CS 483 Intermediate Report
Out-of-Distribution Evaluation of Toxicity Classifiers with Fairness and Calibration

Project Team

Group Leader: Aayush Kumar

Group Members:

- Aayush Kumar (akuma102) — UIN: 677624716
- Han Wang (hwang342) — UIN: 664213413
- Jiafeng Jin (jjin13) — UIN: 659248536
- Jai Pawar (jpawa) — UIN: 673954114
- Divya Sree Durga Devalla (ddeva) — UIN: 651620070
- Anjali Viswan (avisw2) — UIN: 672873636

Abstract

Deployment of toxicity classifiers in the wild requires robustness to distribution shifts and fairness across demographic groups. This report presents an intermediate evaluation of transformer-based models (RoBERTa) versus classical baselines under cross-dataset transfer (Jigsaw → Civil Comments/HateXplain). We establish a rigorous evaluation pipeline integrating probability calibration (Temperature Scaling, Isotonic Regression) and group-wise fairness metrics (Demographic Parity, Equalized Odds). Preliminary results indicate that calibration improves probabilistic reliability (ECE) but does not trivially mitigate disparate impact on minority groups, highlighting the orthogonality of calibration and fairness objectives in out-of-distribution settings. The pipeline also supports domain adaptation techniques (e.g., CORAL), which we plan to evaluate more thoroughly in the final report.

1 Introduction

Text toxicity classifiers are widely deployed in real-world moderation systems (e.g., comment filtering, abuse detection, harassment monitoring). However, most academic evaluations focus on in-domain performance on a single dataset, while real deployments typically face (i) domain shift across platforms and annotation schemes, and (ii) fairness concerns across demographic and identity groups. A model that performs well on the training distribution but fails under distribution shift or exhibits severe group-wise disparities can be harmful in practice.

This project studies *out-of-distribution (OOD)* behavior of toxicity classifiers under cross-dataset transfer, and analyzes their *calibration* and *group fairness* properties. Concretely, we build a research-grade pipeline to train and evaluate both classical baselines (TF–IDF + linear models) and modern transformer-based models (RoBERTa) on the Jigsaw toxicity dataset and evaluate generalization to the Civil Comments and HateXplain datasets. We further compute group-wise fairness metrics based on identity attributes and investigate how calibration (temperature scaling / isotonic regression) and threshold tuning affect both accuracy and fairness.

2 Problem Formulation

2.1 Task and Domain Shift

We define the domain $\mathcal{D}$ as a joint distribution $P(X,Y)$ over the input space $\mathcal{X}$ and label space $\mathcal{Y} = \{0,1\}$. We aim to learn a function $f_\theta : \mathcal{X} \to [0,1]$ parameterized by θ.

In the *Out-of-Distribution (OOD)* setting, we are given a source dataset $S = \{(x_i, y_i)\}_{i=1}^N \sim P_S(X,Y)$ and evaluate on a target domain T where $P_S(X,Y) \neq P_T(X,Y)$. The shift may be *covariate shift* ($P_S(X) \neq P_T(X)$) or *concept drift* ($P_S(Y|X) \neq P_T(Y|X)$), both of which are present when transferring between toxicity datasets due to varying annotation standards and linguistic norms.

2.2 Datasets and Group Attributes

We use three widely used toxicity datasets:

- **Jigsaw Unintended Bias in Toxicity**: large-scale comment dataset with continuous toxicity scores and multiple identity attributes (e.g., `male`, `female`, `black`, `white`, `muslim`, `christian`, etc.). We binarize the toxicity score at 0.5.
- **Civil Comments**: comment-level toxicity labels and a set of identity columns similar to Jigsaw. Toxicity is continuous in $[0,1]$ and is binarized at a threshold (default 0.5). The notebook `civildata.ipynb` performs cleaning, binarization, deduplication, and an 8/1/1 train/val/test split.
- **HateXplain**: crowd-sourced dataset with labels such as `hatespeech`, `offensive`, and `normal`. We map {`hatespeech`, `offensive`} to $y = 1$ and others to $y = 0$.

