, if $\rho_w \geq \rho_w^*$ in the window are attacked, then $\boldsymbol{P}[\kappa_t \leq \bar{\kappa}] \geq 1 - \delta_{\mathrm{det}}$. Since $\rho_w^* = \Theta(1/\sqrt{n_w})$ (the constants $\Delta_{\max}/\Delta_{\min}$ and $\log(1/\delta_{\mathrm{det}})$ being fixed), larger n_w detects weaker attacks but lengthens $\mathbb{E}[\tau_{\mathrm{det}}]$, extending the pre-detection phase.*

(c) Detection delay *under GTSA targeting group $g^\dagger$ with prevalence ρ: $\mathbb{E}[\tau_{\mathrm{det}}] \leq n_w\rho_w^*/\rho^2 + n_w$, where $\tau_{\mathrm{det}} = \inf\{t : \kappa_t \leq \bar{\kappa}\}$.*

Two-phase attack dynamics. Theorem 3 partitions any GTSA trajectory into two regimes with qualitatively different anchor behavior: **(1) Pre-detection ($t < \tau_{\mathrm{det}}$):** trust remains near 1, so $\alpha_t \approx \bar{\alpha}$ and the anchor drifts at the full EMA rate. Proposition 1's suppression factor $\bar{\kappa}^{p_{\min}}$ does *not* yet apply; the attacker accrues drift linearly in the duration of this phase, whose expected length is bounded by Theorem 3(c). **(2) Post-detection ($t \geq \tau_{\mathrm{det}}$):** trust has dropped to $\bar{\kappa}$, activating the suppression regime of Proposition 1. Further anchor shifts now require $W = \Omega(1/(\bar{\alpha}\bar{\kappa}^{p_{\min}}))$ sustained rounds per unit of TV distance, and the adaptive exponent drives $p_t \to p_{\max}$ to tighten the suppression further.

5.2 Clustering correctness

Warm-start initialization. ANCHORFAIR begins with a burn-in phase of T_0 rounds under standard NeuralUCB (with Huber-loss anchor training and CUSUM monitoring active from round 1) to learn a meaningful embedding ϕ before activating the full defense. Real demographic groups may overlap or be non-clusterable. Assumption 3 is therefore a sufficient but not necessary condition; clustering purity degrades gracefully when groups overlap in embedding space, and reduces to a standard robust bandit without group-level guarantees in the limit $\Delta_c \to 0$. The Δ_c-separation with intra-group diameter at most $\Delta_c/4$ is a strong geometric assumption that simplifies clustering analysis. It has two main consequences. First, clustering becomes effectively "easy" after warm-up: the online task is mainly to preserve separation under adversarial drift rather than discover new clusters. Second, the assumption excludes heavy-tailed or multimodal embedding distributions, which naturally arise when demographic groups correspond to non-convex attribute combinations.

Theorem 4 (Clustering correctness under GTSA)**.** *Under Assumptions 2 and 3 with attack window W and safe window $W^* = \Delta_c/(8L_\phi\bar{\alpha}\bar{\kappa}^{p_{\min}}\Psi)$ (proven in F.11) and $W \leq W^*$,*

(a) *Anti-fragmentation: $\boldsymbol{P}[\hat{g}(\mathbf{x}) \neq \hat{g}(\mathbf{x}') \mid F_{\mathrm{g}}(\mathbf{x}) = F_{\mathrm{g}}(\mathbf{x}')] \leq (G-1)\exp(-\Delta_c^2/(32\sigma_c^2))$.*

(b) *Anti-merging: $\boldsymbol{P}[\hat{g}(\mathbf{x}) = \hat{g}(\mathbf{x}') \mid F_{\mathrm{g}}(\mathbf{x}) \neq F_{\mathrm{g}}(\mathbf{x}')] \leq \exp(-9\Delta_c^2/(128\sigma_c^2))$.*

(c) *Overall purity: $\mathbb{E}[\mathcal{P}] \geq 1 - G\exp(-\Delta_c^2/(32\sigma_c^2)) \triangleq 1 - \eta_{\mathrm{cluster}}$.*

The safe window W^* uses $p_{\min}$ as a conservative bound; under active attack, p_t increases toward $p_{\max}$, yielding a longer effective safe window $W_{\mathrm{eff}}^* = \Delta_c/(8L_\phi\bar{\alpha}\bar{\kappa}^{p_{\max}}\Psi) \geq W^*$. A guarantee requires $\sigma_c = \Theta(\Delta_c/\sqrt{\log(GT)})$, where Δ_c is an oracle quantity. In practice, σ_c is set adaptively using the silhouette score and estimate $\hat{\Delta}_c^t = \min_{g\neq g'}\|c_g^t - c_{g'}^t\|$ and set $\sigma_c = \hat{\Delta}_c^t/\sqrt{\log(GT)}$.

5.3 Fairness guarantees

The exploration bonus ensures equitable learning across groups.

Definition 2 (Regret equity)**.** *A policy achieves* $(\epsilon_{\text{eq}}, T)$*-regret equity if* $|R_T^{(g)}/n_g - R_T^{(g')}/n_{g'}| \le \epsilon_{\text{eq}}$ *for all pairs* $g, g' \in [G]$.

Theorem 5 (Regret equity under ANCHORFAIR)**.** *Under budget-constrained GTSA (*$B \le c\sqrt{T}$ *for a universal constant* $c > 0$*) and* $\gamma_0 \ge 4\Delta_{\max} \log T$*,* ANCHORFAIR *achieves* $(\epsilon_{\text{eq}}, T)$*-regret equity with:*

$$\epsilon_{\text{eq}} = O\left(\sqrt{\frac{d\log(KT)}{n_{\min}}} + \frac{B}{n_{\min}\sqrt{n_{\min}}}\right) = \tilde{O}\left(\sqrt{\frac{G}{T}}\right) \quad \textit{when } n_{\min} = \Omega(T/G), \tag{16}$$

where $n_{\min} = \min_{g\in[G]} n_g$.

Regret equity bounds disparities in *learning quality* relative to each group's conditional optimum, not in realized outcomes. The two coincide when groups share the same optimal-arm reward distribution; otherwise, equalizing outcomes requires layering an additional fairness criterion [Grazzi et al., 2022, Joseph et al., 2016] on top, compatible with ANCHORFAIR's regret-gap correction $\gamma_t[\hat{\Delta}_g^t]_+$ but beyond our present scope.

5.4 Group regret bound under ANCHORFAIR

Theorem 6 (Group regret bound under ANCHORFAIR)**.** *Consider a GTSA with budget* $B \le c\sqrt{T}$ *for a universal constant* $c > 0$ *and target group* $g^\dagger$ *with* $n_{g^\dagger} = \Omega(\sqrt{T})$*. Under* ANCHORFAIR *with appropriately tuned parameters* $(\bar{\alpha}, p_{\min}, p_{\max}, \beta, \lambda, \mu, \gamma_0)$*, assuming polynomially decaying NTK eigenvalues so that* $\tilde{\gamma}_T = O(\sqrt{T})$*, the group regret satisfies for all* $g \in [G]$*:*

$$R_T^{(g)} = O\left(\sqrt{n_g T}\cdot d\log(KT) + \frac{B}{\sqrt{n_g}}\cdot d\right). \tag{17}$$

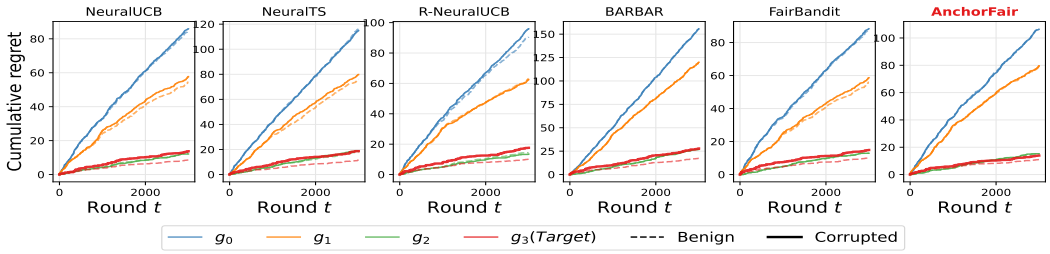
In particular, when $B = O(\sqrt{T})$ *and* $n_g = \Theta(T/G)$*:*

$$R_T^{(g)} = \tilde{O}\left(\frac{T}{\sqrt{G}}\cdot d\right) = \tilde{O}(\sqrt{T}) \quad \textit{for fixed } G, d. \tag{18}$$

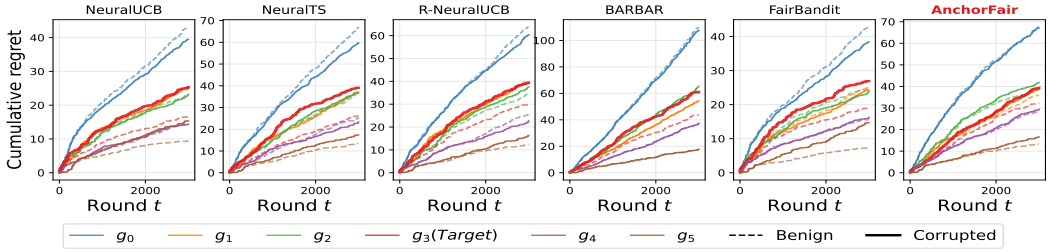
Proof sketch. The proof proceeds in three stages. First, the bounded drift of the anchor (Proposition 1) ensures that the trust score accurately detects sustained perturbations, yielding $\kappa_t \le \bar{\kappa}$ during attack periods with high probability (Theorem 3). Second, representation stability (Proposition 2) guarantees that clustering correctly assigns at least a $(1 - \eta_{\text{cluster}})$-fraction of group g's contexts to the same cluster (Theorem 4). Third, the exploration bonus(Proposition 12) ensures that group g accumulates $\tilde{\Omega}(\sqrt{n_g T})$ exploratory samples, which combined with the neural approximation guarantee yields the stated bound. The budget term $B/\sqrt{n_g}$ captures residual regret from perturbations that evade trust detection and reveals a phase transition. When $B = O(\sqrt{T})$, this term is absorbed into the $\tilde{O}(\sqrt{T})$ rate, so every group matches the unattacked regret up to logarithmic factors. Beyond this threshold, the attacker's residual budget overwhelms trust-mediated suppression, and group regret grows with the excess. This boundary is information-theoretically tight: Theorem 10 shows that *any* defense is vulnerable to super-$\sqrt{T}$ budgets, so ANCHORFAIR neutralizes the strongest attacks that are fundamentally defensible. See Appendix F.8 for the complete argument. □

6 Experiments

Experimental Setup. Baselines. We compare against: (i) undefended NeuralUCB [Zhou et al., 2020] and NeuralTS [Zhang et al., 2021], (ii) robust bandit defenses including R-NeuralUCB Qi et al. [2024] and BARBAR Gupta et al. [2019], and (iii) fair bandit algorithms including FairBandit [Patil et al., 2021]. **Attack strategies.** We implement (i) the proposed GTSA with varying budgets



(a) COMPAS Dataset.



