International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Auditing the Fairness of AI-Detection Tools: A Comparative Study of ESL, Published, and AI-Generated Texts and Their Misclassification Risks

R. Paul Lege

Graduate School of Law, Nagoya University, Japan

Received: 11 Sep Aug 2025, Received in revised form: 09 Oct 2025, Accepted: 14 Oct 2025, Available online: 18 Oct 2025

Abstract

This study investigated the classification fairness at the threshold level of four commercially available AI detection tools on the Internet: Copyleaks, ZeroGPT, Scribbr, and Quillbot Premium. The research included the submission of three distinct chunks of texts (N=1212) of between 400-5oo words for evaluation. The writing texts came from fully AI-generated examples (N=307), prompted between 2024 and 2025, and published human-written texts (N=302), and ESL graduate student texts (N=303) written before 2021. The texts were analyzed using binary classification thresholds to determine how the three free devices (Copyleaks, ZeroGPT, Scribbr) and the one paid service (QPremium) performed when checking for potentially AI-generated material in each of the writing examples. The study employed a performance metrics to illustrate the issue with threshold application in such devices. The research included the use of the Chi-square test of independence as well as other inferential statistics to assess inter-detector consistency and potential bias patterns. The results indicated that such devices perform well in identifying AI-generated text written artificially; however, significant disparities emerged in the misclassification of human texts. In particular, AI detectors disproportionally flagged ESL writing with false positives. Such findings illustrate the importance of such fairness audits in assessing the linguistic sensitivity in such tools, especially in the educational setting, where misclassification can have academic or reputational consequences.

***Keywords*—Fairness Audit, AI-generated Text, AI-detectors, ESL pattern bias**

I. INTRODUCTION

With the advent of AI technology, an increasing number of educational institutes report problems associated with students submitting assignments created or written by artificial intelligence. In turn, this threat to academic integrity has compelled educational institutions and teachers in general to depend on AI-detectors to counter the problem. In a survey of articles about this problem in the USA and UK, Anara (2005) found that over 70 percent of schools at all levels may be turning to the use of such devices out of a desperate attempt to halt student cheating [1]. In Europe and Asia, the concern and tendency for educators to fight the unethical use of AI engines with AI detectors has been pretty much the same [2], [3]. The *Ashai Simbun* reported in 2024 that while many Japanese educators understood the limitations of such detectors, they saw them as the last defense against the increasing problem of students misusing AI for assignments [4]. However, while some institutions have ambiguous polices that allow learners to use ChatGPT engines, they have little to say about the use of detection tools to curtail potential misuse of AI technology. As a result, teachers lack the proper training or comprehension of how such tools work to actually employ them properly [5].

Many studies continue to show troubling trends. One problem concerns the misunderstanding that many educators have regarding the nature of such machines;

that is, these devices are probabilistic and not actual indicators of a learner's possible guilt. Institutions can correct this problem through proper training and policy development [6]. The second problem revolves around how such detectors function, which is the concern of this study. Growing research indicates that such tools are programmed algorithmically in such a way that they produce far too many false positives or false negatives to be used to judge learner outcomes [7], [8], [9]. Furthermore, several studies now indicate that such technology may be unintentionally biased toward ESL writing [10], [11], [12].

As of 2025, there appear to be at least 50 commercially available detection tools on the market that vary according to the type of detection (text, images, video, and multimedia) and in terms of detection methodology (linguistic heuristics), audience (education, publication), as well as transparency and reliability. While there has been some small regulatory pressures for change and improvements in the accuracy of such devices, overall, only a few of these companies have published validation studies, and even fewer offer transparent evidence that have addressed the concerns of ESL bias [13] A few companies in the industry have attempted to respond to such concerns [14], [15], but only superficially and without independent verification.

Since the pedagogical risks are high, the public at large, and educators specifically, must continuously view such corporate internal evaluations with healthy skepticism. The evolution of AI technology, combined with the multiple ways to assess such profit-making tools, will drive a need for further research. As the industry offers many detectors that include a host of manipulative features that can change over time, this will necessitate independent corroborative research. Consumers, for example, should find it noteworthy when a new detection service claims that other competitive devices on the market produce false positives while their service does not [16]. Educators, in particular, must be concerned with how accurate and fair such tools are in assessing whether students generate assignments with AI technology. Thus, there remains a continuous need for accuracy and fairness studies concerning such detection tools. This paper aims to conduct a fairness audit of four available detectors on the market that may misclassify ESL text as AI-generated when it was not.

1.1 About Fairness

As a matter of fairness, this study is primarily concerned with identifying *who* is impacted when educators employ AI tools to evaluate student assignments. In general, fairness refers to the equitable treatment of learners regardless of their linguistic background, proficiency level, or writing style [17]. In this context, fairness is a multi-dimensional concept associated with proper statistical analysis, structural transparency, contextual impartiality, and educational equity. Such tools should minimize any disparities (such as false positives) across all subgroups. Ideally, such tools require contextual sensitivity in which their features do not penalize for linguistic differences. Fairness also requires full and open transparency in terms of defining the thresholds and providing reproducible metrics in the performance of such machines. In the learning environment, fairness means that such devices should not result in disproportionate harm to the student, such as severe discipline or reputational damage [18]. This paper investigates the threshold levels of four machines to confirm or reject the following hypotheses:

> H0: *The proportion of human-written texts misclassified as AI-generated is the same across all four detectors.*

The alternative hypothesis is:

> H1: *The proportion of ESL texts misclassified as AI-generated is higher than other texts, while scores differ across all four detectors*.

A confirmation of the null hypothesis (H0) would mean that any observed differences in false positives (FP) would be due to random variation and not intentional bias. On the other hand, a rejection of the null hypothesis and confirmation of the alternative hypothesis would provide evidence that such tools can misclassify ESL texts that results in unfairness.

1.2 Understanding AI Detectors

As already noted, commercially available AI detectors come in various types and serve different purposes. While consumers may be naturally confused about which to adopt, the important point is that no such device can predict or verify the absolute truth as to whether an element of writing is AI or human-generated. These are probabilistic machines that measure a number of features such as perplexity, burstiness, repetition, semantic richness, entropy, idiomaticity, and syntactic variety (just to name a few). The definitions of these features derive from a cross-section of theories such as Computational Linguistics, Natural Language Processing, Machine Learning, and Information theory

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

[19]. While there are many features, Fig. 1 below shows four key features that relate to this study.

