ASEAN Journal of Open and Distance Learning (AJODL)
https://ajodl.oum.edu.my/
Vol. 18, No. 1, 2026, pp. 12 – 19

False Positives in AI Writing Detection: A Small-Scale Empirical Study Using Authentic Filipino Student Essays

Mhel Cedric D. Bendo

College of Business Administration, Polytechnic University of the Philippines, Cavite, Philippines.

Corresponding author: cedricbends@gmail.com

Paper Info:
Received: 13 Dec 2025 | Revised: 19 Feb 2026
| Accepted: 30 Mar 2026 | Available Online: 1 Apr 2026

DOI: https://doi.org/10.64233/VYVI9613

Citation:
Bendo, M. C. D. (2026). False positives in AI writing detection: A small-scale empirical study using authentic Filipino student essays. *ASEAN Journal of Open and Distance Learning, 18*(1), 12-19, https://doi.org/10.64233/VYVI9613

Abstract

This research note reports a small-scale exploratory study into how two AI writing detectors, ZeroGPT and Copyleaks, classify authentic student essays. A total of ten anonymised college-student essays were analysed to observe misclassification patterns, particularly false positives, where human-written content has been incorrectly flagged as AI-generated. The assessment was conducted focusing on essays produced by Filipino undergraduates whose nonnative English writing may have features that could lead to misclassification by the detectors. Results show that both detectors inconsistently classified the same set of essays: five essays were labelled as "AI-generated," while the other five were labelled as "Human," when in fact all the texts were authentically written by students. These results point to the potential misclassification risks when AI detection tools are used within educational contexts, where it is commonplace for teachers to make important decisions about academic integrity based on the outputs of detectors. The present study underlines the need to validate the outputs of AI detectors with human judgment and advises educators against the use of these tools as sole evidence of misconduct. Implications for ICT-supported assessment practices and policies of academic honesty are discussed, together with recommendations for a more responsible integration of AI detection tools in educational contexts. The results described above have implications for the practice of academic integrity in all countries, but also specifically for those countries with multilingual student bodies whose writing styles may differ from the writing style used to train the detectors.

Keywords: academic integrity, AI writing detection, algorithmic bias, false positives, Filipino students, ICT-supported assessment

ASEAN Journal of Open and Distance Learning
Vol. 18, No. 1,2026, pp. 12 – 19

1. Introduction

The rapid adoption of AI tools in education has changed both student writing practices and assessment methods. The use of AI systems such as ChatGPT has become increasingly common in recent years, and a significant number of learners rely on those tools to assist them in writing or editing school assignments (Kasneci et al., 2023). As a result of this trend, educators have begun using AI writing detection tools to identify whether a piece of writing has been written by a student (Cotton et al., 2024). Although such detectors are sometimes marketed as useful, concerns have been raised regarding their reliability, particularly when used in academic integrity decisions. Recent studies indicate that numerous detectors falsely label text written by humans as produced by AI (Hadra et al., 2026; Liang et al., 2023), and their vocabulary and sentence construction are more likely to be mistaken as less complex to algorithmic systems (Mitchell et al., 2023).

Filipino students may be particularly susceptible to false positives. While English is commonly used, it is heavily influenced by local languages so their writing may differ from the native-English datasets on which commercial detectors are trained, possibly leading to misclassification. Therefore, it is important to examine the reliability of AI writing detectors in classifying the writing of Filipino undergraduates and the frequency with which they misclassify human writing of AI-written texts. Thus, the given research represents a mini-empirical analysis of two popular tools, ZeroGPT and Copyleaks, in terms of their consistency, error behaviour, and the possible implications on the ICT-facilitated assessment practice in the multilingual learning environment.

The purpose of this study is to examine how accurately AI writing detection tools like ZeroGPT and Copyleaks correctly classify real student essays. More precisely, it probes how these detectors evaluate essays from Filipino undergraduate students and how often they may falsely flag legitimate human-written work as AI-generated. This should be documented in the context of the extent and pattern of such errors, especially among students in Southeast Asia with whom English is widely used but not the primary language.