(b) MovieLens Dataset.

Figure 1: Cumulative victim regret for all groups (benign (dashed) corrupted (solid)). The vertical dashed line at round t = 400 marks the end of the GTSA reconnaissance phase, after which context perturbations target the demographic group.

$B \in \{0.1\sqrt{T}, 0.5\sqrt{T}, \sqrt{T}, 2\sqrt{T}\}$, (ii) uniform random perturbations (non-targeted), and (iii) the context perturbation attack of Ma et al. [2018]. **Metrics.** We evaluate: (i) global cumulative regret R_T, (ii) worst-group regret $\max_g R_T^{(g)}$, (iii) regret gap $\max_{g,g'} |R_T^{(g)}/n_g - R_T^{(g')}/n_{g'}|$, (iv) attack detectability via AUROC of aggregate-only vs. group-aware monitors, and (v) cluster purity relative to ground-truth groups. **Datasets.** (i) Semi-synthetic based on COMPAS dataset [2] with injected group structure; (ii) MovieLens dataset [3] with demographic groups (The details for datasets and configurations are explained in Appendix B.).

Results. Figure 1 compares the regret for all demographic subgroups under benign and corrupted contexts. It reveals that AnchorFair effectively protects the sensitive subgroup against GTSA and the damage of the target group ($g^\dagger = g_3$) was less than any baseline. AnchorFair's amplification ratio of ×1.52 stands well below the ×2.07–×2.45 range observed across all other methods, and its per-capita regret gap under attack (0.041) is the smallest by a wide margin compared to R-NeuralUCB (0.096), BARBAR (0.134), and FairBandit (0.076). Notably, R-NeuralUCB exhibits the worst group-level vulnerability, confirming that aggregate robustness can actively harm minorities by down-weighting their corrupted observations as outliers. AnchorFair's three-phase defense (anchor accumulation, CUSUM-triggered selective adaptation, and arm-specific exploration) produces the tightest clustering of group trajectories under attack, demonstrating that its protection generalizes across different group structures and dataset characteristics. The per-group breakdown in Table 2 (Appendix H) reveals robustness differences under GTSA that are invisible at the aggregate level.

Fig. 2 shows the percentage increase in target-group regret under each attack scenario relative to the benign baseline of the same algorithm. Because each bar is normalized by the algorithm's own benign performance, this metric isolates how much harder a given defense is hit by the attacker and removes the confound of differing learning speeds. AnchorFair's amplification stays within roughly -2% to $+20\%$ across all GTSA budgets and the Uniform-attack control on both datasets, while BARBAR and FairBandit exhibit amplifications of $+30\%$ to $+200\%$, quantitative evidence of the "robust-but-not-fair" trap, in which median-of-means and proportional-fair updates inadvertently down-weight or equalize on corrupted minority-group signals. The flat amplification curve of AnchorFair across attack budgets is the empirical counterpart of the phase transition predicted by Theorem 6: as long

[2] https://github.com/propublica/compas-analysis
[3] https://grouplens.org/datasets/movielens/

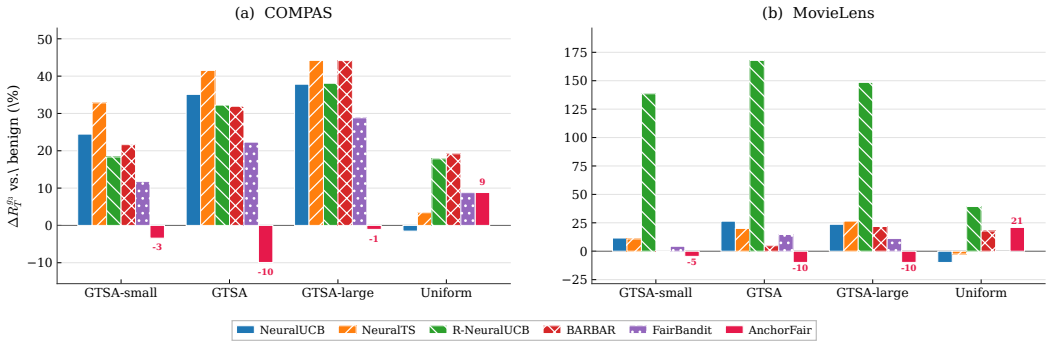


Figure 2: Percentage increase in target-group regret relative to the benign run.

Table 1: Scalability comparison across datasets and numbers of arms.

(a) Wall-clock time per round (ms).

Data	K	NUCB	NTS	R-NUCB	BARBAR	FairB	AF
COMPAS	5	0.08	0.07	0.09	0.18	0.03	0.14
MovieLens	5	0.19	0.20	0.23	0.19	0.05	0.20
COMPAS	10	0.18	0.15	0.19	0.21	0.09	0.30
MovieLens	10	0.41	0.34	0.45	0.34	0.13	0.43
COMPAS	20	0.37	0.27	0.35	0.50	0.21	0.62
MovieLens	20	0.83	0.62	0.83	0.80	0.32	0.90
COMPAS	30	0.65	0.47	0.65	0.94	0.41	1.14
MovieLens	30	1.36	1.06	1.34	1.53	0.63	1.70

(b) Peak memory per round (MB).

Data (G)	K	NUCB	NTS	R-NUCB	BARBAR	FairB	AF
COMPAS (4)	5	0.11	0.11	0.11	0.02	0.01	0.02
MovieLens (6)	5	0.11	0.11	0.11	0.02	0.02	0.03
COMPAS (4)	10	0.11	0.11	0.11	0.03	0.02	0.04
MovieLens (6)	10	0.11	0.11	0.11	0.03	0.02	0.05
COMPAS (4)	20	0.11	0.11	0.11	0.04	0.02	0.07
MovieLens (6)	20	0.11	0.11	0.11	0.04	0.03	0.10
COMPAS (4)	30	0.12	0.12	0.12	0.05	0.03	0.10
MovieLens (6)	30	0.12	0.12	0.12	0.05	0.03	0.15

as $B = O(\sqrt{T})$, the trust-anchored detection-and-suppression pipeline keeps target-group regret essentially indistinguishable from its benign value.

Table 1 summarizes runtime and memory usage as the number of arms increases from $K = 5$ to $K = 30$, with $G = 4$ for COMPAS and $G = 6$ for MovieLens. ANCHORFAIR's runtime increases approximately linearly with K, reaching 1.14 ms and 1.70 ms per round on COMPAS and MovieLens, respectively. Its reported peak memory remains at or below 0.15 MB across all tested settings. These results indicate modest computational costs within the evaluated range, although scalability to larger numbers of groups requires further evaluation.

7 Conclusions

We have identified and formalized group-targeted suppression attacks (GTSA), a class of adversarial threats against neural contextual bandits that exploit the gap between aggregate and group-level performance metrics. Our impossibility result (Theorem 10) demonstrates that aggregate monitoring alone is fundamentally insufficient to detect GTSA. The ANCHORFAIR framework addresses this vulnerability through trust-weighted policy anchoring, prediction-error-aware demographic discovery, and arm-specific demographic-aware exploration, achieving $\tilde{O}(\sqrt{T})$ group regret guarantees for all groups under GTSA (Theorem 6). Our analysis assumes the attacker's group classifier does not adapt to the defender's clustering strategy. Extending ANCHORFAIR to adaptive attackers—for example through randomized clustering, differential privacy mechanisms on cluster centroids, or minimax formulations over the attacker's strategy space—represents an important direction for future work. The GTSA threat model also extends naturally to reinforcement learning settings, where state-dependent demographic structures may create additional opportunities for targeted suppression. Moreover, the proposed defense framework is agnostic to the decision-making formalism and could potentially be extended to contextual Markov decision processes.

8 Acknowledgment

The authors thank the anonymous reviewers for their insightful comments and constructive feedback.

References

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

Jackie Baek and Vivek F. Farias. Fair exploration via axiomatic bargaining. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2026.

Yikun Ban, Yuchen Yan, Arindam Banerjee, and Jingrui He. Ee-net: Exploitation-exploration neural networks in contextual bandits. *10th International Conference on Learning Representations (ICLR)*, 2022.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits. *Advances in Neural Information Processing Systems*, 35:23951–23964, 2022.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

Evgenii Chzhen, Christophe Giraud, Zhen Li, and Gilles Stoltz. Small total-cost constraints in contextual bandits with knapsacks, with application to fairness. *Advances in Neural Information Processing Systems*, 36:46737–46772, 2023.

Matthew Crain. The limits of transparency: Data brokers and commodification. *New media & society*, 20(1):88–104, 2018.

Michael Dang'ana, Yuqiu Zhang, and Hans-Arno Jacobsen. Ksurf-drone: Attention kalman filter for contextual bandit optimization in cloud resource allocation. *IEEE Transactions on Cloud Computing*, 2026.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions: Y. deldjoo et al. *User Modeling and User-Adapted Interaction*, 34(1):59–108, 2024.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 7111–7123. PMLR, 2022.

Dongqi Fu and Jingrui He. Sdg: A simplified and dynamic graph neural network. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2273–2277, 2021.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *34th Advances in Neural Information Processing Systems (NeurIPS)*, 33:14362–14373, 2020.

Riccardo Grazzi, Arya Akhavan, John Isak Texas Falk, Leonardo Cella, and Massimiliano Pontil. Group meritocratic fairness in linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems (NeurIPS)*, 35:34614–34625, 2022.

Benedikt Höltgen and Nuria Oliver. Reconsidering fairness through unawareness from the perspective of model multiplicity. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 241–255, 2025.

Seyed Mohammad Hadi Hosseini, Amir Najafi, and Mahdieh Soleymani Baghshah. Efficient adversarial attacks on high-dimensional offline bandits. In *International Conference on Learning Representations*.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *32nd International conference on machine learning (ICML)*, pages 767–776. PMLR, 2015.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *36th International Conference on Machine Learning (ICML)*, pages 4042–4050. PMLR, 2019.

Guanlin Liu and Lifeng Lai. Action poisoning attacks on linear contextual bandits. *Transactions on Machine Learning Research*, 2023.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. Data poisoning attacks in contextual bandits. In *International Conference on Decision and Game Theory for Security*, pages 186–204. Springer, 2018.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.

Daniel Morales-Brotons, Thijs Vogels, et al. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. arXiv:2411.18704.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron Courville. Language model alignment with elastic reset. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174):1–31, 2021.

