

False Positives in Artificial Intelligence Detectors and Their Impact on Academic Quality Assurance: A Systematic Review

Eliana Geomar Leon Abad¹, Wilson Alexander Zambrano Vélez², Rudy Jonathan Parraga Solorzano¹, Rosa Elena Niola Sanmartín¹, Gioryi Augusto Sornoza Zavala², Eduardo Fabio Henríquez Mendoza²

¹Faculty of Educational Sciences and Languages, Santa Elena Peninsula State University; La Libertad, Ecuador

²Faculty of Social and Health Sciences, Santa Elena Peninsula State University; La Libertad, Ecuador

Correspondence: Wilson Alexander Zambrano Vélez, Faculty of Social and Health Sciences, Santa Elena Peninsula State University, La Libertad, 240204, Ecuador.

Received: August 21, 2026

Accepted: October 20, 2026

Online Published: October 28, 2026

doi:10.11114/jets.v15i1.9451

URL: <https://doi.org/10.11114/jets.v15i1.9451>

Abstract

Generative artificial intelligence has transformed assessment in higher education, accelerating the adoption of automated text classifiers to uphold academic integrity; nonetheless, their diagnostic reliability remains contested due to systemic algorithmic biases. This systematic review synthesizes empirical evidence on the incidence of false positives in AI-generated text detectors, their algorithmic and sociolinguistic drivers, and their broader ramifications for institutional quality assurance. Following the PRISMA 2020 guidelines and eligibility criteria based on the SPICE framework, searches were conducted in Web of Science, Scopus, and SciELO, resulting in a qualitative and narrative synthesis of 30 primary empirical studies; the protocol was registered in OSF. Preliminary results from the studies indicate that current detectors rely heavily on reductionist probabilistic metrics, such as perplexity and burstiness, which frequently misclassify structured academic discourse and non-native English writing as synthetic text. Furthermore, the synthesized evidence points to the possibility that punitive enforcement may disproportionately disadvantage specific linguistic cohorts, potentially exacerbating assessment anxiety and compromising evaluative equity. The review suggests that relying exclusively on automated AI detection represents a documented institutional vulnerability for assessment fairness and academic quality assurance, particularly within English-medium higher education contexts where these tools are most extensively implemented.

Keywords: AI detectors, false positives, academic integrity, quality assurance, artificial intelligence, education, systematic review

1. Introduction

Generative Artificial Intelligence has transformed evaluation processes in higher education, challenging institutional regulations regarding academic integrity (Duah & McGivern, 2024). Given the growing complexity of determining authorship, AI-based text detection tools have emerged, although their reliability remains a subject of intense debate (Weber-Wulff et al., 2023). Technically, these instruments operate as Natural Language Processing (NLP) systems driven by predictive classification architectures (Verma et al., 2024). Such models evaluate the statistical properties of textual sequences to distinguish genuine human discourse from synthetic output (Schaff et al., 2024). Grounded in Shannon's (1948) Information Theory, this approach assumes that informational content can be measured through entropy and the statistical predictability of constituent lexical elements.

Building upon these probabilistic foundations, contemporary detectors rely on operational proxies such as perplexity and burstiness (Xu et al., 2024), which directly precipitates the problem of Type I errors or false positives (Giray, 2024). This misclassification occurs when a detector categorizes authentic human writing as machine-generated text, an operational vulnerability emphasized by Verma et al. (2024). In practice, such diagnostic deviation arises because disciplined academic prose is frequently structured, uniform, and statistically predictable, yielding a probabilistic signature that resembles generative language models (Liang et al., 2023). From an epistemological perspective, Messick's Theory of Validity (1990) maintains that validity is a comprehensive evaluative judgment, where the control of construct-irrelevant variance and social consequences determine the legitimacy of the system.

Under this approach, the uncritical adoption of automated detectors introduces measurement biases by penalizing clarity,

coherence, or disciplinary academic style conventions rather than evaluating effective learning (Doğan & Doğan, 2026). Addressing this limitation, Nikolić & Basta Nikolić (2026) and Amnie (2026) point out that institutions have prioritized rushed technological control, relegating instructional transformation. Likewise, Giray (2024) warns that false plagiarism accusations derived from algorithmic flaws reconfigure quality systems, damaging reputation and fracturing ethical trust. Therefore, academic quality assurance is not limited to the regulatory control of infractions, but is conceptualized as the institutional capacity to safeguard the validity of evaluative processes, sociocultural equity, and the ethical governance of learning (Rudakova, 2025), allowing for the overcoming of the deficiencies of the current punitive approach.