Feature	Definition	Analyzed at	How measured
Perplexity	Language model confidence in predicting word sequences	Word or sub word level	Averaged across the entire text. Lower Perplexity = more Predictability.
Syntactic Complexity	Level of sentence structure	Clause & sentence level	Uses dependency graphs to assess the use of Subordinate clauses or modifiers
Semantic Richness	Depth and diversity of ideas Despite Syntactic complexity	Phrase and sentence level	Embedding models that assess meaning by phrase coherence and sentence level
Lexical Diversity	Variety of unique words in a text	Word-level & document-level	Uses Type-Token Ratio (TTR) counts unique words versus total words across entire text

Fig. 1: Four Features Commonly Measured and Classified by AI Devices

As Fig. 1 above shows, these features are measured at multiple levels (from word to document), then transformed into a classification model, piped into algorithms and thresholds that provide a probability score as to their origin (either human or AI-generated). However, depending on the brand, the thresholds may be too rigid or uncalibrated for under-skilled or ESL writers [20]. Since ESL writers often use simpler clauses and repetitive vocabulary, this could sway the metrics at different levels. Furthermore, such devices may over- or underemphasize perplexity (which is why many studies focus on this issue) and misclassify authentic writing as AI-generated [21]. Finally, semantic richness is particularly sensitive to idiomatic phrasing and cultural context. A Japanese student might write the following sentence: *Although my friend said John was smart, I was surprised to see how heavy John was.* A detector might parse this sentence as syntactically complex (subordinate clause), but lacking diversity (*John* twice and *was* three times), and perhaps logically unclear, so that it is semantically poor [22]. In addition, it may miss the contextual use of the word "smart," which can mean "thin" to some Japanese learners.

Essentially, these devices act in a similar way to airport screening machines. Fig. 2 provides a basic conceptual model of the two levels of diagnostics that includes measurement systems of algorithms (the scan) and thresholds (settings). The full scan requires a four step process: (1) the raw data (blue); is inserted into the machine (2) the scanning (green) occurs with feature extraction (perplexity, lexical diversity) and then assigns as a score; (3) these scores are matched to the threshold settings (orange) and given a binary label; and (4) an output label of "likely" AI or human generated is delivered to the user [23]. As this detecting measurement system involves two levels of assessment (algorithms and thresholds), this means that problems can arise at either level or both. Problems at the algorithmic level can lead to structural bias depending on how well they are "trained" in classifying linguistic variation, such that errors can result in penalizing ESL writing. Even if a well-trained scan provides a reliable score on a feature, a poorly calibrated threshold (ie, a setting that is too high or low) could result in procedural bias that also misclassifies a text. The scope of this study is at the threshold level.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

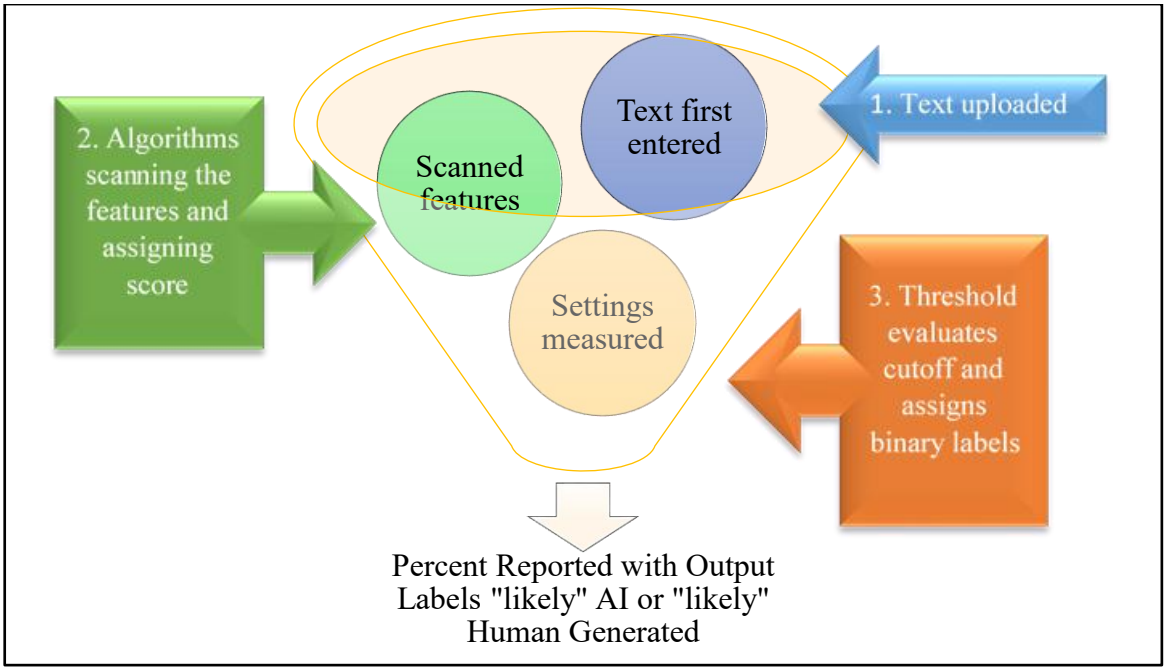

Fig 2: Conceptual Model of AI-detectors as Scanners

II. LITERATURE REVIEW

Gotoman et al. (2024) conducted a systematic literature review of 34 scholarly studies from three online databases in order to assess what the research found regarding commercially available detection devices [24]. They noted eight evaluative approaches, with the main three being concerned with accuracy, reliability, and fairness. The findings revealed that while many detectors achieved above 50% accuracy rates, in total, they remained unreliable. Most studies also indicated that paid or premium machines outperformed free versions. In terms of fairness, the consensus was that such imperfect tools should serve as supportive evidence and not as a final judgment. The authors concluded that such technology needed improvement in terms of transparency, fairness, and the strength of the measurement system. Selectively, the remainder of this review will discuss research that aligns with the aims of this paper associated with investigating how such tools may misclassify human text (especially ESL) as AI-generated.

While the research regarding AI text detection and ESL writing in the educational or multimedia context has expanded, much of it has focused on the algorithmic level. For example, Chaka (2023) reviewed 17 studies by combining corpus analysis and qualitative synthesis to examine how such devices may be misclassifying authentic writing [10]. This evaluation revealed that structural uniformity in the devices triggered false positives in ESL writing, but it was concerned with fairness. The author advocated that educators should triangulate such tools with other devices, along with human judgment.

Meanwhile, Liang et al. (2023) employed stratified sampling and cross-detector benchmarking to evaluate seven commonly used detectors on 40 TOEFL essays [11]. Their study found that these tools misclassified 61% of the ESL writing compared to the high accuracy of essays produced by native authors. This highly cited research revealed that such tools do indeed target linguistic variation common within ESL text. The Liang study maps well with this present study, which is concerned with semantic richness, threshold sensitivity, and stratified fairness auditing.

Echoing some of the same concerns, Price and Sakellarios (2023) sampled 120 essays written by Japanese college students with several commonly available detectors [25]. Their research also found a high number of false positives, especially among lower-skilled learners. They noted that such machines misinterpreted features such as lexical simplicity and syntactic repetition as AI-generated text, and that threshold levels varied widely among the devices. They further concluded that such misclassification can result in pedagogical risks to such learners. This aligns well here because it notes that fairness concerns are dependent on a complete evaluation of both levels of such detection systems.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Li and Wan (2025) produced a large-scale empirical study using 483,360 student essays to benchmark classifiers with six detectors (3 English and 3 multilingual) based on features such as perplexity and lexical richness [26]. The study employed Random Forest models and stratified sampling across academic fields, finding that at this algorithmic level, false positives occurred at a high rate in both categories, which would impact ESL writing. This study is relevant here because it analyzed several detectors, found data suggesting false positives that can be generalized, and suggested a need for adjustments. While this study implied potential problems in authenticating ESL writing, Pratama (2025) used similar devices to analyze 300 scholarly abstracts from both native and non-native authors [12]. The results revealed that such tools disproportionally flagged ESL abstracts as AI-generated. In addition, the aim of the study was to improve educational integrity.

