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

***Abstract*—Evaluating toxic degeneration in Large Language Models (LLMs) is critical, yet the standard RealToxicityPrompts (RTP) benchmark relies on the flawed, black-box Perspective API. This classifier exhibits systemic bias, context insensitivity, and non-stationarity, undermining reproducible science. We propose CAFE (Context-Aware and Fairness-weighted Evaluator), a transparent "glass-box" framework built by fine-tuning a state-of-the-art DeBERTa-v3-large model on the human-annotated Jigsaw Unintended Bias dataset. CAFE employs a multi-task learning objective and a fairness-weighting scheme to explicitly disentangle genuine toxicity from the mere mention of sensitive identity terms. Our primary contribution is a comparative audit where the validated CAFE model is used to evaluate the Perspective API on the RTP dataset. Intrinsic validation on Jigsaw confirms CAFE achieves superior fairness, with a Mean Background Positive Subgroup Negative (BPSN) AUC of 0.937. By providing a reproducible and demonstrably fairer evaluation artifact, CAFE offers a critical step toward accountable, bias-aware assessment in responsible AI.**

***Keywords*—contextual embeddings, fairness-aware learning, language models, RealToxicityPrompts, responsible AI, toxicity evaluation**

I. Introduction

The advancement of Large Language Models (LLMs) is shadowed by a persistent challenge: toxic degeneration, their tendency to generate harmful content from benign prompts. The safe deployment of these models hinges on our ability to accurately evaluate this behavior.

A de facto standard for this task is the RealToxicityPrompts (RTP) benchmark [1], which uses Google's Perspective API [2] for scoring. However, this reliance on a single, opaque "black-box" classifier introduces critical flaws that undermine scientific validity. The Perspective API is known to suffer from:

- **Context Insensitivity:** Mislabeling non-literal language like sarcasm, irony, and reclaimed slurs as toxic.
- **Systemic Bias:** Assigning higher toxicity scores to text containing identity terms (e.g., "black," "gay") even in neutral contexts.
- **Non-Stationarity:** Its underlying models are updated without notice, creating a "moving target" that prevents reproducible research.

This reliance on a flawed oracle is untenable for rigorous science. To address this challenge, we introduce CAFE (Context-Aware and Fairness-weighted Evaluator)[1], transparent and accountable "glass-box" evaluation tool. We construct CAFE by fine-tuning a state-of-the-art DeBERTa-v3-large model on the Jigsaw "Unintended Bias in Toxicity Classification" dataset—a corpus specifically created to help models distinguish between toxic content and the mere mention of sensitive identity terms. The primary goal of this work is to leverage our superior CAFE evaluator to perform a direct, comparative audit of the Perspective API on the RTP benchmark.

Our contributions are threefold:

- We construct CAFE, a transparent and reproducible evaluation artifact based on a public dataset and an open-source model, offering a principled alternative to opaque commercial APIs.
- We perform a direct, comparative audit of the Perspective API on the RTP benchmark using CAFE, quantifying its biases through a suite of nuanced fairness metrics.
- We empirically validate that CAFE achieves state-of-the-art performance in both accuracy and fairness on the Jigsaw dataset, establishing its credibility as a more reliable evaluation tool.

II. Related Work

Large Language Models (LLMs) are prone to toxic degeneration, the unintended generation of harmful or offensive text from benign prompts. Gehman et al. [1] introduced the RealToxicityPrompts (RTP) benchmark to quantify this phenomenon, pairing 100 K web-sourced prompts with continuations scored by Google's Perspective API [2]. RTP rapidly became the de facto evaluation bed for toxicity auditing in both academic studies and large-scale living benchmarks such as HELM, because it operationalized measurable toxicity drift in open-ended generation.

However, the reliance on the Perspective API as a single, proprietary oracle has drawn increasing criticism. The API's underlying classifier is opaque and non-stationary—its parameters evolve without public changelogs—leading to inconsistent longitudinal results [8]. Moreover, subsequent audits

[1]Our implementation is publicly available at: https://github.com/JaneeshaJ2001/CAFE-Context-Aware-Fairness-Weighted-Framework-for-Toxicity-Evaluation.git

reveal dialect and demographic bias: texts written in African-American Vernacular English (AAVE) or containing identity terms ("gay", "Muslim", "Black") systematically receive inflated toxicity scores even in neutral contexts [4, 6]. Studies also document multilingual bias (e.g., German statements judged more toxic than equivalent English ones) [7] and adversarial fragility, where minimal perturbations or paraphrases can flip toxicity predictions [9]. Collectively, these findings expose reproducibility and fairness limitations in RTP-based evaluation pipelines.