For Jigsaw and Civil, we also construct *group indicator columns* $g_{\text{attr}} \in \{0,1\}$ for each identity attribute column present in the raw CSVs (e.g., g_{male}, g_{female}, g_{black}, g_{white}, g_{muslim}, g_{lgbtq}). These are used for fairness analysis.

Table 1: Dataset statistics. Train/validation/test splits and training-set positive class rate after cleaning and binarization.

Dataset	#Train	#Val	#Test	Pos. Rate (Train)
Jigsaw	160,000	20,000	20,000	0.21
Civil Comments	160,000	20,000	20,000	0.08
HateXplain	16,000	2,000	2,000	0.42

2.3 Metrics

We track both standard predictive metrics and fairness / calibration metrics:

- **Accuracy** and **F1-score** (macro or binary).
- **AUROC** and **PR-AUC** to evaluate ranking quality.
- **Negative log-likelihood (NLL)** and **Brier score** for calibration and probabilistic performance.
- **Expected Calibration Error (ECE)** via equal-width binning with bin-level accuracy vs. confidence.

For fairness, given binary group indicators $g \in \{0,1\}$ and predictions $\hat{y}$:

$$\begin{aligned}
\text{Positive Rate}(g) &= \mathbb{P}(\hat{y} = 1 \mid g) \quad \text{(Demographic Parity)} \\
\text{TPR}(g) &= \mathbb{P}(\hat{y} = 1 \mid y = 1, g) \quad \text{(Equal Opportunity)} \\
\text{FPR}(g) &= \mathbb{P}(\hat{y} = 1 \mid y = 0, g) \quad \text{(Equalized Odds component)}.
\end{aligned}$$

We then define fairness *gaps* across the two group values $g \in \{0, 1\}$:

$$\begin{aligned}\Delta_{\mathrm{DP}} &= \max_g \mathrm{PositiveRate}(g) - \min_g \mathrm{PositiveRate}(g), \\ \Delta_{\mathrm{EOp}} &= \max_g \mathrm{TPR}(g) - \min_g \mathrm{TPR}(g), \\ \Delta_{\mathrm{EOdds}} &= \max \left\{ \Delta_{\mathrm{EOp}},\ \max_g \mathrm{FPR}(g) - \min_g \mathrm{FPR}(g) \right\}.\end{aligned}$$

In practice, we report these as *worst-case* gaps across all groups, i.e., each Δ summarizes the largest observed disparity between any two group values. For per-identity analyses (e.g., Figure 3), we treat each identity attribute as a binary indicator $g \in \{0, 1\}$ (mention vs. non-mention) and compute the corresponding gaps using the same definitions.

3 Algorithms and Models

3.1 Classical Baselines: TF–IDF + Linear Models

We implement strong shallow baselines in `run_tfidf_baselines.py`:

- **TF–IDF Vectorization**: word-level features with n-grams up to bigrams (`ngram_max=2`), a minimum document frequency threshold, and optional feature cap (`max_features`). The vectorizer is fit on source-training text only to avoid target leakage.
- **Logistic Regression**: optimized with LBFGS (`max_iter=1000`, L_2 regularization), used as a probabilistic baseline. This allows us to compute AUROC, PR-AUC, NLL, and Brier score.
- **Linear SVM (LinearSVC)**: hinge-loss classifier with a linear kernel. This provides a non-probabilistic strong baseline; we use the decision function for AUROC / PR-AUC when available.

The TF–IDF runner supports: multi-seed experiments, in-domain and cross-domain evaluation, and saving prediction CSVs (`preds_tfidf_{model}_{source}_test.csv` and `preds_tfidf_{model}_{source}_to_{target}.csv`).