Pudasaini et al. (2025) tested five detectors with 1,500 writing samples from academic, journalistic, and evasive LLM outputs (ie, camouflaged AI-generated text) [13]. Such a strategy is related to robustness studies rather than a fairness audit. With such testing, they indeed found that thresholds degrade under real conditions. While such a study centers around robustness and accuracy, its findings support fairness-oriented research by showing how such devices perform weakly with linguistic variability in academic and multilingual writing samples.

Together, these studies form a layered map of the linguistic feature analysis, the scope, the approaches, sampling, and methodology that can be used in studying such devices. For comparison, Fig. 3 below provides a summary of the review as each of the studies aligns with this paper. The present study builds on this foundation by integrating semantic richness and cohesion metrics, modeling threshold sensitivity, and visualizing bias through stratified confusion matrices as well as advancing a reproducible framework for ESL-aware fairness auditing.

Study	Scope	Features Measured	Sample Size & Type	# Detectors Analyzed	Approach	Main Methods
Chaka (2023)	Algorithmic	Grammar interference	40 L1 & L2 essays	30	Fairness	Comparison across platforms
Liang et al (2023)	Algorithmic	Stylistic shifts, reviewer behavior	91 TOEFL & 88 L1 essays	7	Adversarial	Corpus-level estimation of LLM influence
Price et al (2023)	Thresholds	Grammar confounders	Japanese university 12 essays	5 Free detectors	Fairness	Manual vs. automated detection comparison
Le & Wan (2024)	Algorithmic	Perplexity false positives	480,000+	6	Adversarial	Inverse perplexity-weighted ensemble
Pratama (2025)	Thresholds	Detection metrics, disciplinary bias	71 academic articles	3	Fairness	Accuracy vs. bias trade-off analysis
Pudasaini (2025)	Algorithmic	Paraphrasing robustness, evasion tactics	6,000 human 6,000 AI texts	3	Adversarial	Benchmarking detectors with adversarial inputs
Lege (2025)	Thresholds	Threshold recalibration	1212 Stratified texts (Japanese university context)	4 (3 free, 1 paid)	Fairness	Performance metrics, Stat. residuals

Fig. 3: Chronological Map of Key Studies on AI-Generated Detection

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

III. METHODS

In terms of inputs, this study investigated and performed an interdependent fairness audit on three free versions of such tools (Copyleaks, ZeroGPT, and Scribbr) and one paid version (Quillbot-Premium) to compare and detect whether written texts (N=1212) were AI-generated or human-produced. These tools were selected due to ease of access, because they are marketed toward education and publications, and the research suggested low to moderate issues with false positives The writing examples consisted of chunks of texts (400-500 words) produced from three classes or groups: actual AI-generated texts (n=307), portions of scholarly published articles (n=302) available on the internet that predate the arrival of Open AI technology, and sections of text from ESL theses (n=303) written at Nagoya University in Japan prior 2021. Thus, with a 100% grounded truth value, this study employed descriptive and inferential analysis to audit the performance of fairness in four AI tools that many educators presently use to evaluate student writing.

Articles by Chaka (2023) and Gotoman et al. (2024) uncovered at least seven methodological approaches to the study of such devices, including adversarial tests, accuracy and error analysis, content obfuscation sensitivity, cross-domain generalization, fairness audits, human-AI discrimination, and watermark detection. The approaches to such studies are typically divided between adversarial (the *why* and *how*) or a fairness (*who* is impacted and *why*). As the column on the scope shows in Fig. 3, several known studies have explored similar risks with such devices, but looked at either algorithmic scanning or the threshold settings. Ideally, both should be done to ascertain a full understanding of the problems associated with such devices, but practical considerations restrict such studies to a single approach.

While two of the studies did look at threshold levels (Price et al. and Pratama), they relied on descriptive statistics to explore *who* was impacted, but did not conduct performance metrics to address *why* this occurs. This present study expands beyond the typical fairness study by investigating *why* such devices may be misclassifying scores. Typically, the adversarial approach incorporates methods to investigate technical weaknesses with the scan (robustness) rather than social vulnerabilities (as to who is impacted) [26]. However, when the case involves unintentional bias, then a fairness audit might be of use to probe how a threshold can fool a device when settings err when scoring demographic groups [27]. Such threshold instability can result in misclassification and occurs when there is over-reliance on algorithmic features that may be exaggerated, flawed, or biased. Though the scope of this study cannot fully explore the scanning level (algorithmic features), inferential statistical analysis will provide clues as to potential problems at that level.

Going back to the airport scanner analogy, if any of the pixels within the lens (algorithmic features) are smudged or misaligned, then this could result in a blurred image or misleading result (threshold score). In this instance, each lens represents one algorithmic feature set. If the scan distorts the scan of any feature, then the algorithm bias the threshold scoring. Such distortion occurs due to poor designing, training, or functioning of the algorithms and is equivalent to a coarse or grainy quality in the scan. This, in turn, can result in algorithmic bias. Meanwhile, the threshold slides up and down between sensitivity and specificity settings that bring a macro or global view of the features [28]. Thus, as a scanner device requires care and calibration, AI-detectors need properly selected features to make fair decisions, particularly with ESL texts.

The methods applied to assist with a fairness audit at these two levels fall under the toolboxes of recalibration (of thresholds) and stylometric profiling (of the algorithmic features) [29]. The main focus in this study is recalibration. As Bellamy et al. (2019) noted, recalibration methods involve post-hoc techniques that assist in showing the strengths and weaknesses of a device while adjusting the scores of probabilistic classifiers (such as detectors) to show possible improvements in correct evaluations [30]. As a form of reverse engineering, investigating threshold calibration helps to identify unfair outcomes, such as false positive rates. This first level of investigation involves using a confusion metric to help show potential misclassification, performance metrics to identify optimal thresholds, a Levene's Test to justify whether threshold instability exists, and a Chi-square analysis to establish statistical justification that there is significant misclassification across the three groups.

With the assistance of this recalibration approach and inferential statistical analysis, residual data will provide clues that problems exist at the algorithmic level, thereby hinting at the reason for the outcome errors. As an example, the study examines four linguistic traits to understand which stylometric signals may be unintentionally causing bias. The four features include perplexity, syntactic-semantic complexity, and

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

lexical diversity in educational equity. While there are indeed many features, these four are selected because they align with the four aspects of fairness (perplexity to statistical analysis, syntactic-richness to structural transparency, semantic complexity to contextual impartiality, and lexical diversity to educational equity) [31]. Furthermore, these features are relevant because research indicates that ESL writing diverges from native writing with all four, and they since they allow for a visual quantification of where the detectors may be misclassifying due to linguistic bias [32]. Thus, the bridge from recalibration or threshold tuning to stylometric analysis allows the study to help educators see the limitations in such devices to ensure that they are applied equitably across diverse writing populations.