Yunzhe Qi, Yikun Ban, Arindam Banerjee, and Jingrui He. Robust neural contextual bandit against adversarial corruptions. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 19378–19446, 2024.

Sina Shaham, Arash Hajisafi, Minh K Quan, Dinh C Nguyen, Bhaskar Krishnamachari, Charith Peris, Gabriel Ghinita, Cyrus Shahabi, and Pubudu N Pathirana. Privacy and fairness in machine learning: A survey. *IEEE Transactions on Artificial Intelligence*, 6(7):1706–1726, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ray Telikani and Amir H Gandomi. Learning to attack: A bandit approach to adversarial context poisoning. *arXiv preprint arXiv:2603.00567*, 2026.

Matilde Tullii, Solenne Gaucher, Nadav Merlis, and Vianney Perchet. Improved algorithms for contextual dynamic pricing. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 126088–126117, 2024.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Ziyan Wang, Xiaoming Huo, and Hao Wang. Towards domain adaptive neural contextual bandits. *13th International Conference on Learning Representations (ICLR)*, 2025.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR, 2021.

Qirun Zeng, Eric He, Richard Hoffmann, Xuchuang Wang, and Jinhang Zuo. Practical adversarial attacks on stochastic bandits via fake data injection. *arXiv preprint arXiv:2505.21938*, 2025.

Lin Zhang et al. EMA policy gradient: Taming reinforcement learning for LLMs with EMA anchor and top-k KL, 2026. arXiv:2602.04417.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021.

Chen Zhao, Feng Chen, and Bhavani Thuraisingham. Fairness-aware online meta-learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2294–2304, 2021.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: a survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–28, 2025.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning (ICML)*, pages 11492–11502. PMLR, 2020.

A Notations

Symbol	Meaning	Symbol	Meaning
T, K, G	Horizon, arms, groups	$\bar{\alpha}$, α_t	Max/adaptive anchor learning rate
$\mathbf{x}_{t,i}$, $\tilde{\mathbf{x}}_{t,i}$	True/perturbed context	κ_t, $\bar{\kappa}$	Trust score / threshold
$h(\cdot)$, ξ_t	Reward function / noise	p_t	Adaptive sensitivity exponent in t
π_t, π_a^t	Main / anchor policy	β, λ, μ	Exploration parameters
$\phi(\cdot)$, d_a	Anchor embedding / dim.	γ_t, γ_0, η_γ	Regret-gap correction params
$F_{\mathrm{g}}(\cdot)$	Group mapping function	σ_c, ω_t	Clustering temp. / error weight
$g^\dagger$, B, ϵ	Target group, budget, magnitude	Δ_c	Inter-group separation
n_g, ρ_g	Group sample size / prevalence	η_{cluster}	Clustering error rate
$\Delta_{\min}$, $\Delta_{\max}$	Gap bounds	τ_r	reward threshold
C_ϕ, L_ϕ, Ψ	Embedding constants	δ_{CUSUM}, h_{CUSUM}	CUSUM parameters
W	Attack window duration	n_w	Trust score sliding window size
W^*	Safe window for clustering	$p_{\min}, p_{\max}$	Adaptive sensitivity bounds
$\mathcal{R}$	RKHS norm bound		

B Datasets and Setup

Datasets. (1) **COMPAS** is a popular commercial algorithm used by judges and parole officers for scoring criminal defendants' likelihood of reoffending (recidivism). The COMPAS dataset contains 6167 entries of defendants from Broward County, FL, USA, for which COMPAS risk scores were available Bao et al. [2021]. (2) **MovieLens Dataset** Harper and Konstan [2015] consists of 25 million ratings between 1.6×10^5 users and 6×10^4 movies.

Data Preprocessing. For *MovieLens* dataset, we follow the experimental setup conducted in Ban et al. [2022]. A rating matrix was constructed by selecting the top 2,000 users and the top 10,000 restaurants (or movies), and singular value decomposition (SVD) was applied to obtain 10-dimensional latent feature vectors for both users and items. The bandit learner aims to identify restaurants (or movies) associated with poor ratings. **Context Construction.** At each round t, we randomly sample one user and form a candidate set of $K = 10$ items. The set consists of one item with reward 1 (rating < 2 stars) and nine items with reward 0 (rating ≥ 2 stars). The context for each arm is the concatenation of user and item latent vectors $\mathbf{x}_{t,a} \in \mathbb{R}^{20}$. For the *Disin* dataset, we followed Fu and He [2021] to convert the text into vector representations, with each article encoded as a 300-dimensional vector. In each round, a 10-arm pool was constructed by randomly sampling 9 real news articles and 1 fake news article. Selecting the fake news article yielded a reward of 1, while selecting any real news article resulted in a reward of 0.

Victim Algorithm Configuration. For **NeuralUCB and NeuralTS** Zhou et al. [2020], Zhang et al. [2021], we use a neural network with two hidden layers of size $[100, 100]$ and ReLU activations. Training is performed using SGD with the learning rate $\eta = 0.01$, batch size 32, and 100 gradient steps per round. The confidence scaling parameter is set to $\nu = 1.0$. **R-NeuralUCB:** the architecture matches NeuralUCB, augmented with a trust-weight mechanism. The trust threshold is $\tau_{\text{trust}} = 0.1$. Gradient covariance is estimated using an exponential moving average with decay $\alpha = 0.95$. **BARBAR** Gupta et al. [2019]: We set $\lambda = 1024$ for the number of pulls per arm per epoch and $\delta = 1/T$ for confidence parameter. **FairBandit** Patil et al. [2021]: We use fairness fraction $r = 0.05$ and unfairness tolerance $\alpha = 1.0$. **AnchorFair Hyperparameters.** We set EMA learning rate $\bar{\alpha} = 0.05$, sensitivity exponent range $p_{min} = 1$, $p_{max} = 5$, Sliding window size $n_w = 100$, Max number of groups $G_{max} = 6$, and Anchor-disagreement factor $\mu = 0.5$.

Computational Environment. All experiments were conducted on a machine equipped with a 20-core Intel CPU running at 3.9GHz, 64GB of RAM, and a NVIDIA GeForce RTX5070 GPU with 12GB of memory, ensuring efficient training of neural networks. The software environment included Python3.10, PyTorch2.5.1, and NumPy2.1.3.

C Related Works

Adversarial attacks on bandits. Adversarial attacks on bandit algorithms have been studied through reward poisoning [Garcelon et al., 2020, Liu and Shroff, 2019, Ding et al., 2022], action manipulation [Liu and Lai, 2023], and context perturbation [Ma et al., 2018, Zeng et al., 2025]. These works uniformly target global regret degradation. Hosseini et al. recently proposed efficient context-space attacks for neural bandits, but again with the objective of maximizing aggregate regret. The notion that an attacker might selectively target a subpopulation while preserving aggregate metrics has not been formalized in the bandits literature.

Robust contextual bandits. Robust bandit methods defend against various forms of corruption, including heavy-tailed noise [Medina and Yang, 2016], Huber-contamination models [Gupta et al., 2019], and adversarial reward corruption with bounded budget [Lykouris et al., 2018, Qi et al., 2024, He et al., 2022]. However, existing robustness guarantees are stated in terms of global regret and do not address group-level vulnerabilities. We demonstrate that global robustness guarantees are fundamentally insufficient to prevent group-targeted harm.

Fairness in bandits. Fairness-aware bandit algorithms address disparities in treatment quality across groups [Joseph et al., 2016, Patil et al., 2021, Zhao et al., 2021, Li et al., 2019]. Grazzi et al. [2022] introduced *group meritocratic fairness* in linear contextual bandits, defining fairness via relative rank within each demographic group and proving $\tilde{O}(\sqrt{dT})$ fair pseudo-regret for a greedy policy. Baek and Farias [2026] studied the distribution of exploration cost across groups through Nash bargaining, showing that regret-optimal policies are arbitrarily unfair and proposing proportionally fair UCB policies with bounded price of fairness. These works typically assume a benign environment and focus on algorithmic bias or exploration cost-sharing rather than adversarial manipulation. Our work is complementary: we show that even a fair bandit algorithm can be rendered discriminatory by a targeted adversary, motivating the need for adversarial robustness in fairness-critical deployments.

Adversarial robustness and fairness. The intersection of adversarial robustness and fairness has been explored in supervised learning [Xu et al., 2021, Nanda et al., 2021], where adversarial examples disproportionately affect underrepresented groups. Our work extends this concern to sequential decision-making under bandit feedback, introducing novel challenges related to adaptive exploration, non-stationary trust, and delayed group discovery.

D Main Policy Sensitivity to Context Perturbation

Proposition 7 (Main policy sensitivity to context perturbation)**.** *Let π_t be updated via online SGD with learning rate η_{lr}:*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_{\mathrm{lr}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \tilde{\mathbf{x}}_{t,a_t}, r_{t,a_t}), \tag{19}$$

where ℓ is the training loss evaluated on the (potentially perturbed) context. Suppose $\nabla_{\boldsymbol{\theta}}\ell$ is L_∇-Lipschitz in the context argument, i.e., $\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; \mathbf{x}, r) - \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; \mathbf{x}', r)\| \le L_\nabla \|\mathbf{x} - \mathbf{x}'\|_\infty$ for all $\boldsymbol{\theta}, r$. Then under a GTSA perturbation with $\|\boldsymbol{\delta}_t\|_\infty \le \epsilon$, a single attacked round shifts the main policy parameters by $\|\boldsymbol{\theta}^{\mathrm{att}}_{t+1} - \boldsymbol{\theta}^{\mathrm{clean}}_{t+1}\| \le \eta_{\mathrm{lr}} \cdot L_\nabla \cdot \epsilon$, and over W consecutive attacked rounds, the cumulative drift satisfies $\|\boldsymbol{\theta}^{\mathrm{att}}_{t_0+W} - \boldsymbol{\theta}^{\mathrm{clean}}_{t_0+W}\| \le \eta_{\mathrm{lr}} \cdot L_\nabla \cdot \epsilon \cdot W$.