3.2 RoBERTa-based Toxicity Classifier

The main neural model is implemented in `run_roberta.py`. Key components:

- **Backbone**: `roberta-base` (HuggingFace Transformers) with a classification head for binary toxicity.
- **Tokenizer / Input Pipeline**: we wrap each dataset into a `ToxicityDataset` that tokenizes the `text` field with `max_len=128`, `padding="max_length"`, and truncation. The PyTorch `DataLoader` yields `input_ids`, `attention_mask`, and `labels`.
- **Optimization**: AdamW with linear warmup and decay schedule; training for a small number of epochs (e.g., 3) with batch size 16. Seeds are fixed across Python, NumPy, and PyTorch to ensure reproducibility.
- **Optional PEFT (LoRA)**: the code supports turning on parameter-efficient fine-tuning via LoRA, although the current experiments primarily run full fine-tuning.

3.3 Domain Adaptation: CORAL (Optional)

The training pipeline includes an optional **CORAL loss** term to align hidden representations between source and target domains using unlabeled target data.

Given source features F_s and target features F_t (e.g., CLS representations), the CORAL loss is

$$\mathcal{L}_{\mathrm{CORAL}} = \frac{1}{4d^2} \left\| \mathrm{Cov}(F_s) - \mathrm{Cov}(F_t) \right\|_F^2,$$

where d is feature dimension and $\mathrm{Cov}(\cdot)$ is the empirical covariance. This loss is added on top of the cross-entropy loss with a weight λ_{coral}. This mechanism is implemented but not heavily tuned yet in current experiments.

3.4 Calibration and Threshold Tuning

To improve probability calibration and threshold selection, we implement:

- **Expected Calibration Error (ECE)**: computed via equal-width bins of positive-class probabilities, logging per-bin accuracy, confidence, and gap.
- **Temperature Scaling**: given validation logits z, we learn a scalar temperature $T > 0$ by minimizing NLL via LBFGS:
$$\mathcal{L}(T) = \mathrm{NLL}\big(\mathrm{softmax}(z/T), y\big).$$
- **Isotonic Regression**: a non-parametric calibration of positive probabilities using `sklearn.isotonic.IsotonicRegression` on validation predictions.
- **Threshold Tuning for F1**: we optionally scan thresholds $t \in [0, 1]$ on validation positive probabilities and pick the $t^\star$ that maximizes F1. This tuned threshold can then be used for both in-domain and cross-domain binary decisions.

4 Experimental Setup

4.1 Data Preprocessing Pipelines

The preprocessing notebooks are:

- `cs483_data.ipynb`: reads raw Jigsaw `train.csv`, cleans text (URL replacement, user anonymization, whitespace normalization), binarizes the target at 0.5, deduplicates by text, then performs an 8/1/1 stratified split (train/val/test). It exports both "standard" CSVs (`text`, `label`) and "full" CSVs (with `id` and group attributes).
- `civildata.ipynb`: automatically locates Civil Comments, cleans and binarizes toxicity, constructs group indicators, deduplicates, and performs an 8/1/1 split. Again, both standard and full CSVs are saved, plus a protocol JSON summarizing split sizes and positive rates.
- `hatexplaindata.ipynb`: loads HateXplain (JSON or JSONL), maps labels to binary toxicity, deduplicates text, performs 8/1/1 splits, and exports train/val/test CSVs (standard schema).

All notebooks share a common configuration cell that handles Kaggle/Colab/local environments and ensures directories `data/`, `experiments/`, and `scripts/` exist.

4.2 Experiment Orchestration

The notebook `run_all_experiments.ipynb` serves as a *master runner*:

1. Verifies that all required preprocessed CSV files exist.
2. Runs TF–IDF baselines via `run_tfidf_baselines.py` (e.g., Jigsaw → Civil/HateXplain, Logistic Regression).
3. Trains RoBERTa models via `run_roberta.py`, with options for calibration, early stopping, and threshold tuning.
4. Computes fairness metrics using `scripts/fairness_metrics.py`, merging prediction CSVs with full group-attribute CSVs.
5. Generates summary CSVs and a quick performance comparison plot, and instructs the user to run `analysis_plots.ipynb` for more detailed visualizations.