Empirically, Liang et al. (2023), documented a marked disparity in error margins, reporting that over 50% of essays authored by non-native English speakers were misclassified as synthetic text. According to Za'rour & Altkhaine (2026), the problem occurs because organized academic writing produces low levels of perplexity and burstiness, metrics that detectors confuse with styles created by algorithms, with this configuration persisting even in recent models (Sun et al., 2025). Despite this, Jiang et al. (2024) demonstrated that through adequate design and representative data, it is possible to build detectors free of linguistic bias. Quantitatively, Stone (2025) reveals that 11% of the student body has been the victim of false accusations, while 72% manifests latent concern, operating as a trigger for intellectual anxiety and the disruption of academic identity (Delina & Del Rosario Rondilla, 2026).

Within psychological and affective dimensions, Kirsanov et al. (2026) define the resulting student distress as sanction anxiety, exacerbated by perceived procedural vulnerability. At the pedagogical level, Giray et al. (2025) observe that substituting instructional guidance with automated scrutiny erodes institutional trust, straining the relationship between educators and learners (Deep et al., 2025). Furthermore, the inefficiency of punitive policies and the severity of governance gaps confirm that blind trust in defective software exacerbates regulatory loopholes (Giray et al., 2025). In response to the coercive environment, Perkins et al. (2024) demonstrated that students adopt defensive writing strategies, altering syntax or concealing personal styles to evade algorithms, a tactic supported by Kirsanov et al. (2026), who observed the strategic omission of reporting technological use out of fear of academic retaliation.

Given the complexity of these dynamics, empirical evidence remains fragmented. Much of the technical literature omits the human consequences of classification errors (Chaka, 2024b; Gritsai et al., 2024), whereas studies under a punitive logic exclusively profile the infringing student, ignoring the phenomenon of unfounded accusations (Bittle & El-Gayar, 2025; Sozon et al., 2025). Although recent research recognizes biases against non-native writers and student anxiety, these contributions remain at a descriptive-statistical level (Caner, 2025; Ndacyayisenga et al., 2026), without delving into sociolinguistic mechanisms or identity damages. Therefore, the objective of this systematic review lies in analyzing the empirical evidence on the incidence of false positives (Type I error) in AI-generated text detectors, their underlying algorithmic and sociolinguistic causes, and the impact of these biases on academic integrity regulations and quality assurance in higher education institutions. To address the general objective, the following research questions were formulated:

1. What is the empirical prevalence of false positives (Type I error) of AI detectors in human-authored academic texts, and how do sociolinguistic variables impact these rates?
2. What technical, stylistic, and design characteristics of the algorithms constitute the primary causes of classification failures in genuine texts?
3. What normative and institutional repercussions do these false accusations have on quality assurance systems, evaluative ethics, and educational equity?

2. Method

The research was carried out using the systematic review method, which involves the comprehensive and structured collection of relevant research findings on a specific topic (Rossi, 2023). Under the PRISMA 2020 guidelines (Page et al., 2021), the study followed a transparent, comprehensive, and high methodological rigor process. The sequential stages of identification, screening, and inclusion were mapped using the PRISMA 2020 flow diagram template developed by Haddaway et al. (2022); the elements of the PRISMA 2020 checklist were included in the supplementary materials (Table S1 and S2), while the protocol was pre-registered on the Open Science Framework (OSF) portal (available at: <https://doi.org/10.17605/OSF.IO/3EP9R>).

2.1 Databases and Search Strategy

The search strategy was executed on July 27, 2026, across three multidisciplinary and indexed bibliographic platforms: Web of Science (Core Collection), Scopus, and SciELO. These databases were selected because their indexing overlaps substantially with specialized educational repositories, including ERIC, thereby ensuring comprehensive access to peer-reviewed journals specializing in higher education, educational technology, and academic integrity. Grey literature—including non-peer-reviewed preprints, commercial white papers, and vendor technical documentation—was systematically

excluded from the review protocol. Because this synthesis required rigorous empirical verification of Type I error rates against validated human-authored benchmarks (ground truth), restricting the corpus to peer-reviewed indexed publications ensured methodological transparency and permitted standardized quality appraisal using the MMAT framework. The search formulation combined controlled terms and keywords related to the study topic: "False positives", "AI detectors", and "academic quality assurance", utilizing Boolean operators (AND/OR), truncation marks (*), and database-specific search tags (Table 1).

