CAFE: A Context-Aware and Fairness-Weighted Framework for Toxicity Evaluation in Language Models

August 23, 2025

Wickramasinghe J.J.
210706H
Department of Computer Science and Engineering
University of Moratuwa

Table of Contents

1. Introduction

Language models (LMs) have achieved remarkable success across diverse Natural Language Processing (NLP) tasks. However, their deployment in real-world systems remains constrained by their tendency to generate toxic, biased, or harmful text. The RealToxicityPrompts (RTP) dataset [1] provides a benchmark for assessing such behavior by scoring prompts with the Perspective API [2]. Gehman et al. [1] highlighted that even seemingly harmless prompts can trigger toxic generations. Existing evaluation frameworks such as the Perspective API provide a starting point for toxicity measurement, but they suffer from critical flaws such as context insensitivity (e.g. mislabeling sarcastic text as toxic) and social bias.

To address these limitations, this project proposes **CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator)** - an enhanced framework that integrates contextual embeddings, fairness-aware loss, and scalable augmentation of toxicity prompts to deliver more reliable and equitable toxicity assessment.

2. Problem Statement & Motivation

2.1 Problem Statement

Current toxicity evaluators rely on simplistic classification models such as the Perspective API, which fail to capture nuanced linguistic context (e.g., sarcasm, slang) and exhibit systemic bias by disproportionately flagging texts associated with minority groups (e.g., African American Vernacular English). These shortcomings arise from its single-objective training, which optimizes primarily for toxicity classification accuracy without incorporating explicit fairness or context-awareness constraints. Moreover, the RealToxicityPrompts (RTP) dataset, widely used in toxicity evaluation, was released in 2020 might not reflect evolving linguistic trends.

2.2 Motivation

Reducing unfair toxicity detection is crucial for ethical AI deployment, since biased or context-insensitive models can negatively impact marginalized populations and weaken trust in AI systems. A key requirement is context sensitivity, i.e., distinguishing between literal and non-literal expressions (e.g., interpreting *"That's sick!"* as praise rather than an insult). Equally important is robustness, which demands moving beyond outdated datasets by incorporating adversarial and paraphrased prompts to capture the diversity of contemporary online discourse. By addressing these gaps, the proposed CAFE framework offers a novel approach that integrates fairness and contextual awareness, facilitating a more fair and accurate evaluation and advancing the broader field of responsible AI.

3. Literature Review

3.1 Toxic Degeneration in LMs and the RTP Benchmark

Large LMs can "degenerate" into toxic continuations even from seemingly innocuous prompts. Gehman et al. introduced RTP: 100k naturally occurring web prompts with Perspective API toxicity scores to standardize evaluation of toxic generation and compare detoxification methods (e.g., bad-word lists, adaptive pretraining) across models (GPT-2/CTRL, etc.) [1]. RTP became the de facto evaluation bed in both research papers and living benchmarks (e.g., HELM scenarios) because it captures real-world triggers for toxic drift and supports scoring.

Google Jigsaw's **Perspective API** scores toxicity and related attributes and underpins most RTP-based evaluations. Foundational work from Jigsaw introduced **unintended-bias metrics** (e.g., Subgroup AUC, BPSN, BNSP) on the Civil Comments dataset [3] to quantify systematic over-flagging around identity mentions and to study mitigation strategies [4][5]. Subsequent studies show dialect-linked disparities - especially for African American English (AAE) and highlight how annotator insensitivity to dialect/context propagates model bias [6]. More recent audits report multilingual biases (e.g., German receiving higher toxicity than equivalent English), underscoring limits of a single, static classifier as a ground truth oracle across cultures and time [7].

Because production APIs evolve, toxicity scores are not stationary. Pozzobon et al. show that rescoring prior RTP evaluations with a newer Perspective release changes model rankings, warning against apples-to-apples comparisons over time and calling for versioning and transparent rescoring protocols [8]. Earlier work also demonstrated that black-box toxicity classifiers are adversarially fragile (small perturbations can suppress scores), complicating robust evaluation design [9].