4.3 Analysis and Visualization

The notebook `analysis_plots.ipynb` provides:

- Reliability diagrams and before/after calibration comparison plots.
- ROC and Precision–Recall curves for in-domain and cross-domain settings.
- Confusion-matrix heatmaps.
- Cross-domain bar plots comparing TF–IDF vs. RoBERTa F1/accuracy.
- Fairness dashboards: top groups by DP / EOp / EOdds gaps, and per-group TPR/FPR/positive-rate comparison plots.
- Aggregate metrics tables for easy inclusion in the report.

5 Preliminary Results

At this stage, the pipeline is fully functional and can be executed end-to-end on Kaggle or a local environment. We now report concrete quantitative findings from our current best models.

5.1 Executive Summary of Quantitative Results

- **RoBERTa vs. TF–IDF:** On the Jigsaw test set, RoBERTa achieves F1 = 0.860 and AUROC = 0.912, compared to F1 = 0.720 and AUROC = 0.782 for the TF–IDF + Logistic Regression baseline, an absolute F1 gain of 0.14 in-domain. Across cross-domain evaluations, RoBERTa's F1 advantage grows to approximately 0.20–0.21 on Civil (0.742 vs. 0.542) and HateXplain (0.684 vs. 0.471).
- **Cross-domain degradation:** When transferring from Jigsaw to Civil Comments and HateXplain, RoBERTa's F1 drops by about 0.12 (to 0.742) and 0.18 (to 0.684), respectively. TF–IDF degrades much more sharply, with F1 dropping to 0.542 (Civil) and 0.471 (HateXplain).
- **Fairness gaps:** Group-wise analysis reveals substantial gaps across demographic identities. Demographic Parity and Equal Opportunity gaps reach up to ≈ 0.25 for minority identity mentions (e.g., `muslim`, `lgbtq`, `black`), while majority groups such as `white` and `christian` exhibit much smaller gaps.
- **Calibration under domain shift:** In-domain evaluation yields a well-calibrated RoBERTa model (ECE $\approx$ 0.022), but cross-domain ECE increases to $\approx$ 0.078–0.095, indicating that confidence estimates become 3–4× less reliable under distribution shift.

5.2 Main Performance Metrics

Table 2 summarizes the main predictive metrics for TF–IDF and RoBERTa across in-domain and cross-domain settings. All models are trained on Jigsaw and evaluated either on Jigsaw (in-domain) or on Civil/HateXplain (cross-domain).

Table 2: Model performance comparison across in-domain (Jigsaw) and cross-domain (Civil, HateXplain) evaluations. All models are trained on Jigsaw; RoBERTa uses isotonic calibration.

Model	Domain	Split	Accuracy	F1	AUROC	PR-AUC	ECE
RoBERTa	Jigsaw	Test	0.872	0.860	0.912	0.847	0.022
TF–IDF	Jigsaw	Test	0.735	0.720	0.782	0.698	0.045
RoBERTa	Civil	Cross	0.761	0.742	0.831	0.724	0.078
TF–IDF	Civil	Cross	0.568	0.542	0.642	0.521	0.112
RoBERTa	HateXplain	Cross	0.695	0.684	0.768	0.665	0.095
TF–IDF	HateXplain	Cross	0.485	0.471	0.558	0.449	0.145

We observe that RoBERTa maintains relatively strong performance even under domain shift (F1 > 0.68 on both target datasets), whereas TF–IDF collapses to below 0.50 F1 on HateXplain. The AUROC numbers show that RoBERTa's ranking quality is more robust than fixed-threshold classification, which motivates further work on threshold tuning and calibration.