Table 1. Search Equation and Initial Results

Database	Search Equation
Web of Science	TS = (("false positive*" OR "type I error*" OR "detection error*" OR "misclassification" OR "bias*" OR "accuracy" OR "reliability" OR "vulnerability" OR "evasion" OR "false accusation*") AND ("AI detector*" OR "AI text detection" OR "LLM detection" OR "machine-generated text" OR "detecting AI" OR "Turnitin" OR "GPTZero" OR "automated text classification")) AND ("higher education" OR "academic integrity" OR "quality assurance" OR "university" OR "college" OR "assessment" OR "plagiarism"))
Scopus	TITLE-ABS-KEY (("false positive*" OR "type I error*" OR "detection error*" OR "misclassification" OR "bias*" OR "accuracy" OR "reliability" OR "vulnerability" OR "evasion" OR "false accusation*") AND ("AI detector*" OR "AI text detection" OR "LLM detection" OR "machine-generated text" OR "detecting AI" OR "Turnitin" OR "GPTZero" OR "automated text classification")) AND ("higher education" OR "academic integrity" OR "quality assurance" OR "university" OR "college" OR "assessment" OR "plagiarism"))
SciELO	(ti:("false positive*" OR "falso positivo*" OR "error tipo I" OR sesgo* OR bias* OR "falsa acusacion*" OR "misclassification" OR ab:("false positive*" OR "falso positivo*" OR "error tipo I" OR sesgo* OR bias* OR "falsa acusacion*" OR "misclassification")) AND (ti:("AI detector*" OR "detector* de IA" OR "inteligencia artificial" OR Turnitin OR GPTZero OR "deteccion de texto") OR ab:("AI detector*" OR "detector* de IA" OR "inteligencia artificial" OR Turnitin OR GPTZero OR "deteccion de texto")) AND (ti: ("higher education" OR "educacion superior" OR university OR universidad* OR "integridad academica" OR "academic integrity") OR ab:("higher education" OR "educacion superior" OR university OR universidad* OR "integridad academica" OR "academic integrity"))

2.2 Eligibility Criteria

During the eligibility assessment phase, inclusion and exclusion criteria were established and applied to determine the final selection of studies included in the systematic review, as detailed in Table 2.

Table 2. Eligibility criteria matrix (inclusion and exclusion) based on the SPICE methodological framework and cross-cutting parameters

SPICE Component / Cross-Cutting Criterion	Inclusion Criteria	Exclusion Criteria
S (Setting)	Research situated exclusively in Higher EducationStudies in K-12 basic/secondary education, corporate and its academic publishing ecosystem.	sector, journalism, or other contexts.
P (Perspective)	Studies focused on the impact or perspective ofResearch on AI from a purely technical, algorithmic, or university students, teachers, evaluators, or qualityengineering approach, without a direct connection to the committees.	educational user.
I (Intervention)	Empirical evaluation of the performance, accuracy,Articles focused solely on traditional plagiarism or biases of AI text detection tools (e.g., Turnitin AI,detection or exclusively on the pedagogical benefits of GPTZero).	generative AI.
C (Comparison)	Contrast of automated detectors against a valid reference standard (verified 100% human texts / Ground Truth) or rigorous human evaluation.	Studies without a control group of legitimate human texts.
E (Evaluation)	Reporting of empirical quantitative metrics of Type IArticles that omit the specific false positive metric or are Error (False Positive Rate - FPR, biases) orlimited to theoretical, philosophical, and ethical debates qualitative impact of false accusations.	without empirical data.
Methodology (Cross-Cutting Criterion)	Primary empirical studies (quantitative, qualitative,Previous reviews (systematic, narrative, or scoping), or mixed methods).	diffuse methodologies, or purely anecdotal case studies.
Document Type	Peer-reviewed indexed scientific journal articles or high-impact conference proceedings (no year restrictions).	Industrial/commercial reports, unreviewed preprints, editorials, letters, book chapters, and grey literature.
Language	Publications in any language to ensure the elimination of geographic and linguistic biases.	No exclusion by language.

Description: Ground Truth = Empirically verified human-authored control texts.

2.3 Study Selection and Identification Process

The study selection process is presented in the flow diagram of Figure 1:

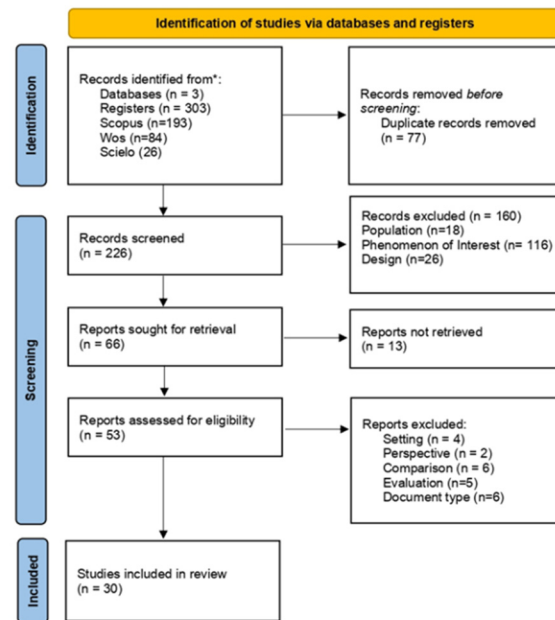


Figure 1. PRISMA flow diagram. Source: Adapted from the template by Page et al. (2021)

Identification: The initial search strategies implemented in the high-impact databases WoS, Scopus, and SciELO yielded a total of 303 scientific articles. Subsequently, the results underwent an automated and manual process to identify and remove duplicate records, through which 77 articles were discarded, obtaining a total of 226 records for the screening stage.