The study is important because educators increasingly rely on AI detection outputs in academic integrity decisions. In providing preliminary empirical evidence of misclassification, this highlights the need for cautious use of AI detectors within the ICT-supported assessment practices. In open and distance learning (ODL) environments, where assessment is frequently conducted without continuous face-to-face supervision, educators often rely on technological mediation to support evaluation processes (Laurillard, 2012), making detector reliability a critical concern.

To address the objectives of the study, the following research questions guided the investigation:

1. How do ZeroGPT and Copyleaks classify ten authentic student essays written by Filipino undergraduates?
2. What proportion of these essays are incorrectly flagged as AI-generated by the detection systems?
3. What implications do these misclassification patterns have for ICT-supported assessment practices in multilingual educational environments?

2. Literature Review

AI writing detectors have become a trendy product that is sold to differentiate between human- and AI-written texts. Nevertheless, studies are also indicating that these systems are yet to be reliable and demonstrate different forms of algorithmic bias. Hadra et al. (2026) and Weber-Wulff et al. (2023) suggest that the accuracy of the detectors varies with the genre of

the text, the level of linguistic skills, and the style of writing, which restricts their use in the actual educational practice. Most of the detectors are based on statistical indicators like perplexity and stylometric consistency, yet these indicators are not consistent with true human authorship (Fraser et al., 2024).

There exists also a considerable amount of literature that proves that non-native writers of the English language are misclassified at unproportionately elevated levels. Liang et al. (2023) discovered that multilingual authors tend to be identified since their vocabulary scope, linguistic decisions, and writing proficiency are not similar to the native-English information that the detectors are conditioned to. Similar findings were made by Mitchell et al. (2023) who discovered that the students who have developing English fluency are punished due to the detectors confusing linguistic simplicity and machine writing. Of particular concern are such problems in Southeast Asia, where English is profusely spoken but influenced by the local languages.

In addition to bias, there are also the issues of inconsistency and obscurity of detectors. The research by Sadasivan et al. (2023) indicates that even with repetitive testing of the same or slightly modified text the detectors can provide conflicting outputs. Cotton et al. (2024) also highlight the overdependence on the outputs of these detectors as they undermine academic trust and expose students to the threat of false accusations. Altogether, the literature points out that AI detectors should be critically considered before incorporating them into the assessment systems, particularly in multilingual settings. This work is a contribution to this evidence, as it investigates the performance of detectors in the real application to actual undergraduate writing of Filipinos.

Beyond reporting bias and inconsistency, the prior literature can be interpreted through a common explanatory mechanism. Most AI writing detectors rely on probabilistic language modeling and stylometric regularities learned from large corpora dominated by native-English text. Rather than identifying authorship directly, the systems evaluate how closely a text conforms to statistical expectations of “typical human writing” within their training distribution. Consequently, writing that deviates from these norms — such as multilingual or developing English writing — may be interpreted as machine-generated. This phenomenon may be understood as a distributional mismatch problem, where detectors operationalise linguistic typicality rather than genuine human authorship. Under this perspective, false positives are not random errors but systematic outcomes of the modeling assumptions underlying AI detection tools.

Taken together, prior studies consistently indicate three related concerns: detector inconsistency, bias against non-native English writing, and overreliance on automated outputs in educational decisions. While these issues have been documented largely using synthetic or Western datasets, limited empirical evidence exists using authentic student writing from multilingual Southeast Asian contexts. This gap motivates the present study, which examines detector behaviour using real Filipino undergraduate essays to provide context-specific evidence relevant to ICT-supported assessment environments.

3. Research Method

This study employed a descriptive small-scale evaluation design to examine the behaviour of AI writing detectors when applied to authentic student work. The design did not aim to establish causal relationships but to document how detectors classified human-written essays. As an exploratory pilot investigation, the study provides preliminary evidence intended to inform future, larger-scale research on the reliability of AI detection tools in educational contexts. The study context reflects assessment conditions common in digitally

mediated and open and distance learning environments, where automated tools are used to supplement instructor judgement.