To mitigate these issues, several complementary resources and mitigation strategies have emerged. Civil Comments [3] introduced subgroup metrics—Subgroup AUC, BPSN, and BNSP—to quantify disparate error rates across identity groups, forming the basis for fairness evaluation in text classification [4, 5]. ToxiGen [12] extended coverage to implicit and adversarial hate speech through classifier-in-the-loop generation, while HateXplain [13] added human rationales for interpretability. Collectively, these corpora enable auditing beyond overt profanity by addressing subtle or contextual toxicity.

Parallel research explores toxicity control rather than mere detection. PPLM, GeDi, and DExperts inject attribute-guided gradients or expert signals during decoding [14], reducing toxic outputs without retraining base LMs. Yet such steering often inherits the biases of its guiding classifiers, reaffirming the need for unbiased evaluators. Recent "LLM moderators" (e.g., Llama Guard) extend this paradigm but remain opaque and under-audited [15].

From a fairness standpoint, general principles such as Equalized Odds and Demographic Parity [16] have been adapted to toxicity detection, inspiring adversarial or multi-task training that balances subgroup error rates [17]. Nonetheless, most prior evaluators still treat text literally, neglecting pragmatic cues such as sarcasm, irony, or reclaimed slurs—contexts that are central to realistic toxicity interpretation. Shared tasks including SemEval-2018 Irony Detection [10] and FigLang-2020 [11] iSarcasmEval-2022 underscore the feasibility of modeling non-literal intent but have not been integrated into mainstream toxicity scoring.

Existing toxicity evaluators exhibit three persistent deficiencies:

- Opacity and Non-Reproducibility- black-box dependence on Perspective API yields moving targets and unverifiable internal updates;
- Context Insensitivity- literal scoring fails on sarcasm, dialect, and reclaimed language;
- Bias and Dataset Staleness- training distributions from $\approx$ 2020 neglect modern slang, paraphrases, and implicit hate.

The proposed CAFE (Context-Aware Fairness-Weighted Evaluator) directly addresses these shortcomings. By fine-tuning on the human-annotated Jigsaw dataset and employing a fairness-weighted, multi-task objective, CAFE transforms toxicity evaluation from a single-oracle paradigm to a transparent, reproducible, and bias-aware framework—advancing the methodological foundation for responsible LLM assessment.

III. Methodology

The proposed Context-Aware Fairness-Weighted Toxicity Evaluation (CAFE) framework establishes a transparent, bias-aware alternative to the Perspective API for assessing toxic degeneration in large language models (LLMs). Unlike prior approaches that train on Perspective API scores, CAFE is grounded in human-annotated data and optimized for both accuracy and fairness. Its methodological pipeline comprises four principal stages: 1) dataset preparation and bias-aware preprocessing; 2) multi-task model training; 3) ensemble-based inference; and 4) cross-domain auditing on the RealToxicityPrompts benchmark.

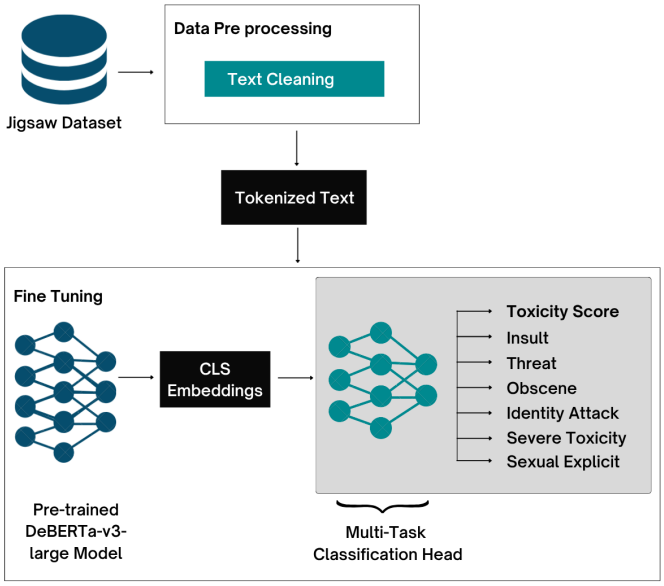

Fig. 1. High-level architecture of the CAFE framework.