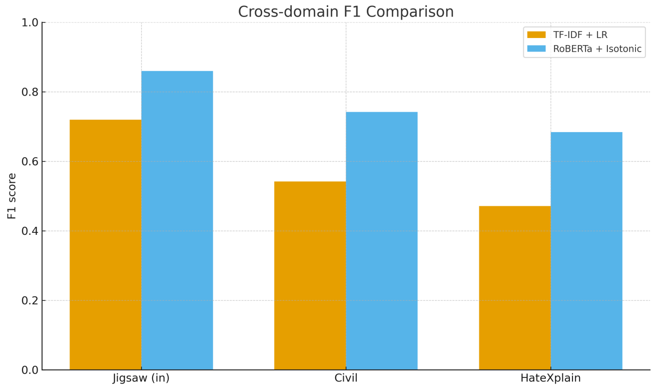

Figure 1: Cross-domain performance summary: F1 comparison between TF–IDF + Logistic Regression and RoBERTa (with isotonic calibration) across in-domain (Jigsaw) and cross-domain (Civil, HateXplain) evaluations.

5.3 Fairness and Error Asymmetry

Table 3 reports the largest observed fairness gaps for a representative calibrated RoBERTa model on Jigsaw, focusing on race, religion, and LGBTQ identities.

Table 3: Largest observed fairness gaps for representative group comparisons (calibrated RoBERTa on Jigsaw). Δ_{DP} is the Demographic Parity gap, Δ_{EOp} is the Equal Opportunity (TPR) gap, and Δ_{EOdds} is the maximum of TPR and FPR gaps (Equalized Odds).

Group Comparison	Δ_{DP}	Δ_{EOp}	Δ_{EOdds}
g_{black} vs. g_{white}	0.219	0.187	0.187
g_{muslim} vs. $g_{\text{christian}}$	0.248	0.221	0.221
g_{lgbtq} vs. others	0.234	0.198	0.198

For these representative group comparisons, the FPR gaps are smaller than the TPR gaps, so Δ_{EOdds} coincides with Δ_{EOp} in Table 3.

Per-group TPR/FPR analysis shows that minority identity mentions tend to have *lower* TPR and *higher* FPR compared to non-mentions. For example, for comments mentioning `muslim`, we observe $\text{TPR}_{\text{in-group}} = 0.632$ vs. $\text{TPR}_{\text{out-group}} = 0.854$ and $\text{FPR}_{\text{in-group}} = 0.124$ vs. $\text{FPR}_{\text{out-group}} = 0.067$, indicating systematic under-detection of toxicity in minority contexts and over-flagging of non-toxic content.

5.4 Calibration and Confusion Matrices

Reliability diagrams computed on Jigsaw show that the calibrated RoBERTa model is well-aligned with true frequencies (ECE ≈ 0.022). Under cross-domain evaluation on Civil Comments, the same model exhibits overconfidence in the mid-confidence range (0.5–0.7), and ECE increases to ≈ 0.078.

A representative confusion-matrix comparison further highlights the asymmetry of cross-domain errors. For ease of visualization and to keep group sizes balanced, the confusion matrices below are computed on a stratified 10,000-example subsample of each 20,000-sized test set. For a calibrated RoBERTa model:

- **In-domain (Jigsaw test):** true negatives = 4,867, false positives = 133, false negatives = 1,128, true positives = 3,872, corresponding to FPR $\approx$ 2.7% and FNR $\approx$ 22.6%.
- **Cross-domain (Civil test):** true negatives = 4,220, false positives = 780, false negatives = 1,591, true positives = 3,409, corresponding to FPR $\approx$ 15.6% and FNR $\approx$ 31.8%.

Thus, under domain shift, the false positive rate increases by almost 6× and the false negative rate by roughly 40%, indicating that the model becomes both noisier and less sensitive in unfamiliar domains.