3.1. Participants and Data

The dataset contained ten essays written by college students attending a general education course at a state university in the Philippines. The essays were routine class submissions completed without instructions related to AI use. All texts were argumentative essays representing a single academic writing genre. Each essay ranged approximately between 600 and 900 words and addressed comparable course-related argumentative prompts assigned within the same instructional period. This ensured relative consistency in genre and academic level across the dataset.

Purposive sampling was used because the objective was not statistical representativeness but examination of typical student writing and detector response. All essays were then manually reviewed to ensure they were genuinely human-written and not generated or heavily edited by AI systems. Authenticity of the essays was verified through instructor review, submission timestamps within the learning management system, and comparison with students' prior in-class writing samples to ensure consistency in writing style. No evidence of AI-generated content was detected during manual review. The texts were anonymised prior to the analysis: all names, dates, identifiers, and personal information were removed in order to protect the students' privacy. Demographic information was omitted to maintain confidentiality and was unrelated to the objectives of the current investigation. The dataset was intentionally limited because the study aimed to document detection patterns rather than produce statistically generalizable estimates.

3.2. AI Detectors Used

Two AI writing detection tools, ZeroGPT and Copyleaks, were evaluated. The testing of all ten essays was conducted in October 2025. Both are widely used by educators, are freely available online, and are commonly cited in discussions of AI detection in academic contexts. Each employs a unique proprietary algorithm and yields a final classification which classifies a text as "AI-generated," "Human," or in some instances "Mixed." To ensure consistency and comparability across the experiment, the study recorded only the final categorical label given by each tool. Detector percentage scores were excluded since such scores differ across systems, meaningful interpretive standards are lacking, and they may mislead users about confidence. This reflects typical classroom use, where instructors rely primarily on categorical outputs rather than probability scores.

3.3. Procedure

The analysis followed four steps:

1. Each essay was submitted separately to ZeroGPT and Copyleaks to avoid cross-session interference.
2. The classification label produced by each detector was recorded in a spreadsheet together with the essay identifier and verified authorship status (human-written).
3. Any essay labelled "AI-generated" was recorded as a false positive.
4. Frequencies and percentages were calculated to summarize misclassification rates.

Descriptive statistics were used because the objective was pattern documentation rather than inferential testing. All procedures were documented as part of the study audit trail, and no automated scripts or external software were used. Classifications were collected manually to reflect typical instructor use of AI detection tools.

3.4. Ethical Considerations

This study followed standard practices for minimal-risk educational research using anonymised coursework data. Prior to inclusion in the study, permission to use the essays was obtained from student participants through a printed written consent form, which they signed to authorise the use of their work for anonymised research purposes. Participants were informed of the study's purpose and assured that participation would not affect their academic standing. No personally identifiable information was collected or stored, and the study did not evaluate student performance. The investigation focused solely on the behaviour of AI detection tools and posed no foreseeable risk to participants. Because the study involved voluntary anonymised coursework and no intervention, formal institutional ethics approval was not required under institutional policy.

4. Findings and Discussion

4.1. Findings

The data were analysed descriptively to examine how the two AI writing detectors categorised the collection of ten genuine student essays. As shown in Table 1, ZeroGPT and Copyleaks provided conflicting classifications for several of the same texts. All essays were written by human participants; however, each detector incorrectly marked five essays as AI-generated. The outcomes are summarised in Table 2, which presents the number and percentage of false positives. Both detection tools recorded a misclassification rate of 50%, meaning that half of the authentic human-written essays were flagged as AI-generated.