The recalibration analysis first incorporates descriptive raw data in the form of a contingency table to assist in showing the averages, mean scores, and standard deviation in the three submitted forms of writing (AI-generated, published text, and student text), which helps identify *who* is most likely affected by such misclassifications. A Levene's test was used to check the significance of some of the variances of the standard deviations. The next step includes the use of performance metrics to illustrate the impact of thresholds in assigning false positives in both the raw data and with recalibration. The study follows with a Post Hoc Chi-square test to compare the significance of the assignment of false positives by each of the devices. To strengthen the results of the Chi-square comparison, the study includes a standardized residual test that hints the problem is not simply at the threshold level but is occurring at the algorithmic level (that is, the *why*), and such misclassifications may be occurring.

IV. RESULTS

The recalibration approach provides insight into how such mechanical devices use thresholds for classification. Using raw data, descriptive and inferential statistics, and adjustments to the threshold, it is possible to reweigh prediction probabilities and visualize *who* may be affected by such scores. With the help of a confusion matrix, performance metrics, Levene's Test, and Chi-square, the study will establish that ESL writers were most likely to be given a false positive score with such tools. Alone, such a method can assist with aligning outputs that improve FP parity. The assistance of a standard residual test will hint that the reason why such scoring occurs may be found at the feature or scanning level as well.

Table 1 represents a form of contingency table or summary matrix showing all the raw data collected from the results of the devices as they were assessed for AI-generation. Here, if any device scored a text (even 1%), then it is listed as being flagged for AI. Each machine evaluated 1212 texts (307 AI-generated, 302 published texts, and 303 ESL texts). As the table shows, all the machines detected automated texts correctly (scores ranged 88-100%), suggesting a large number of true positives (TP). On the other hand, these tools flagged published text 33.6% of the time (406/1208) and, more importantly, ESL text at 69.9% (809/1212). Though not shown, the range for the actual scores for each text was 88-100 for AI-generated text, 1-32 for published, and 1-52 for ESL, indicating a strong true positive rate for the AI texts, and some degree of false positive assignment for the human written texts.

Table 1: Summary Matrix of Raw Data across Four Devices Flagging for AI-generation

Tool	**Binary # Flagged Scores**			
	AI-text (TP)	**Published(FP)**	**ESL (FP)**	**Total**
Copyleaks	307/307	127/302	232/303	1212
Zero	307/307	104/302	219/303	1212
Scribbr	307/307	101/302	211/303	1212
Q-Premium	307/307	74/302	177/303	1212
Totals	1228/1228	406/1208	809/1212	

Continuing with the presentation of the raw data, the apparent disparity between the strength of the devices in terms of recognizing AI-generated text while struggling with some aspects of identifying human texts justifies further investigation. Because the human-generated texts predate the onset of commercially available AI technology, the observed texts have a grounded truth value of 100%. Even though the actual threshold settings for the devices are unknown (could be set between 20-80%), as this tends to be proprietary

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

information, having a strong grounded truth value (knowing that observed values are true) is critical to building performance metrics [33].

At this stage, the main disparity in the raw data between the two human texts hints at potential bias toward ESL writing. The fact that such tools flagged published texts with actual false positives (33.6%) at all is surprising, but the more than double rate of labeling ESL (69.9%) raises even further concerns. This disparity exhibits a systemic fairness issue, suggesting that many of the current detection models on the market may incorrectly conflate linguistic variations in writing differences. Such findings reinforce the need for an evaluation of their actual capabilities. The next step required an examination of the raw continuous scores (1-100) of the devices assigned to each individual text to establish the extent of the difference in scoring between the two human-written texts.

Table 2: Average of the Continuous (1-100) AI-Generation Scores per Writing Text

Group	# Articles	Copyleaks	Zero	Scribbr	QP
AI-gene.	307	96.8	97.1	98.3	98.5
ESL	303	17.2	15.6	15.4	9.4
Published	302	3.9	4.5	5.9	3.1

Table 2 above is a descriptive summary of the average continuous scores that each of the devices gave after submitting them for AI evaluation. As the table shows, all four devices largely detected the actual veracity of AI-generated texts (true positives), though imperfectly. Each of these devices appears quite capable of identifying true positive scoring for AI-text, with Copyleaks performing the weakest, with scores averaging 96.8% and Q-premium the best at 98.5%. Since these devices were not perfect, this raises a question as to the acceptable amount of error (true negative) that would be acceptable. Typically, such an acceptable error rate would depend on the purpose of use and could be less than 2-5% in terms of legal or policy development or for defending academic integrity [34]. The average for true negatives (TN) for the AI-texts in this instance ranged between 1.5 (Q-premium) and 3.2 (Copyleaks), which is calculated by subtracting the average scores from 100. Therefore, the error rate here (total TN average of all four devices), while questionable, is within the standard acceptable margin at 2.34%. Thus, such a device appears strong at correctly identifying AI-produced text, but there could be some issues.

However, Table 2 also shows that the average amount of error or false positive rates (FP) for the human texts (published and ESL) shows averages above the acceptable rates. The range of averages for FP for the published texts is 3.1 (Q-premium) and 5.9 (Scribbr); meanwhile, the range of the average FP scores for the ESL texts is 9.4 (Q-premium) and 17.2 (Copyleaks). While all four devices assigned false positive (FP) scores for all of the human texts, there exists an obvious difference between how the tools evaluated the published texts (total average of 4.15 FP) and the ESL texts (total average of 14.4). As the table indicates, these devices scored the ESL with higher FP scores by 3-4 times relative to the published texts. At this stage, more analysis is needed to confirm that the ESL texts were subjected to unintentional or systematic bias.

Since the error rate of the FP for the published text (4.15) is closer to the TN averages of error for the AI-text (2.34%), a t-test is needed to understand more clearly if the devices are tagging published texts closer to the acceptable TN rate for the AI-generated texts. The results from a t-test compared the TN and FP rates across the four tools revealed a consistent misclassification bias with a difference between the means of 2.025. While the results did not reach an actual statistical significance of $\alpha = .05$ level ($t(3) = 2.66$, $p \approx .08$), the magnitude of the difference in the mean average suggests a practical difference (keeping in mind that TN is an error even at the smallest rate). While these tools did flag some AI-text as human-written, though at small rates, they simultaneously over-flagged human texts that appear to disproportionally affect ratings for ESL writing. Such findings support the need for recalibration, as current thresholds may be misaligned with linguistic diversity. A further look at the mean and standard deviation for each device may illuminate these differences.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Table 3: Mean Scores and Standard Deviations for the Continuous Scoring

	AI-generated		ESL		Published	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Copyleaks	96.5	3.3	17.2	12.1	3.8	4.2
Zero	96.8	4.5	15.6	13.7	4.6	5.1
Scribbr	98.3	5.8	15.2	14.5	5.5	6.3
Q-Pre	98.4	2.6	9.3	10.2	3.2	4.0

Table 3 above presents the mean scores and standard deviations for each device as they relate to the three different writing forms. As shown, the high means for the AI-generated texts, along with the tight clustering of the SDs, illustrate that these devices are quite adept at identifying when a text is fully AI-written. In contrast, such tools assign FP scores to ESL texts at a much higher rate than native published writing. The table shows much higher means (up to 17.2) and greater variability (SD 14.5) than with the published group, which indicates much lower means and smaller variability. In addition, for the three free versions that measured ESL writing, the SDs had a wider spread but were still more clustered than with the published text. When such tools show low SDs and high false positives, then this is an indicator of algorithms set toward rigid heuristics (targeting unusual grammar patterns, for example) [35]. In general, then, these descriptive findings raise questions about the accuracy and fairness of such detection tools in analyzing ESL writing outcomes.