A major failure mode is context insensitivity: sarcasm, figurative language, and reclaimed slurs confuse literal toxicity models. Shared tasks like **SemEval-2018 (Irony Detection)** and **FigLang-2020** established benchmark corpora and methods for sarcasm with context (Twitter/Reddit), and **iSarcasmEval-2022** targeted *intended* sarcasm with author-provided labels to reduce noisy supervision [10][11]. These threads inform CAFE's plan to incorporate contextual embeddings and sarcasm signals into toxicity judgments rather than treating text in isolation.

3.2 Alternative / Complementary Datasets Beyond RTP

To move past dated distributions and overt toxicity, researchers introduced **ToxiGen** [12], a large adversarial/implicit hate dataset about minority groups, generated via classifier-in-the-loop prompting and human validation, useful for stress-testing implicit and adversarial toxicity that standard datasets miss. **Civil Comments** (basis of Jigsaw's

unintended-bias work) [3] remains crucial for identity-aware evaluation and subgroup metrics. **HateXplain** adds target labels and human rationales, enabling explanation-aware training/evaluation and bias analyses beyond raw labels [13].

Several decoding-time control methods aim to steer LMs away from toxicity without retraining: **PPLM** (classifier- or bag-of-words-guided gradients at inference), **GeDi** (generative discriminators that bias token probabilities towards desired attributes) [14], and **DExperts** (product-of-experts combining expert / anti-expert models). These improve safety while preserving fluency, but none are fail-safe and can introduce distribution shifts or fairness trade-offs if the steering signal is itself biased [14]. Recent "guard" LLM moderators (e.g., Llama Guard family) illustrate a trend toward LLM-based safety classifiers, yet external evaluations report mixed results and continuing fairness/robustness gaps, motivating context- and fairness-aware evaluators rather than single-source oracles [15].

General fairness work (e.g., Equalized Odds/Opportunity) provides group-based criteria that can be incorporated during training or post-processing to balance error rates across protected attributes - principles increasingly adopted in toxicity research and in Jigsaw's subgroup-AUC-style metrics [16]. In toxicity classification, multi-task and adversarial objectives have been explored to reduce identity-term shortcutting while keeping overall accuracy competitive, but these methods often lack explicit modeling of context (sarcasm, stance, pragmatics), leaving residual gaps in framework targets [17].

To mitigate dataset staleness and improve robustness, NLP commonly uses back-translation and EDA (synonym replacement, swaps, deletions) to diversify training distributions. These strategies can generate paraphrases and adversarial paraphrases for stress-testing toxicity detectors beyond 2020 web snapshots.

3.3 Summary of Gaps CAFE Framework Addresses

1. **Over-reliance on a single black-box model** (Perspective API) leads to moving targets and cultural/language biases.
2. **Context insensitivity** (sarcasm, reclaimed language, dialect) remains a primary error source which is not captured by literal toxicity thresholds.
3. **Stale evaluation distributions** (RTP ~2020) under-represent current discourse, adversarial phrasing, and implicit toxicity.

CAFE is well-positioned to contribute,

(i) fuse contextual embeddings
(ii) incorporate fairness-weighted loss/metrics aligned to subgroup performance
(iii) augment prompts (adversarial/paraphrase/multilingual)

for a modern, stress-tested evaluation approach.

4. Baseline

4.1 Perspective API (Baseline Evaluator)

The Perspective API serves as the baseline toxicity evaluator. It assigns probabilities for categories including *toxicity, severe toxicity, insult, threat, identity attack, profanity,* and *sexually explicit content*. Despite its popularity, Perspective is a black-box classifier with documented weaknesses [1]. It tends to over-flag minority dialects (e.g., African American English), misinterpret sarcasm, and treat toxicity dimensions independently without fairness constraints.