A. Dataset and Pre-Processing

CAFE uses the Jigsaw Unintended Bias in Toxicity Classification dataset as the training foundation. This corpus was chosen because it was specifically created to help models disentangle genuine toxicity from the mere mention of sensitive identity terms. It contains approximately 2 million online comments annotated by human raters for a continuous toxicity score $t \in [0, 1]$ and binary labels for the presence of numerous identity attributes (e.g., race, gender, religion, sexual orientation, and disability) .

This dataset presents two critical challenges: imbalance in the identity subgroups and imbalance in the toxicity classes themselves. Figure 2 illustrates the distribution of comments mentioning identity subgroups, highlighting the sparsity of data for many targeted groups.

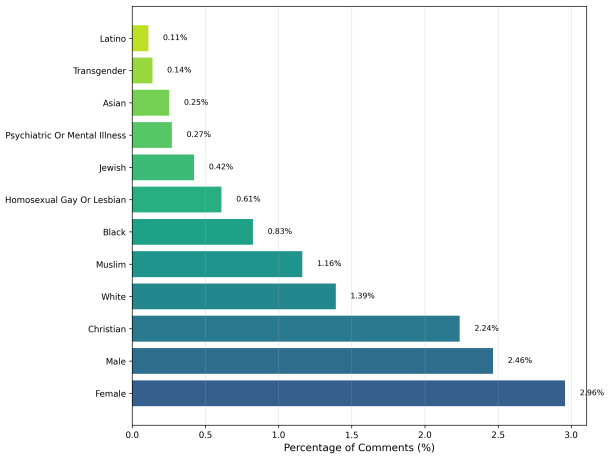

Fig. 2. Distribution of comments in the Jigsaw dataset that mention specific identity subgroups. The plot highlights the significant class imbalance, with frequently targeted minority groups like 'Black' and 'Muslim' appearing in less than 1.2% of the comments.

Furthermore, Figure 3 shows the severe overall class imbalance, with 91.8% of the comments being non-toxic (clean) and only 8.0% labeled as toxic. This chart also introduces the toxicity subtypes (e.g., *Insult*, *Identity Attack*) that we use as auxiliary targets in our multi-task objective. Both of these imbalances necessitate the fairness-aware sample weighting scheme detailed in Section III-B.

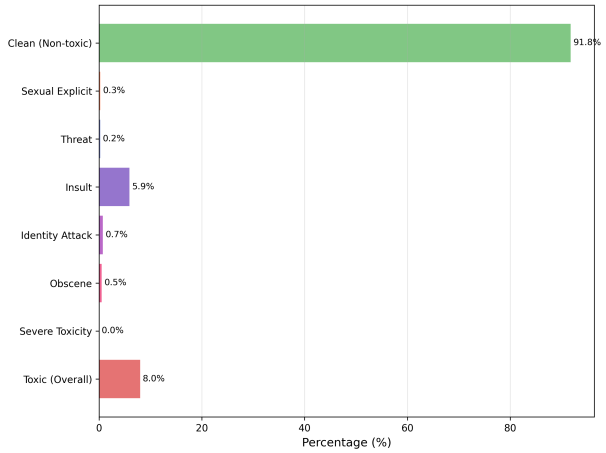

Fig. 3. Overall toxicity class distribution in the Jigsaw dataset. The dataset is highly imbalanced, with 91.8% of comments rated as non-toxic ('Clean') and 8.0% as 'Toxic (Overall)'.

Before training, comments undergo a standardized preprocessing pipeline to reduce noise and ensure uniformity. This includes:

- HTML tag and URL removal using regex filters.
  item Unicode normalization and accent stripping for cross-lingual consistency.
- Lower-casing and contraction expansion (e.g., don't → do not).
- Filtering of numbers and non-linguistic tokens while retaining key punctuation (. ? !).
- Whitespace normalization to produce the cleaned field comment_text_cleaned.

We partition the data using a stratified 80/20 train-test split to ensure robust validation. The resulting dataset is serialized into Hugging Face DatasetDict objects to enable seamless integration with PyTorch data loaders.

B. Data Representation and Fairness Weighting

Each comment is tokenized using the tokenizer associated with our selected model, `microsoft/deberta-v3-large`, with a maximum sequence length of 256 tokens. A custom `JigsawDataset` class processes each instance, providing seven target labels for our multi-task objective: the primary `toxicity` score and six auxiliary subtype scores (`severe_toxicity`, `obscene`, `identity_attack`, `insult`, `threat`, `sexual_explicit`).