Proof. Let $\boldsymbol{\theta}^{\mathrm{att}}_{t+1} = \boldsymbol{\theta}_t - \eta_{\mathrm{lr}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \tilde{\mathbf{x}}_{t,a_t}, r)$ and $\boldsymbol{\theta}^{\mathrm{clean}}_{t+1} = \boldsymbol{\theta}_t - \eta_{\mathrm{lr}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \mathbf{x}_{t,a_t}, r)$. Then $\|\boldsymbol{\theta}^{\mathrm{att}}_{t+1} - \boldsymbol{\theta}^{\mathrm{clean}}_{t+1}\| = \eta_{\mathrm{lr}} \|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \tilde{\mathbf{x}}, r) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \mathbf{x}, r)\| \le \eta_{\mathrm{lr}} \cdot L_\nabla \cdot \epsilon$. The cumulative bound follows by the triangle inequality over W rounds. □

Corollary 8 (Anchor advantage under attack)**.** *Under a sustained GTSA of duration W, the per-round parameter drift rates of the main policy (Proposition 7) and anchor (Proposition 1) satisfy:*

$$\underbrace{\eta_{\mathrm{lr}} \cdot L_\nabla \cdot \epsilon}_{\textit{main policy drift/round}} \quad \textit{vs.} \quad \underbrace{\bar{\alpha}\bar{\kappa}^{p_t} \cdot \|\Delta \mathbf{w}\|_\infty}_{\textit{anchor drift/round}}. \tag{20}$$

When the trust score drops ($\bar{\kappa} \to 0$), the anchor drift vanishes as $O(\bar{\kappa}^{p_t})$ while the main policy continues to drift at rate $\Theta(\eta_{\mathrm{lr}} \cdot L_\nabla \cdot \epsilon)$ per round, independent of trust. Under attack, the adaptive $p_t \to p_{\max}$ further strengthens this asymmetry.

E Impossibility of Aggregate Detection

Definition 3 (Aggregate monitor)**.** *An aggregate monitor $\mathcal{M}$ is a function that, at each round t, observes only global summary statistics—the running average reward $\bar{r}_t = \frac{1}{t}\sum_{s=1}^{t} r_{s,a_s}$, the running regret estimate $\hat{R}_t$, and the empirical reward distribution $\hat{P}_t^{\text{global}}$—and outputs a binary decision $\mathcal{M}_t \in \{0,1\}$ indicating whether an attack is detected. Critically, $\mathcal{M}$ has no access to group labels or group-conditional statistics.*

Lemma 9 (Budget-feasible targeted perturbation)**.** *Let π be any contextual bandit policy and let group $g^\dagger$ have prevalence $\rho = n_{g^\dagger}/T$. Suppose the reward function h is L_h-Lipschitz and the suboptimality gaps satisfy $\Delta_{\min} \le \Delta_{t,i} \le \Delta_{\max}$. Then there exists a perturbation strategy $\{\boldsymbol{\delta}_{t,i}\}$ satisfying $\|\boldsymbol{\delta}_{t,i}\|_\infty \le \epsilon$ that forces the learner to select a suboptimal arm on every round $t \in \tau_{g^\dagger}$, provided $\epsilon \ge \Delta_{\min}/(2L_h)$, with total budget consumption*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \|\boldsymbol{\delta}_{t,i}\|_\infty \le K \cdot \epsilon \cdot n_{g^\dagger} = K\epsilon\rho T. \tag{21}$$

Proof. Let $a_t^* = \arg\max_{i\in[K]} h(\mathbf{x}_{t,i})$ be the optimal arm and let $a_t^\dagger \neq a_t^*$ be the attacker's target arm. Since h is L_h-Lipschitz w.r.t. ℓ_∞, for any $\mathbf{x} \in \mathbb{R}^d$ and any $\epsilon > 0$, there exist directions $\boldsymbol{\delta}^+, \boldsymbol{\delta}^- \in \mathbb{R}^d$ with $\|\boldsymbol{\delta}^\pm\|_\infty \le \epsilon$ such that

$$h(\mathbf{x}+\boldsymbol{\delta}^+) - h(\mathbf{x}) \ge L_h\epsilon, \qquad h(\mathbf{x}+\boldsymbol{\delta}^-) - h(\mathbf{x}) \le -L_h\epsilon, \tag{22}$$

by the definition of the Lipschitz constant as a supremum over displacement directions (this holds without requiring differentiability of h). The attacker sets $\boldsymbol{\delta}_{t,a_t^\dagger} = \boldsymbol{\delta}^+$ (boosting the target arm) and $\boldsymbol{\delta}_{t,a_t^*} = \boldsymbol{\delta}^-$ (suppressing the optimal arm), with $\boldsymbol{\delta}_{t,i} = \mathbf{0}$ for $i \notin \{a_t^*, a_t^\dagger\}$. The attacked ranking flips when $2L_h\epsilon \ge \Delta_{t,a_t^\dagger}$, which is ensured by $\epsilon \ge \Delta_{\min}/(2L_h)$. For rounds $t \notin \tau_{g^\dagger}$, no perturbation is applied, giving total budget $K\epsilon n_{g^\dagger} = K\epsilon\rho T$. □

Theorem 10 (Impossibility of aggregate detection)**.** *Let π be any contextual bandit policy achieving sublinear global regret $R_T^{\text{clean}} = O(\sqrt{T})$ in the unattacked setting. Suppose the target group $g^\dagger$ has prevalence $\rho = n_{g^\dagger}/T$ and the attacker operates under Definition 1 with per-round magnitude $\epsilon \ge \Delta_{\min}/(2L_h)$ and total budget B. Then, provided $B \ge K\epsilon\rho T$, there exists a GTSA strategy such that:*

(i) The global regret under attack satisfies $R_T^{\text{att}} \le R_T^{\text{clean}} + \rho T \cdot \Delta_{\max} = O(\sqrt{T} + \rho T\Delta_{\max})$.

(ii) The target group suffers linear group regret: $R_T^{(g^\dagger)} \ge n_{g^\dagger} \cdot \Delta_{\min} = \Omega(n_{g^\dagger})$.

(iii) For any aggregate monitor $\mathcal{M}$ (Definition 3) and any significance level $\alpha > 0$:

$$\boldsymbol{P}\big[\mathcal{M}_T = 1 \mid \textit{GTSA active}\big] \le \alpha + \rho. \tag{23}$$

In particular, when $B = \Omega(\sqrt{T})$ and $\rho = B/(K\epsilon T)$, the attack achieves $R_T^{(g^\dagger)} = \Omega(B\Delta_{\min}/(K\epsilon))$ group regret while maintaining detection power at most $\alpha + B/(K\epsilon T) \to \alpha$ as $T \to \infty$.

Proof. **Part (i).** Decompose $R_T^{\text{att}} = \sum_{t\notin\tau_{g^\dagger}} \mathbb{E}[\Delta_{t,a_t}] + \sum_{t\in\tau_{g^\dagger}} \mathbb{E}[\Delta_{t,a_t}]$. Unperturbed rounds contribute $\le R_T^{\text{clean}}$; targeted rounds contribute $\le n_{g^\dagger} \cdot \Delta_{\max}$. When $\rho = O(T^{-1/2})$, this gives $R_T^{\text{att}} = O(\sqrt{T})$.

Part (ii). By Theorem 9, the attacker forces a suboptimal arm on every $t \in \tau_{g^\dagger}$, giving $R_T^{(g^\dagger)} \ge n_{g^\dagger} \cdot \Delta_{\min}$.

Part (iii). The empirical reward distributions under clean and attacked settings differ only on the ρ-fraction of targeted rounds, so $\text{TV}(\hat{P}_T^{\text{att}}, \hat{P}_T^{\text{clean}}) \le \rho$. By Le Cam's method [Le Cam, 2012], the sum of Type I and Type II errors is at least $1-\rho$. Since the monitor's false positive rate is at most α, detection power is bounded by $\alpha + \rho$. □

Remark 2 (Concentration-based refinement)**.** *For mean-reward monitors specifically, the per-round shift is $\rho\Delta_{\max}$, and by Hoeffding's inequality: $\boldsymbol{P}[|\bar{r}_T^{\text{att}} - \bar{r}_T^{\text{clean}}| > \rho\Delta_{\max}/2] \leq 2\exp(-T\rho^2\Delta_{\max}^2/(8\sigma^2))$. When $\rho = O(T^{-1/2})$, the exponent is $O(1)$, confirming that even the most natural scalar statistic fails to detect the attack.*

F Supporting Theoretical Analysis

F.1 Proof of Proposition 1 (Bounded Anchor Drift)

Proof. Recall the anchor update: $\mathbf{w}_a^{t+1} = (1-\alpha_t)\mathbf{w}_a^t + \alpha_t \mathbf{w}_a^{\pi_t}$ with $\alpha_t = \bar{\alpha}\kappa_t^{p_t}$. Since $\kappa_t \leq \bar{\kappa} < 1$ during the attack window and $p_t \geq p_{\min}$, we have $\kappa_t^{p_t} \leq \kappa_t^{p_{\min}} \leq \bar{\kappa}^{p_{\min}}$ (because $\kappa_t \in [0,1]$ and raising to a higher power makes it smaller). Thus $\alpha_t \leq \bar{\alpha}\bar{\kappa}^{p_{\min}}$. Telescoping over $[t_0, t_0+W]$:

$$\mathbf{w}_a^{t_0+W} - \mathbf{w}_a^{t_0} = \sum_{s=t_0}^{t_0+W-1} \alpha_s(\mathbf{w}_a^{\pi_s} - \mathbf{w}_a^s). \tag{24}$$

Taking the ℓ_∞ norm:

$$\|\mathbf{w}_a^{t_0+W} - \mathbf{w}_a^{t_0}\|_\infty \leq \sum_{s=t_0}^{t_0+W-1} \bar{\alpha}\bar{\kappa}^{p_{\min}} \|\mathbf{w}_a^{\pi_s} - \mathbf{w}_a^s\|_\infty \leq \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot W \cdot \|\Delta\mathbf{w}\|_\infty. \tag{25}$$

To convert to TV distance: for the softmax $\sigma : \mathbb{R}^K \to \Delta^{K-1}$, the Jacobian satisfies $\|\nabla\sigma(\mathbf{v})\|_{\ell_1/\ell_\infty} \leq 1/2$. The logits shift by at most $\|\phi(\mathbf{x})\|_1 \cdot \|\mathbf{w}_a^{t_0+W} - \mathbf{w}_a^{t_0}\|_\infty \leq C_\phi \cdot \bar{\alpha}\bar{\kappa}^{p_{\min}} W \|\Delta\mathbf{w}\|_\infty$ where $C_\phi = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|_1$. Since TV equals half the ℓ_1 distance on probability vectors, the mean value theorem gives:

$$\text{TV}(\pi_a^{t_0+W}(\cdot|\mathbf{x}), \pi_a^{t_0}(\cdot|\mathbf{x})) \leq \frac{K}{2} \cdot C_\phi \cdot \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot W \cdot \|\Delta\mathbf{w}\|_\infty. \tag{26}$$

Absorbing $1/2$ into C_ϕ and taking the supremum over $\mathbf{x}$ yields the stated bound. □

F.2 Proof of Theorem 2 (Representation Stability)