The lower SDs for the three free versions compared to their mean for the free versions (Copyleaks, Zero, Scribbr) show strong enough clustering of scores that AI may be unintentionally targeting ESL writing for two reasons. First, the means for the FP scores for the published text are small (in fact, closer to the means of true negatives for the AI-generated text). Second, the SDs show less clustering, which may suggess erratic classification rather than bias. Indeed, five of the standard deviations are greater than their means, suggesting either a statistical anomaly or something more subtle. The standard deviation for the Q-premium (10.2) scores for the ESL texts was slightly higher than the mean (9.6), perhaps reflecting the possibility that premium models calibrate to be more sensitive to false negatives than false positives [11], [28].

Furthermore, all the standard deviations for the devices that evaluated the published articles (4.2, 5.1, 6.3, & 4.0) in Table 3 were slightly higher than their corresponding means (3.6, 4.6, 5.5, & 3.2). Essentially, this means that scores were smaller and spread more widely across the published texts group. Compared to the ESL group, the difference suggests several possible things, for example, a wider dispersion of scores due to variations in native writing skills, some form of internal calibration bias, or just a fluke. Thus, these variances between ESL and the native writers require an inferential test for significance.

The present differences in the case of the larger SDs shown in Table 3 could suggest three main things. First, if the writer's scores actually show lower variance, this might suggest bias in the detection devices. Second, a greater variance might hint at inconsistent treatment. Third, if the variance is moderate, then this could reflect various subtleties, measurement error, or statistical noise. To clarify this issue, this study employed a Levene's Test, often used in educational research, to assess the variance of differences across groups [36]. While such a test cannot isolate which of the groups (ESL or native) was subject to biased treatment, it can show that the variance of scores targeted at least one of the groups.

Table 4: Results of Levene's Test across the Four Devices for the ESL and Published Texts

Device	Levene's F score	p-value	Interpretation
Copyleaks	28.37	<.0001	Scores have significant variance
Zero	31.22	<.0001	Strong evidence of unequal dispersion
Scribbr	26.45	<.0001	Scores vary widely
Q-Pre.	19.88	<.0001	Score shows a tighter but erratic spread

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

The results from Levene's Test in Table 4 indicate a high degree of variance for at least one of the writing groups. This is a global test that signals one group may have a wider variance or spread in scores, but cannot identify which group. This test revealed significant variance differences between the two groups across all four detectors (F range: 19.88–31.22, p< .0001), indicating heteroscedasticity (variance is not uniform), which could imply potential fairness concerns. An F-score above 10 is considered quite high and, combined with a p-value of <.0001, suggests that overall score dispersion is unequal between ESL and native writers. Essentially, one group has more clustered FP scores relative to the other. While clustering suggests potential uniform bias, a wide spread in the scores implies an inconsistent application of the scoring when attempting to judge a text as A-generated. As such, when detectors show such inconsistency when assigning false Positives across devices, then this indicates systematic bias [12].

By combining Levene's Test with the higher ESL means and SDS, it becomes apparent that the devices allotted higher FP rates to this group. While such a variance measure in the raw data can signal instability in a tool, it cannot reveal how accurate or fair such a device may classify texts across different groups [37]. As such, the next step is to utilize performance metrics to evaluate the practical implications of the disparity found above. With the assistance of metrics such as precision, recall, and false positive rates, this stage will quantify the extent to which ESL writers are misclassified and assess whether the devices meet equitable standards of reliability and fairness.

Table 5: Confusion Matrix for each Device using ≥30% Threshold

Actual	**Devices and Predictability (AI or Human) AI≥30%**							
	Copyleaks		**Zero**		**Scribbr**		**Q-Pre**	
	P-AI	**P-H**	**P-AI**	**P-H**	**P-AI**	**P-H**	**P-AI**	**P-H**
307 (AI)	307 (TP)	— (TN)	307	—	307	—	307	—
303 (ESL)	37 (FP)	266 (TN)	28 (FP)	275 (TN)	19 (FP)	284 (TN)	2 (FP)	301 (TN)
302 (Published)	1 (FP)	301 (TN)	—	302 (TN)	1	302 (TN)	—	302 (TN)

Table 5 provides the results of a standard confusion matrix set with an idealized threshold of 30%. While typically the standard for such machines is supposed to be 50%, such settings for commercial devices tends to be a trade secret and independent research can only infer such threshold settings[38]. Since the raw data showed many scores less than 50%, the assumption is that these actual market tools were set somewhere between 20-50.% Recalibrating at the 30% threshold, in this context, the classification occurs by designating all scores above 30% as FP. As the table shows, at the 30% threshold, all the detectors accurately identified the AI-generated texts as true positives (TP) with scores well above 30% (88-100). Moreover, only two of the published next showed a score over 30% (Copyleaks 32; Scribber, 32), which indicates that at this level the devices only misclassified two texts with a false positive. On the other hand, the devices misclassified 86 of the ESL writing with false positives (Copyleaks, 37; Zero 27; Scribbr, 19, Q-premium, 2).

As a form of descriptive analysis, the confusion matrix provides only limited insight into the four scoring classifications (TP, TN, FP, and FN). However, as Neeley and Englehart (2025) noted, this form of analysis helps to transition to an even more powerful metric that measures accuracy, precision, recall, specificity, and F1 score [39]. These performance metrics normalize raw counts across group sizes and allow for meaningful comparisons between detectors, especially when assessing fairness toward the ESL writers. For example, precision helps quantify false positives, while recall ensures that AI-generated texts are detected reliably. Performance metrics also help audit the detector calibration, sensitivity, and bias, which is critical in any decision regarding their use, especially in education [40].

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Fig. 4 below highlights the essential formulas and meaning for these metrics as drawn from the confusion matrix. For example, when calculating the accuracy of the Copyleaks device, the numbers from the binary classification are plugged into the following formula: (TP +TN) ÷ (TP +TN+FP+FN) to provide a value. After submitting the input values from the confusion matrix, the accuracy percent for Copyleaks would be 95.95% calculated from (307 +569) ÷ (307+569+37+0). Thus, the next step.

