Enhancing Fairness in Open-Source AI-Generated Text Detectors for Non-Native English Writers

Astalaxmi Dhanaseelan
University of Pittsburgh
ASD153@pitt.edu

1 Problem Statement

Automated detection of AI-generated text has become increasingly important in educational, professional, and content moderation contexts. While prior research has demonstrated that domain-specific detectors,such as models trained exclusively to distinguish human-written and AI-generated GRE essays,can achieve high accuracy with low bias (Zhong et al., 2024), these approaches often rely on large proprietary datasets and substantial computational resources, limiting their accessibility.

In contrast, widely available open-source, general-purpose detectors (e.g., RoBERTa classifiers, DetectGPT) exhibit significant bias against non-native English writers. Studies have shown that essays written by non-native speakers, including TOEFL submissions, are misclassified as AI-generated at rates approaching 60% (Liang et al., 2023). This bias arises from the overrepresentation of native English text in training datasets and the sensitivity of detectors to linguistic simplicity or reduced stylistic diversity, common in non-native writing.

Such misclassifications pose fairness concerns in educational assessment and hinder the broader adoption of AI detection tools among diverse populations. This research addresses these challenges by focusing on improving the fairness and accessibility of general-purpose GPT detectors, rather than developing new domain-specific models. By employing strategies such as accessible domain-adaptive fine-tuning, threshold calibration and ensemble methods, the project aims to enhance detection accuracy while reducing bias across diverse English writing styles, creating AI detection tools that are both practical and equitable.

2 Literature Review

While AI detection tools are intended for fairness and disclosing LLMs' use, their evaluation is based on pattern recognition only (Pratama et al., 2025). Non-native speakers' writing tends to exhibit lower linguistic variability, which makes it susceptible to being unfairly flagged as generated text. (Liang et al., 2023) Recent work has highlighted a significant bias in GPT detectors against non-native English writers, with TOEFL essays and other non-native writings being misclassified as AI-generated at rates approaching 60%. (Liang et al., 2023)

Among debiasing strategies, dataset balancing has been shown to be effective for demographic detection-based biases (Han et al., 2022). Some balancing strategies include explicit balancing through sample adaptation (Zhao et al., 2018; Wang et al., 2019). Nevertheless, these are not effective enough, as linguistic features reveal demographic information (Han et al., 2022).

Recent work demonstrates that class imbalance remains a fundamental challenge in machine learning, with traditional solutions like synthetic data augmentation often introducing new problems. Group aware threshold calibration setting different decision thresholds for different demographic groups has been shown to provide superior robustness compared to synthetic data generation methods. Specifically, group-specific thresholds achieve 1.5–4% higher balanced accuracy than SMOTE and CT-GAN augmented models while also improving worst-group balanced accuracy. (Gittlin, 2025).

Ensemble methods also offer a promising approach to mitigating detector bias while maintaining robustness. Prior work shows that aggregating multiple classifiers improves group fairness leveraging bias-variance. (Gupta et al., 2022)

This focus on model-level strategies comple-

ments analyses of the text itself, where linguistic variation, examined through vocabulary bands, syntactic complexity, sentence and word length, and other features, is widely used to assess proficiency in Automated Essay Scoring (Vajjala, 2017; Dong and Zhang, 2016; Guo et al., 2013). Therefore, since complexity and perplexity are considered the root of non-native writer bias, we also aim to analyze differences between proficiency levels and bias outcomes.

3 Approach and Experimental Setup

3.1 Dataset Preparation

To evaluate and improve the fairness of AI-generated text detectors, a diverse dataset representing native, non-native, and AI-generated essays will be constructed.

3.2 Native English Essays

Collected from publicly available datasets, including:

- **Louvain Corpus of Native English Essays (LOCNESS):** Contains high-quality academic essays written by native English speakers.

3.3 Non-Native English Essays

Sourced from international and non-native writing datasets, including:

- **ETS Corpus of Non-Native Written English (TOEFL Dataset):** Standardized test essays from non-native English speakers.
- **ICNALE: The International Corpus Network of Asian Learners of English:** Essays by Asian learners with varied proficiency levels.

3.4 AI-Generated Essays

Produced using GPT-3.5 and GPT-4 with standard prompts covering similar topics as the human-written essays. These essays simulate machine-generated content for detector testing.

3.5 Mitigation Strategies

To improve the fairness and robustness of AI-generated text detectors, the following mitigation strategies will be implemented:

3.5.1 Accessible Domain-Adaptive Fine-Tuning

Purpose: Retrain existing general-purpose detectors on publicly collected datasets reflecting different compositions of native, non-native, and AI-generated essays.

Configurations:

- Equal distribution: 33% native, 33% non-native, 33% AI
- Non-native and AI: 50% non-native, 50% AI
- Native and AI: 50% native, 50% AI
- Realistic global ratio: 24% native, 76% non-native (AI-generated essays included proportionally to reflect this ratio)

Note: Unlike fully domain-specific models trained from scratch on proprietary datasets (Jiang et al., 2024), this approach is computationally feasible, accessible, and deployable for broad research and educational use.

3.5.2 Ensemble Methods

Multiple ensemble approaches will be applied to enhance robustness and reduce biases from individual models:

- **General Ensemble:** Combines all detectors, including open-source and fine-tuned models, to leverage complementary strengths.
- **Open-Source Ensemble:** Groups only unmodified detectors to assess the performance of off-the-shelf models collectively.
- **Domain-Adaptive Fine-Tuning Ensemble:** Combines models fine-tuned on different dataset compositions, as described above.
- **Individual Fine-Tuned Models:** Each configuration will also be evaluated separately to understand the contribution of different training compositions.