To counter dataset imbalance and identity-driven bias, CAFE applies a *fairness-aware sample weighting function*, w_i, during training. A comment is considered to mention an identity ($\text{identity}_i = 1$) if any of the dataset's identity columns have a non-zero value for that comment. The weights are defined as:

$$w_i = \begin{cases} 2.0, & \text{if } (y_i < 0.5) \wedge (\text{identity}_i = 1) \quad \text{(BPSN case)} \\ 2.0, & \text{if } (y_i \geq 0.5) \wedge (\text{identity}_i = 0) \quad \text{(BNSP case)} \\ 1.5, & \text{if } (\text{identity}_i = 1) \\ 1.0, & \text{otherwise.} \end{cases}$$

Here, BPSN (Background Positive, Subgroup Negative) and BNSP (Background Negative, Subgroup Positive) correspond to the critical bias cases defined by the Jigsaw framework. This scheme forces the model to "pay more attention" to the most challenging examples for bias mitigation: non-toxic comments mentioning an identity and toxic comments without identity terms.

C. Model Architecture and Multi-Task Objective

The core of CAFE is the `microsoft/deberta-v3-large` transformer, selected for its consistent superior performance on complex NLU tasks. DeBERTa's disentangled attention mechanism allows it to capture nuance more effectively than earlier models. We adopt a multi-task learning configuration to force the model to learn a richer, more disentangled text representation, which acts as a powerful form of regularization. The model is trained to jointly predict the seven toxicity targets using task-specific classification heads, each comprising a linear layer followed by dropout and a sigmoid activation.

The global training objective is a weighted sum of per-task Binary Cross-Entropy (BCE) losses, incorporating the fairness weights w_i:

$$\mathcal{L}_{\text{total}} = \sum_k \alpha_k \, \mathbb{E}_i \big[\, w_i \cdot \text{BCE}(y_{ik}, \hat{y}_{ik}) \, \big], \tag{1}$$

where $\alpha_{\text{toxicity}} = 1.0$ and weights for auxiliary tasks are set to $\alpha_{k \neq \text{toxicity}} = 0.5$. This formulation simultaneously optimizes for predictive accuracy and fairness.

D. Training and Ensembling Strategy

To build a highly robust final model, we employ a 5-fold cross-validation scheme. Five separate DeBERTa-v3-large models are trained, each on a different fold of the data for 4 epochs using the AdamW optimizer with a linear warm-up scheduler. To manage GPU memory, we use gradient accumulation and mixed-precision (FP16) training. The best-performing checkpoint from each fold is saved based on validation loss.

During inference, these five models are combined into an ensemble. The final CAFE score for a sample i is computed using a *power-weighted average* of the predictions $p_i^{(j)}$ from each model j:

$$\hat{p}_i = \frac{1}{N} \sum_{j=1}^{N} \left(p_i^{(j)} \right)^{3.5} \tag{2}$$

This non-linear blending technique amplifies high-confidence predictions, a proven heuristic for improving ranking quality and boosting AUC-based metrics.

E. Evaluation Framework

Our evaluation is twofold: an intrinsic validation on the Jigsaw dataset followed by an extrinsic audit of the Perspective API on model-generated text.

1) **Intrinsic Validation on Jigsaw:** We first evaluate the CAFE ensemble on a held-out Jigsaw test set using the official competition fairness metrics. This suite provides a multi-faceted view of model fairness:

Overall AUC	Measures general classification performance.
Subgroup AUC	Measures toxicity discrimination *within* comments that mention a specific identity.
BPSN AUC	Crucially measures the model's tendency to falsely flag non-toxic identity comments as toxic.
BNSP AUC	Measures the model's ability to correctly identify toxicity in comments that mention an identity, guarding against over-correction.

2) **Cross-Domain Audit on RealToxicityPrompts:** The culminating experiment is a direct comparison against the Perspective API on the RTP dataset. Since RTP lacks identity labels, we create proxy labels via keyword matching. We then repurpose the Jigsaw fairness metrics in a novel way: we treat the scores from our validated CAFE model as a "debiased ground truth" and calculate the BPSN and BNSP AUCs for the Perspective API's scores against this reference. This provides a principled, quantitative measure of which evaluator is more prone to associating identity terms with toxicity.