Table 1

Classification of Authentic Essays by AI Detectors

Essay No.	Actual Classification	ZeroGPT Output	Copyleaks Output
1	Human	AI	AI
2	Human	Human	Human
3	Human	AI	Human
4	Human	Human	AI
5	Human	AI	Human
6	Human	Human	Human
7	Human	AI	Human
8	Human	Human	AI
9	Human	AI	Human
10	Human	Human	AI

Table 2

Misclassification Rates

Detector	Essays Flagged as AI	Percentage
ZeroGPT	5	50%
CopyLeaks	5	50%

The 50% false-positive rate is not only numerically high but educationally significant, as it implies that in a typical classroom setting half of genuinely written submissions could be wrongly suspected as AI-generated. This pattern suggests that detector outputs should be interpreted as probabilistic indicators rather than definitive judgements. The inconsistency observed between the two systems further indicates that classification depends on

algorithmic sensitivity to linguistic structure rather than actual authorship, reinforcing concerns about fairness in multilingual assessment contexts.

4.2. Discussion

The findings support the distributional mismatch explanation proposed in the literature review, indicating that the detectors respond to conformity with learned linguistic patterns rather than actual authorship. The observed 50% false-positive rate suggests that detector outputs may not reliably distinguish between authentic human writing and AI-generated text in this context. This aligns with prior research indicating that non-native linguistic patterns, including simpler syntactic structures or limited lexical variation, are sometimes misinterpreted as machine-generated features (Liang et al., 2023; Mitchell et al., 2023).

The inconsistency between ZeroGPT and Copyleaks further reflects detector instability reported in earlier studies (Sadasivan et al., 2023). When two widely used systems produce conflicting classifications on the same authentic texts, the reliability of such tools for academic integrity adjudication becomes questionable. In high-stakes educational settings, a false accusation based on automated classification may carry serious academic and reputational consequences, particularly when automated outputs are perceived as authoritative despite known limitations (Kreps et al., 2022). From an institutional perspective, a 50% false-positive rate would be unacceptable if used as a primary evidentiary basis in misconduct investigations. AI detection outputs should therefore be interpreted as preliminary indicators rather than definitive judgments. Especially in ICT-supported and open and distance learning (ODL) environments—where instructors may rely more heavily on automated systems—robust human verification procedures are essential.

More broadly, the results highlight an educational equity concern. If detectors are trained predominantly on native-English datasets, multilingual students whose writing reflects legitimate linguistic variation may be disproportionately affected. The issue thus extends beyond technical accuracy to questions of fairness and policy governance in AI-supported assessment systems.

4.3. Limitations

This study is limited by its small-scale exploratory design and the restricted dataset of ten essays drawn from a single institutional context. While the results indicate notable misclassification patterns, they should be interpreted as preliminary evidence rather than statistically generalisable conclusions. The purpose of the investigation was to document detector behaviour in authentic classroom writing rather than to estimate population-level accuracy. Collectively, these constraints define the study as context-bounded evidence rather than a generalisable performance evaluation. In addition, only two commercially available AI detection tools (ZeroGPT and Copyleaks) were examined. Other systems may operate using different training data and detection architectures and therefore may produce different classification outcomes. Prior research has shown that detector outputs vary depending on linguistic complexity and writing style (Liang et al., 2023; Mitchell et al., 2023) and may produce inconsistent results across repeated evaluations (Sadasivan et al., 2023). Consequently, the present findings should not be interpreted as universal performance indicators for all AI detection technologies.

Furthermore, the analysis relied on final categorical outputs (“AI-generated” or “Human”) rather than internal probability scores or algorithmic features. This approach reflects typical classroom use but does not allow examination of the specific linguistic variables influencing classification. Future studies may incorporate larger multilingual datasets, multiple detectors, and qualitative linguistic analysis to better understand how writing features interact with detection mechanisms. Beyond sample size and tool scope, several contextual constraints

must also be acknowledged. The essays analysed were produced within a specific academic discipline and institutional environment, which may shape linguistic patterns, rhetorical structures, and instructional expectations. As a result, detector performance observed in this study may be influenced by contextual writing norms rather than inherent classifier bias alone.