4.2 Limitations of the Baseline

While RTP and Perspective API provide a starting point for benchmarking toxicity, they exhibit several shortcomings that motivate this project:

- **Bias**: Overestimation of toxicity in texts associated with marginalized groups, creating fairness gaps.
- **Context Insensitivity**: Inability to distinguish non-literal or culturally specific expressions (e.g., sarcasm).
- **Dataset Staleness**: RTP was collected in 2020, and may not reflect current online discourse, adversarial phrasing, or evolving linguistic trends.
- **Single-Objective Optimization**: Toxicity is scored in isolation, without explicit multi-objective trade-offs for fairness and context.

These limitations establish RTP + Perspective as the baseline framework against which the proposed CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator) is designed to demonstrate improvements.

5. Objectives & Research Questions

5.1 Objectives

This research is guided by three primary objectives:

1. **Develop CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator):** Design and implement a novel evaluator that integrates contextual embeddings and fairness-aware modeling to overcome the limitations of the Perspective API.

2. **Enhance toxicity evaluation with data augmentation and multi-objective loss optimization:** Introduce paraphrased and adversarial prompt variants to extend the RealToxicityPrompts dataset, and optimize evaluation performance through a multi-objective loss that balances toxicity accuracy, and context sensitivity.

3. **Demonstrate measurable improvements over the baseline:** Empirically validate CAFE against the Perspective API baseline using standard and task-specific metrics

(e.g., F1 score, Expected Maximum Toxicity), ensuring both statistical and practical significance.

5.2 Research Questions

To achieve these objectives, the study seeks to answer the following research questions:

1. **Contextual Embeddings:** Can contextual embeddings (e.g., RoBERTa) improve the detection of nuanced toxicity such as sarcasm, slang, or non-literal expressions that are frequently misclassified by the Perspective API?

2. **Fairness:** Can a fairness-aware loss function reduce systematic bias against minority dialects (e.g., African American Vernacular English) and identity-related expressions, thereby narrowing the fairness gap?

3. **Robustness:** How can dataset augmentation, via paraphrasing and adversarial crafting, extend the coverage of toxicity types and improve the robustness of toxicity evaluation beyond the original 2020 RTP dataset?

4. **Multi-Objective Optimization:** Does a multi-objective optimization framework—balancing toxicity accuracy, fairness, and context-awareness—yield measurable performance improvements across both standard classification metrics (e.g., F1) and RTP-specific metrics (e.g., Expected Maximum Toxicity, Toxicity Probability)?

6. Proposed Methodology

6.1 Data Augmentation

The first stage of the methodology is dataset preparation. The baseline RealToxicityPrompts (RTP) dataset contains 100,000 prompts paired with continuations, each annotated with Perspective API scores across multiple toxicity dimensions (e.g., *toxicity, insult, identity attack, flirtation*). While widely adopted, this dataset is limited by its age (collected in 2020) and coverage gaps. To improve representativeness and robustness, we extend RTP with two augmentation strategies:

- **Paraphrasing:** Prompts are reworded using transformer-based paraphrasing models (e.g., T5), generating semantically equivalent but lexically distinct inputs. This increases lexical diversity while preserving intent.
- **Adversarial Crafting:** Additional prompts are created to stress-test evaluators by introducing sarcasm, slang, and cultural references that frequently confuse existing classifiers.

These augmentations are expected to contribute approximately 10,000 new prompts, thereby broadening coverage of underrepresented linguistic patterns and increasing fairness sensitivity in downstream evaluation.

6.2 Model Architecture

The core of the proposed CAFE (Context-Aware Fairness-Weighted Toxicity Evaluator) is a fine-tuned RoBERTa-base model, chosen for its strong contextual embedding capabilities. Each prompt–continuation pair is concatenated into a single input string (e.g., *"The politician was accused of corruption"*), tokenized with a maximum sequence length of 128 tokens, and encoded into a 768-dimensional embedding (CLS token).

The embedding is passed through a two-layer multilayer perceptron (MLP) classification head to produce a toxicity score between 0 and 1. The model is fine-tuned on the augmented RTP dataset using an 80/20 train–test split with the Adam optimizer, a learning rate of 2e-5, batch size of 16. This process adapts RoBERTa's weights to the toxicity evaluation task while retaining its pre-trained contextual understanding.

