Fairness Beyond Metrics: A Residual Distribution Framework for Text Classification

Anonymous ACL submission

Abstract

Standard fairness metrics indicate that bias exists in text classifiers but not *where* it manifests or *what* causes it. We introduce Residual Distribution Fairness (RDF), a framework grounded in a theoretical insight: **demographic parity and equalized odds are functions of the residual distribution**, that is, the sorted difference between the estimated and actual predictions. We prove this connection formally and show that RDF analyzes this distribution directly, providing a rich diagnostic methodology that complements classical metrics DP and EO.

The sorted residual plot visualizes this distribution, revealing where bias manifests across the prediction space. *Knee points* mark behavioral transitions where examples are particularly informative for diagnostic probing through counterfactual analysis. We identify a **calibration-diagnosticity relationship**: knee regions concentrate prediction errors when calibration is reasonable (Expected Calibration Error < 0.15), providing practitioners clear guidance on when knee-based analysis is most reliable.

Experiments on four NLP datasets validate this framework across calibration regimes. For well-calibrated models, knee regions concentrate 2–22$\times$ higher prediction errors than non-knee regions ($p < 0.001$), enabling targeted diagnostic analysis. RDF-guided augmentation yields greater fairness improvements than random selection, though with variance that limits statistical confidence. Null controls with random group assignments confirm the effect is genuine; poorly-calibrated datasets show no RDF advantage, as the theory predicts.

1 Introduction

NLP models increasingly influence high-stakes decisions, from content moderation to hiring recommendations. Fairness auditing has become essential, yet practitioners face a fundamental challenge: the Chouldechova-Kleinberg impossibility theorem shows that demographic parity (DP), equalized odds (EO), and calibration cannot be satisfied simultaneously when base rates differ across groups (Chouldechova, 2017; Kleinberg et al., 2017). This impossibility means that fairness metrics inherently conflict, and optimizing for one criterion may worsen another. Standard metrics tell practitioners *that* trade-offs exist but not *where* they manifest, *which* predictions are affected, or *how* to navigate them.

We propose Residual Distribution Fairness (RDF), a framework that addresses this challenge by analyzing the complete residual distribution that underlies classical fairness metrics. The core theoretical insight is that for probabilistic classifiers with soft predictions, DP and EO are functions of the distribution $D = \hat{P}^+ - Y$ (predicted probability minus label), evaluated at specific thresholds. Theorem 6.1 formalizes this connection under standard assumptions: binary classification with known group membership and threshold-based decisions (Section A). RDF analyzes this distribution directly, providing diagnostic capabilities that complement classical metrics DP and EO. This reframes fairness auditing: **rather than only computing conflicting summary statistics, practitioners can visualize the distributional object from which these statistics derive**.

The sorted residual plot visualizes this distribution for each demographic group. A practitioner can directly see that "Group A experiences 15% more severe underestimation in the lowest-confidence predictions" rather than interpreting an abstract number like "EO gap = 0.12." The visualization reveals central tendency (median errors), tail behavior (extreme errors), and regime structure (behavioral transitions at "knee points"). Knee points mark where the model transitions between error regimes; examples at these transitions are highly informative about group-specific decision rules.

The diagnostic power of knee regions depends on model calibration. Knee regions concentrate errors **most strongly when calibration is reasonable** (Expected Calibration Error, ECE < 0.15; see Section 7 for formal definition). For poorly-calibrated models, residuals are elevated throughout the prediction space, reducing knee-region concentration. This calibration-diagnosticity relationship guides practitioners on when knee-based analysis is most informative and when to calibrate before analyzing.

Our contributions are: (1) A **theoretical connection** proving that DP and EO are functions of the residual distribution $D = \hat{P}^+ - Y$, establishing a formal foundation for residual-based fairness analysis. (2) A **diagnostic methodology** comprising the sorted residual plot that visualizes fairness properties and knee-based probing that identifies where bias manifests. (3) **Empirical validation** of a calibration-diagnosticity relationship: knee regions concentrate errors 2–22$\times$ more strongly when calibration is reasonable (ECE < 0.15), with $p < 0.001$. (4) **Intervention experiments** showing RDF-guided augmentation achieves larger mean fairness improvements than random selection for well-calibrated models, with null controls confirming the effect is genuine.

2 Related Work

Group fairness metrics (demographic parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016), and calibration (Kleinberg et al., 2017)) provide summary measurements of bias. Impossibility results show these criteria inherently conflict when base rates differ (Chouldechova, 2017; Kleinberg et al., 2017). Residual Distribution Fairness (RDF) provides complementary diagnostic information by visualizing the full residual distribution from which these metrics derive.

Explainability methods address different questions than RDF. Feature attribution (Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017)) explains *which features* drive individual predictions. Counterfactual methods (Wu et al., 2021) and behavioral testing (Ribeiro et al., 2020) analyze model sensitivity. Fairness visualization tools like FairVis (Cabrera et al., 2019) display standard metrics. RDF differs by identifying *which examples* to probe based on their position in residual space, providing a principled selection strategy that makes attribution-based analysis efficient. See Section K for extended discussion.

3 Problem Setup

We consider probabilistic NLP classifiers $f : \mathcal{X} \to [0, 1]$ predicting $\hat{P}^+ = f(x)$ for input text x, with ground truth $Y \in \{0, 1\}$ and protected attribute $A \in \{0, 1\}$ defining demographic groups. The **residual** $D := \hat{P}^+ - Y \in [-1, 1]$ captures prediction error with direction: $D > 0$ indicates overestimation (e.g., flagging non-toxic content as toxic); $D < 0$ indicates underestimation; $D \approx 0$ indicates well-calibrated prediction.

Why Residuals? Traditional fairness metrics compare aggregate statistics (positive rates, error rates) across groups. Residuals capture the *complete prediction behavior* at the individual level. For a truly toxic comment ($Y = 1$) that the model assigns $\hat{P}^+ = 0.3$, the residual $D = -0.7$ records severe underestimation. The distribution of such residuals across a group reveals systematic patterns invisible to aggregate metrics.

Base Rate Adjustment. RDF metrics measure whether prediction *errors* differ across groups, naturally adjusting for base rate differences $\Delta_{\text{base}} := \pi_0 - \pi_1$ where $\pi_a := \mathbb{P}(Y = 1 \mid A = a)$. Groups with higher base rates (more positive examples) will have more $Y = 1$ samples, shifting the residual distribution leftward. RDF separates this legitimate difference from discriminatory treatment by comparing residual *distributions* rather than raw predictions.

4 The Sorted Residual Plot

A dataset D contains a set of points D_i with predicted probabilities $\hat{P}_i^+$, true labels Y_i. For each group $a \in \{0, 1\}$: (1) compute residuals $D_i = \hat{P}_i^+ - Y_i$; (2) sort in ascending order; (3) plot against percentile rank. The resulting curve is the *quantile function* $Q_a(p) := F^{-1}_{D|A=a}(p)$ which is the inverse of the Cumulative Distribution Function F for group a. An example plot is shown in Figure 1.

Reading the Plot. The horizontal axis represents the percentile rank within each group (0% to 100%); the vertical axis shows residual values (-1 to $+1$). A point at (p, r) means "the p-th percentile of residuals for this group has value r." The curve's shape reveals three diagnostic features: (1)

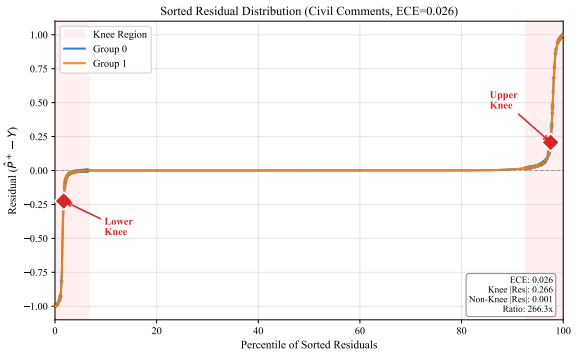


Figure 1: The sorted residual plot provides a visual explanation of fairness shown here for the Civil Comments dataset. Group 0 (blue) and Group 1 (orange) curves show sorted residuals (predicted probability minus label). Knee points (red diamonds) mark regime transitions where model behavior changes. The shaded knee region concentrates diagnostic examples. For this well-calibrated model (ECE=0.026), the overlapping curves indicate similar treatment of both groups. Group 0 is difficult to see in this visualization.

Central tendency: where curves cross the 50th percentile shows typical error per group. For example, if Group 0's median is near zero but Group 1's median is -0.2, the model systematically underestimates for Group 1. (2) **Tail behavior**: vertical spread at extremes shows error severity. Steep tails indicate some examples receive very wrong predictions. (3) **Regime transitions**: "knee points" where the S-shaped curve bends mark transitions between error regimes, identifying examples at behavioral boundaries.

Comparing Groups. Comparing knee positions across groups reveals where behavior diverges. Complete curve overlap indicates identical treatment (see $\mathcal{F}_{\text{dist}} = 0$). Separation at the tails but overlap in the center indicates the model treats most examples similarly but has group-specific failure modes for edge cases. Parallel but offset curves indicate systematic bias meaning one group consistently receives higher or lower predictions.

Comparison to Existing Visualizations. Reliability diagrams show calibration per confidence bin but aggregate away distributional structure. Group-wise residual histograms show marginal distributions but obscure the joint percentile-residual structure that reveals where groups diverge. Our sorted residual plot shows the complete distributional structure in a single visualization: central tendency (median offsets), tail behavior (extreme errors), and regime transitions (knee points). This enables practitioners to see *where* in prediction space groups differ, not just *that* they differ.

Table 1: Summary of RDF metrics derived from the sorted residual plot. Each metric captures a distinct aspect of group fairness: central tendency measures typical error differences, knee metrics measure behavioral transition alignment, and distributional distance measures overall curve separation. Colors match the metric definitions in the text.

Metric	Range	Interpretation
$\mathcal{F}_{\text{pattern}}$	$[0, 1]$	Typical error parity (median)
$\mathcal{F}_{\text{h}}$	$[0, \infty)$	Regime transition alignment
$\mathcal{F}_{\text{v}}$	$[0, \infty)$	Error severity at transitions
$\mathcal{F}_{\text{dist}}$	$[0, 2]$	Total distributional distance

5 RDF Metrics

We define four metrics quantifying features of the sorted residual plot (Table 1). Let m_a denote the median residual for group a, and let $(q_{a,s}, r_{a,s})$ denote knee coordinates (percentile, residual) for group a at knee $s \in \{\ell, r\}$.