Proof. Let $\phi^t(\mathbf{x}) = \phi(\mathbf{x}; \boldsymbol{\theta}_\phi^t)$. The embedding parameters evolve via $\boldsymbol{\theta}_\phi^{t+1} = (1-\alpha_t)\boldsymbol{\theta}_\phi^t + \alpha_t \boldsymbol{\theta}_\phi^{\pi_t}$. By the same argument as Section F.1, using $\alpha_t \leq \bar{\alpha}\bar{\kappa}^{p_{\min}}$:

$$\|\boldsymbol{\theta}_\phi^{t_0+W} - \boldsymbol{\theta}_\phi^{t_0}\|_\infty \leq \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot W \cdot \Psi, \tag{27}$$

where $\Psi = \max_s \|\boldsymbol{\theta}_\phi^{\pi_s} - \boldsymbol{\theta}_\phi^s\|_\infty$. Since ϕ is L_ϕ-Lipschitz in its parameters, for any $\mathbf{x}$: $\|\phi^{t_0+W}(\mathbf{x}) - \phi^{t_0}(\mathbf{x})\| \leq L_\phi\bar{\alpha}\bar{\kappa}^{p_{\min}} W\Psi$. For same-group contexts $\mathbf{x}, \mathbf{x}'$, the triangle inequality gives:

$$\|\phi^{t_0+W}(\mathbf{x}) - \phi^{t_0+W}(\mathbf{x}')\| \leq \|\phi^{t_0}(\mathbf{x}) - \phi^{t_0}(\mathbf{x}')\| + 2L_\phi\bar{\alpha}\bar{\kappa}^{p_{\min}} W\Psi. \tag{28}$$

For the augmented features $\tilde{\mathbf{z}}_t = [\mathbf{z}_t; \omega_t \bar{e}_g^t]$, the error component $\bar{e}_g^t$ is a running average with bounded increments $O(1/n_g^t)$, so $\|\tilde{\mathbf{z}}^{t_0+W} - \tilde{\mathbf{z}}^{t_0}\|$ is controlled by the embedding drift plus $O(\omega_0/n_g^{t_0})$, preserving stability. □

F.3 Proof of Theorem 3 (Trust Score Detection)

Proof. **Part (a).** Without attack, $D_{\text{KL}}(\pi_t \| \pi_a^t) \in [0, \log K]$. By Hoeffding's inequality on the windowed average:

$$\boldsymbol{P}\left[\frac{1}{n_w}\sum_s D_{\text{KL}}(\pi_s\|\pi_a^s) > \log(1/\bar{\kappa}) + \epsilon'\right] \leq \exp\left(-\frac{2n_w\epsilon'^2}{(\log K)^2}\right). \tag{29}$$

For the reward indicator, $\hat{\mu}_w^t - \hat{\mu}_w^{t-n_w}$ is the difference of two sample means of n_w σ-sub-Gaussian variables:

$$\boldsymbol{P}\left[|\hat{\mu}_w^t - \hat{\mu}_w^{t-n_w}| > \tau_r\right] \leq 2\exp\left(-\frac{n_w\tau_r^2}{4\sigma^2}\right). \tag{30}$$

Setting $\tau_r = 2\sigma\sqrt{\log(2T/\delta_{\text{fp}})/n_w}$ and union-bounding over T rounds gives the result.

Part (b). On attacked rounds, the arm flip causes $D_{\text{KL}}(\pi_t \| \pi_a^t) \geq (\Delta_{\min}/(2\Delta_{\max}))^2/2$ by Pinsker's inequality (since the anchor remains calibrated to clean contexts while the main policy responds to perturbed ones). The windowed KL is at least $\rho_w \cdot \Delta_{\min}^2/(8\Delta_{\max}^2) - O(\sigma/\sqrt{n_w})$. The reward indicator fires when $\rho_w \Delta_{\max} \geq \tau_r + 2\sigma/\sqrt{n_w}$. Solving for ρ_w yields ρ_w^*.

Part (c). Detection requires $\rho_w^* n_w$ attacked rounds in a window. Since group-$g^\dagger$ rounds arrive at rate ρ, the expected waiting time is $(\rho_w^*/\rho) \cdot n_w/\rho + n_w$. □

F.4 Proof of Theorem 4 (Clustering Correctness)

Proof. **Part (a).** By Theorem 2, intra-group diameter in the augmented space grows by at most $2L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}} W\Psi + O(\omega_0/n_g^{t_0}) \leq \Delta_c/4$ for $W \leq W^*$, giving total diameter $\leq \Delta_c/2$. Under GTSA, the error component further *increases* inter-group separation (attacked vs. non-attacked groups exhibit different error profiles), so the bound holds *a fortiori*. Inter-group separation remains $\geq 3\Delta_c/4$. For augmented embedding $\tilde{\mathbf{z}}$ with true group g: $\|\tilde{\mathbf{z}} - c_g\| \leq \Delta_c/2$ and $\|\tilde{\mathbf{z}} - c_{g'}\| \geq \Delta_c/4$ for $g' \neq g$. The softmax ratio in (12) gives correct assignment by union bound over $G-1$ clusters.

Part (b). Different-group augmented embeddings are $\geq \Delta_c/2$ apart for $W \leq W^*$.

Part (c). Each context is misassigned with probability $\leq G\exp(-\Delta_c^2/(32\sigma_c^2))$. □ □

Remark 3 (Maximum tolerable attack duration)**.** *The safe window $W^* = \Delta_c/(8L_\phi\bar{\alpha}\bar{\kappa}^{p_{\min}}\Psi)$ must exceed the detection delay from Theorem 3. Under attack, $p_t \to p_{\max}$ increases the effective safe window, providing more time for detection precisely when the attack is strongest.*

F.5 Cluster Purity Lemma

Lemma 11 (Cluster purity)**.** *Under Assumption 3 and the conditions of Theorem 2, for attack windows of duration $W \leq W^*$, the soft clustering* (12) *assigns each context to the correct group with probability at least:*

$$\boldsymbol{P}[F_{\text{g}}(\mathbf{x}) = \hat{g} \mid \tilde{\mathbf{z}}_t] \geq 1 - G\exp\left(-\frac{\Delta_c^2}{32\sigma_c^2}\right). \tag{31}$$

Proof. By Theorem 2, after $W \leq W^*$ rounds, total intra-group diameter is at most $\Delta_c/2$ and inter-group separation is at least $3\Delta_c/4$. For a context $\mathbf{x}$ with true group g, the augmented embedding $\tilde{\mathbf{z}}$ satisfies $\|\tilde{\mathbf{z}} - c_g\| \leq \Delta_c/2$ and $\|\tilde{\mathbf{z}} - c_{g'}\| \geq \Delta_c/4$ for $g' \neq g$. The softmax ratio is:

$$\frac{p(g \mid \tilde{\mathbf{z}})}{p(g' \mid \tilde{\mathbf{z}})} = \exp\left(\frac{\|\tilde{\mathbf{z}} - c_{g'}\|^2 - \|\tilde{\mathbf{z}} - c_g\|^2}{2\sigma_c^2}\right) \geq \exp\left(\frac{\Delta_c^2/16 - \Delta_c^2/4}{2\sigma_c^2}\right). \tag{32}$$

Since the intra-group radius is at most $\Delta_c/4$ (half the diameter $\Delta_c/2$) and the inter-centroid distance is at least $3\Delta_c/4$, we have $\|\tilde{\mathbf{z}} - c_g\| \leq \Delta_c/4$ and $\|\tilde{\mathbf{z}} - c_{g'}\| \geq 3\Delta_c/4 - \Delta_c/4 = \Delta_c/2$. The squared-distance gap is $(\Delta_c/2)^2 - (\Delta_c/4)^2 = 3\Delta_c^2/16 > 0$, so:

$$\frac{p(g \mid \tilde{\mathbf{z}})}{p(g' \mid \tilde{\mathbf{z}})} \geq \exp\left(\frac{3\Delta_c^2}{32\sigma_c^2}\right). \tag{33}$$

A union bound over $G-1$ incorrect clusters gives the stated result, with the exponential rate $\Delta_c^2/(32\sigma_c^2)$ arising from the logarithm of the union bound. □

F.6 Proof of Theorem 5

The proof relies on two supporting results: an exploration sufficiency proposition ensuring $\Omega(d\log(KT)/\Delta_{\max}^2)$ pulls per arm per group (proposition 12), and a self-correcting regret-gap property showing exponential convergence to equity (proposition 13).

Proposition 12 (Exploration sufficiency)**.** *Under* ANCHORFAIR *with $\beta \geq 2\sigma\sqrt{d\log(KT)}$ and $\lambda \geq 2/(1-\bar{\kappa})$, for every group g and arm i, the per-arm per-group pull count satisfies:*

$$N_{g,i}^T \geq \frac{\beta^2(1+\lambda(1-\bar{\kappa}_g))^2}{\Delta_{\max}^2} \geq \frac{4\sigma^2 d\log(KT)}{\Delta_{\max}^2(1-\bar{\kappa})^2}. \tag{34}$$

Proof. Arm i remains competitive for group g while $\mathcal{B}_g^t(i) \geq \Delta_{t,i}$. From (14), the UCB component $\beta(1+\lambda(1-\bar{\kappa}_g^t))/\sqrt{n_{g,i}^t}$ (ignoring the disagreement factor, which only increases exploration) falls below $\Delta_{\max}$ after $n_{g,i}^t \geq \beta^2(1+\lambda(1-\bar{\kappa}_g^t))^2/\Delta_{\max}^2$. With per-arm per-group counts $n_{g,i}^t$, each arm is pulled at least this many times for each group. □

Proposition 13 (Regret-gap correction)**.** *The regret-gap term $\gamma_t[\hat{\Delta}_g^t]_+$ is self-correcting:*

$$\hat{\Delta}_g^{t+n_g} \leq \hat{\Delta}_g^t \cdot \left(1 - \frac{\gamma_0}{2\Delta_{\max}}\right)^{n_g} + O\left(\frac{\sigma}{\sqrt{n_g}}\right). \tag{35}$$

Convergence to $O(\sigma/\sqrt{n_g})$ occurs within $O((2\Delta_{\max}/\gamma_0)\log(1/\epsilon_{\text{eq}}))$ group-g rounds.