Screening and Retrieval: The 226 unique records were examined during title and abstract screening, resulting in the exclusion of 160 records based on predefined eligibility criteria: population outside the higher education ecosystem ($n = 18$), phenomenon of interest not aligned with automated text classifier accuracy or Type I errors ($n = 116$), and non-empirical methodological design ($n = 26$). Subsequently, 66 reports were sought for full-text retrieval, of which 13 could not be accessed despite interlibrary requests and direct author inquiries; the complete inventory of these unretrieved documents, along with the operational retrieval barriers encountered, is detailed in Table S3. Reports not retrieved and specific retrieval barriers (see Supplementary Material). The remaining 53 full-text articles were rigorously scrutinized against the SPICE framework, leading to the exclusion of 23 reports with documented methodological reasons: setting outside higher education ($n = 4$), perspective lacking focus on educational stakeholders ($n = 2$), absence of a verified human benchmark or Ground Truth comparison ($n = 6$), evaluation omitting empirical quantification of Type I errors ($n = 5$), and document type consisting of non-peer-reviewed commentaries, editorials, or secondary reviews ($n = 6$). The complete bibliographic list of these articles and their individualized operational reasons for exclusion are comprehensively systematized in Table S4. Full text excluded studies and reasons for exclusion (see Supplementary Material).

Inclusion and Unit of Analysis Verification: A final corpus of 30 primary empirical studies fulfilled all eligibility parameters and was incorporated into the narrative synthesis. To prevent unit-of-analysis bias and ensure dataset independence, all included investigations were audited to confirm that no textual corpora or participant samples were double-counted across publications. In cases where multiple included articles shared common authorship, cross-examination verified that each report assessed distinct text corpora, independent evaluation cohorts, or non-overlapping tool parameters, thereby ensuring that each of the 30 included reports contributes unique empirical observations to this review.

2.4 Data Extraction and Synthesis

For procedural transparency, two researchers independently extracted the data into an Excel matrix, resolving discrepancies through consensus. This matrix integrated bibliographic variables (such as author, year, title, country, and sample), methodological and analytical variables (design, instruments, evaluated detection software, false positive prevalence, sociolinguistic variables, algorithmic causes, institutional impact, results, and conclusions). A quantitative statistical pooling (meta-analysis) was neither feasible nor methodologically justifiable due to substantial procedural and clinical heterogeneity across the primary studies, in accordance with PRISMA 2020 guidelines and the Synthesis Without Meta-analysis (SWiM) framework (Campbell et al., 2020; Page et al., 2021). Specifically, the primary literature exhibited

pronounced disparities in: (a) diagnostic decision thresholds across proprietary software (ranging from continuous likelihood indexes to divergent binary cut-off points such as 20%, 50%, or 80%); (b) divergent units of evaluation (word, paragraph, or holistic document level); (c) compositional variations in reference standards (historical pre-2021 academic corpora, authentic student coursework, or English as a Foreign Language compositions); and (d) rapid closed-source software versioning across observational intervals. Consequently, evidence was synthesized through a structured narrative synthesis complemented by descriptive range summaries of Type I error rates across comparable tool classes and linguistic conditions.

2.5 Risk of Bias and Methodological Quality Assessment

To ensure the methodological rigor and the robustness of the evidence in this systematic review, the MMAT 2018 tool was employed to evaluate five key dimensions: the sampling strategy, population representativeness, measurement precision, control of biases due to data loss, and the pertinence of the executed analyses (Hong et al., 2018). With the purpose of safeguarding the objectivity of the scrutiny process, the evaluation of each scientific article was carried out independently by two reviewers under a blinding scheme. Likewise, any initial discrepancy in the assigned scores was resolved through dialogue and direct consensus between both experts, resorting to a third external researcher as a definitive arbiter in those methodological scenarios where disagreements persisted.

The evaluated literature demonstrated predominantly high to medium methodological quality. This includes descriptive and non-experimental quantitative studies (AL-Rawas et al., 2025, 2026; Asiri et al., 2025; Bordalejo et al., 2025; Chaka, 2024a; Cheng et al., 2025; Cingillioglu, 2023; Erol et al., 2025; Fiedler & Döpke, 2025; Gosling et al., 2024; Hadra et al., 2026; Harwood et al., 2024; K. Ibrahim, 2023; K. H. S. Ibrahim et al., 2025; Kaya Ergen, 2025; Liu et al., 2024; Makhmutova et al., 2026; Odri & Yoon, 2023; Ozkara et al., 2025; Paullet et al., 2024, 2025; Richards et al., 2024; Van Vlasselaer et al., 2026; Zhao et al., 2024), as well as complex mixed-methods (Arvidsson, 2025; Hyatt et al., 2025; Mizumoto, 2026) and qualitative works (Alexander et al., 2023; Angeles et al., 2024; Weber-Wulff et al., 2023), providing a solid scientific backing for AI detection and plagiarism analysis.