3.5.3 Threshold Calibration

Purpose: Reduce disparities in misclassification rates between native and non-native English essays by adjusting decision thresholds for AI detection.

Method: Using validation datasets, the false positive rate (FPR) and false negative rate (FNR) are computed separately for native and non-native groups. Based on these error rates, the decision

threshold for classifying text as AI-generated is adjusted for non-native essays to narrow the FPR gap. This group-specific postprocessing decreases the threshold for non-native essays as needed, while native essays retain the standard threshold.

Benefit: Threshold calibration is lightweight, ethical, and accessible. It reduces unfair penalization of non-native essays while preserving overall detection accuracy. This method can be applied to both supervised classifiers (e.g., RoBERTa) and score-based detectors (e.g., DetectGPT), making it a versatile mitigation strategy for improving fairness across diverse writing styles.

3.6 Bias Measurement

Following detector evaluation, bias will be measured to quantify disparities in performance between native and non-native English essays. The following steps will be implemented:

- **Subgroup Analysis:** Examine detector performance across different levels of English proficiency for non native speakers to identify patterns of bias.
- **Impact of Mitigation Strategies:** Assess how threshold calibration, accessible domain-adaptive fine-tuning, and ensemble methods affect FPR and FNR for both native and non-native essays.
- **Correlation Analysis:** Use chi-square tests and ROC/AUC metrics to explore relationships between lexical diversity, syntactic complexity, and detector outcomes.

This analysis will provide a comprehensive understanding of the fairness and robustness of each mitigation strategy and inform best practices for deploying AI-generated text detectors in diverse settings.

3.7 Statistical Analysis

To rigorously evaluate detector performance and the effectiveness of mitigation strategies, statistical analyses will be conducted as follows:

- **ROC Curve:** The Receiver Operating Characteristic (ROC) curve illustrates how the *true positive rate* (correctly identified AI essays) versus the *false positive rate* (human essays misclassified as AI) changes as the detection threshold varies. This allows us to observe how bias and detection accuracy fluctuate across different thresholds.
- **AUC (Area Under the ROC Curve):** The AUC provides a single, threshold-independent measure of overall detector quality. It enables fair comparison between different mitigation methods.
- **Chi-Square Tests:** Perform chi-square tests to evaluate whether differences in false positive and false negative rates between native and non-native essays are statistically significant.
- **Correlation Analysis:** Examine correlations between lexical diversity, syntactic complexity, and detector outcomes to identify features that influence bias.
- **Comparative Analysis of Mitigation Approaches:** Analyze the impact of each mitigation strategy on detector fairness and accuracy.
- **False Negative Rate (FNR):** The proportion of AI-generated essays incorrectly classified as human. Monitoring FNR alongside FPR highlights the trade-off between over-flagging non-native essays and under-detecting AI text.
- **False Positive Rate (FPR):** The proportion of human essays incorrectly classified as AI-generated. High FPR disproportionately harms non-native writers, making it a critical fairness metric.

These statistical analyses will provide quantitative evidence of bias reduction, robustness, and fairness improvements across different essay types and mitigation strategies.

3.8 Expected Outcomes

Based on the proposed mitigation strategies, we anticipate the following outcomes:

- **Reduced Bias:** Balanced and domain-adaptive fine-tuning, along with threshold calibration, will significantly reduce false positives for non-native essays while maintaining detection accuracy for AI-generated text.
- **Improved Detection Accuracy:** Multi-detector ensembles combining open-source and fine-tuned models will provide robust detection across native and non-native essays.

- **Insight into Feature Influence:** Analysis of lexical diversity, syntactic complexity, and perplexity will reveal which textual features most strongly affect detector bias and performance.

- **Actionable Guidelines:** The project will provide practical, accessible recommendations for improving open-source AI text detectors, ensuring fairness for non-native writers without requiring proprietary datasets or large-scale compute resources.

These outcomes will demonstrate that accessible mitigation strategies can meaningfully reduce bias and improve the fairness and robustness of general-purpose GPT detectors.

Together, these strategies will create a more fair and effective AI detection framework.

References

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Hunter Gittlin. 2025. Beyond synthetic augmentation: Group-aware threshold calibration for robust balanced accuracy in imbalanced learning.

Ying Guo, Scott A. Crossley, and Danielle S. McNamara. 2013. Modeling text quality using cognitive and linguistic features. *Journal of Second Language Writing*, 22(4):383–400.

Neha Gupta, Jamie Smith, Ben Adlam, and Zelda Mariet. 2022. Ensembling over classifiers: a bias-variance perspective.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Jiang, Jiangang Hao, Michael Fauss, and Chen Li. 2024. Detecting chatgpt-generated essays in a large-scale writing assessment: Is there a bias against non-native english speakers? *Computers & Education*, 217:105070.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers.

Ahmad R. Pratama et al. 2025. The accuracy-bias trade-offs in ai text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science*, 11.

Sowmya Vajjala. 2017. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations.

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, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Yang Zhong, Jiangang Hao, Michael Fauss, Chen Li, and Yuan Wang. 2024. Evaluating ai-generated essays with gre analytical writing assessment.