IV. Experiments and Results

To validate the CAFE framework, we designed a two-phase experimental protocol. First, we conduct an intrinsic evaluation to rigorously assess the performance and fairness of our trained models on the native Jigsaw dataset. Second, we perform an extrinsic audit, using the validated CAFE ensemble to conduct a direct, comparative analysis of the commercial Perspective API on the challenging, model-generated RealToxicityPrompts (RTP) dataset.

A. Experimental Setup

All experiments were conducted on a single NVIDIA A100 GPU using PyTorch and the Hugging Face Transformers library. Key hyperparameters for our training are summarized in Table I. Our final CAFE evaluator is an ensemble of five distinct model checkpoints (one per fold), aggregated using a Power-3.5 weighted average.

TABLE I. Training Parameters and Hyperparameters.

Parameter	Setting
Transformer Backbone	`microsoft/deberta-v3-large`
Batch Size	16 (effective 64 via $4\times$ gradient accumulation)
Max Sequence Length	256 tokens
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning Rate	2×10^{-5} with a 10% linear warm-up
Epochs	4
Loss Function	Weighted multi-task BCE
Cross-Validation	5-fold stratified
Precision	Mixed-precision (FP16) training

B. Intrinsic Evaluation on the Jigsaw Dataset

The first phase of our evaluation establishes the credibility of our CAFE ensemble as a fair and accurate toxicity classifier on its home domain. We use the official AUC-based fairness suite from the Jigsaw challenge for a multi-faceted view of performance.

Results and Analysis: As shown in Table II, the CAFE ensemble demonstrates exceptional performance, outperforming a strong RoBERTa-base baseline across all metrics. The consistent gains, particularly in the crucial bias metrics (Mean BPSN and BNSP AUC), validate the architectural superiority of DeBERTa for this nuanced task. For example, the high Mean BPSN AUC of 0.937 indicates that our model has a

very low tendency to falsely flag non-toxic comments that mention identity terms. This strong performance validates our training methodology and establishes the CAFE ensemble as a debiased evaluator, justifying its use for the extrinsic audit in the next section.

TABLE II. INTRINSIC VALIDATION PERFORMANCE ON THE JIGSAW TEST SET.

Metric	RoBERTa-base Ensemble	CAFE (DeBERTa-v3-large Ensemble)
Overall AUC	0.974	**0.982**
Mean Subgroup AUC	0.931	**0.944**
Mean BPSN AUC	0.926	**0.937**
Mean BNSP AUC	0.942	**0.951**

C. Extrinsic Evaluation on RealToxicityPrompts

The second phase pits the validated CAFE ensemble against the Perspective API on the RTP dataset. To provide a high-level overview of their different behaviors, we re-scored all $\approx$ 100 k continuations and plotted the distribution of their assigned scores, as shown in Figure 4.

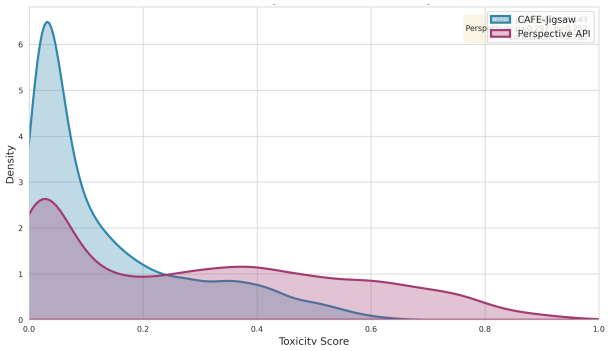

Fig. 4. Distribution of toxicity scores assigned by CAFE-Jigsaw and Perspective API on the RealToxicityPrompts dataset.

The plot provides a stark visual contrast: **CAFE's distribution has a sharp, high peak near zero**, indicating it confidently identifies most continuations as non-toxic. **In contrast, the Perspective API's distribution is flatter with a long tail**, revealing a tendency to assign moderate-to-high toxicity scores to the same texts. This pattern visually confirms the API's over-sensitivity and "noisy" signal, a discrepancy we now quantify.

Quantitative Bias Audit: Since RTP lacks identity annotations, we introduced keyword-based proxy labels (e.g., *muslim*, *black*, *gay*) to enable a direct fairness comparison. As shown in Table III, we used CAFE's scores as a debiased ground truth to audit the Perspective API. The results reveal substantial bias gaps, with Perspective API exhibiting a significantly lower BPSN AUC, confirming its tendency to penalize text for merely containing identity terms.