2.6 Declaration on the Use of Generative Artificial Intelligence (GenAI)

Generative artificial intelligence and assistive technologies were employed strictly for language refinement and preliminary conceptual scoping. Specifically, Scopus AI supported preliminary exploratory keyword discovery during initial query structuring, whereas Google Gemini was utilized exclusively for stylistic copyediting, grammar verification, and English translation refinement. Crucially, these artificial intelligence tools did not participate in, influence, or alter the primary data collection, study screening, methodological quality appraisal (MMAT), narrative synthesis, or interpretive intellectual conclusions. In full accordance with Redfame Publishing guidelines and international publication ethics standards (COPE, 2019), the authors independently designed the review protocol, conducted all eligibility screenings, performed data extraction, formulated the conceptual discussion, and assume comprehensive responsibility for the integrity, validity, and authenticity of the final manuscript.

3. Results

Table 3 shows a marked geographical concentration among the 30 selected studies, with North America and Europe accounting for 50% of the corpus, led by the United States (30%). Asia and the Middle East contributed nearly 40%, while Africa and Latin America remained substantially underrepresented. Methodologically, observational designs predominated (56%), mainly involving cross-sectional comparisons of algorithmic performance. Experimental studies represented 36.6%, emphasizing controlled manipulation of textual variables. Qualitative approaches were scarce, reflecting a predominant emphasis on quantitative validation. Regarding sample size, only 13.3% of studies used macro-corpora exceeding 1,000 documents. In contrast, 86.6% relied on controlled samples, generally involving fewer than 500 texts. Technologically, the evaluation landscape showed strong concentration among a limited number of commercial detectors. GPTZero was the most frequently evaluated tool (63%), followed by Turnitin (50%). ZeroGPT and Copyleaks accounted for 30% and 26%, respectively, while specialized classifiers had marginal representation. This convergence facilitates comparability across studies but also reveals recurrent limitations in detector performance. These limitations become particularly evident when systems encounter obfuscation strategies or authentic human-generated writing.

Table 3. General characteristics of the final corpus

Author / Year	Country	Sample	Methodological Design	AI Detection Tools Software Evaluated	Main Results
(Van Vlasselaer et al., 2026)	Belgium	160 synthetic articles and 1,163 real theses.	Mixed design (experimental and observational).	GPTZero, Pangram, Copyleaks, Turnitin.	Pangram was the most accurate (65% strict / 97.5% inclusive); detected AI in 45.5% of real theses (1.9% with >80%).
(Makhmutova et al., 2026)	Uzbekistan and Kazakhstan	116 texts (1,044 manual evaluations).	Experimental (within-subjects, Friedman, Wilcoxon, PCA).	Turnitin, ZeroGPT, Detecting-AI, Copyleaks, Originality.ai, GPTZero, QuillBot, Grammarly, Sapling.	Copyleaks and Originality achieved 100% baseline accuracy; Grammarly was the least reliable (70.6%).
(Fiedler & Döpke, 2025)	Germany	63 instructors evaluating 38 short excerpts.	Experimental (survey experiment, Fisher, Cochran, OLS).	GPTZero, Gwriter, Tom Tlok's tool, Radar Tester.	Instructors and detectors identified AI text barely better than chance (instructors: 57% AI, 64% human).