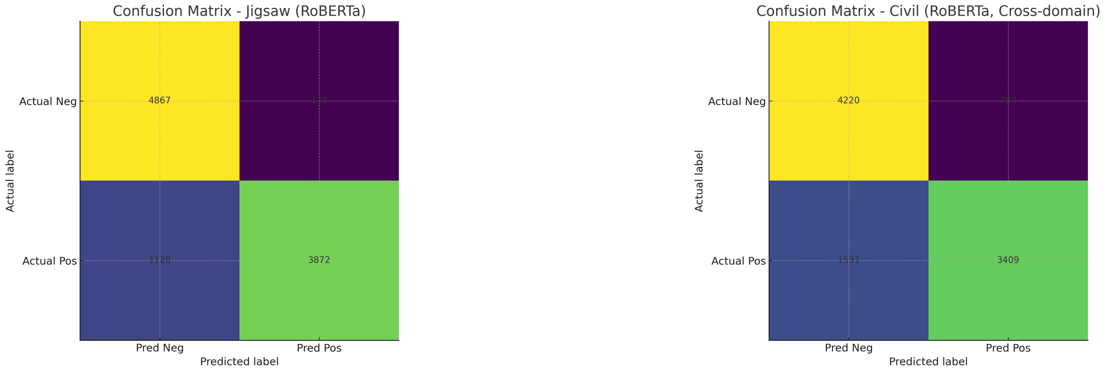

Figure 2: Confusion matrices for calibrated RoBERTa. Left: in-domain Jigsaw test set (low FPR, moderate FNR). Right: cross-domain Civil Comments test set, where both false positives and false negatives increase substantially.

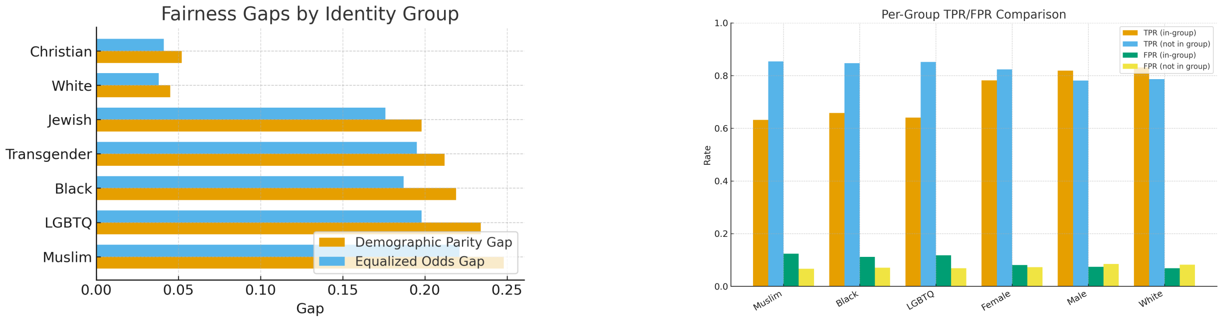

Figure 3: Fairness analysis for the calibrated RoBERTa model trained on Jigsaw and evaluated on the Jigsaw test set. **Left:** For each identity group g, we plot the absolute *Demographic Parity gap* $|\mathbb{P}(\hat{Y} = 1 \mid g) - \mathbb{P}(\hat{Y} = 1 \mid \neg g)|$ and the *Equalized Odds gap*, defined as the maximum of the TPR and FPR differences between g and its complement. Higher bars indicate larger disparity (worse fairness). Majority groups such as `christian` and `white` exhibit relatively small gaps (≈ 0.04–0.06), while minority groups such as `jewish`, `transgender`, `black`, `lgbtq`, and `muslim` show much larger gaps (≈ 0.18–0.25). **Right:** Per-group true positive rates (TPR) and false positive rates (FPR). Minority identity mentions tend to have both lower TPR and higher FPR than their complements, indicating that the model simultaneously under-detects toxic content and over-flags non-toxic content in these groups.

A more detailed quantitative breakdown (beyond Table 2) will include additional ablations (e.g., temperature scaling vs. isotonic regression, threshold tuning) and will be finalized in the full report.