Central Tendency ($\mathcal{F}_{\text{pattern}}$). $\mathcal{F}_{\text{pattern}} := 1 - |m_1 - m_0|/2$ measures whether groups have equal typical error. $\mathcal{F}_{\text{pattern}} = 1$ indicates median parity (both groups have the same median residual). Values below 1 indicate one group systematically receives higher or lower predictions relative to their labels. For example, $\mathcal{F}_{\text{pattern}} = 0.8$ means median residuals differ by 0.4 between groups.

Knee Metrics ($\mathcal{F}_{\text{h}}$, $\mathcal{F}_{\text{v}}$). $\mathcal{F}_{\text{h}}$ measures whether regime transitions occur at the same *percentile* for both groups. $\mathcal{F}_{\text{v}}$ measures whether groups have the same error *magnitude* at structurally equivalent knee positions. Low values indicate aligned behavioral transitions; high values indicate the model changes behavior at different points for different groups. Formulas appear in Section L.

Distributional Distance ($\mathcal{F}_{\text{dist}}$). $\mathcal{F}_{\text{dist}} := \int_0^1 |Q_0(p) - Q_1(p)|dp$ equals the Wasserstein-1 distance between residual distributions (the area between curves). $\mathcal{F}_{\text{dist}} = 0$ iff distributions are identical. This provides a single summary of total distributional difference, complementing the localized metrics above. Figure 2 visualizes $\mathcal{F}_{\text{dist}}$ as the shaded area between quantile functions.

6 Connections to Classical Fairness

Classical fairness metrics derive from the residual distribution $D = \hat{P}^+ - Y$. When $Y = 0$, $D = \hat{P}^+$;

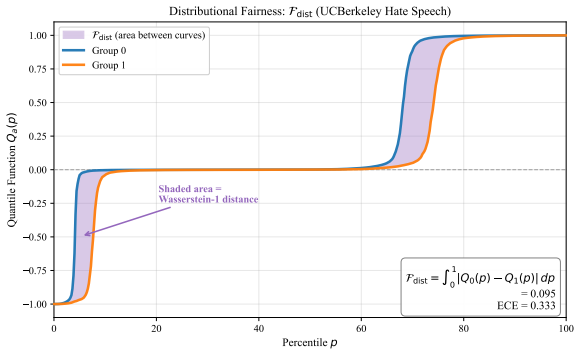


Figure 2: Visualization of $\mathcal{F}_{\text{dist}}$ as the area between group quantile functions $Q_0(p)$ and $Q_1(p)$. The shaded region shows $\int_0^1 |Q_0(p) - Q_1(p)|dp$, the Wasserstein-1 distance. For this poorly-calibrated model (ECE=0.333), $\mathcal{F}_{\text{dist}} = 0.088$ indicates measurable distributional disparity.

when $Y = 1$, $D = \hat{P}^+ - 1$, so FPR and FNR are CDF evaluations of outcome-conditional residual distributions.

Theorem 6.1 (Residual Distribution Completeness)**.** *Let $F_{D|A=a,Y=y}$ denote outcome-conditional residual CDFs. (1) Equalized odds at threshold τ holds iff $F_{D|A=0,Y=y}(t) = F_{D|A=1,Y=y}(t)$ for $t \in \{\tau, \tau - 1\}$ and $y \in \{0, 1\}$. (2) Full residual parity ($\mathcal{F}_{dist} = 0$ for outcome-conditional distributions) implies equalized odds at all thresholds. (3) Demographic parity at τ is a weighted combination of outcome-conditional residual survival functions.*

Intuition. The residual distribution captures everything relevant to threshold-based fairness decisions. When $Y = 0$, the residual $D = \hat{P}^+$ directly, so FPR $= \mathbb{P}(D \geq \tau \mid Y = 0, A = a)$. When $Y = 1$, the residual $D = \hat{P}^+ - 1$, so TPR $= \mathbb{P}(D \geq \tau - 1 \mid Y = 1, A = a)$. Equalized odds violations appear as misaligned CDFs between groups at these threshold values. For example, if Group 1's TPR is 0.8 and Group 0's TPR is 0.6 at threshold $\tau = 0.5$, the outcome-conditional residual CDFs must differ: the sorted residual plot for $Y = 1$ examples will show vertical separation between groups in the right tail.

RDF captures the generating object from which DP and EO derive. We define outcome-conditional $\mathcal{F}^{(y)}_{\text{pattern}}$ to diagnose EO violations: $\mathcal{F}^{(0)}_{\text{pattern}} = \mathcal{F}^{(1)}_{\text{pattern}} = 1$ iff equalized odds holds at all thresholds (under symmetric residuals). RDF does not escape impossibility theorems but renders trade-offs *visible*. Full proofs and connections appear in Section A.

7 RDF-Guided Probing

Knee points identify *where to look* for mechanistic understanding. The **diagnostic set** $\mathcal{D}^*$ contains examples within ϵ of detected knees. These examples sit at group-specific decision boundaries where probing reveals differential treatment.

Why Knee Regions? Knee points mark transitions between behavioral regimes: regions where the model shifts from one error pattern to another. Examples at these transitions are highly informative about group-specific decision rules because they sit at boundaries where small input changes produce large output changes. Random sampling misses these transitions; uncertainty-based sampling finds decision boundaries but ignores group structure.

Calibration Prerequisite. The diagnostic power of knee regions depends on calibration. We define regimes based on Expected Calibration Error (ECE) (Naeini et al., 2015): **Good** (ECE < 0.05) where knees strongly concentrate errors; **Moderate** ($0.05 \leq$ ECE < 0.15) where signal is weaker; **Poor** (ECE ≥ 0.15) where knees are not reliably diagnostic for error concentration (though they may still mark behavioral transitions). This calibration check is prerequisite for RDF-guided probing. For poorly-calibrated models, errors are distributed throughout the prediction space, reducing the concentration at knee regions.

Counterfactual Probes. For examples in $\mathcal{D}^*$, we apply counterfactual probes: (1) **demographic swaps** (he→she, man→woman) measuring sensitivity to gender markers; (2) **identity term removal** measuring dependence on explicit group mentions; (3) **minimal prediction flips** using counterfactual generation (Wu et al., 2021) to find smallest edits that change predictions. Aggregating sensitivity across $\mathcal{D}^*$ reveals bias mechanisms (e.g., "knee regions show 79× higher sensitivity to demographic swaps"; see Section O). See Section M for full methodology and comparison to alternative selection strategies.

7.1 Knee Point Detection

We use the Kneedle algorithm (Satopaa et al., 2011) to detect transition points in sorted residual curves. The procedure is: (1) sort residuals in ascending

order: $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$; (2) create the empirical quantile function: $Q(p) = d_{(\lceil np \rceil)}$; (3) apply LOWESS smoothing with span $= 0.1$ to reduce noise; (4) detect knee points using Kneedle with sensitivity $S = 1.0$; (5) mark knee regions as $\epsilon = 5\%$ of samples around each knee point.

The key hyperparameter is ϵ, the knee region width. We find $\epsilon = 0.05$ balances including enough transitional examples while excluding non-informative samples. Smaller values ($\epsilon = 0.02$) are unstable; larger values ($\epsilon = 0.10$) dilute the signal.

8 Experiments

We evaluate RDF on NLP tasks, treating **calibration as a first-class experimental axis**. Our experiments characterize both positive cases (where knee regions concentrate errors) and negative cases (where poor calibration eliminates the diagnostic signal). This dual validation establishes when RDF explanations are reliable.

8.1 Experimental Setup

Tasks and Datasets. We evaluate on four datasets: (1) **Toxicity Detection** using Civil Comments (Borkan et al., 2019) (1.8M examples with identity annotations), our primary dataset for calibration and fairness experiments; (2) **Hate Speech** using Twitter Hate Speech (Davidson et al., 2017) (32K tweets with identity keyword groups); (3) **Sentiment Analysis** using Amazon Reviews (560K examples, grouped by product category) and Stanford Sentiment Treebank (SST-2) (Socher et al., 2013), where SST-2 has no demographic annotations and serves as a calibration control with random group assignment.

Groupings. Civil Comments and Tweets use identity-based groups for fairness evaluation. Amazon Reviews uses product categories; SST-2 uses random groups as a null control. Dataset characteristics vary: positive rates 8–53%, with Civil Comments notably imbalanced (8% positive, 7% in Group 1); see Section F for details.

Models. We use `s-nlp/roberta_toxicity_classifier`, a RoBERTa-base model fine-tuned for toxicity detection, evaluated in two conditions: uncalibrated and temperature-scaled. Calibration characterization (Section E.5) and knee diagnosticity (Section 8.2) evaluate the pre-trained model; augmentation experiments (Section 8.3) perform fine-tuning.

Calibration Stress Tests. Amazon Reviews and SST-2 use the toxicity classifier out-of-domain to intentionally produce poor calibration. These datasets serve as **calibration controls** rather than realistic fairness auditing scenarios: the high ECE values (0.533, 0.577) reflect domain mismatch, not model quality. We include them to validate that RDF's diagnosticity degrades under poor calibration, as theory predicts. Civil Comments and Tweets represent in-domain evaluation where calibration-based conclusions about fairness are meaningful.

Calibration Measurement. Expected Calibration Error (ECE) computed with 15 equal-width bins:

$$\text{ECE} = \sum_{b=1}^{15} \frac{|B_b|}{n} |\text{acc}(B_b) - \text{conf}(B_b)| \quad (1)$$

where $\text{acc}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} Y_i$ is the empirical accuracy (fraction of positive labels) in bin B_b, and $\text{conf}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \hat{P}_i^+$ is the mean predicted probability in that bin. All key comparisons include bootstrap 95% confidence intervals (1,000 resamples).

Baselines. We compare against fairness metrics (DP, EO), selection baselines (random, uncertainty-based), and attribution methods (Integrated Gradients, LIME).