Metric Analyzed	Formula	Meaning
Accuracy	(TP +TN)÷(TP +TN+FP+FN)	Overall correctness of the device
Precision	(TP) ÷(TP+FP)	Number actually AI-generated
Recall	(TP) ÷ (TP+FN)	Number of actual AI correctly identified
Specificity	(TN) ÷ (TN +FP)	Number of actual Human texts correct
F1 score	2 x (P x R) ÷ (P +R)	Harmonic mean that balances precision and recall

Fig. 4: Clarifying Performance Metrics

The total performance metrics for all four devices at a 30% threshold are presented in Table 6 below. Most notably, the free versions showed lower scores in all the major categories relative to the premium. That is, they were less accurate, less precise, given to assign more FP, and were less balanced (again suggesting thresholds were not set at 50%). The recall for all devices was 100% indicating that such tools could identify an actual AI-generated text at this threshold (no false negatives). The term specificity refers to the degree to which the devices recognized when a text was actually human-generated, and here, Q-premium scored the highest at 99.68% while Copyleaks lagged at 93.98%. The F1 score establishes the degree of balance or harmony between precision and recall. As Table 6 shows, the Q-premium was more balanced (99.34%) than the free versions, suggesting their thresholds may be set between 30-50%.

Table 6: Performance Metrics for all Four Detectors

Tool	Accuracy	Precision	Recall	Specificity	F1 score
Copyleaks	95.95%	89.2%	100%	93.98%	94.29%
Zero	96.77%	91.84%	100%	95.82%	95.73%
Scribbr	97.56%	94.17%	100%	97%	97.44%
Q-premium	99.78%	99.35%	100%	99.68%	99.34%

At a 30% threshold, the performance metrics above demonstrate that while all four tools exhibit strong capability to identify actual AI-generated texts; however, there remains some variability in the treatment of human-generated texts, especially for the ESL writers. Such metrics provide a better picture of the severity of the scoring patterns. Essentially, the devices tend to target ESL writing across the board, but scoring above 30% occurs with less frequency. While these metrics assist in quantifying the overall behavior of the digital tools, they do not test whether the observed differences across writing groups and detectors are statistically significant. As such, the need for statistical significance warrants the use of a Chi-square test of independence. This type of test can assess whether the apparent variation in the rates of FP occurs from systemic bias within the digital technology or if this is due to random chance. This shift from descriptive metrics to inferential statistical analysis strengthens the fairness audit.

This study ran a Chi-square test of independence from the performance metrics threshold of 30% from the observed totals in the recalibration for the four devices. Even at this threshold, the test revealed a significant relationship between the digital tools and the ESL false positives. The degree of freedom (df) was 3, and the critical value was 7.81. The test result exceeded the critical value at 32.03 ($p < .001$), which indicates that the assigned FP scores for the ESL writing were not due to chance but rather the result of calibration bias within the detectors. Thus, these findings reject the null hypothesis and support the hypothesis that the threshold settings in AI-detectors can impact fairness outcomes for ESL writers.

However, the chi-square test above only produced a statistically significant connection between

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

the ESL false positive rates and the four digital tools, but did not clarify which of these devices contributed most to the apparent bias. As Shan and Gerstenberger (2017) opined, a Post Hoc Chi-square comparison would be applicable here as a way to isolate where the significant differences lie between such devices [41]. Such a step is important to help identify if a specific detector may be impacting the overall effect by comparing the tools with each other. Such a comparison provides a more subtle look at the significant difference in bias patterns that would make these results more useful in system calibrations and for educators endeavoring to use the tools in a triangulated way.

Table 7: Post Hoc Chi-square Comparison of the Four Devices for False Positives

Compared Pair	Chi-Square	Deg. of Freedom	P-value	Significance
Copyleaks v. Zero	1.35	1	p<0.245	None
Copy v. Scribbr	6.91	1	P<0.009	Highly
Copy v. Q-pre.	30.91	1	p<0.00001	Extremely
Zero v. Scribbr	2.49	1	P<0.114	None
Zero v. Q-pre.	18.23	1	p<0.0001	Highly
Scribbr v. Q. pre.	9.52	1	p<0.002	Highly

Table 7 provides results from the Post Hoc Chi-square comparison. The point of this test was to isolate which of the technological tools may have skewed the disparity in FP for all of the devices. The table reveals that all of the free versions contrast significantly with Q-Premium, indicating that these three devices were more prone to assigning false positives. Since there were only two categorical variables in each case, the degree of freedom (df) was 1, making the critical value 3.84 At this juncture, the comparison identifies that the bias is significantly concentrated in the free versions, which may inform educators about which of these tools (or a combination) they might adopt or avoid in the learning environment. .The most extreme comparison occurred between Copyleaks and Q-premium (x^2 =30.91, p<.00001), suggesting that, in this instance, the free version was essentially much more biased. However, since the comparative performances between Copyleaks and ZeroGPT, as well as Zero and Scribbr, show no significant difference, another test may be necessary to clarify the residual effects and help to see if more study is necessary at the algorithmic level.

Sharpe (2015) recommended using a standardized residual test to further strengthen the Post Hoc comparison [42]. This additional test assesses how each of the devices' observed values may have deviated from the expected count under the null hypothesis. This type of test represents a micro-level diagnostic that can reveal which of the tools had the largest impact on the overall chi-square signal, thereby improving the fairness audit of such devices. Much like z-scores, standardized residuals measure the degree to which an observed count deviates from its expected count, scaled by the standard deviation (±2). The formula for such a calculation is Standard Residual = (Observed Cell minus Expected Cell) ÷ √ Expected Cell.

Since Copyleaks appeared as the most isolated in the Post Hoc comparison, the data from this device will serve as an example to show how the calculations work and can actually be done by hand. Fig. 5 below illustrates the two essential steps: first, calculating the expected cell values, and second, calculating the residual score. As the figure shows, after calculating the expected values from the 30% threshold contingency table for FP, it is now possible to obtain the standardized residual by plugging the inputs into the formula. The results show that the standardized residual is +3.27, which exceeds the scaled standard deviation of ±2. This, in turn, indicates statistical over-prediction on the part of this tool. By running standardized residuals for all four of the devices, it is possible to confirm the extent to which this device was the main contributor in assigning FP.

Scribbr ESL False Positives		
Step 1. Calculation of expected values		
	Cells	Amount
	Observed FP (O)	37

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

	Total ESL Texts	303
	Total Devices	4
303 x 4 =1212	Total ESL Prediction	1212
(84 ÷ 1212) x 303=21	Expected FP (E)	21
Step 2. Calculation of residual		
(36-21) ÷ √21 = (15) ÷ 4.58 = +3.27	Standardized Residual	+3.27

Fig 5: Illustrating the Calculation of the Copyleaks Device for Standard Residual

Table 8 below displays the results from the calculation of the 2x4 standardized residuals (FP and TN) for the four devices. This dual-row examination offers educators and researchers several insights. First, this approach establishes the importance of how Chi-square tests work and why full contingency tables are essential for interpreting bias in algorithmic systems. Second, the contrast in the results illuminates which of the devices contributes to potential bias. Third, since many teachers are concerned about the pedagogical risks associated with false negatives [7], such an approach illustrates the see-saw effect or inversion between FP and TN when considering such a tool. In this instance, the table clearly shows that while Copyleaks is more sensitive to FP (+3.27) when evaluating ESL writing, it is the least sensitive to TN (-0.39). The opposite is true for Q-premium, which may be setting their algorithms to be more sensitive to TN (+.5), while overcompensating for FP (-4.15). Of the four devices, Table 8 shows the free version of Scribbr to be the most balanced with respect to evaluating ESL writing.