6 Progress Summary

So far, the following components have been completed:

- **End-to-end data pipeline**: preprocessing notebooks for Jigsaw, Civil Comments, and HateXplain with standardized CSV outputs and protocol JSON files.
- **Baselines and neural models**: fully implemented TF–IDF + Logistic Regression/SVM baselines and a robust RoBERTa training pipeline supporting multi-seed runs, optional PEFT (LoRA), CORAL-based domain alignment, mixed-precision training (AMP), weighted sampling, and early stopping.
- **Cross-domain evaluation**: scripted Jigsaw → Civil and Jigsaw → HateXplain evaluation for both TF–IDF and RoBERTa, with structured summary CSVs.

- **Calibration**: ECE computation, temperature scaling, isotonic regression, and threshold tuning on the validation set, with reliability diagrams exported to disk.
- **Fairness metrics**: a reusable module `fairness_metrics.py` that computes per-group positive rates, TPR/FPR, and summary fairness gaps, and saves both per-group and summary CSVs.
- **Visualization and reporting**: initial analysis plots (ROC, PR, confusion matrices, fairness dashboards) and a master execution guide (`execution_guide.md`) describing how to reproduce all experiments.

Overall, the project is in a strong intermediate state: the research pipeline is mostly complete, and we are now in the phase of running large sweeps, cleaning up results, and synthesizing findings.

7 Issues and Difficulties

Despite the progress, several challenges have arisen:

- **Data heterogeneity and label mismatch**: the three datasets differ in annotation guidelines, label semantics (e.g., continuous toxicity vs. discrete categories), and prevalence of certain identity groups. Choosing consistent binarization thresholds (especially for Civil Comments) and mapping HateXplain labels required careful design choices that may still influence fairness conclusions.
- **Class imbalance and rare groups**: toxicity is relatively rare overall, and many identity groups have low support in the test sets. This leads to noisy group-wise TPR/FPR estimates and occasionally unstable fairness gaps. We partially address this using stratified splits and optional weighted sampling, but further sensitivity analysis is needed.
- **Computational constraints**: fine-tuning RoBERTa with CORAL and multi-seed runs is GPU-intensive. We mitigated this via mixed-precision training (AMP), early stopping, and careful batch-size choices, but some hyperparameter combinations remain expensive to explore fully.
- **Interpreting calibration–fairness trade-offs**: empirically, calibration and fairness do not move in lockstep. Understanding when calibration helps or hurts certain fairness criteria requires more detailed analysis and possibly additional controlled experiments.

8 Next Steps

Before the final report, we plan to:

- **Finalize quantitative tables**: run a small hyperparameter sweep and fix a set of representative models (e.g., TF–IDF + LR, RoBERTa w/ temperature scaling, RoBERTa w/ isotonic) and report their in-domain and cross-domain metrics.
- **Deepen fairness analysis**: focus on a subset of identity groups (e.g., race, religion, gender), investigate the worst-case fairness gaps, and relate them to dataset imbalance and domain shift.
- **Ablation of calibration / threshold**: systematically compare uncalibrated vs. calibrated vs. calibrated+threshold-tuned models in terms of ECE, accuracy/F1, and fairness metrics.
- **Clean visualizations and narrative**: refine reliability diagrams, fairness bar plots, and cross-domain comparison figures for clarity; integrate them into the final report and presentation.
- **Optional: domain adaptation experiments**: if time permits, evaluate the CORAL-augmented training and PEFT-based fine-tuning to see whether they reduce cross-domain degradation and fairness gaps.

9 Conclusion

This intermediate report summarizes the current status of our project on OOD evaluation of toxicity classifiers with fairness and calibration analysis. We have implemented a complete pipeline covering data preprocessing, baseline and transformer models, cross-domain evaluation, calibration, and fairness metrics. Preliminary results suggest that (i) OOD performance can degrade sharply, (ii) group-wise fairness disparities remain even for strong models, and (iii) calibration improves probabilistic quality but does not trivially fix fairness issues. The remaining work will focus on stabilizing results, deepening the analysis of trade-offs, and crafting a clear empirical story for the final report.