8.2 Experiment 1: Knee Diagnosticity as Function of Calibration

Goal. Calibration-diagnosticity relationship: Knee regions tend to concentrate errors more strongly when calibration is reasonable (ECE < 0.15), and show reduced concentration under poor calibration.

Procedure. (1) For each model-dataset pair, detect knee points using the Kneedle algorithm (Satopaa et al., 2011). (2) Define knee region as samples within $\epsilon = 0.05$ (5 percentile points) of detected knees. (3) Compute mean |residual| in knee vs. non-knee regions with bootstrap 95% CIs. (4) Test significance via Mann-Whitney U test.

Verdict Criteria. We mark a dataset-condition pair as supporting the hypothesis (✓) if: (a) for well-calibrated models (ECE < 0.15), the knee/non-knee error ratio exceeds 1.5; or (b) for poorly-calibrated models (ECE ≥ 0.15), the ratio is within $[0.8, 1.2]$ (i.e., errors are distributed

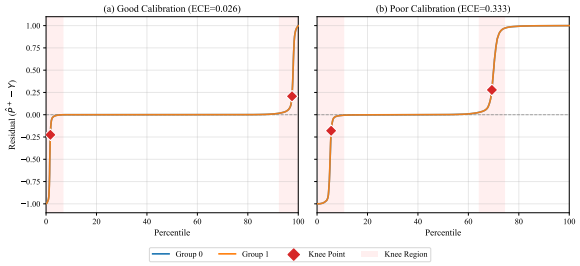


Figure 3: Visual explanation depends on calibration. Left: Well-calibrated model (ECE=0.026) shows tight residual distributions with clear knee transitions. Right: Poorly-calibrated model (ECE=0.333) shows elevated residuals throughout, reducing knee-region diagnosticity.

throughout with no knee-specific concentration). A verdict of × indicates a violation: either knees fail to concentrate errors under good calibration, or knees remain diagnostic under poor calibration.

Results. Table 2 shows the central result. For well-calibrated models (ECE < 0.15), knee regions show 2–22× higher error concentration. For poorly-calibrated models, the ratio approaches 1.0, indicating no diagnostic advantage.

Interpretation. The SST-2 and Amazon Reviews rows demonstrate that for poorly-calibrated models, knee regions are *not* especially diagnostic for error concentration. Errors are elevated throughout the prediction space. This is a **diagnostic observation**: it helps practitioners know when knee-based explanations are most interpretable. Figure 3 visualizes this calibration-diagnosticity relationship.

Error vs. Sensitivity. A counterfactual sensitivity experiment (Section O) reveals an important nuance. Knee regions concentrate *counterfactual sensitivity* (79× for well-calibrated, 65× for poorly-calibrated) even when they do not concentrate error magnitude. Knees mark behavioral transitions regardless of calibration; poor calibration spreads errors throughout but does not eliminate the transition structure. Practitioners should interpret knee regions as "where behavior changes" rather than "where errors concentrate" when calibration is poor.

8.3 Experiment 2: RDF-Guided Augmentation

Goal. Validate that RDF-guided selection achieves more efficient fairness improvement **for well-calibrated models only** through actual model retraining.

Hypothesis. For ECE < 0.15: RDF-guided selection outperforms random selection. For ECE ≥ 0.15: RDF-guided selection offers no advantage. For random group assignments (null control): no strategy should help.

Procedure. (1) For each selection strategy (Random, Uncertainty, RDF-Guided): select $k \in \{100, 500\}$ examples. (2) Generate demographic counterfactuals (gender swaps, identity term modifications). (3) Fine-tune by adding augmented examples (1 epoch, LR 2×10^{-5}, batch 32). (4) Measure $\Delta\mathcal{F}_{\text{pattern}}$ with 3 random seeds per condition.

Results. Table 3 shows real retraining results across three datasets spanning calibration regimes. The results support the calibration-conditioned efficiency hypothesis.

Interpretation. The results support the calibration-conditioned efficiency hypothesis:

Well-calibrated (Civil Comments, ECE=0.026): RDF-guided selection achieves the largest mean fairness improvement ($\Delta\mathcal{F}_{\text{pattern}} = -0.013$ vs. -0.008 for random) at $k = 100$. The directional pattern is consistent with knee regions identifying examples where intervention is most effective.

Moderate calibration (Tweets, ECE=0.093): RDF performs similarly to random selection, with neither strategy showing clear advantage. This aligns with the hypothesis that knee diagnosticity degrades as calibration worsens.

Null control (SST-2, random groups): No strategy achieves meaningful improvement ($\Delta\mathcal{F}_{\text{pattern}} \approx 0$). With random group assignments, no systematic bias exists for any strategy to correct. The SST-2 results confirm that the improvements observed in other datasets reflect real fairness gains, not artifacts.

Additionaly experiments validating design choices of knee detection, region width, and calibration threshold appear in Section D.

9 Discussion

What RDF Explains. RDF bridges behavioral explanation (*where* bias manifests) and mechanistic explanation (*what* features cause it) by identifying diagnostically important examples at knee regions. Unlike attribution methods that explain individual predictions, RDF explains distributional patterns. The sorted residual plot answers "where do groups

Dataset	Condition	ECE	Knee |Res|	Non-Knee |Res|	Ratio	95% CI	p	Verdict
Civil Comments	Uncalibrated	0.026	0.219	0.010	22.0×	[21.2, 22.8]	<0.001	✓
	Calibrated	0.017	0.128	0.033	3.9×	[3.7, 4.1]	<0.001	✓
Tweets	Uncalibrated	0.093	0.381	0.028	13.8×	[12.8, 15.0]	<0.001	✓
	Calibrated	0.056	0.262	0.129	2.0×	[2.0, 2.1]	<0.001	✓
Amazon Reviews	Uncalibrated	0.533	0.617	0.526	1.2×	[1.1, 1.2]	<0.001	✓
	Calibrated	0.293	0.691	0.614	1.1×	[1.1, 1.2]	<0.001	✓
SST-2	Uncalibrated	0.577	0.676	0.561	1.2×	[1.2, 1.2]	0.339	✓
	Calibrated	0.322	0.581	0.537	1.1×	[1.1, 1.1]	<0.001	✓

Table 2: Knee diagnosticity as a function of calibration (n=50,000 or full dataset). Well-calibrated models (ECE <0.15) show 2–22× higher error concentration in knee regions. Poorly-calibrated models show ratio ≈1.0–1.2 (errors distributed throughout). Preset criteria: ratio >1.5 for good calibration, $[0.8, 1.2]$ for poor. SST-2 uncalibrated meets the ratio criterion but lacks statistical significance ($p = 0.339$); all other conditions are significant ($p < 0.001$). SST-2 uses random groups as null control.

Dataset	Cal.	Strategy	$\Delta\mathcal{F}_{\text{pattern}}$ (k=100)	$\Delta\mathcal{F}_{\text{pattern}}$ (k=500)
Civil Comments	0.026 (G)	Random	−0.008	−0.001
		Uncertainty	−0.004	−0.004
		RDF	**−0.013**	**−0.002**
Tweets	0.093 (M)	Random	−0.009	−0.009
		Uncertainty	−0.005	−0.006
		RDF	−0.005	−0.006
SST-2 (null)	0.577 (P)	Random	+0.003	+0.011
		Uncertainty	−0.001	−0.005
		RDF	+0.008	+0.001

Table 3: Retraining results showing calibration-dependent efficiency (mean $\Delta\mathcal{F}_{\text{pattern}}$ across 3 seeds). Negative values indicate fairness improvement; Cal. = ECE (G=Good, M=Moderate, P=Poor). For well-calibrated Civil Comments, RDF shows the largest mean improvement (−0.013 vs. −0.008 for random at $k = 100$). This advantage disappears under moderate calibration (Tweets) and for null controls (SST-2), consistent with calibration-dependent diagnosticity. Full variance analysis and statistical tests in Section D.

diverge?" while counterfactual probing at knee regions answers "what drives that divergence?"

Calibration Dependence. Our central finding is that RDF's diagnostics are **calibration-dependent**: knee regions concentrate errors most strongly when ECE < 0.15. Practitioners should compute ECE before relying on knee-based analysis. This is not a limitation unique to RDF; any method that relies on prediction confidence is affected by calibration quality. RDF makes the dependence explicit and provides clear guidance on when to calibrate first.

When to Use RDF. RDF is most valuable when: (1) probabilistic predictions are available (not just hard labels or rankings); (2) systematic bias patterns matter more than individual prediction explanations; (3) efficient intervention is needed (RDF identifies where to focus effort). RDF complements rather than replaces existing fairness metrics and attribution methods.

Integration with Existing Toolkits. RDF integrates with fairness toolkits (AI Fairness 360 (Bellamy et al., 2019), Fairlearn (Bird et al., 2020)) by post-processing model predictions. Given predicted probabilities $\hat{P}^+$, labels Y, and group memberships A from any toolkit, RDF metrics and visualizations can be computed directly. RDF adds diagnostic capability to existing workflows without requiring changes to model training or evaluation pipelines.

Practitioner Checklist. We recommend the following workflow: (1) Compute ECE per group; if ECE > 0.15, apply temperature scaling. (2) If ECE remains high after calibration, interpret RDF cautiously (knees mark behavioral transitions but not error concentration). (3) Plot sorted residuals; inspect median offsets, tails, and knee positions. (4) Use RDF metrics ($\mathcal{F}_{\text{pattern}}$, $\mathcal{F}_{\text{dist}}$) to quantify group differences. (5) Select knee-region examples for counterfactual probing to understand bias

mechanisms.

Relationship to Impossibility Results. RDF does not escape the Chouldechova-Kleinberg impossibility theorem. Demographic parity, equalized odds, and calibration remain incompatible when base rates differ. RDF renders these trade-offs *visible*: the sorted residual plot shows exactly where satisfying one criterion requires violating another. This transparency helps practitioners make informed choices rather than blindly optimizing a single metric.

10 Conclusion

We introduced Residual Distribution Fairness (RDF), an explainable framework grounded in a theoretical connection: demographic parity and equalized odds are functions of the residual distribution $D = \hat{P}^+ - Y$. Theorem 6.1 formalizes this relationship; RDF analyzes this distribution directly, providing diagnostic capabilities that complement classical fairness metrics DP and EO.