Additionally, the study design did not experimentally manipulate variables such as prompt type, essay length, revision history, or level of language proficiency, all of which may affect detection outcomes. Future research should adopt multi-institutional and cross-disciplinary sampling frameworks, incorporate controlled experimental conditions, and examine detector performance across varying proficiency levels and drafting stages. Longitudinal designs may also help determine whether misclassification patterns remain stable over time or shift as detection systems evolve. Such extensions would provide stronger empirical grounding for policy recommendations regarding the pedagogical use of AI detection tools.

5. Conclusion

This small-scale assessment provides early but meaningful evidence that AI writing detectors may not reliably classify authentic student essays. The findings indicate a particular risk for Filipino undergraduate writers, as the detectors appear biased against features common in non-native English writing. Although all ten essays were authored by students, both ZeroGPT and Copyleaks incorrectly flagged half of them as AI-generated, raising serious concerns for educators who may rely on such tools in high-stakes academic integrity decisions. The results suggest that detectors are sensitive to writing characteristics often found in student work, including straightforward vocabulary, simple sentence structures, and limited elaboration, which may not align with algorithmic expectations of complex human-like language. The study also draws attention to a broader issue in ICT-supported assessment, namely the assumption that automated tools are neutral, objective, and error-free. In practice, AI detectors rely on probabilistic and opaque models, producing outputs that teachers cannot reasonably be expected to interpret fully. When used without caution, these systems risk introducing unintended bias against students who write in a direct or less sophisticated manner. For institutions, this implies that AI detection results should be treated as preliminary indicators rather than definitive evidence of misconduct. While the study's findings are limited by a small sample size, they offer useful preliminary insights into current detector performance. Educators are encouraged to combine detector outputs with human judgment, evidence of the writing process, and dialogue with students. Future research using larger and more diverse datasets, as well as additional detection tools, would provide a stronger empirical basis for evaluating reliability. Overall, while AI detectors are increasingly promoted as convenient solutions to academic integrity challenges, this study cautions against their uncritical adoption and emphasizes that automated tools should support, rather than replace, ethical, fair, and informed decision-making.

Funding: This research received no external funding.

Acknowledgement: The author thanks the student participants for permitting the use of their anonymised coursework. The author also acknowledges the constructive feedback and invaluable guidance provided by the editor and peer reviewers, which significantly enhanced the quality of this work. Finally, the author would like to thank the ASEAN Journal of Open and Distance Learning (AJODL) for considering this research for publication and for their continued commitment to advancing open and distance education.

Conflict of Interest Statement: The author declares no conflict of interest.

Author contributions statement: The author contributed to conceptualisation, methodology, investigation, data curation, formal analysis, software, validation, visualisation, interpretation of results, and manuscript preparation (writing – original draft and review & editing).

Data availability statement/ supplementary data: The data used in this study are not publicly available due to ethical and privacy restrictions related to student coursework.

Ethics Statement: Formal institutional ethics approval was not required for this small-scale study involving anonymised coursework submissions; however, informed consent was obtained from all participating students. No personally identifiable or sensitive data were collected or stored.

References

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2024). Detecting AI-generated text: Factors influencing detectability with current methods*. Journal of Artificial Intelligence Research, 82*(2025), 2233-2278. https://doi.org/10.1613/jair.1.16665

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity, 22:4*(2026), 1-18. https://doi.org/10.1007/s40979-026-00213-1

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kuhn, J., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103:102274*. https://doi.org/10.1016/j.lindif.2023.102274

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology* (1st ed.) Routledge. https://doi.org/10.4324/9780203125083

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7). https://doi.org/10.1016/j.patter.2023.100779

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv.* https://doi.org/10.48550/arXiv.2301.11305

Sadasivan, A., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI detectors detect AI-generated text? A study of false positives in AI writing detection. *arXiv.* https://doi.org/10.48550/arXiv.2303.11156

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltynek, T., Gipp, B., Glendinning, I., Kravchenko, O., Kunnari, I., Moten, A., Rajić, V., Štula, M., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19:26*(2023). https://doi.org/10.1007/s40979-023-00146-z