Table 8: Comparison of Standardized Residuals for the Four Devices (FP and TN)

Detector	**Residual (FP)**	**Residual (TN)**
Copyleaks	+3.27	-0.39
Zero	+1.45	-0.17
Scribbr	-0.17	+0.02
Q-Premium	-4.15	+0.5

V. DISCUSSION

In general, the findings suggest that while these tools perform quite well with actual AI-written outputs, their capability to classify human writing diverges significantly. The lack of variation in the scoring of the AI-generated writing establishes that such detectors are fairly effective in identifying synthetic content [43]. On the other hand, the significant variability in FP of the human writing raises concerns about consistency and reliability. The most striking result was the large number of misclassifications of ESL text at both the grounded truth level and the recalibrated 30% threshold. Free versions of Copyleaks and ZeroGPT flagged ESL writing at high rates. In total, this suggests that these tools may erroneously tag linguistic features common in ESL writing as AI-written. The results of the study reject the null hypothesis that the proportion of human-written texts misclassified as AI-generated is the same across all four detectors.

The results clearly support the hypothesis that the proportion of ESL texts misclassified as AI-generated is higher than other texts, while scores differ across all four detectors. To help test the hypothesis, the analysis represented a recalibration of the threshold aspect of these detectors which assist in identifying *who* is more impacted by such scanning tools. The threshold settings assign the outcome based on a scale between sensitivity and specificity (false positives and false negatives) based on the clarity of the algorithmic measures. Commercially available detection tools do not calibrate on grounded truth values, and they do not publish at what level they set thresholds. So, even with a fairness audit (ie, moving the threshold measure between 30-70%), users of such devices cannot assume that such tools reliably confirm that learners have used AI technology for assignments. Misinterpreting how these devices function risks reinforcing assumptions about ESL writing because this learning group may receive allotted FP scores (even at higher thresholds) due to linguistic features similar to AI-generated writing patterns, despite being originally written.

In addition, the fact that there were so many lower scores at the grounded truth level compared to

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

the 30% threshold does not guarantee fairness. Any device that flags writing at a 17.2 mean score (Copyleaks) may still misclassify authentic ESL writing at higher threshold settings, particularly if the algorithmic canning amplifies the markers in the linguistic features. Without transparent threshold settings and clear validation of the features, users of such devices should interpret findings cautiously. These outcome statements represent only one possible signal that requires contextual review and are not definitive proof of misconduct [44].

For educational institutes and educators, the results here suggest that they should resist the urge to treat scores from such thresholds as hard evidence rather than as one probabilistic signal. Instead, such scores should prompt those in education to engage more with student writing, including pre-and post-diagnostics, revision history, assignment scaffolding, and continuous dialogue. As such, fairness in AI detection is not simply about accuracy but ensuring that educators are not penalizing ESL writing that reflects their linguistic background [18]. While further study is needed, the use of the standardized residual comparison suggests that problems are occurring at the algorithmic level as well.

2 CONCLUSION

In the educational and evaluative environment, teachers and administrators are increasingly dependent on AI detection tools to protect academic integrity. Such devices are being employed to verify the veracity of student assignments; however, the reliability of such tools varies depending on the thresholds and linguistic characteristics of the inputs. Fairness audits can assist in comprehending the issue at the thresholds, while accuracy audits would analyze the algorithms. A study of both would require more written space as a practical matter.

As a fairness audit, this paper demonstrated that while detectors are effective at identifying fully AI-generated texts, such tools show inconsistent and biased classification of authentic human writing, especially from ESL students. While the sample size could always be larger and more diverse, the amount of text here was sufficient in establishing significance in the findings, especially since the grounded truth value was very strong. As such, the significant number of false positive rates for the ESL writing, even after applying the performance metrics, suggests that such tools may confuse non-native linguistic writing patterns with AI features, which in turn can result in the misclassification of ESL writing.

The findings highlight the importance of fair thresholds while refining detection algorithms to reduce potential bias. The point is that problems with such devices can occur at both levels of the scan. As such, Educators and institutions intending to use such devices must approach these tools with care to ensure that they are applied equitably. Future research should expand the scope of a study to analyze both thresholds and algorithms (specific features) while exploring mitigation strategies that promote responsible and fair use of such detection technology.

REFERENCES

[1] Anara, M.K. (2025, July 15). *AI in education statistics: How artificial intelligence is transforming higher education*. Anara. https://anara.com/blog/ai-in-education-statistics (Accessed 6 August 2025).

[2] Holmes, W. (2023). *Asian and European teachers' perspectives on AI and education*. Asia-Europe Foundation (ASEF). pp 1-56. https://asef.org/publications/asian-and-european-teachers-perspectives-on-ai-and-education (Accessed 1 August 2025).

[3] Son, J., Ružić, N. & Philpott, A. (2025). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning*, *5*(1), 94-112. https://doi.org/10.1515/jccall-2023-0015

[4] Kano, K. & Takahama, Y. (2024, July 3). *Educators fear rise in AI-created essays as tools for detection lag*. The Asahi Shimbun. https://www.asahi.com/ajw/articles/15302691 (Accessed 27 July 2025).

[5] Dwyer, M, and Laird, E. (2023*).* Up in the air: Educators juggling the potential of generative AI with detection, discipline, and distrust. *Center for Democracy and Technology*. (2023, March). https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf (Accessed 6 July 2025).

[6] Gustilo, L, Ong, E, and Lapinid, MR (2024). Algorithmically-driven writing and academic integrity: Exploring educators' practices, perceptions, and policies in the AI era. *International Journal for Educational Integrity, 20*(3). https://doi.org/10.1007/s40979-024-00153-8

[7] Dalalah, D. & Dalalah, OMA. (2023, July).The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT, *The International Journal of Management Education, 21*(2). https://doi.org/10.1016/j.ijme.2023.100822

[8] Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *Journal of Academic Integrity and Technology Ethics, 9*(2), 134–147. https://www.researchgate.net/publication/386998010

[9] Walters, W. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*(1). https://doi.org/10.1515/opis-2022-0158

[10] Chaka, C. (2024). Accuracy pecking order: How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning and Teaching,* *7*(1). https://doi.org/10.37074/jalt.2024.7.1.33

[11] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns,* *4*(8). https://doi.org/10.1016/j.patter.2023.100779

[12] Pratama, R. (2025). The accuracy–bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science, 11*, e2953. https://doi.org/10.7717/peerj-cs.2953

[13] Pudasaini, A., Zhang, Y., & Lee, J. (2025). *Benchmarking AI text detection: Assessing detectors against new datasets, evasion tactics, and enhanced LLMs*. Proceedings of the 2025 Conference on Generative AI Detection (GenAIDetect), 1(4): 45–62. https://aclanthology.org/2025.genaidetect-1.4/ (Accessed 15 August 2025).