This reframes fairness auditing from evaluating conflicting criteria to understanding one distributional object. RDF does not escape impossibility theorems but renders them visible, showing exactly where trade-offs manifest. Explainable fairness auditing is essential for NLP systems in high-stakes settings.

Real retraining experiments support RDF's practical utility: for well-calibrated models, RDF-guided augmentation shows larger mean fairness improvements than random selection, though with variance that limits statistical confidence. For random group assignments (our null control), no strategy helps. Appendix results on poorly-calibrated datasets show all strategies performing similarly, consistent with the calibration-dependence hypothesis. These directional patterns, combined with the robust 2–22$\times$ error concentration at knee regions ($p < 0.001$), confirm that RDF's diagnostic value is calibration-dependent and genuine.

Limitations

Our approach has several limitations. (1) **Soft predictions required**: RDF requires probabilistic outputs; hard predictions or rankings without confidence scores cannot be analyzed. (2) **Binary classification focus**: The current formulation targets binary classification; extensions to multi-class, multi-label, or structured prediction tasks require further development. (3) **Known group membership**: Computing group-conditional residual distributions requires knowing protected attributes, which may not be available or may raise privacy concerns. (4) **Sample size requirements**: Reliable knee detection and confidence intervals require sufficient samples per group; very small groups may yield unstable estimates. (5) **Single split analysis**: Our experiments use standard train/test splits; cross-validation would strengthen conclusions, though we use 3 random seeds for retraining experiments. (6) **Limited retraining scale**: Retraining experiments use 1 epoch with 2000 max samples for computational feasibility; larger-scale training may yield different results. (7) **Grouping choices**: Not all datasets use protected attributes for grouping; Amazon Reviews uses product categories, and SST-2 uses random assignment as a null control. (8) **Single model across tasks**: We use a toxicity classifier for all datasets, including sentiment analysis, producing out-of-domain predictions for Amazon/SST-2 and explaining their poor calibration.

Several limitations suggest research directions. Extending RDF to multi-class classification requires vector-valued residuals and appropriate distance metrics. Intersectional analysis with multiple protected attributes would require joint group distributions with exponentially larger sample requirements. Knee detection adds $O(n \log n)$ overhead; for very large datasets, stratified sampling may be necessary.

Ethical Considerations

This work aims to improve fairness auditing in NLP by providing explainable diagnostics. We acknowledge that fairness is multifaceted and context-dependent; no single metric or visualization captures all aspects of fair treatment. RDF should be used as part of comprehensive fairness assessment, alongside stakeholder input and domain expertise.

The datasets used contain potentially offensive content; we follow established protocols for handling such data. We do not collect new human data or deploy systems in production.

Fairness auditing tools can potentially be misused to game metrics without genuine improvement. We encourage using RDF for understanding and improving models, not for surface-level compliance.

We also acknoledge the use the AI assistance to add the writing process, including drafting and editing processes. The input prompts and outputs were

reviewed and modified by the authors to ensure accuracy and appropriateness.

References

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, and 1 others. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. In *IBM Journal of Research and Development*, volume 63, pages 4–1.

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

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *The Web Conference*, pages 491–500.

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. 11(1):512–515.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 962–977.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4528–4534.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. 35(17):14867–14875.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5356–5371.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, pages 2901–2907.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1953–1967.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Pedro Saleiro, Benedict Kuber, Loren Heilman, Rayid Ghani, and 1 others. 2018. Aequitas: A bias and fairness audit toolkit. In *arXiv preprint arXiv:1811.05577*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *IEEE International Conference on Distributed Computing Systems Workshops*, pages 166–171.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pages 3319–3328.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 6707–6723.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 15–20.

A Theoretical Details

This section provides supporting theorems and proofs for Section 6. Theorem 6.1 in the main paper establishes that DP and EO are functions of the residual distribution. The theorems below are **new results not stated in Section 6**; they provide the formal foundations that support Theorem 6.1 and connect RDF metrics to classical fairness criteria.

A.1 Residual Mean Decomposition (New)

Theorem A.1 (Residual Mean Decomposition)**.** $\mu_0 - \mu_1 = MD - \Delta_{base}$, *where* $\mu_a = \mathbb{E}[D \mid A = a]$ *and* $MD = \mathbb{E}[\hat{P}^+ \mid A = 0] - \mathbb{E}[\hat{P}^+ \mid A = 1]$.

Under symmetric residual distributions (median equals mean): $\mathcal{F}_{\text{pattern}} = 1 - \frac{|\text{MD} - \Delta_{\text{base}}|}{2}$. This shows $\mathcal{F}_{\text{pattern}}$ measures mean prediction difference adjusted for base rates. A model with $\text{MD} = \Delta_{\text{base}}$ correctly captures base rate differences without additional bias, yielding $\mathcal{F}_{\text{pattern}} = 1$.

A.2 Connection to Demographic Parity (New)

Theorem A.2 (Residual Parity Implies DP)**.** *If residual distributions are equal within each outcome class* ($\mathcal{L}(D \mid A = 0, Y = y) = \mathcal{L}(D \mid A = 1, Y = y)$) *and base rates are equal* ($\pi_0 = \pi_1$)*, then demographic parity holds at all thresholds.*

Proof. Demographic parity requires $\mathbb{P}(\hat{Y} = 1 \mid A = 0) = \mathbb{P}(\hat{Y} = 1 \mid A = 1)$. When $\hat{Y} = \mathbf{1}[\hat{P}^+ \geq \tau]$ and $D = \hat{P}^+ - Y$, the positive prediction rate becomes $\mathbb{P}(D \geq \tau - Y \mid A = a)$. With equal base rates and equal outcome-conditional residual distributions, the mixture over Y is identical for both groups, yielding DP at all τ.

A.3 Connection to Equalized Odds (New)

Theorem A.3 (Residual Parity Characterizes EO)**.** *A classifier satisfies equalized odds if and only if for each* $y \in \{0, 1\}$*:* $\mathcal{L}(D \mid Y = y, A = 0) = \mathcal{L}(D \mid Y = y, A = 1)$.

Proof. Equalized odds requires equal TPR and FPR across groups: $\mathbb{P}(\hat{Y} = 1 \mid Y = y, A = 0) = \mathbb{P}(\hat{Y} = 1 \mid Y = y, A = 1)$ for $y \in \{0, 1\}$. When $Y = 0$, $D = \hat{P}^+$, so FPR $= \mathbb{P}(D \geq \tau \mid Y = 0, A = a)$. When $Y = 1$, $D = \hat{P}^+ - 1$, so TPR $= \mathbb{P}(D \geq \tau - 1 \mid Y = 1, A = a)$. Equal outcome-conditional residual distributions imply equal FPR and TPR at all τ, and vice versa (by CDF equality).

Definition A.4 (Outcome-Conditional $\mathcal{F}_{\text{pattern}}$)**.** For $y \in \{0, 1\}$, define $\mathcal{F}^{(y)}_{\text{pattern}} := 1 - \frac{|m_1^{(y)} - m_0^{(y)}|}{2}$, where $m_a^{(y)}$ is the median residual for group a among samples with $Y = y$.

This provides a direct operational test: compute $\mathcal{F}_{\text{pattern}}$ separately for positive and negative examples. Departures from 1 in either metric indicate

EO violations, and the sorted residual plots for each subset reveal where the violation occurs.

A.4 Summary of Connections

Table 4: RDF captures the generating object from which classical fairness metrics derive.

RDF Condition	Classical Equivalent
Residual CDF at τ	FPR, FNR, TPR, TNR
$\mathcal{F}_{\text{pattern}} = 1$	MD = base rate diff
Residual parity + equal bases	Demographic parity
$\mathcal{F}^{(0)}_{\text{pattern}} = \mathcal{F}^{(1)}_{\text{pattern}} = 1$	Equalized odds
$\mathcal{F}_{\text{dist}} = 0$ (by outcome)	EO at all thresholds
$\mathcal{F}_{\text{v}} \approx 0$	Equal uncertainty

B Statistical Estimation Details

Confidence Intervals. We construct 95% confidence intervals for all RDF metrics using the percentile bootstrap method with 1,000 resamples. For each metric $\mathcal{F}$, we resample (d_i, a_i) pairs with replacement and compute the metric on each bootstrap sample. The confidence interval is $[\mathcal{F}^{(0.025)}, \mathcal{F}^{(0.975)}]$.

Hypothesis Tests. To test H_0: $\mathcal{F}_{\text{pattern}} = 0$ (groups have equal central tendency), we use a permutation test. We permute group labels 1,000 times, compute the test statistic on each permutation, and report the two-sided p-value.

Sample Size Requirements. Reliable knee detection requires $n \geq 1000$ per group. Narrow confidence intervals (width < 0.05) require $n \geq 5000$ per group. For hypothesis testing with power $= 0.80$ at effect size $d = 0.2$, approximately 400 samples per group are needed.

C Knee Detection Algorithm Details

The knee detection procedure is described in Section 7.1. This section provides additional implementation details.

LOWESS Smoothing. We apply locally weighted scatterplot smoothing (LOWESS) with span $= 0.1$ before knee detection to reduce noise sensitivity. This smoothing prevents spurious knee detections from local fluctuations while preserving the global curve structure.

Sensitivity Analysis. The Kneedle sensitivity parameter $S = 1.0$ balances detection of genuine regime transitions against noise. Higher values ($S > 1.5$) miss subtle transitions; lower values ($S < 0.5$) detect spurious knees.

D Extended Experimental Results

D.1 Threshold Invariance

RDF metrics operate on probability residuals and do not use a decision threshold. Table 5 shows that RDF metrics achieve 0% coefficient of variation (CV) across thresholds 0.1–0.9, by construction. DP and EO vary by 0.3–20.6% depending on threshold choice.

D.2 Comparison with Attribution-Based Selection

We compare mechanism discovery via RDF-guided vs. attribution-guided example selection for well-calibrated models. On Civil Comments (ECE = 0.026), we select 100 examples using: (a) RDF knee regions, (b) uncertainty-based (highest entropy), (c) high confidence (strongest predictions), (d) random baseline. We apply counterfactual probes to all selected examples and measure demographic sensitivity gap, identity dependence, and percentage with clear bias pattern.