(Asiri et al., 2025)	Saudi Arabia and India	1 systematic review and abstracts evaluated by 10 experts.	Exploratory descriptive (blinded qualitative-quantitative).	Writer AI, Grammarly, Writefull X, Content Detector AI, ZeroGPT, GPTZero.	Humans identified 68% AI / 86% human. Global sensitivity of detectors on rewritten texts: 23%.
(AL-Rawas et al., 2025)	Malaysia and Iraq	160 dental abstracts evaluated by 4 academics.	Cross-sectional observational (blinded comparative).	GPT-2 Output, Writefull, GPTZero, Turnitin.	GPTZero was the most accurate (92.6%), outperforming GPT-2 (46.8%) and Writefull (50%).
(Hyatt et al., 2025)	United States	190 STEM students (99 essays and 48 blindly evaluated).	Experimental (blind comparison and survey).	GPTZero, Originality AI, Detect GPT, Copyleaks.	Best detectors reached 93–98% accuracy. Consensus reclassified 100% of human false positives.
(K. H. S. Ibrahim et al., 2025)	Kuwait	480 academic texts (EFL humans, AI, paraphrased, and "humanized").	Descriptive comparative observational (Mann-Whitney U, confusion matrices).	Turnitin AI Content Detector (RoBERTa fine-tuned).	Human EFL: 99% specificity. Pure AI: 85% accuracy. Humanized AI: accuracy dropped to 60%.
(Bordalejo et al., 2025)	Canada	3 pre-AI texts of different genres and styles.	Exploratory experimental (textual transformations and blind evaluation).	Turnitin, GPTZero, Grammarly, EditPad, Writefull, ChatGPT, Microsoft Bing Copilot.	Turnitin did not detect paraphrasing tool interventions (0%); Copilot and Writefull altered texts without being detected.
(Arvidsson, 2025)	United States	797 physics lab reports (2019-2023).	Descriptive-comparative and mixed observational (Mann-Whitney U).	Turnitin (AI module) and GPTZero (v2023-11-21).	7.7% of post-ChatGPT reports showed AI language (19% of students). Turnitin demonstrated higher specificity (97.7% vs 51.1% GPTZero).
(Paulet et al., 2025)	United States	100 text samples (human, AI, and mixed).	Quantitative experimental (blind validation and error matrix).	GPTZero, QuillBot AI Detector, Polygraf.AI.	Polygraf.AI stood out with highest accuracy, identifying 36/36 humans, 34/34 AI, and 30/30 mixed with minimal error margin.
(Cheng et al., 2025)	Canada, UK, Australia	US, 150 texts derived from 30 medical introductions in 5 conditions.	Experimental (quantitative repeated measures).	ZeroGPT, PhraslyAI, Grammarly AI Detector (GPTZero excluded in pilot).	Human accuracy: 19%. Tools distinguished the 5 conditions (p<0.001); ZeroGPT and Phrasly with excellent concordance (ICC 0.96).
(Angeles et al., 2024)	Philippines	20 Information Technology students.	Qualitative observational (holistic case study, Zoom interviews, thematic analysis).	GPT-2 Output Detector, GPTZero, Turnitin, Plagscan, Copyleaks.	Mixed attitudes: recognize detection prevents AI dependency, but fear unfair evaluations and loss of feedback.
(Harwood et al., 2024)	Canada	Study 1: 107,805 applicants; Study 2: 138 participants.	Mixed (longitudinal observational and inter-subject experimental).	GPTZero (via Python API).	ChatGPT use produced minimal improvements (+0.53 on 1-9 scale); GPTZero identified only 1.99% of AI texts and labeled 55.78% as mixed.
(Liu et al., 2024)	Hong Kong and US	50 original articles, 50 from ChatGPT, and 50 paraphrased.	Cross-sectional study (observational).	Originality.ai, Turnitin, ZeroGPT, GPTZero, Content at Scale, GPT-2 Output Detector.	Originality.ai was the most accurate (100% in AI and paraphrased). ZeroGPT detected 96% direct and 88% paraphrased.
(Richards et al., 2024)	United Kingdom	90 scripts (60 from students and 20 from ChatGPT).	Observational and experimental (double-blind interviews, and detection analysis).	GPT-2 Detector (v1.5B) and Turnitin AI Detection.	ChatGPT scripts passed undergraduate level (>40%), postgraduate failed (<50%). Turnitin identified 100% of ChatGPT with 0% FP.