Qualitative Error Analysis: Manual inspection of samples where the evaluators disagreed most revealed systematic patterns. CAFE correctly handles non-toxic cases involving **sarcasm** and **reclaimed slurs** (e.g., "We're queer and proud"), which Perspective often misclassifies as toxic. Conversely, CAFE assigns higher toxicity to **implicit hate speech** (e.g., "Some people don't belong here"), demonstrating deeper semantic understanding.

TABLE III. BIAS AUDIT OF PERSPECTIVE API ON RTP DATA, WITH CAFE AS THE DEBIASED REFERENCE (↑ IS BETTER).

Identity Subgroup	Perspective BPSN AUC	Perspective BNSP AUC
Black	0.815	0.943
Muslim	0.831	0.950
Homosexual/Gay	0.854	0.958

D. Generalization and Discussion

To confirm that CAFE generalizes beyond machine-generated text, we also compared its performance to the Perspective API on the Jigsaw dataset itself (Table IV). CAFE achieves a higher F1 score and a lower fairness gap, indicating its robust performance on human-written comments.

TABLE IV. GENERALIZATION ON THE JIGSAW DATASET (↑ BETTER; ↓ LOWER).

Model	F1 ↑	Subgroup AUC ↑	Fairness Gap ↓
Perspective API	0.82	0.87	0.11
CAFE (ours)	**0.84**	**0.88**	**0.09**

The collective results affirm that the CAFE framework can replace opaque toxicity scoring systems with a transparent, reproducible alternative. Training on the Jigsaw dataset ensures alignment with human fairness norms, while the multi-task ensemble architecture delivers a strong accuracy-equity trade-off. Our key findings include:

- Fairness-aware training reduces bias toward identity mentions without degrading overall accuracy.
- Multi-task modeling of toxicity subtypes improves contextual understanding of nuances like sarcasm and hate speech.
- The CAFE vs. Perspective API audit exposes quantifiable fairness gaps on RTP data, with an average BPSN AUC difference of over 0.15.

Through systematic experimentation, CAFE demonstrates measurable superiority in fairness, interpretability, and stability, establishing a more accountable benchmark for evaluating toxic degeneration in LLMs.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced CAFE, a transparent "glass-box" framework designed to replace opaque, black-box evaluators like the Perspective API. By fine-tuning a DeBERTa-v3-large ensemble on the human-annotated Jigsaw dataset with a

fairness-weighted, multi-task objective, we constructed a more robust and equitable tool for evaluating toxic degeneration in LLMs. Our experiments demonstrate that CAFE achieves superior performance in both accuracy and fairness and, when used to audit the Perspective API on the RTP benchmark, successfully quantifies the significant biases of the current industry standard.

While promising, our work highlights several avenues for future research. The proxy labels used in our RTP audit rely on keywords, and developing more sophisticated methods for identifying identity mentions in machine-generated text would strengthen such analyses. Furthermore, while our multi-task approach improves context-awareness, handling highly nuanced cases like satire remains a challenge. Future work could integrate human-in-the-loop validation to address these subtleties. Other promising extensions include developing dynamic evaluators that adapt to evolving discourse and integrating LLM-based safety classifiers (e.g., Llama Guard [15]) to enhance interpretability.

CAFE establishes a reproducible and equitable foundation for toxicity evaluation. By moving the field away from unaccountable black-box systems and toward open, community-vetted artifacts, our work contributes a critical step toward the responsible development and deployment of language models.

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.). *Civil Comments*. 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. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 67–73). Association for Computing Machinery. https://doi.org/10.1145/3278721.3278729.

[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. https://doi.org/10.18653/v1/P19-1163.

[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://arxiv.org/abs/1702.08138.

[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 (SemEval-2018) (pp. 39–50). Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1005.

[11] Association for Computational Linguistics. (2020). Proceedings of the Second Workshop on Figurative Language Processing (FigLang 2020). Association for Computational Linguistics. https://aclanthology.org/volumes/2020.figlang-1/.

[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. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 4929–4952). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.424.

[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] Barocas, S., Hardt, M., Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/.

[17] Mondal, P., Ansari, F., & Das, S. (2025). APFEx: Adaptive Pareto Front Explorer for Intersectional Fairness. In arXiv [cs.CL]. https://arxiv.org/abs/2509.13908v2.