

METHODOLOGICAL PAPER

# Ethical Guardrails and Algorithmic Bias Mitigation in Automated Formative Essay Feedback Systems

Katherine Adams, Nadia Zahra, Claire Dupont

1. Dept. of Language Arts & Ethics in AI, Universitas Tadulako; 2. Center for Equitable Assessment; 3. European EdTech Observatory.

Received: 12 August 2026 • Revised: 28 September 2026 • Accepted: 10 October 2026 • Published Online: 30 September 2026

## Abstract

*Auditing automated writing evaluation models for lexical bias against non-native English speakers. Implementing counterfactual data augmentation reduced false grammar penalties by 74% across multilingual cohorts.*

**Keywords:** Automated Essay Scoring, Algorithmic Bias, Formative Feedback, Ethical AI, Multilingual Assessment

## 1. Introduction

Automated essay grading and feedback platforms often penalize valid dialectical variations and idiomatic structures used by Second Language (L2) writers. Ensuring demographic parity and linguistic equity is essential prior to institutional deployment.

## 2. Method

An algorithmic audit evaluated three commercial automated essay scoring engines on an ethnically diverse corpus of 4,500 secondary school essays. Counterfactual demographic text augmentation and adversarial debiasing layers were subsequently integrated and tested.

## 3. Results and Discussion

Original baseline models exhibited a 22% score penalty against Indonesian and Southeast Asian ESL essay syntax. Implementing the debiasing layer eliminated false syntactic penalties by 74% while maintaining 0.92 correlation with human master raters.

## 4. Conclusion

Algorithmic equity auditing and counterfactual retraining must become mandatory compliance standards for all educational AI writing assessment systems.

## Declarations

**Funding:** This study was supported by institutional research development grants.

**Conflicts of Interest:** The authors declare that they have no competing financial or personal interests.

**Data Availability:** The dataset and analysis matrices associated with this study are available in the public data repository.

**Author Contributions:** All authors contributed equally to conceptualization, investigation, formal analysis, and manuscript preparation.

## References

Adams, K., & Khuzaimah, A. U. (2024). Mitigating algorithmic linguistic prejudice in automated formative essay feedback. *International Journal of Educational Technology in Higher Education*, 21(1), 54–72.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092.