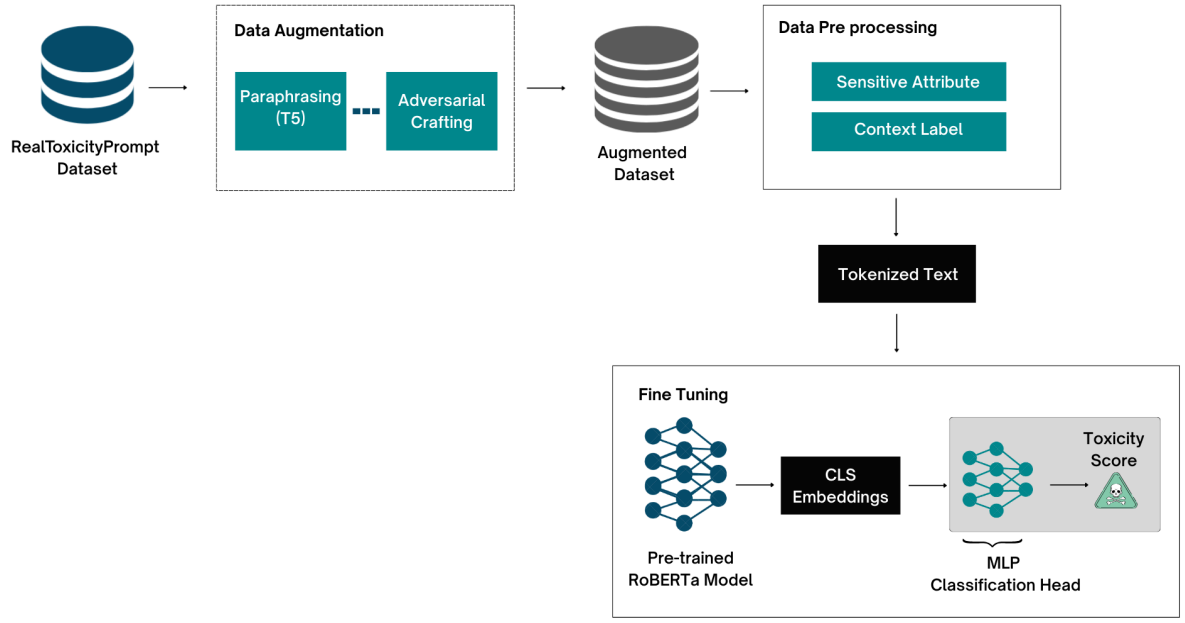

Fig 1: CAFE Architecture

6.3 Multi-Objective Loss Design

The novelty of CAFE lies in its multi-objective loss function, which simultaneously optimizes for toxicity accuracy, fairness, and context-awareness - in contrast to the single-objective optimization of Perspective API. The total loss is formulated as a weighted sum of three components (Equation 1):

- **Toxicity Loss :** Ensures accurate predictions by minimizing the mean squared error (MSE) between CAFE's predicted scores and the RTP's Perspective API scores.

$$\text{Toxicity Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(Predicted_i - True_i\right)^2 \quad \text{.....(1)}$$

where N is the batch size, $Predicted_i$ is CAFE's score, and $True_i$ is the Perspective API's toxicity score. The fairness loss promotes unbiased predictions across demographic groups.

- **Fairness Loss:** Encourages unbiased predictions by penalizing disparities in average toxicity scores between sensitive and non-sensitive groups (e.g., texts marked with identity-related flags).

$$\text{Fairness Loss} = \left| \frac{1}{N_0}\sum_{i \epsilon Group_0} Predicted_i - \frac{1}{N_1}\sum_{i \epsilon Group_1} Predicted_i \right| \quad \text{.....(2)}$$

where N_0 and N_1 are the number of texts in each group. This loss penalizes CAFE if it over-predicts toxicity for sensitive groups, addressing Perspective's bias issues. The context loss ensures embeddings capture non-literal meanings.