Table 3 (Continued)

(Paulet et al., 2024)	United States	100 writing samples (39 human, 27 AI, 34 mixed).	Experimental and blind validation (controlled samples).	GPTZero (Burstiness, Perplexity, GPTZeroX, Shield, Deep Learning).	GPTZero successfully identified 100 samples (67 with 99% accuracy, 24 with 95-98%, 7 with 90-94%, 2 mixed >82%).
(Chaka, 2024a)	South Africa	40 university essays (English L1 and L2, 2nd and 3rd year).	Exploratory and non-experimental quantitative (confusion matrices).	30 free online AI detectors (Copyleaks, Undetectable AI, ZeroGPT, Turnitin, etc.).	Copyleaks and Undetectable AI tied for 1st place (100% human samples detected, TNR = 1.0, FPR = 0); 9 detectors at the bottom.
(Weber-Wulff et al., 2023)	Germany, Latvia, Sweden, Czechia, Mexico, UK	54 test cases (756 individual tests across 14 tools).	Experimental and quantitative with error analysis.	14 tools (Check For AI, Compilatio, Content at Scale, Crossplag, Turnitin, GPT Zero, etc.).	No tool exceeded 80% global accuracy. Turnitin obtained the highest (81%). Accuracy dropped with manual editing (42%) and paraphrasing (26%).
(Alexander et al., 2023)	Cyprus	6 ESL instructors evaluating 4 C1 essays.	Qualitative and exploratory (independent essay analysis).	Turnitin, OpenAI Detector, GPTZero, Crossplag.	Detectors precise on 100% artificial/human texts, ineffective on mixed ones (Turnitin detected only 12% in combined). Instructor accuracy: 33%–66%.
(K. Ibrahim, 2023)	Kuwait	240 essays (120 pre-ChatGPT humans and 120 generated by GPT-3.5).	Observational / descriptive (group comparative).	GPT-2 Output Detector Demo and Crossplag AI Content Detector (RoBERTa).	Both classifiers discriminated with average accuracy of ~89% (89.2% GPT-2 and 88.75% Crossplag).
(Cingillioglu, 2023)	Australia	230 training essays and 150 test essays.	Experimental and quantitative (n-gram BOWs and Machine Learning).	Own ML algorithms (SVM, Naive Bayes, Random Forest, etc.), OpenAI Detector, GPTZero, Copyleaks.	SVM reached 100% accuracy in humans (0 FNs, 92.7% global accuracy, F2 97.2%). Commercial detectors generated between 2 and 5 FNs.
(Hadra et al., 2026)	Oman	192 texts (EFL essays, XSum, AI, and mixed).	Observational / quantitative (confusion matrix, Chi-square, Fisher).	Turnitin and Originality.	Originality best global precision (0.69 vs 0.61). Both failed to identify hybrid texts (recall close to zero).
(Mizumoto, 2026)	Japan	142 first-year students (142 essays of 200–300 words).	Observational / proof-of-concept (local prototype and interviews).	AI-Likeness Inspector (fixed threshold of 3.44 errors per 100 words).	System flagged 1.4% of essays. Interviews revealed both used ChatGPT asking it to add subtle errors to mimic humans.
(AL-Rawas et al., 2026)	Malaysia and Iraq	150 dental abstracts evaluated by 6 young academics.	Cross-sectional study (observational / analytical).	GPT-2 Output, Writefull GPT, GPTZero, Turnitin.	Young academics struggled without differences by university type. GPTZero achieved highest accuracy (90.0%).
(Erol et al., 2025)	Turkey	1,000 texts (250 pre-ChatGPT human and 750 from ChatGPT).	Observational / cross-sectional (ROC and ANOVA analytical comparison).	GPTZero, ZeroGPT, Corrector App, Plagiarism Detector.	Recent ChatGPT versions obtained higher similarity scores ($p < 0.01$). GPTZero with perfect AUC (1.00).
(Kaya Ergen, 2025)	Turkey	100 texts (50 original introductions and 50 from ChatGPT-4o).	Observational and comparative (Mann-Whitney U, Friedman, ROC).	GPTZero, ZeroGPT, CorrectorApp, Writer, Plagiarism Detector, QuillBot.	GPTZero demonstrated highest accuracy (AUC 1.000, 100% sensitivity, 96% specificity). Paraphrasing reduced detection in GPTZero (100% to 23%).
(Ozkara et al., 2025)	US, Canada, Belgium, Italy	32 radiology editorials (16 human and 16 from GPT-4).	Observational and comparative (blind expert evaluation).	ZeroGPT, GPTZero (premium plan), iThenticate.	Unpaired environment: 58% correctly classified; paired: 70%. iThenticate marked median similarity of 3% in AI.
(Gosling et al., 2024)	United States	305 responses (145 from students and 160 from ChatGPT-3.5).	Observational / cross-sectional (analytical text comparison).	Turnitin (AI module).	ChatGPT obtained significantly higher scores ($M = 36.69$) than students ($M = 0.00$). False negatives in AI: 5.3%.
(Zhao et al., 2024)	United States	3,755 LORs and 1,973 SOIs from Fordham University graduate admissions.	Observational / analytical (supervised classification with transformers).	Domain-specific classifier models (LR, NB, BERT, DistilBERT).	Models reached close to 100% accuracy in generated content. BERT and DistilBERT exceeded 99% in specific domain.
(Odri & Yoon, 2023)	France	1 synthetic text (GPT-4) and 1 historical article (1960).	Observational and comparative (tool evaluation and evasion).	11 online detectors (Originality, ZeroGPT, Writer, Copyleaks, Crossplag, GPTZero, etc.).	7 of 11 detectors initially cataloged GPT-4 as human. Modifications reduced detection down to 100% human.

3.1 Empirical Prevalence of False Positives and Incidence of Sociolinguistic Variables

The empirical prevalence of false positives (Type I error) in AI detectors varies substantially depending on algorithmic architecture, decision thresholds, and textual complexity. While investigations by Van Vlasselaer et al. (2026) and Gosling et al. (2024) observe negligible or zero error rates in established commercial systems under pristine baseline conditions, broader evidence indicates marked diagnostic instability across diverse textual contexts. Substantial misclassification risks emerge when evaluating texts produced by non-native English speakers (NNES) or compositions marked by highly structured academic rhetoric. Although K. H. S. Ibrahim et al. (2025) reported 99% specificity in controlled ESL essays, Makhmutova et al. (2026) and Weber-Wulff et al. (2023) observed accuracy degradation when evaluating translated or non-native phrasing, which raised classification errors to 11.1%. This vulnerability appears exacerbated when authentic human submissions incorporate automated grammar and stylistic checkers, a practice documented to yield false-positive classifications in up to 33.0% of verified human manuscripts (Bordalejo et al., 2025).