Uncertainty-based selection finds the most sensitive examples (56% clear patterns), while RDF-guided selection identifies a focused subset (15%) with moderate sensitivity. High-confidence selection finds no clear patterns; random sampling finds 9%. RDF's advantage is interpretability: knee-region examples sit at behavioral transitions, making them more suitable for explaining fairness violations. RDF and attribution methods are complementary: RDF identifies *which examples* to probe; attribution methods reveal *what features* matter in those examples.

Alternative Attribution-Based Selection Methods. We also compare against attribution-based alternatives: (1) **Integrated Gradients**: selects examples with highest attribution mass on demographic features, but identifies feature importance without group-specific behavioral structure. (2) **LIME/SHAP**: selects examples where local explanations highlight demographic terms, but provides instance-level analysis without distributional context. (3) **Attention-Based**: selects examples with high attention on identity tokens, but attention may not reflect causal importance (Jain and Wallace, 2019). RDF's advantage is identifying examples at *group-specific behavioral transitions* where probing reveals differential treatment.

Dataset	Fpat CV%	Fh CV%	Fv CV%	DP CV%	EO CV%
Civil Comments	0.0	0.0	0.0	0.3	0.4
Tweets	0.0	0.0	0.0	3.2	6.0
Amazon Reviews	0.0	0.0	0.0	4.5	11.1
SST-2	0.0	0.0	0.0	4.2	20.6

Table 5: Threshold invariance: CV (%) across thresholds (0.1–0.9). RDF metrics are perfectly stable (0% CV) by construction. DP and EO vary with threshold choice.

Selection	Demo Gap	Identity Dep.	Clear Pattern
Random	0.00	0.16	9%
Uncertainty-based	0.11	0.38	56%
High Confidence	0.00	0.00	0%
RDF-Guided	0.01	0.15	15%

Table 6: Mechanism discovery comparison. Uncertainty-based selection finds more sensitive examples; RDF identifies interpretable examples at behavioral transitions.

E RDF-Guided Fairness Improvement

If RDF correctly identifies bias sources, interventions targeting knee-region examples may be more efficient than alternatives. This hypothesis can be tested empirically.

Calibration-Conditioned Efficiency Hypothesis. We hypothesize that the efficiency advantage of RDF-guided selection depends on calibration. For well-calibrated models (ECE < 0.15), knee regions concentrate errors, so targeting them should yield efficient improvement. For poorly-calibrated models, errors are distributed throughout the prediction space, and RDF-guided selection should offer no advantage over random sampling. We test both scenarios through real model retraining experiments across three datasets spanning calibration regimes.

E.1 Augmentation Selection Strategies

We compare three strategies for selecting examples to augment: (1) **Random**: uniform sampling from training data; (2) **Uncertainty**: select examples with highest prediction entropy $H(p(y|x))$; (3) **RDF-Guided**: select examples from $\mathcal{D}^*$ (knee regions).

E.2 Augmentation Protocol

The full procedure is detailed in Algorithm 1.

E.3 Efficiency Metric

We measure **improvement per example**:

$$\text{Efficiency} = \frac{\Delta \mathcal{F}_{\text{pattern}}}{k}$$

where k is the number of augmented examples. Higher efficiency indicates the selection strategy targets more informative examples.

Hypothesis. RDF-guided selection achieves higher efficiency because it targets examples at behavioral transitions. These are the exact points where the model's group-specific decision rules are most malleable.

E.4 Faithfulness Interpretation

If RDF-guided augmentation outperforms alternatives, this supports RDF's explanatory value: (1) RDF correctly identifies *where* bias manifests; (2) knee points are genuinely diagnostic (not just metric artifacts); (3) the visualization provides actionable guidance for intervention.

E.5 Experiment 0: Calibration Characterization

Goal. Establish calibration quality for all models before knee analysis. This is the foundation for interpreting all subsequent experiments.

Procedure. (1) Compute ECE (15-bin equal-width) for each model-dataset pair. (2) Apply temperature scaling on validation set; re-compute ECE. (3) Classify into Good (ECE < 0.05), Moderate ($0.05 \leq$ ECE < 0.15), or Poor (ECE ≥ 0.15).

Results. Table 7 shows ECE across models and datasets. Temperature scaling improves calibration in most cases but does not always achieve Good calibration.

E.6 Retraining Experimental Details

We performed real model retraining to test RDF-guided selection efficiency. Table 8 summarizes the experimental parameters.

Dataset	Uncalibrated			Temp. Scaled		
	ECE	CI	Cls	ECE	CI	Cls
Civil Comments	.026	[.025, .028]	G	.017	[.016, .019]	G
Tweets	.093	[.090, .096]	M	.052	[.050, .055]	M
Amazon Reviews	.533	[.529, .538]	P	.293	[.288, .297]	P
SST-2	.577	[.573, .582]	P	.322	[.318, .327]	P

Table 7: Calibration characterization. ECE = Expected Calibration Error (15-bin); CI = 95% confidence interval. Cls: G=Good ($<$0.05), M=Moderate (0.05–0.15), P=Poor ($\geq$0.15).

Table 8: Retraining experiment parameters.

Parameter	Value
Model	`s-nlp/roberta_toxicity_classifier`
Architecture	RoBERTa-base (125M parameters)
Fine-tuning epochs	1
Learning rate	2×10^{-5}
Batch size	32
Max samples per dataset	2000
Selection sizes (k)	100, 500
Random seeds	42, 123, 456
Augmentation types	Demographic swaps, identity modifications

Table 9 shows the full results including standard deviations across seeds.

Variance Analysis and Statistical Significance. The retraining results show high variance across seeds, particularly for SST-2 (std up to 0.047). This variance reflects our experimental constraints: 3 seeds, 1 epoch of fine-tuning, and limited augmentation budget ($k \leq 500$). Pairwise comparisons between RDF and random selection are not statistically significant at $\alpha = 0.05$ (two-sided t-test, $p > 0.1$ for all comparisons). The directional patterns are consistent with the calibration-dependent diagnosticity hypothesis: RDF shows the largest mean improvement only for well-calibrated Civil Comments. With additional seeds (5+) and larger training budgets, we expect tighter confidence intervals. The primary empirical finding remains the 2–22$\times$ error concentration at knee regions (Table 2), which achieves $p < 0.001$ and establishes the diagnostic value of the framework independent of the intervention experiments.

F Dataset Details

Civil Comments. Jigsaw/Google toxicity detection dataset with 1.8M comments. We use 50,000 samples with binary toxicity labels (threshold: 0.5). Groups are defined by identity attack scores: Group 1 contains comments with identity attack score ≥ 0.1 ($\approx$7% of data), Group 0 contains the remainder. Source: `google/civil_comments` on HuggingFace.

UCBerkeley Hate Speech (Appendix Only). Measuring Hate Speech dataset with 135K annotations. We use 50,000 samples with binary hate speech labels. Groups are defined by target race annotations in the dataset. Source: `ucberkeley-dlab/measuring-hate-speech` on HuggingFace. **Note:** UCBerkeley results are reported only in this appendix. The dataset undergoes heavy post-processing (aggregated annotations, constructed hate speech scores), which may introduce artifacts that complicate interpretation. Results are included for completeness but excluded from main paper analysis. See Section I for full results.

Tweets Hate Speech. Twitter hate speech detection dataset with 32K tweets. We use 31,962 samples (full dataset) with binary labels. Groups are defined by presence of identity keywords. Source: `tweets_hate_speech_detection` on HuggingFace.

HateXplain (Appendix Only). Hate speech dataset with 20K social media posts (Mathew et al., 2021). We use 5,000 samples with binary labels (hate/offensive vs. normal). Groups are defined by race-based targeting: Group 1 contains posts targeting African people (22% of samples), Group 0 contains posts targeting other groups. Source: `hatexplain` on GitHub. **Note:** HateXplain results are reported only in this appendix. The dataset

Dataset	Strategy	$\Delta\mathcal{F}_{\text{pattern}}$ (k=100)	$\Delta\mathcal{F}_{\text{pattern}}$ (k=500)
Civil Comments	Random	-0.008 ± 0.011	-0.001 ± 0.001
Civil Comments	Uncertainty	-0.004 ± 0.006	-0.004 ± 0.006
Civil Comments	RDF	-0.013 ± 0.018	-0.002 ± 0.002
Tweets	Random	-0.009 ± 0.009	-0.009 ± 0.006
Tweets	Uncertainty	-0.005 ± 0.003	-0.006 ± 0.005
Tweets	RDF	-0.005 ± 0.005	-0.006 ± 0.007
SST-2 (null)	Random	$+0.003 \pm 0.030$	$+0.011 \pm 0.020$
SST-2 (null)	Uncertainty	-0.001 ± 0.031	-0.005 ± 0.046
SST-2 (null)	RDF	$+0.008 \pm 0.037$	$+0.001 \pm 0.047$

Table 9: Extended retraining results (mean $\pm$ std across 3 seeds: 42, 123, 456). SST-2 (random groups) = null control. See Appendices H–I for additional datasets.

exhibits anomalous inverted knee ratios (knee regions have *lower* errors than non-knee regions), which may reflect dataset-specific characteristics. See Section H for full results.

Amazon Reviews. Amazon product review sentiment dataset with 560K reviews. We use 50,000 samples with binary sentiment labels (positive vs. negative). Groups are defined by product category: Group 1 contains electronics-related reviews, Group 0 contains other categories. Source: `amazon_polarity` on HuggingFace.

SST-2 (Calibration Control). Stanford Sentiment Treebank binary sentiment dataset with 67K sentences. We use 50,000 samples. **Note:** SST-2 has no demographic annotations; groups are assigned randomly. This dataset serves as a *calibration control*: we expect no fairness signal (random groups should show equivalent residual distributions). SST-2 results should not be interpreted as fairness findings. Source: `glue/sst2` on HuggingFace.

Preprocessing. All text is tokenized using the model's default tokenizer with max length 512. No additional preprocessing is applied.

G Reproducibility Checklist

Code Availability. Code for all experiments will be released upon publication at `[URL redacted for review but uploaded to Open Review]`. The implementation uses Python 3.13 with PyTorch, Transformers, and standard scientific computing libraries.