- **Context Loss :** Promotes context awareness by clustering embeddings of non-literal texts (e.g., sarcastic, playful, or flirtatious) through cosine similarity with a reference embedding of context-labeled examples.

$$\text{Context Loss} = 1 - \frac{1}{N}\sum_{i=1}^{N} cos(Embedding_i, Reference\ Embedding) \quad \text{.....(3)}$$

where cos is the cosine similarity, and the reference embedding is the mean of embeddings for context label = 1 (context label is 0 for literal and 1 for sarcastic/non-literal)

The final objective is:

$$L_{total} = \alpha . L_{toxicity} + \beta . L_{fairness} + \gamma . L_{context} \quad \text{.....(4)}$$

where α, β, and γ are tunable weights calibrated through experimentation to balance accuracy, fairness, and context sensitivity.

A critical limitation of this setup is that Perspective API scores are biased, yet they serve as the primary supervisory signal in RTP. In CAFE, we address this by treating Perspective scores only as proxy labels, while explicitly counterbalancing their flaws through fairness and context-aware loss terms. This strategy reduces the risk of overfitting to Perspective's inherent biases and moves the evaluator closer to an equitable scoring system.

6.4 Implementation Feasibility

Training over the full prompt–continuation pairs in RTP is computationally expensive. To ensure feasibility, we train on a subset of 10,000 prompts, sampled to preserve diversity in toxicity labels and sensitive attributes. This balances computational tractability with experimental validity, allowing for reproducibility on accessible hardware (e.g., Google Colab GPUs).

7. Evaluation

7.1 CAFE Expected Outcomes

To rigorously assess the effectiveness of CAFE, we employ several standard classification metrics and compare its performance with the Perspective API, using Jigsaw as the benchmark dataset. The table below summarizes the key evaluation metrics and the expected performance of CAFE.

Metric	*Description*	*CAFE Expected Outcome*
F1 Score	The harmonic mean of precision and recall, measuring the balance between correctly identified toxic texts and minimization of false positives.	CAFE is expected to achieve a notably higher F1 score, reflecting better accuracy in toxicity classification.
Fairness Gap	Quantifies demographic bias by computing the absolute difference in average predicted toxicity scores between sensitive (e.g., identity-flagged) and non-sensitive groups.	Significant reduction in fairness gap, indicating improved equity in scoring across subgroups.
Expected Maximum Toxicity (EMT)	For each prompt, the maximum toxicity score across 25 continuations is taken; the average across prompts reflects worst-case degeneration risk.	CAFE is expected to reduce EMT substantially, indicating improved robustness against toxic generations.
Toxicity Probability	Measures the fraction of continuations per prompt that exceed a toxicity threshold (e.g., ≥0.5).	CAFE is expected to reduce the toxicity probability, indicating better mitigation of harmful content generation.

Table 1: Evaluation metrics and the expected performance of CAFE

7.2 Validation Techniques

To ensure robust and reproducible evaluation, we adopt multiple validation strategies:

- **Benchmark Datasets**:CAFE is evaluated using the "Jigsaw Unintended Bias in Toxicity Classification" dataset to avoid bias and ensure real-world data-based comparisons, assessing model performance and fairness.
- **Train/Test Split:** An 80/20 split on the benchmark dataset ensures sufficient training data while preserving test diversity.
- **Cross-Validation:** K-fold cross-validation is used to evaluate variance in performance and mitigate overfitting.
- **Ablation Studies:** To isolate the contribution of each component, ablations will remove (i) fairness loss, (ii) context loss, and (iii) augmentation. This validates the necessity of CAFE's design choices.

8. Research Timeline

The project timeline as depicted in Fig 2 for the CAFE framework spans from August 14 to October 10, 2025, divided into five phases: planning, experiment setup, evaluation, analysis & writing and finalization, ensuring a structured approach.