[14] Lavergne, T. (2023, May 24). *AI detectors: Addressing the challenge of false positives*. Winston AI.https://gowinston.ai/ai-detectors-addressing-the-challenge-of-false-positives/ (Accessed 29 July 2025).

[15] Tian, E. (2023, October 24). *ESL bias in AI detection is an outdated narrative*. GPTZero. https://gptzero.me/news/esl-and-ai-detection (Accessed 6 September 2025).

[16] Emi, B., & Spero, M. (2024). *Technical report on the Checkfor.ai AI-generated text classifier* (Version 2) [Technical report]. arXiv. https://arxiv.org/abs/2402.14873v2 (Accessed 1 August 2025).

[17] González-Sendino, R., Serrano, E., Bajo, J., & Novais, P. (2024). A review of bias and fairness in artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence,* *9*(1). https://doi.org/10.9781/ijimai.2023.11.001

[18] Woelfel, K. (2023, December 18). *Late applications: Disproportionate effects of generative AI detectors on English learners* [Policy brief]. Center for Democracy & Technology. https://cdt.org/insights/brief-late-applications-disproportionate-effects-of-generative-ai-detectors-on-english-learners/ (Accessed 16 July 2025).

[19] Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed., draft version). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf (Accessed 29 July 2025).

[20] Chiusano, F. (2023, February 21). *Two minutes NLP: Perplexity explained with simple probabilities*. Medium. https://medium.com/nlplanet/two-minutes-nlp-perplexity-explained-with-simple-probabilities-6cdc46884584 (Accessed 20 August 2025).

[21] Colla, D., Delsanto, M., Agosto, M., Vitiello, B., & Radicioni, D. P. (2025). *Semantic coherence markers: The contribution of perplexity metrics* [Preprint]. University of Turin. https://iris.unito.it/bitstream/2318/1875282/1/colla2022semantic_preprint.pdf (Accessed 13 September 2025).

[22] Khan, A. (2023). *Mastering perplexity AI: A comprehensive guide to understanding and using perplexity in AI and NLP* [Kindle edition]. Amazon Digital Services LLC.

[23] Yeung, S. (2025, March 3). A comparative study of rule-based, machine learning, and large language model approaches in automated writing evaluation (AWE) (pp. 984-991). Lak`25: Proceedings of the 15th International Learning Analytics and Knowledge Conference. https://dl.acm.org/doi/proceedings/10.1145/3706468 (Accessed 15 August 2025).

[24] Gotoman, J. E. J., Luna, H.L.T., Sangria, J.C.S., Santiago, C.S., Barbuco, D. D. (2024). Accuracy and reliability of AI-generated text detection tools: A literature review. *American Journal of Interdisciplinary Research and Development,* *3*(2). https://journals.e-palli.com/home/index.php/ajirb/article/view/3795

[25] Price, G., & Sakellarios, M. D. (2023). The effectiveness of free software for detecting AI-generated writing. *International Journal of Teaching, Learning and Education, 2*(6), 31–38. https://doi.org/10.22161/ijtle.2.6.4

[26] Li, J., & Wan, X. (2025). *Who writes what: Unveiling the impact of author roles on AI-generated text detection*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 26620–26658). Vienna, Austria: Association for Computational Linguistics. https://aclanthology.org/2025.acl-long.1292/ (Accessed 30 July 2025).

[27] Buchert, J.-M. (2025, March 10). *The 6 best AI detectors based on objective studies & usage*. Intellectual Lead. https://intellectualead.com/best-ai-detectors-guide/. (Accessed 2 July 2025).

[28] Sallami, D., & Aïmeur, E. (2024). Fairframe: A fairness framework for bias detection and mitigation in news. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00568-6

[29] Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientific articles. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 327–337. Association for Computational Linguistics.

https://aclanthology.org/N12-1033/ (Accessed 2 August 2025).

[30] Bellemy, R.K.E., Hind, M., Hoffman, S.C., Houde, S., & Kannan, K. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5). https://doi.org/10.1147/JRD.2019.2942287

[31] Ramzan, M., & Alahmadi, T. S. (2024). The impact of explicit syntax instruction on ESL learners' writing complexity. *World Journal of English Language, 14*(2), 25103. https://doi.org/10.5430/wjel.v14n2p25103

[32] André, Q. C., Ghosh, S., & Khatri, C. (2023). Detecting AI-generated abstracts using linguistic features: A comparative analysis. *CEUR Workshop Proceedings, 3551*, 18–25. https://ceur-ws.org/Vol-3551/paper3.pdf (Accessed 20 July 2025).

[33] Krig, S. (2016). *Computer vision metrics: Survey, taxonomy, and analysis*. Textbook edition. Springer

[34] Alexander, J., Alghamdi, A., & Alzahrani, M. (2023). ESL lecturers' perceptions of AI-generated writing: A deficit model in disguise? *Teaching English with Technology, 23*(2): 45–61. https://doi.org/10.56297/BUKA4060/XHLD5365

[35] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1). https://doi.org/10.1007/s40979-023-00146-z

[36] Nordstokke, D. W., & Zumbo, B. D. (2007). A cautionary tale about Levene's tests for equal variances. *Journal of Educational Research and Policy Studies, 7*(1): 1–14. https://files.eric.ed.gov/fulltext/EJ809430.pdf (Accessed 1 July 2025).

[37] Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys, 52*(6): 1–39. https://doi.org/10.1145/3345317

[38] Hind, M. (2019). *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*. IBM Journal of Research and Development, 63(4/5). https://doi.org/10.1147/JRD.2019.2942287

[39] Neeley, T., & Englehart, T. (2025, January). AI vs human: Analyzing acceptable error rates using the confusion matrix. *Harvard Business School Technical Note* (No. 425-049). https://www.hbs.edu/faculty/Pages/item.aspx?num=66718 (Accessed 9 Sept. 2025).

[40] Padilla, R., Netto, S. L., and da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niteroi, Brazil, 2020, pp. 237-242. https://doi:10.1109/IWSSIP48289.2020.9145130

[41] Shan, G., & Gerstenberger, D. (2017). Fisher's exact approach for post hoc analysis of a chi-squared test. *PLOS ONE, 12*(12). https://doi.org/10.1371/journal.pone.0188709

[42] Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation, 20*(8), 1–10. https://doi.org/10.7275/tbfa-x148

[43] Samue, A. (2024, November 30). *7 best AI reference finder tools in 2025: A comprehensive review for researchers*. Tenorshare. https://ai.tenorshare.com/comparisons-and-reviews/ai-reference-finder.html (Accessed 16 August 2025).

[44] Kar S.K., Bansal T., Modi S., Singh A. (2024). How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. *Indian Journal of Psychological Medicine 47(*3): 275-278. https://doi:10.1177/02537176241247934