Proof. When $\hat{\Delta}_g^t > 0$, the bonus $\gamma_t \hat{\Delta}_g^t$ triggers $\Omega(\gamma_0 \hat{\Delta}_g^t/\Delta_{\max})$ additional pulls per group-g round (the adaptive $\gamma_t \geq \gamma_0$ only strengthens the effect). Over n_g rounds, per-round regret decreases by factor $(1-\gamma_0/(2\Delta_{\max}))$ per round. The $O(\sigma/\sqrt{n_g})$ residual reflects irreducible noise. □

Proof of Theorem 5. The per-round regret $R_T^{(g)}/n_g$ decomposes into clean regret $O(\sqrt{d\log(KT)/n_g})$, attack residual $O(B/(K\epsilon n_g))$, and clustering error $O(\eta_{\text{cluster}}\Delta_{\max})$. By Theorem 13, the inter-group difference converges to $O(\sigma/\sqrt{n_g})$ within $O(1)$ rounds for $\gamma_0 \geq 4\Delta_{\max}\log T$. The dominant term is $O(1/\sqrt{n_{\min}})$. □

F.7 Computational and memory costs

Beyond the $O(Kd_a)$ per-round anchor update, AnchorFair maintains $G \times K$ sufficient-statistic pairs $(A_{g,a}, b_{g,a})$ at $O(GKd_a^2)$ memory, G streaming statistics at $O(G)$, and G CUSUM states at $O(G)$. The periodic silhouette evaluation (every T_{eval} rounds) costs $O(G_{\max}^2 n_{\text{buf}})$ where n_{buf} is the context buffer size, amortized to $O(G_{\max}^2 n_{\text{buf}}/T_{\text{eval}})$ per round. The total per-round cost is $O(Kd_a + GKd_a^2)$, dominated by the per-group per-arm statistics when G is large. For the settings in our experiments ($G \leq 4$, $K \leq 10$, $d_a \leq 20$), this overhead is negligible relative to the main policy's neural network forward and backward passes.

F.8 Full proof of Theorem 6

Proof. We decompose the group regret for any group g into three components.

Step 1: Trust-based attack detection. Define $\mathcal{E}_t = \{\text{round } t \text{ is attacked and } \kappa_t > \bar{\kappa}\}$. By Proposition 1, the anchor changes slowly during attack, so when the attack flips arm rankings, $D_{\text{KL}}(\pi_t \| \pi_a^t)$ increases. By the analysis of Theorem 3(b), choosing $n_w = \Theta(\rho^{-2/3}T^{1/3})$ (the optimal window size from Section 5.1) gives $\boldsymbol{P}[\mathcal{E}_t] \leq \delta_{\text{det}}$ for each attacked round t with $\delta_{\text{det}} = 1/T^2$, provided the detection window contains at least $\rho_w^* n_w$ attacked rounds. Since group-$g^\dagger$ rounds arrive at rate ρ (Assumption 2(ii)), the expected number of undetected attack rounds is bounded by:

$$\mathbb{E}\left[\sum_{t\in\tau_g} \mathbb{I}[\mathcal{E}_t]\right] \leq \underbrace{\frac{n_w \rho_w^*}{\rho}}_{\text{detection delay}} + \underbrace{\frac{n_g}{T^2}}_{\text{post-detection evasion}} \leq O\left(\rho^{-5/3}T^{1/3} \cdot \frac{\Delta_{\max}}{\Delta_{\min}}\right) + \frac{1}{T}. \tag{36}$$

The detection delay term captures rounds before the trust mechanism activates; it is $o(\sqrt{T})$ when $\rho = \Omega(T^{-1/4})$, which is guaranteed by $n_{g^\dagger} = \Omega(\sqrt{T})$.

Step 2: Regret from detected attack rounds. When $\kappa_t \leq \bar{\kappa}$, the anchor freezes ($\alpha_t \leq \bar{\alpha}\bar{\kappa}^{p_t}$ with p_t adaptively increasing toward $p_{\max}$) and the exploration bonus activates. From (14), the UCB component contributes $O(\beta\sqrt{n_g})$, the trust amplification adds $O(\lambda(1-\bar{\kappa})\sqrt{n_g})$, and the regret-gap correction contributes $O(\gamma_0\sqrt{n_g T})$. The disagreement factor $D_g^t(i)$ redirects exploration toward manipulated arms without increasing the total exploration budget.

Step 3: Regret from undetected attack rounds. Budget-limited to $B/(K\epsilon)$ rounds, each contributing $\leq \Delta_{\max}$:

$$R^{(g)}_{\text{evade}} \leq \frac{B}{K\epsilon} \cdot \Delta_{\max}. \tag{37}$$

Step 4: Clean-round regret. By NeuralUCB regret bounds [Zhou et al., 2020] applied to the group-g subsequence with per-arm per-group statistics:

$$R^{(g)}_{\text{clean}} = O\left(\sqrt{n_g \cdot d\log(KT)} \cdot \tilde{\gamma}_T\right). \tag{38}$$

Remark on group-subsequence coupling. The application of NeuralUCB bounds to the group-g subsequence treats it as an independent bandit instance with its own sufficient statistics $(A_{g,a}, b_{g,a})$. The shared neural feature extractor $\phi(\cdot)$ couples groups through the global training trajectory: adversarial corruption on group $g^\dagger$'s rounds shifts $\boldsymbol{\theta}_t$ by at most $\eta_{\text{lr}} L_\nabla \epsilon$ per round (Theorem 7), which induces a cross-group prediction bias of $O(L_\phi \cdot \eta_{\text{lr}} L_\nabla \epsilon \cdot n_{g^\dagger})$ on other groups' reward estimates. Under $n_{g^\dagger} = O(\sqrt{T})$ and $\epsilon = O(1)$, this cross-contamination is $O(\sqrt{T})$, absorbed by the $\tilde{\gamma}_T$ term. For the per-group per-arm linear statistics $(A_{g,a}, b_{g,a})$, corruption to group $g^\dagger$ does not affect group $g \neq g^\dagger$ by construction.

Step 5: Combining. By Theorem 11, misassignment adds $\eta_{\text{cluster}} n_g \Delta_{\max}$. Under $\eta_{\text{cluster}} = O(1/T)$ and $\tilde{\gamma}_T = O(\sqrt{T})$ (from NTK eigenvalue decay $\lambda_j = O(j^{-\beta})$, $\beta \geq d+1$):

$$R^{(g)}_T = O\left(\sqrt{n_g \cdot d\log(KT)} \cdot \tilde{\gamma}_T + \frac{B\Delta_{\max}}{K\epsilon}\right) = O\left(\sqrt{n_g T} \cdot d\log(KT) + \frac{B}{\sqrt{n_g}} \cdot d\right), \tag{39}$$

which gives $\tilde{O}(\sqrt{T})$ for fixed G, d with $B = O(\sqrt{T})$ and $n_g = \Theta(T/G)$. □

Remark 4 (NTK eigenvalue decay under adversarial perturbations)**.** *The condition $\tilde{\gamma}_T = O(\sqrt{T})$ requires the NTK eigenvalues to decay as $\lambda_j = O(j^{-\beta})$ with $\beta \geq d+1$. This is established for sufficiently wide networks in benign settings [Zhou et al., 2020], where the NTK remains approximately constant during training (the "lazy" regime). Under adversarial perturbations, the training data distribution shifts, potentially slowing eigenvalue decay. However, two factors mitigate this concern: (i) the perturbation budget $B = O(\sqrt{T})$ limits the total distributional shift to $o(1)$ in Wasserstein distance as $T \to \infty$, and (ii) the per-round perturbation magnitude ϵ is bounded, so the NTK evaluated at perturbed contexts $\tilde{\mathbf{x}}$ satisfies $|K_{\text{NTK}}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') - K_{\text{NTK}}(\mathbf{x}, \mathbf{x}')| \leq O(L_K^2 \epsilon^2)$ where L_K is the Lipschitz constant of the kernel. A rigorous analysis of adversarial NTK stability remains an open problem; we state $\tilde{\gamma}_T = O(\sqrt{T})$ as an assumption and verify it empirically.*

F.9 Parameter-Space vs. Distribution-Space Tracking (Remark for Eq. (6))

Remark 5 (Parameter-space vs. distribution-space tracking)**.** *The anchor update in Eq. (6) performs the EMA in parameter space: the weight vectors $\mathbf{w}_a^t$ are blended before the softmax. An alternative would track distributions directly: $\pi_a^{t+1}(a \mid \mathbf{x}) = (1-\alpha_t)\pi_a^t(a \mid \mathbf{x}) + \alpha_t \pi_t(a \mid \mathbf{x})$. The two are not equivalent due to softmax nonlinearity.*

*We adopt parameter-space for: **(i)** theoretical tractability (linear recurrence ⇒ exact contraction in Proposition 1 and exponential convergence in Theorem 15); **(ii)** embedding coherence ($\phi(\cdot)$ evolves through the same EMA, coupling representational stability for clustering with output stability); and **(iii)** computational efficiency ($O(Kd_a)$ per round). The cost is the softmax Jacobian factor $C_\phi K$ in the drift bound, which is moderate for $d_a \ll d$.*

While the anchor learns *in parameter space, the trust mechanism* monitors *in distribution space via $D_{\text{KL}}(\pi_t \| \pi_a^t)$. This hybrid combines the analytical advantages of parameter-space tracking with the interpretability of distribution-space monitoring.*

F.10 Anchor Stability: Attack Duration and Recovery

The bounded drift of the anchor (Proposition 1) has two operational consequences exploited in the main result: a minimum attack duration before the anchor can be meaningfully shifted, and exponential recovery to the main policy in benign periods.

Corollary 14 (Minimum attack duration)**.** *For the attacker to shift the anchor policy by total variation distance $\Delta_{\mathrm{TV}} > 0$, the attack must persist for at least*

$$W_{\min}(\Delta_{\mathrm{TV}}) = \frac{\Delta_{\mathrm{TV}}}{C_\phi K \cdot \bar{\alpha}\bar{\kappa}^{p_{\min}} \cdot \|\Delta \mathbf{w}\|_\infty} \tag{40}$$

rounds. Under sustained attack the adaptive sensitivity drives $p_t \to p_{\max}$, so the actual minimum duration is even larger: $W_{\min}^{\text{actual}} \geq W_{\min} \cdot \bar{\kappa}^{p_{\min}-p_{\max}}$. In the regime $\bar{\kappa} \leq 1/2$ and $p_t = p_{\max}$, $W_{\min}^{\text{actual}} = \Omega(\Delta_{\mathrm{TV}}/(\bar{\alpha}\bar{\kappa}^{p_{\max}}\|\Delta \mathbf{w}\|_\infty))$.