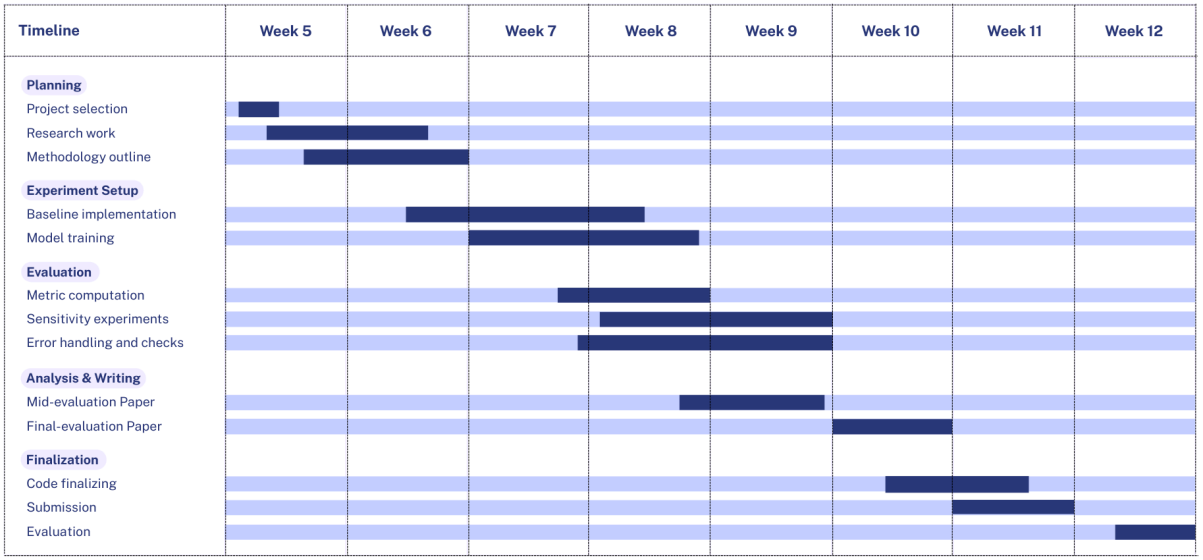

Fig 2: Research Timeline

9. References

[1] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *arXiv [cs.CL]*. http://arxiv.org/abs/2009.11462

[2] *Perspective API*. (n.d.). Perspectiveapi.com. Retrieved August 23, 2025, from https://perspectiveapi.com/

[3] *TensorFlow datasets*. (n.d.). TensorFlow. Retrieved August 23, 2025, from https://www.tensorflow.org/datasets/catalog/civil_comments

[4] Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

[5] Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *arXiv [cs.LG]*. http://arxiv.org/abs/1903.04561

[6] Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics.

[7] Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., & Giordano, S. (2023). Toxic bias: Perspective API misreads German as more toxic. In *arXiv [cs.SI]*. http://arxiv.org/abs/2312.12651

[8] Pozzobon, L., Ermis, B., Lewis, P., & Hooker, S. (2023). On the challenges of using black-box APIs for toxicity evaluation in research. In *arXiv [cs.CL]*. http://arxiv.org/abs/2304.12397

[9] Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API built for detecting toxic comments. In *arXiv [cs.LG]*. https://labs.ece.uw.edu/nsl/papers/view.pdf

[10] Van Hee, C., Lefever, E., & Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 39–50). Association for Computational Linguistics.

[11] *Proceedings of the second workshop on figurative language processing*. (2020). Association for Computational Linguistics.

[12] Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *arXiv [cs.CL]*. http://arxiv.org/abs/2203.09509

[13] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). HateXplain: A benchmark dataset for explainable hate speech detection. In *arXiv [cs.CL]*. http://arxiv.org/abs/2012.10289

[14] Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., & Rajani, N. F. (2021). GeDi: Generative discriminator guided sequence generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*.

[15] Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for Human-AI conversations. In *arXiv [cs.CL]*. http://arxiv.org/abs/2312.06674

[16] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *arXiv [cs.LG]*. http://arxiv.org/abs/1610.02413

[17] Vaidya, A., Mai, F., & Ning, Y. (2020). Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 683–693. https://doi.org/10.1609/icwsm.v14i1.7334