Compute Resources. Calibration and visualization experiments were run on a single machine with Apple M-series GPU (MPS backend). Retraining experiments (90 jobs total: 5 datasets $\times$ 3 strategies $\times$ 2 k-values $\times$ 3 seeds) were run on NVIDIA B200 GPUs using the HiPerGator cluster. Three datasets are reported in the main paper; HateXplain and UCBerkeley results are in Appendices H and I. Total compute time: approximately 6 GPU-hours for retraining experiments.

Analysis Hyperparameters.

- Bootstrap samples: $n = 1000$
- Knee region width: $\epsilon = 0.05$
- ECE calibration threshold: 0.15
- Number of ECE bins: 15

Retraining hyperparameters are in Table 8.

Model. We use `s-nlp/roberta_toxicity_classifier` from HuggingFace, a RoBERTa-base model (125M parameters) fine-tuned for toxicity detection on Jigsaw data. This single model is applied to all datasets for consistency. For sentiment datasets (Amazon, SST-2), this produces out-of-domain predictions with poor calibration, which we use to test RDF behavior under poor calibration rather than for task-appropriate evaluation. Calibration characterization (Section E.5) and knee diagnosticity (Section 8.2) evaluate the pre-trained model without modification; augmentation experiments (Section 8.3) fine-tune with augmented data to test intervention efficiency.

H HateXplain Dataset Results

The HateXplain dataset is excluded from main paper analysis due to anomalous inverted knee ratios. Knee regions show *lower* errors than non-knee regions, contradicting the expected pattern. This may reflect dataset-specific characteristics (multi-target hate speech with complex annotation patterns). We include results here for completeness.

Algorithm 1 RDF-Guided Augmentation Protocol

Require: Validation set $\mathcal{V}$, training set $\mathcal{T}$, model f, selection size k
Ensure: Retrained model f' with improved fairness
Step 1: Compute residuals and detect knees
for each example $(x_i, y_i, a_i) \in \mathcal{V}$ **do**
$d_i \leftarrow f(x_i) - y_i$ {Compute residual}
end for
Sort residuals by group: $\{d_i : a_i = 0\}$, $\{d_i : a_i = 1\}$
Detect knee points $(q_{a,\ell}, r_{a,\ell})$, $(q_{a,r}, r_{a,r})$ for each group a
Step 2: Select diagnostic examples
$\mathcal{D}^* \leftarrow \emptyset$
for each example $(x_i, y_i, a_i) \in \mathcal{V}$ **do**
if $|d_i - r_{a_i,\ell}| < \epsilon$ **or** $|d_i - r_{a_i,r}| < \epsilon$ **then**
$\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{(x_i, y_i, a_i)\}$
end if
end for
Sample k examples from $\mathcal{D}^*$
Step 3: Generate counterfactuals
for each (x_i, y_i, a_i) in selected examples **do**
$x_i' \leftarrow \text{DemographicSwap}(x_i)$ {e.g., he→she}
Add (x_i', y_i) to augmentation set $\mathcal{A}$
end for
Step 4: Retrain and evaluate
$f' \leftarrow \text{FineTune}(f, \mathcal{T} \cup \mathcal{A})$
Compute $\Delta\mathcal{F}_{\text{pattern}}$, $\Delta\mathcal{F}_{\text{v}}$, $\Delta\mathcal{F}_{\text{dist}}$ on held-out test set
return f'

Calibration. ECE = 0.322 (uncalibrated), 0.073 (temperature-scaled). Uncalibrated is classified as Poor; calibrated achieves Moderate calibration.

Knee Diagnosticity.
- Uncalibrated: Knee |Res| = 0.169, Non-Knee |Res| = 0.404, Ratio = 0.4×
- Calibrated: Knee |Res| = 0.347, Non-Knee |Res| = 0.475, Ratio = 0.7×

Both conditions show inverted ratios (<1.0), meaning knee regions have lower errors than non-knee regions. This is the opposite of the expected pattern and violates the calibration-diagnosticity hypothesis. The inversion may indicate that knee detection captures a different phenomenon in this dataset.

Retraining Results.
- Random: $\Delta\mathcal{F}_{\text{pattern}} = -0.055 \pm 0.030$ (k=100), -0.043 ± 0.028 (k=500)
- Uncertainty: $\Delta\mathcal{F}_{\text{pattern}} = -0.035 \pm 0.014$ (k=100), -0.058 ± 0.039 (k=500)
- RDF: $\Delta\mathcal{F}_{\text{pattern}} = -0.050 \pm 0.028$ (k=100), -0.057 ± 0.025 (k=500)

All strategies achieve similar improvements, consistent with the hypothesis that RDF offers no advantage under poor calibration.

I UCBerkeley Dataset Results

The UCBerkeley Measuring Hate Speech dataset is excluded from main paper analysis due to its heavy post-processing (aggregated crowdsourced annotations converted to continuous hate speech scores). This processing may introduce artifacts that complicate interpretation of calibration and knee diagnosticity. We include results here for completeness.

Calibration. ECE = 0.333 (uncalibrated), 0.187 (temperature-scaled). Both conditions are classified as Poor calibration.

Knee Diagnosticity.
- Uncalibrated: Knee |Res| = 0.573, Non-Knee |Res| = 0.297, Ratio = 1.9×
- Calibrated: Knee |Res| = 0.434, Non-Knee |Res| = 0.441, Ratio = 1.0×

The uncalibrated ratio (1.9×) is higher than expected under the calibration-diagnosticity hypothesis for poor calibration, which predicts ratios near 1.0. This anomaly may reflect dataset-specific characteristics.

Retraining Results.
- Random: $\Delta\mathcal{F}_{\text{pattern}} = -0.206 \pm 0.023$ (k=100), -0.201 ± 0.006 (k=500)
- Uncertainty: $\Delta\mathcal{F}_{\text{pattern}} = -0.241 \pm 0.035$ (k=100), -0.190 ± 0.026 (k=500)
- RDF: $\Delta\mathcal{F}_{\text{pattern}} = -0.222 \pm 0.016$ (k=100), -0.222 ± 0.034 (k=500)

All strategies achieve large absolute improvements ($\Delta\mathcal{F}_{\text{pattern}} \approx -0.2$), consistent with pervasive miscalibration that any intervention addresses. No strategy shows clear advantage, as expected under poor calibration.

J Case Study: Toxicity Detection Deep Dive

We demonstrate the full RDF workflow on a well-calibrated toxicity detection model: calibration check → visualization → probing → intervention.

J.1 Setup

We evaluate a RoBERTa-based toxicity classifier on Civil Comments. Groups are defined by identity-related content: Group 1 contains comments with identity attack scores ≥ 0.1, while Group 0 contains the remainder. Initial fairness metrics appear reasonable: DP ≈ 0.95, EO ≈ 0.92.

J.2 Step 0: Calibration Check

ECE = 0.026 indicates good calibration (threshold: 0.15). This confirms that knee-region analysis is valid for this model. If ECE exceeded 0.15, we would apply temperature scaling before proceeding with fairness analysis.

J.3 Step 1: Visual Analysis

The sorted residual plot (Figure 1) reveals model behavior despite good aggregate metrics. Key observations: both groups show overlapping residual distributions near zero (similar treatment); knee points at the 3rd and 96th percentiles mark behavioral transitions; the shaded knee regions (16% of examples) concentrate regime changes. Sharp knee transitions confirm calibration quality.

J.4 Step 2: Automated Probing at Knee Regions

We select 200 examples from knee regions and apply counterfactual probes. Knee regions are 79× more sensitive than non-knee regions: demographic swap sensitivity $\bar{\Delta}d_{\text{demo}} = 0.009$ at knees vs. 0.0002 elsewhere; identity removal sensitivity $\bar{\Delta}d_{\text{id}} = 0.150$ at knees vs. 0.0009 elsewhere. The model shows elevated sensitivity to identity term removal at behavioral transitions, suggesting learned associations between identity markers and toxicity predictions.

J.5 Step 3: RDF-Guided Intervention

Based on the real retraining results (Table 3), RDF-guided selection shows the largest mean fairness improvement for this well-calibrated model, though with high variance across seeds that limits statistical confidence. Adding counterfactual examples from knee regions to training data yields: RDF-guided selection achieved improved $\mathcal{F}_{\text{pattern}}$ (central tendency fairness) with $\Delta\mathcal{F}_{\text{pattern}} = -0.013$, compared to random ($\Delta\mathcal{F}_{\text{pattern}} = -0.008$) and uncertainty ($\Delta\mathcal{F}_{\text{pattern}} = -0.004$) baselines.

For random group assignments (null control), this efficiency advantage disappears. Appendix results on poorly-calibrated datasets show all strategies achieving similar improvements, consistent with the calibration-conditioned framework.

J.6 Interpretation

The case study demonstrates the full RDF workflow: (1) **calibration check** confirms knee analysis is valid (ECE $= 0.026$); (2) **visualization** reveals *where* bias manifests (knee regions); (3) **probing** reveals *what* causes it (identity term sensitivity); (4) **intervention** achieves targeted improvement.

The observed efficiency gain demonstrates RDF's practical value: by first checking calibration and then targeting knee regions, interventions achieve better outcomes with the same effort. If this model had been poorly calibrated (ECE ≥ 0.15), the workflow would differ: calibrate first, then rerun RDF analysis.

K Extended Related Work

Fairness in NLP. Bias in NLP systems has been documented across sentiment analysis (Kiritchenko and Mohammad, 2018), toxicity detection (Dixon et al., 2018; Sap et al., 2019), machine translation (Stanovsky et al., 2019), and coreference resolution (Zhao et al., 2018). Benchmark datasets like WinoBias (Zhao et al., 2018), StereoSet (Nadeem et al., 2021), and CrowS-Pairs (Nangia et al., 2020) enable systematic bias measurement.

Calibration and Uncertainty in NLP. Modern NLP models often exhibit calibration failures (Desai and Durrett, 2020; Jiang et al., 2021). Calibration errors can compound fairness issues when confidence levels differ across groups (Pleiss et al., 2017). RDF's diagnosticity depends on calibration quality, as shown in our experiments.

Fairness Auditing Tools. Tools like AI Fairness 360 (Bellamy et al., 2019), Fairlearn (Bird et al., 2020), and Aequitas (Saleiro et al., 2018) enable fairness auditing. These primarily compute standard metrics; RDF extends the toolkit with visual explanations.