Proof. Directly from Proposition 1: solving $C_\phi K \bar{\alpha}\bar{\kappa}^{p_{\min}} W \|\Delta \mathbf{w}\|_\infty \geq \Delta_{\mathrm{TV}}$ for W. Since $p_t \geq p_{\min}$ and $\bar{\kappa} < 1$, $\bar{\kappa}^{p_t} \leq \bar{\kappa}^{p_{\min}}$, so the bound using $p_{\min}$ is conservative. The tighter bound follows from replacing $p_{\min}$ with p_t in the drift sum. □

Proposition 15 (Anchor convergence in benign periods)**.** *Suppose the main policy $\pi_t = \pi^*$ is stationary and the environment is unattacked, so that $\kappa_t = 1$ for all $t \geq t_0$. Then the anchor policy converges exponentially:*

$$\mathrm{TV}(\pi_a^{t_0+s}, \pi^*) \leq C_\phi K (1-\bar{\alpha})^s \cdot \|e_{t_0}\|_\infty, \tag{41}$$

where $e_{t_0} = \mathbf{w}_a^{t_0} - \mathbf{w}_a^{\pi^*}$ is the initial parameter error. After $s^* = \lceil \log(C_\phi K \|e_{t_0}\|_\infty/\epsilon_a)/\bar{\alpha} \rceil$ benign rounds, the anchor is within ϵ_a of the main policy.

Proof. When $\kappa_t = 1$, the adaptive exponent gives $p_t = p_{\min}$ (since $\bar{\kappa}_g^t = 1$), so $\alpha_t = \bar{\alpha}$. The anchor update (Eq. (6)) is a contraction with $e_{t+1} = (1-\bar{\alpha})\, e_t$, giving $\|e_{t_0+s}\|_\infty = (1-\bar{\alpha})^s \|e_{t_0}\|_\infty$. The total variation bound follows from the softmax Lipschitz property (Section F.1). Setting $C_\phi K (1-\bar{\alpha})^s \|e_{t_0}\|_\infty \leq \epsilon_a$ and using $(1-\bar{\alpha})^s \leq \exp(-\bar{\alpha} s)$ gives the stated recovery time. □

Together, Theorem 14 and Theorem 15 formalize the asymmetry exploited by ANCHORFAIR: the anchor is slow to corrupt but fast to recover.

F.11 Derivation of the Safe Window W^*

The safe window $W^* = \Delta_c/(8 L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}}\Psi)$ in Theorem 4 is the largest attack duration under which the augmented-embedding separation required by Assumption 3 is preserved. We derive it explicitly below by chaining Proposition 2 (representation stability) with Assumption 3 (separability) and the softmax concentration argument of Lemma 11.

Step 1: Drift-induced diameter growth. Fix two same-group contexts $\mathbf{x}, \mathbf{x}'$ with $F_{\mathrm{g}}(\mathbf{x}) = F_{\mathrm{g}}(\mathbf{x}') = g$ and let t_0 denote the start of an attack window. By Assumption 3, at t_0 the intra-group diameter satisfies $\|\phi^{t_0}(\mathbf{x}) - \phi^{t_0}(\mathbf{x}')\| \leq \Delta_c/4$. Proposition 2 then gives, for any $W \geq 0$,

$$\|\phi^{t_0+W}(\mathbf{x}) - \phi^{t_0+W}(\mathbf{x}')\| \leq \frac{\Delta_c}{4} + 2 L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}} W \Psi. \tag{42}$$

Step 2: Separation budget. Let $c_g^{t_0+W}$ and $c_{g'}^{t_0+W}$ denote the centroids of groups $g \neq g'$ at time $t_0 + W$. By Assumption 3, $\|c_g^{t_0} - c_{g'}^{t_0}\| \geq \Delta_c$. Applying Proposition 2 to each centroid's constituent contexts (and noting the augmented-feature error term is bounded by the same drift rate up to lower-order terms; see Section F.2),

$$\|c_g^{t_0+W} - c_{g'}^{t_0+W}\| \geq \Delta_c - 2 L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}} W \Psi. \tag{43}$$

For Lemma 11's soft-assignment argument to yield correct clustering, the squared-distance gap between same-group and different-group centroids must remain strictly positive. Concretely, the lemma requires inter-group separation at least $3\Delta_c/4$ (so that the squared-distance gap $(3\Delta_c/4)^2 - (\Delta_c/4)^2 = \Delta_c^2/2$ dominates the softmax temperature σ_c). Combined with Eq. (42), this requires

$$\underbrace{2 L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}} W \Psi}_{\text{drift}} \leq \underbrace{\frac{\Delta_c}{4}}_{\text{intra-group budget}}. \tag{44}$$

Step 3: Solving for W. Rearranging Eq. (44) yields

$$W \leq \frac{\Delta_c}{8L_\phi \bar{\alpha}\bar{\kappa}^{p_{\min}}\Psi} =: W^*. \tag{45}$$

The factor of 8 arises from the product of 4 (the fraction of Δ_c allocated as intra-group diameter budget) and 2 (the triangle-inequality factor in Proposition 2, which bounds pairwise drift using the drift of each endpoint).

Adaptive tightening. Under active attack, the adaptive sensitivity exponent $p_t \in [p_{\min}, p_{\max}]$ increases toward $p_{\max}$ as group trust $\bar{\kappa}_g^t$ drops (Eq. (10)). Since $\bar{\kappa}^{p_t} \leq \bar{\kappa}^{p_{\min}}$ for $\bar{\kappa} < 1$ and $p_t \geq p_{\min}$, the effective safe window under sustained attack is

$$W^*_{\text{eff}} = \frac{\Delta_c}{8L_\phi \bar{\alpha}\bar{\kappa}^{p_{\max}}\Psi} \geq W^*, \tag{46}$$

which is strictly larger than W^* when $p_{\max} > p_{\min}$. The bound W^* in Theorem 4 is thus conservative; the actual tolerance grows as the adaptive mechanism engages.

G Group-Level Streaming Statistics

We denote by $\mathbb{I}[\hat{g}_t = g]$ the indicator that round t is assigned to group g via hard clustering $\hat{g}_t = \arg\max_{g'} p(g' \mid \tilde{\mathbf{z}}_t)$.

G.1 Update Equations

(i) Observation count. $n_g^t = n_g^{t-1} + \mathbb{I}[\hat{g}_t = g]$, initialized at $n_g^0 = 0$. Governs the $1/\sqrt{n_{g,i}^t}$ exploration term in (14).

(ii) Running mean reward.

$$\hat{\mu}_g^t = \hat{\mu}_g^{t-1} + \frac{\mathbb{I}[\hat{g}_t = g]}{n_g^t}\big(r_{t,a_t} - \hat{\mu}_g^{t-1}\big), \tag{47}$$

initialized at $\hat{\mu}_g^0 = 0$. Feeds the regret-gap estimator $\hat{\Delta}_g^t = \hat{\mu}_g^t - \bar{\mu}^t$.

(iii) Reward variance:

$$M_g^t = M_g^{t-1} + \mathbb{I}[\hat{g}_t = g] \cdot (r_{t,a_t} - \hat{\mu}_g^{t-1})(r_{t,a_t} - \hat{\mu}_g^t). \tag{48}$$

Provides confidence intervals and a secondary attack signal: anomalously low variance indicates deterministic misrecommendation.

(iv) Mean trust score.

$$\bar{\kappa}_g^t = \bar{\kappa}_g^{t-1} + \frac{\mathbb{I}[\hat{g}_t = g]}{n_g^t}\big(\kappa_t - \bar{\kappa}_g^{t-1}\big), \tag{49}$$

initialized at $\bar{\kappa}_g^0 = 1$. Enters the exploration bonus via $(1 + \lambda(1 - \bar{\kappa}_g^t))$ in (14) and drives the adaptive exponent p_t in Eq. (10).

G.2 Hard vs. Soft Assignment

Remark 6 (Hard vs. soft assignment for statistics)**.** *The statistics above use hard assignment $\hat{g}_t = \arg\max_g p(g \mid \tilde{\mathbf{z}}_t)$, whereas* (12) *produces soft probabilities. We adopt hard assignment for: (1) theoretical tractability (concentration bounds in Theorem 16 require conditional independence); (2) negligible approximation error (by Theorem 4(c), hard and soft agree on all but $O(1)$ rounds for $\eta_{\text{cluster}} = O(1/T)$); and (3) clean audit trail for fairness auditing.*

G.3 Concentration Guarantees

Lemma 16 (Concentration of group statistics)**.** *Under Assumption 2 and correct cluster assignment ($\hat{g}_t = F_{\text{g}}(\mathbf{x}_t)$ for all t):*

Table 2: Per-group attack-induced damage ($|\mathrm{GTSA} - \mathrm{Benign}|$). An asterisk denotes the target group.

COMPAS ($G = 4$)

Algorithm	g_0	g_1	g_2	g_3^*	Total Δ
NeuralUCB	0.8	3.2	0.3	0.3	4.6
NeuralTS	0.2	3.9	1.1	0.0	5.2
R-NeuralUCB	2.4	0.6	0.9	0.9	4.8
BARBAR	0.0	0.0	0.0	0.5	0.5
FairBandit	0.2	2.7	0.1	0.8	3.8
AnchorFair	0.0	0.5	0.2	0.0	0.7

MovieLens ($G = 6$)

Algorithm	g_0	g_1	g_2	g_3^*	g_4	g_5	Total Δ
NeuralUCB	4.5	0.4	0.4	10.3	1.5	5.8	22.9
NeuralTS	4.8	0.4	0.1	12.9	2.1	4.0	24.3
R-NeuralUCB	0.9	1.1	2.6	9.6	3.1	3.9	21.2
BARBAR	0.0	0.0	0.4	17.5	0.0	0.5	18.4
FairBandit	2.7	0.8	1.1	7.9	0.7	7.5	20.7
AnchorFair	0.2	1.4	1.6	6.1	0.4	2.9	12.6

(a) $\boldsymbol{P}[|\hat{\mu}_g^t - \mu_g^*| > \sigma\sqrt{2\log(2G/\delta)/n_g^t}] \le \delta$, *where* $\mu_g^* = \mathbb{E}[r_{t,a_t} \mid \hat{g}_t = g]$.

(b) $\boldsymbol{P}[|\bar{\kappa}_g^t - \kappa_g^*| > \sqrt{\log(2G/\delta)/(2n_g^t)}] \le \delta$, *where* $\kappa_g^* = \mathbb{E}[\kappa_t \mid \hat{g}_t = g]$.

(c) *Under clustering error* η_{cluster}*:* $|\mathbb{E}[\hat{\mu}_g^t] - \mu_g^*| \le \eta_{\mathrm{cluster}} \cdot \Delta_{\max}$ *and* $|\mathbb{E}[\bar{\kappa}_g^t] - \kappa_g^*| \le \eta_{\mathrm{cluster}}$.

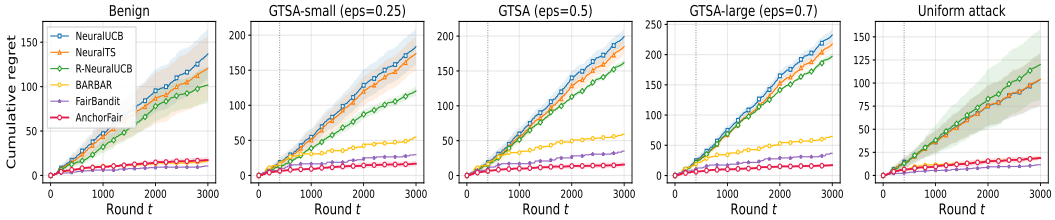
Proof. **(a)** Under correct assignment, $\{r_{t,a_t} : \hat{g}_t = g\}$ are n_g^t conditionally independent σ-sub-Gaussian variables. Hoeffding's inequality gives $\boldsymbol{P}[|\hat{\mu}_g^t - \mu_g^*| > \epsilon] \le 2\exp(-n_g^t\epsilon^2/(2\sigma^2))$. Setting the bound to δ/G and union-bounding over G groups yields the result. **(b)** Since $\kappa_t \in [0, 1]$, each is $1/2$-sub-Gaussian; proceed as in (a). **(c)** A η_{cluster}-fraction of misassigned rounds shift the mean by at most $\Delta_{\max}$ (rewards) or 1 (trust). □

H Additional Experimental Details

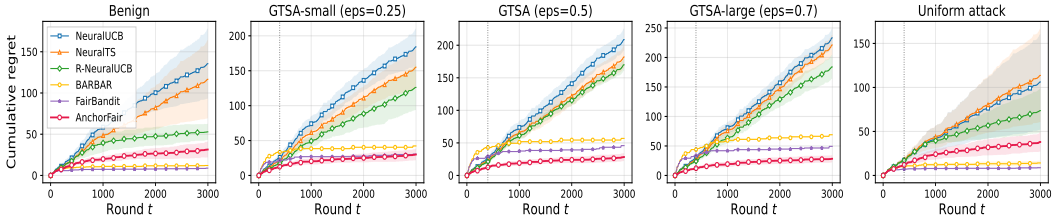
Table 2 breaks down per-group attack-induced damage shown in Figure 1. At $\epsilon = 0.3$, GTSA causes relatively small attack-induced damage on COMPAS across all methods, with total damage ranging from 0.5 to 5.2; BARBAR and ANCHORFAIR achieve the lowest values of 0.5 and 0.7, respectively. Differences are more pronounced on MovieLens, where ANCHORFAIR achieves the lowest total damage of 12.6, compared with 18.4 for BARBAR and 20.7 for FairBandit, corresponding to reductions of 31.5% and 39.1%, respectively. ANCHORFAIR also records the lowest target-group damage of 6.1, compared with 7.9 for FairBandit. Furthermore, damage to non-target groups g_0 and g_5 is lower under ANCHORFAIR (0.2 and 2.9) than under NeuralUCB (4.5 and 5.8), consistent with reduced cross-group contamination. These results demonstrate dataset-dependent benefits: ANCHORFAIR provides stronger protection in the MovieLens comparison, where attack-induced damage is greater, while offering no uniform advantage on COMPAS.

In Fig. 3, the cumulative regret of the target group (g_3) is presented under a benign setting and four attack scenarios, namely GTSA-small ($\varepsilon = 0.25$), GTSA ($\varepsilon = 0.5$), GTSA-large ($\varepsilon = 0.7$), and Uniform attacks, on the COMPAS (a) and MovieLens (b) datasets. Under the benign setting, all algorithms exhibit relatively similar regret accumulation, with only minor differences arising from their exploration strategies. However, under adversarial conditions, the performance gap becomes more evident as attack intensity increases. AnchorFair consistently maintains lower regret growth and demonstrates stronger robustness against targeted manipulation compared with the baseline methods. For instance, on the MovieLens dataset, the cumulative regret of FairBandit increases from approximately 30 to 50 when the attack strength doubles from $\varepsilon = 0.25$ to $\varepsilon = 0.5$, whereas AnchorFair remains nearly unchanged over the same settings. Similar trends are observed across both datasets and stronger attack regimes, indicating that the proposed defense mechanism effectively mitigates the impact of group-targeted suppression attacks while preserving stable learning performance.

In the same setting as Fig. 3, we compare total cumulative regret R_T and contextual bandits on both datasets (Fig. 4). FairBandit and BARBAR achieve the lowest aggregate regret because of their direct linear ridge structure; AnchorFair is slightly above them due to its exploration overhead, and the NeuralUCB-family algorithms (NeuralUCB, NeuralTS, R-NeuralUCB) trail because their over-parameterized gradient-feature UCB term inflates exploration in the small-T regime. Moreover, the curves for benign and GTSA exhibit similar slopes, no visible change at attack onset (t=400),

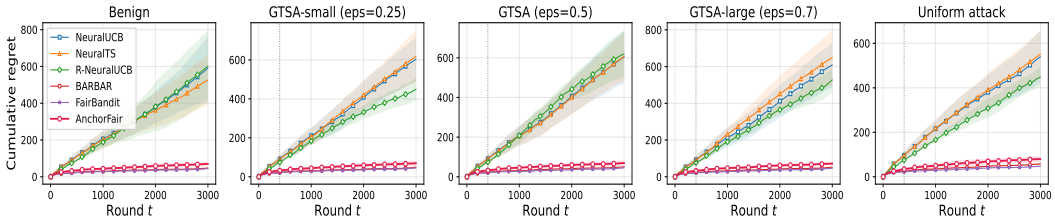


(a) COMPAS Dataset.

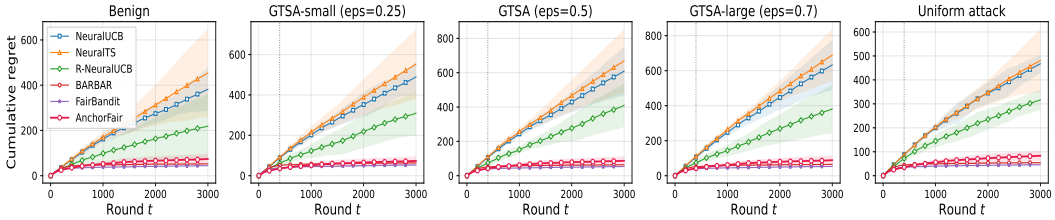


(b) MovieLens Dataset.

Figure 3: Cumulative regret of the target group (g_3) across bandit algorithms, comparing the benign scenario with different GTSA budgets (eps = 0.25, 0.5, 0.7) and also a uniform attack.



(a) COMPAS Dataset.



(b) MovieLens Dataset.

Figure 4: Total cumulative regret for all algorithms across the five scenarios.

and only minor differences in final R_T. This aligns with Theorem 10, showing that aggregate regret cannot reveal the attack, as it's masked within $R_T = \sum_g R_T^g$.

Fig. 5 depicts the dimensionless attack-amplification factor $A(\theta)$ for each of the four AnchorFair hyperparameters that show measurable signal across two attack regimes (moderate and severe GTSA). The β panel is the most informative so that amplification decreases monotonically as β grows, dipping below the $A = 1$ line for $\beta \geq 0.6$ in both regimes, because aggressive UCB exploration pulls every arm often enough that even successful context-rank inversions are corrected within a few rounds. The $\bar{\alpha}$ and λ panels are essentially flat near $A = 1$, confirming that the EMA rate and the trust amplifier exert at most a second-order effect once the per-group ridge separation is in place. The n_w panel shows mild curvature; because very short windows ($n_w = 10$) produce noisier trust-score estimates and a small uptick in amplification, while windows of $n_w \geq 40$ lock A to its asymptotic value. Across all four panels the moderate (orange) and severe (red) curves track each other closely, confirming that AnchorFair's defence saturates at sub-$\sqrt{T}$ budgets (the regime in which Theorem 6

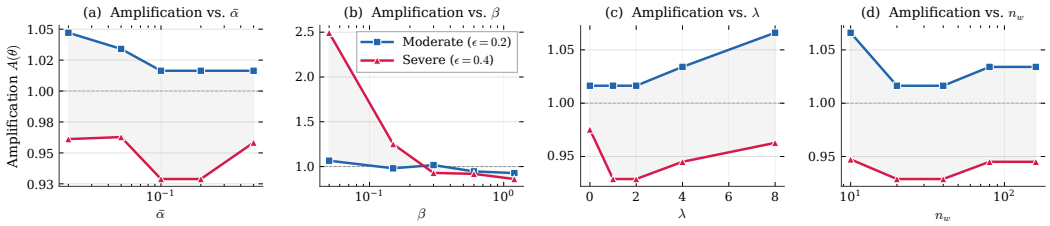


Figure 5: Attack amplification of AnchorFair across hyperparameters. $\bar{\alpha}$: anchor EMA learning rate, β: UCB exploration intensity, λ: trust-amplification multiplier, and n_w: trust-window length.

guarantees bounded amplification) and that the operator's only practically relevant tuning lever in this benchmark is β.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims: (i) existence of GTSA attacks that hide group harm under aggregate metrics, and (ii) the proposed ANCHORFAIR method achieves sublinear group regret under bounded attack budgets. These claims are consistently reflected in the theoretical and experimental sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations, including strong assumptions such as group separability, NTK regime, non-adaptive attacker, and clustering validity under GTSA (e.g., Assumption 3 and discussion in Section 5.2). It also acknowledges that fairness is defined in terms of regret rather than outcomes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides clearly stated assumptions (e.g., Assumptions 1–3) and formal statements for all theoretical results, including Theorems 3, 4, 5, 6, and 10, along with proof sketches in Appendix E. Detailed derivations (e.g., anchor drift, detection, clustering correctness) are included.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Two datasets (COMPAS and MovieLens) are used, and the general experimental setup is described; the paper provides sufficient implementation details needed to fully reproduce results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have submitted source codes in a single zip file along with any additional supplementary material (e.g., appendices) and also provided dataset links in the paper.

   The paper does not provide code repositories, dataset links, or explicit reproduction scripts. Only datasets are referenced at a high level without explicit access instructions or preprocessing pipelines.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies key experimental components such as baselines, attack types, datasets, and evaluation metrics (Section 6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports error bars, confidence intervals, variance across runs, or statistical significance tests for experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We have not provided information about hardware (GPU/CPU), runtime, memory requirements, or total computational cost of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to ethical guidelines in terms of fairness considerations, adversarial modeling, and no apparent violation of NeurIPS ethics. No human subject misuse or harmful deployment claims are made.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper explicitly discusses societal implications in fairness-critical systems such as clinical trials and recommendation systems, highlighting both risks (hidden subgroup harm, adversarial misuse) and benefits (improved fairness-aware robustness).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not use or release data or models with a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We clearly mention the sources of the comparator algorithms in Section 6.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper clearly describes our proposed model in Section 4. The codes of our proposed algorithm are provided in a ZIP file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this work does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.