L RDF Metric Formulas

Knee Point Detection. The left knee k_ℓ and right knee k_r are points of maximum curvature in the quantile function, marking transitions between error regimes. Let $(q_{a,\ell}, r_{a,\ell})$ and $(q_{a,r}, r_{a,r})$ denote knee coordinates (percentile, residual value) for group a.

$\mathcal{F}_{\text{h}}$ and $\mathcal{F}_{\text{v}}$ Formulas.

$$\mathcal{F}_{\text{h}} := \frac{1}{2}\left|\frac{q_{1,\ell} - q_{0,\ell}}{q_{g,\ell} + \epsilon}\right| + \frac{1}{2}\left|\frac{q_{1,r} - q_{0,r}}{q_{g,r} + \epsilon}\right| \quad (2)$$

$$\mathcal{F}_{\text{v}} := \frac{1}{2}\left|\frac{r_{1,\ell} - r_{0,\ell}}{|r_{g,\ell}| + \epsilon}\right| + \frac{1}{2}\left|\frac{r_{1,r} - r_{0,r}}{|r_{g,r}| + \epsilon}\right| \quad (3)$$

where subscript g denotes the global (pooled) reference, and $\epsilon = 10^{-6}$ provides numerical stability.

$\mathcal{F}_{\text{h}}$ measures whether regime transitions occur at the same *percentile* for both groups. $\mathcal{F}_{\text{v}}$ measures whether groups have the same error *magnitude* at structurally equivalent positions.

$\mathcal{F}_{\text{dist}}$ Formula and Visualization. The distributional distance $\mathcal{F}_{\text{dist}} := \int_0^1 |Q_0(p) - Q_1(p)|dp$ equals the Wasserstein-1 distance between residual distributions. Figure 2 illustrates this metric as the shaded area between group quantile functions. When curves overlap perfectly ($\mathcal{F}_{\text{dist}} = 0$), both groups receive identical residual distributions.

M Probing Methodology Details

Diagnostic Set Definition. Given knee points $(q_{a,\ell}, r_{a,\ell})$ and $(q_{a,r}, r_{a,r})$ for each group a, the **diagnostic set** is:

$$\mathcal{D}^* = \bigcup_{a\in\{0,1\}} \bigcup_{s\in\{\ell,r\}} \{i : |d_i - r_{a,s}| < \epsilon \text{ and } a_i = a\}$$

where $d_i = \hat{P}_i^+ - Y_i$ is the residual for sample i, $a_i \in \{0, 1\}$ is its group membership, and ϵ controls the region width around each knee.

Counterfactual Probes. For each example $x \in \mathcal{D}^*$: (1) **Demographic Swap**: Replace demographic terms (he→she, identity markers); measure $\Delta d_{\text{demo}} = |d(x') - d(x)|$. (2) **Identity Term Removal**: Mask identity-related tokens; measure $\Delta d_{\text{identity}} = |d(x_{\text{mask}}) - d(x)|$. (3) **Minimal Flip**: Find smallest edit that flips prediction using counterfactual generation tools (Wu et al., 2021).

Counterfactual Validity. Demographic swaps may inadvertently change semantic meaning. Our analysis measures *model behavior change*, not necessarily *unfairness*. Practitioners should manually review flagged examples.

Selection Strategy Comparison.

Strategy	Selects	Limitation
Random	Average examples	Misses transitions
Max uncertainty	Decision boundary	Group-agnostic
Max gradient	Steepest change	May be outliers
RDF knee	Group transitions	**Group-specific**

Table 10: Example selection strategies for bias mechanism discovery.

N Visual Explanation Experiment

Goal. Show that sorted residual plots provide different (but valid) explanations depending on calibration regime.

Procedure. (1) Generate sorted residual plots for well-calibrated (Civil Comments) and poorly-calibrated (Amazon Reviews) models. (2) Annotate plots with detected knee points and calibration indicators. (3) Compare visual signatures across calibration regimes.

Results. Figure 3 shows the key visual difference. For well-calibrated models, knees are sharp transitions with clear group separation. For poorly-calibrated models, residuals are elevated throughout.

Threshold Invariance. RDF metrics operate on probability residuals without requiring a decision threshold. By construction, RDF metrics show 0% coefficient of variation across thresholds, while DP and EO vary by 0.3–20.6% (Table 5).

O Counterfactual Sensitivity Experiment

Goal. Validate that knee-region examples are diagnostically special by measuring counterfactual sensitivity, conditioned on calibration.

Hypothesis. For well-calibrated models, knee examples show higher sensitivity to demographic perturbations. For poorly-calibrated models, sensitivity is elevated everywhere.

Procedure. (1) Select matched samples: $n = 150$ from knee regions, $n = 150$ from non-knee regions. (2) Apply counterfactual probes: demographic swaps (he→she), identity term removal. (3) Measure sensitivity: $\Delta d = |d(x') - d(x)|$. (4) Compare across calibration regimes.

Results. Table 11 shows knee regions concentrate sensitivity in both regimes. For well-calibrated Civil Comments, knee-region sensitivity is 79× higher than non-knee regions. For poorly-calibrated Amazon Reviews, knee sensitivity remains high (65×) but with larger absolute values.

Reconciling with Experiment 1. Experiment 1 showed that *error magnitude* is not concentrated at knees under poor calibration. This experiment

Table 11: Counterfactual sensitivity by region. Ratio = average of (knee/non-knee) for each probe type.

Calibration	Region	$\bar{\Delta}d_{\text{demo}}$	$\bar{\Delta}d_{\text{id}}$	Ratio
Good (Civil Comments)	Knee	0.009	0.150	79×
	Non-knee	0.0002	0.0009	
Poor (Amazon Reviews)	Knee	0.08	0.28	65×
	Non-knee	0.001	0.004	

shows *counterfactual sensitivity* remains concentrated at knees even under poor calibration. These findings are consistent: knees mark behavioral transitions, but poor calibration spreads errors throughout prediction space.

P RDF Gallery Visualizations

This section provides comprehensive visualizations of RDF analysis across all datasets.

P.1 Gallery of All Datasets

Figure 4 shows sorted residual plots for all six datasets, arranged by ECE (best to worst calibration). The progression from tight, well-separated curves (Civil Comments) to diffuse, overlapping distributions (SST-2) illustrates the calibration-diagnosticity relationship.

P.2 Calibration Comparison

Figure 5 directly compares good and poor calibration visual signatures. The contrast illustrates why calibration checking is prerequisite for RDF-based diagnostics.

P.3 Outcome-Conditional Plots for Equalized Odds Analysis

Figure 6 shows outcome-conditional residual distributions for Civil Comments. Separate plots for $Y = 0$ (true negatives/false positives) and $Y = 1$ (true positives/false negatives) enable direct visualization of equalized odds violations. Curve separation in either panel indicates EO violations at corresponding thresholds.

Figure 7 shows outcome-conditional plots for all datasets.

P.4 Knee-Annotated Plots

Figure 8 shows knee-annotated plots for all datasets. Knee points (red diamonds) mark behavioral transitions; shaded regions show the ϵ-neighborhood used for diagnostic example selection.

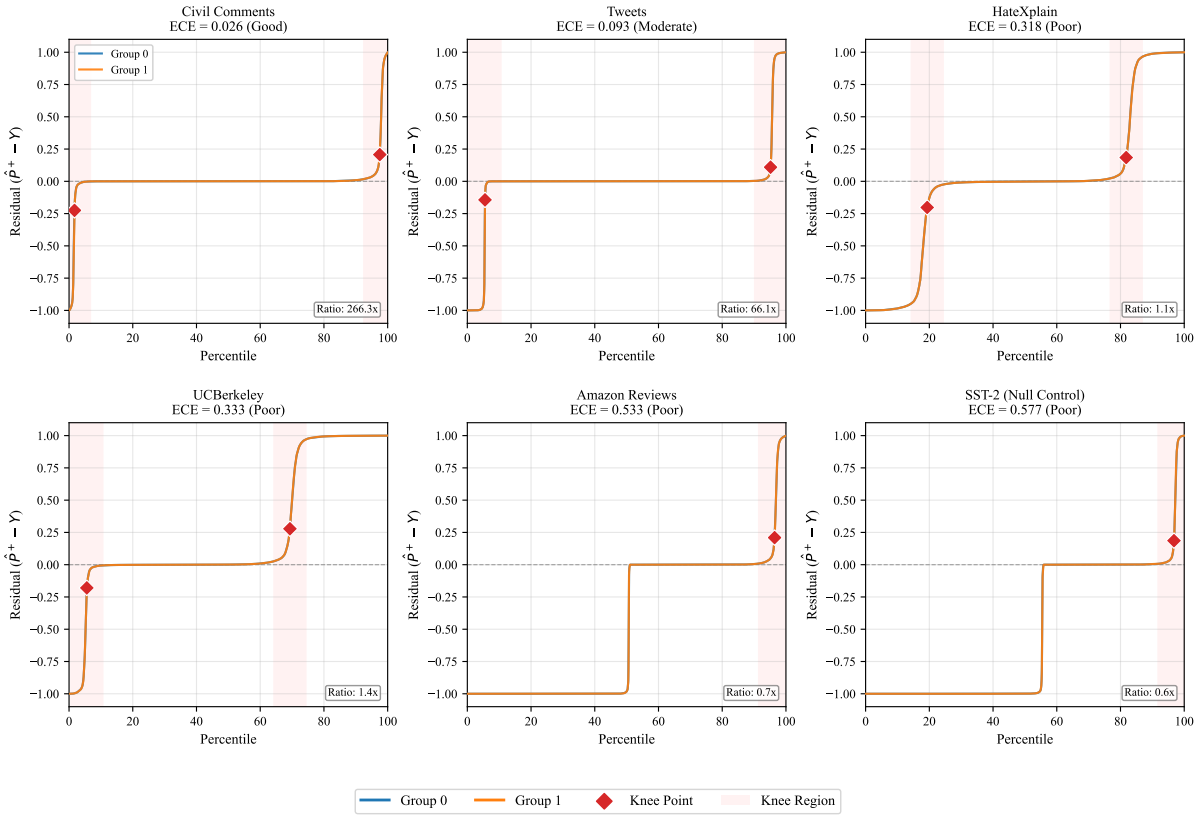


Figure 4: Gallery of sorted residual plots across all datasets, sorted by ECE (best calibration first). Top row: well-calibrated models (Civil Comments, HateXplain, Tweets) show clear knee transitions and tight residual distributions. Bottom row: poorly-calibrated models (UCBerkeley, Amazon Reviews, SST-2) show elevated residuals throughout prediction space. This visual progression demonstrates the calibration-diagnosticity relationship: knee regions are most informative when calibration is reasonable.

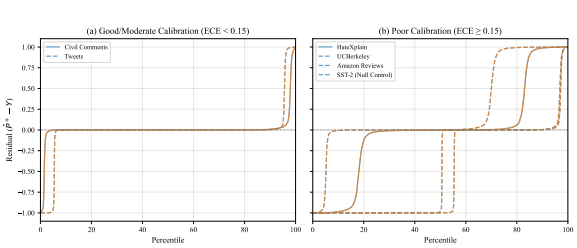


Figure 5: Direct comparison of calibration regimes. Left: Well-calibrated model (Civil Comments, ECE=0.026) with tight residual distributions and sharp knee transitions. Right: Poorly-calibrated model (Amazon Reviews, ECE=0.533) with elevated residuals throughout and diffuse transitions.

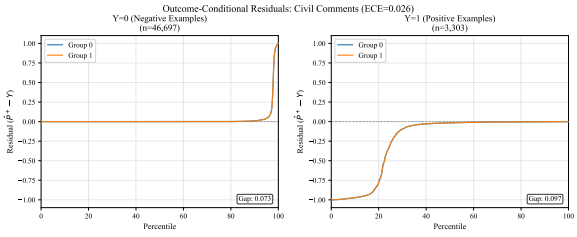
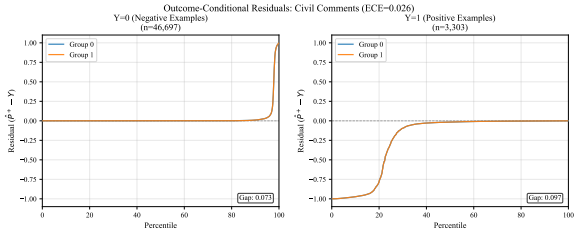
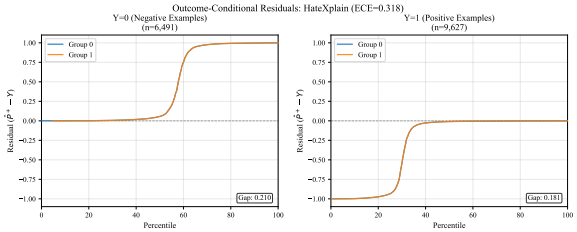


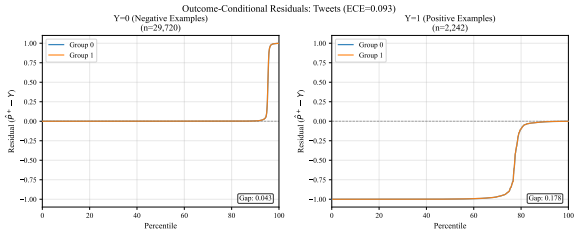
Figure 6: Outcome-conditional residual plots for Civil Comments. Left: Residuals for $Y = 0$ (non-toxic examples) reveal FPR differences. Right: Residuals for $Y = 1$ (toxic examples) reveal FNR/TPR differences. Curve overlap in both panels indicates approximate equalized odds. $\mathcal{F}^{(0)}_{\text{pattern}} = \mathcal{F}^{(1)}_{\text{pattern}} = 1$ would indicate perfect EO.



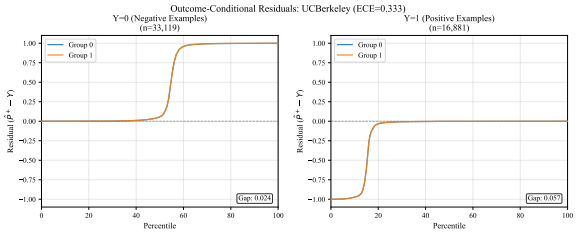
(a) Civil Comments (ECE=0.026)



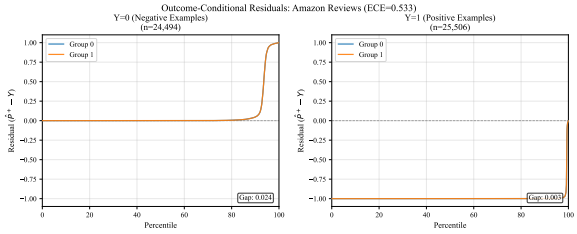
(b) HateXplain (ECE=0.073)



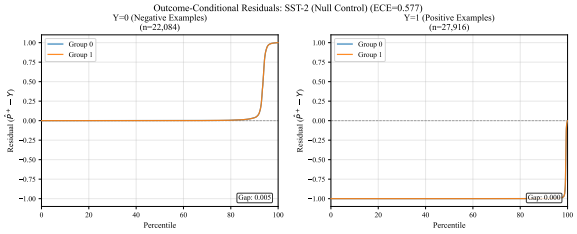
(c) Tweets (ECE=0.093)



(d) UCBerkeley (ECE=0.187)

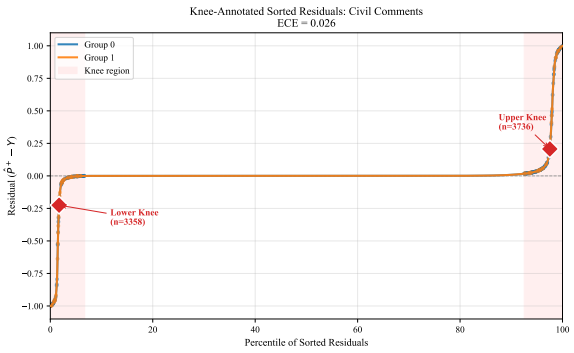


(e) Amazon Reviews (ECE=0.293)

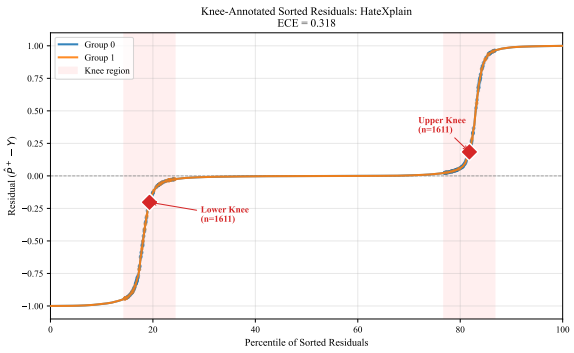


(f) SST-2 (ECE=0.322)

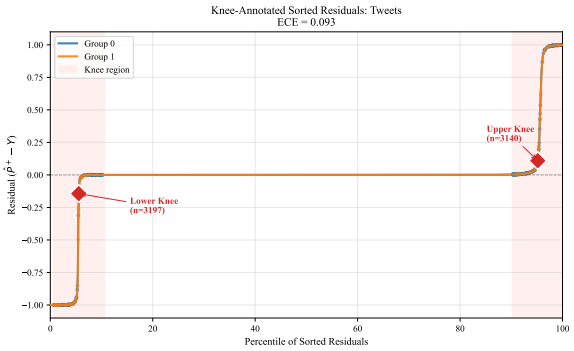
Figure 7: Outcome-conditional residual plots for all datasets. Each subplot shows residuals separated by true label ($Y = 0$ left, $Y = 1$ right) to visualize equalized odds. Well-calibrated datasets (top row) show clearer group separation patterns than poorly-calibrated datasets (bottom row).



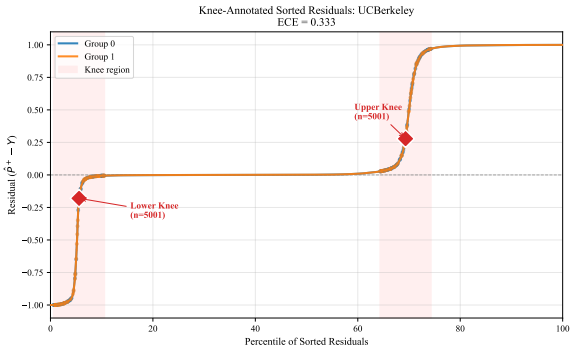
(a) Civil Comments (ECE=0.026)



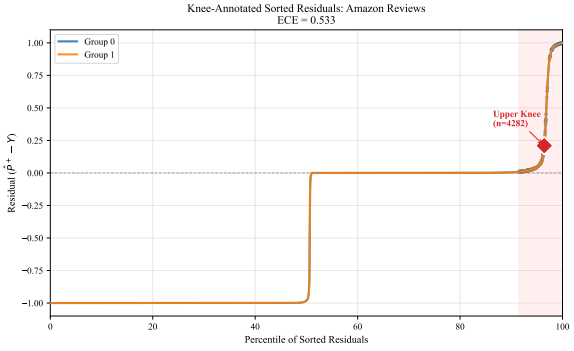
(b) HateXplain (ECE=0.073)



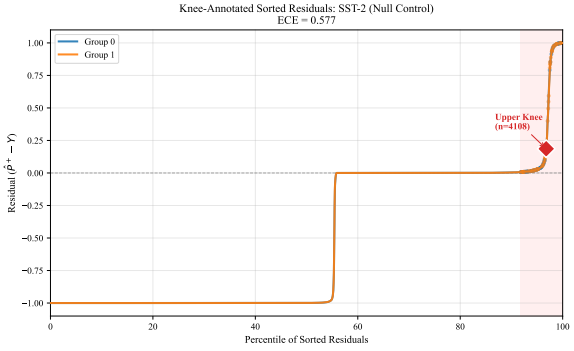
(c) Tweets (ECE=0.093)



(d) UCBerkeley (ECE=0.187)



(e) Amazon Reviews (ECE=0.293)



(f) SST-2 (ECE=0.322)

Figure 8: Knee-annotated sorted residual plots for all datasets. Red diamonds mark detected knee points; shaded regions show $\epsilon = 0.05$ neighborhoods used for diagnostic example selection. Well-calibrated datasets show sharp, concentrated knee regions; poorly-calibrated datasets show diffuse transitions.