Across comparable experimental evaluations, the primary literature documents distinct false-positive rate (FPR) intervals across tool categories. Commercial enterprise platforms, most notably Turnitin AI, consistently yield the lowest baseline false-positive ranges, spanning between 0.0% and 2.3% on standardized, native-English undergraduate coursework (Gosling et al., 2024; K. H. S. Ibrahim et al. 2025; Richards et al., 2024); nevertheless, this margin broadens to 11.1%–33.0% when processing non-native English syntax or writing altered by lexical enhancement tools (Bordalejo et al., 2025; Makhmutova et al., 2026; Weber-Wulff et al., 2023). Similarly, specialized detectors such as Copyleaks and Originality.ai exhibit bounded baseline FPRs between 0.0% and 4.0% on direct human benchmarks (Chaka 2024a; Makhmutova et al., 2026), though their diagnostic reliability diminishes in hybrid compositions. In parallel, multi-metric classifiers such as GPTZero demonstrate wider performance dispersion: while controlled evaluations on disciplinary abstracts and longer essays document false-positive rates between 0.0% and 7.4% (AL-Rawas et al., 2025, 2026; Erol et al., 2025; Hyatt et al., 2025; Poullet et al., 2024), this range expands substantially to 30.0%–48.9% when evaluating short excerpts or historical pre-ChatGPT control corpora (Arvidsson, 2025; Poullet et al., 2025). At the most volatile tier, open-access and consumer-grade classifiers (ZeroGPT, Corrector App, and unvetted online detectors) exhibit severe diagnostic instability, displaying widely dispersed false-positive rates ranging from 4.0% to 100.0%, with classification failures disproportionately concentrated among formulaic academic formats and second-language writers (Chaka, 2024a; Odri & Yoon, 2023).

In comparative evaluations between automated software and human evaluators, notable divergences also characterize authorship attribution. Fiedler & Döpke (2025) and Asiri et al. (2025) indicate that academic instructors mistakenly classify up to 35.7% of genuine human abstracts as synthetic, at times exceeding algorithmic error margins due to subjective evaluative heuristics. Concurrently, historical pre-ChatGPT texts confirm that disciplined scholarly writing frequently incurs unfair penalties due to rigid probabilistic metrics such as perplexity, yielding misclassification rates of up to 41.18% on tools such as Corrector (Arvidsson, 2025; Liu et al., 2024; Odri & Yoon, 2023). Furthermore, textual length and model domain specialization represent critical mediating factors: textual samples under eighty words typically preclude reliable probabilistic classification in GPTZero (Poullet et al., 2025), whereas supervised domain-specific transformer models maintain higher specificity across specialized corpora (Zhao et al., 2024). Overall, the analyzed evidence indicates that within predominantly English-language academic settings, these false-positive classifications pose an emerging challenge to linguistic equity and conventional quality assurance frameworks (Alexander et al., 2023; Mizumoto, 2026).

3.2 Technical, Stylistic, and Design Characteristics as Causes of Classification Failures

The findings indicate that technical failures in AI detection originate from the reliance on traditional probabilistic metrics such as perplexity and burstiness, which prove fragile when confronted with real human writing (Chaka, 2024a; Erol et al., 2025; Gosling et al., 2024). Research points out that current language models surpass these metrics, whereas tools with limited parameters drastically reduce their accuracy when comparing small detectors against massive generators (AL-Rawas et al., 2025; K. Ibrahim, 2023; Van Vlasselaer et al., 2026). Similarly, classifier instability and the algorithmic conflation of human and automated patterns introduce recurring detection failures (Angeles et al., 2024), notably, detectors built on transformer architectures such as RoBERTa appear susceptible to syntactic permutations (K. H. S. Ibrahim et al., 2025). Furthermore, formulaic register conventions and bounded domain lexicons constrain natural linguistic variation, occasionally leading to model failure (Asiri et al., 2025; Zhao et al., 2024). Structural elements characteristic of academic manuscripts—including mathematical expressions, tabular data, and enumerated lists—alter baseline perplexity curves, prompting algorithms to mislabel authentic human prose as machine-generated text (Arvidsson, 2025; Ozkara et al., 2025; Richards et al., 2024). In a comparable manner, token predictability in homogeneous sentence designs systematically confuses institutional analysis tools (Cheng et al., 2025).

The effectiveness of commercial detectors is also altered by automated paraphrasing tools and superficial modifications (Kaya Ergen, 2025; Makhmutova et al., 2026). Software such as Wordtune or EditPad artificially raises perplexity,

reducing detection capacity (Bordalejo et al., 2025; Liu et al., 2024), whereas evasion strategies like intentional errors or character substitutions destabilize online platforms (Mizumoto, 2026; Odri & Yoon, 2023; Weber-Wulff et al., 2023). Added to this is the vulnerability derived from discrepancies in decision thresholds across various programs (AL-Rawas et al., 2026). Simultaneously, corpus architecture and document length impact classification stability (Hadra et al., 2026). High sensitivity to formatting parameters and structured test designs triggers a massive volume of false alerts, a situation aggravated in mixed texts where human and synthetic paragraphs intermingle (Harwood et al., 2024; Poullet et al., 2024, 2025). Given this imbalance between precision and recall, it is imperative to adjust cutoff thresholds (Cingillioglu, 2023) or implement triad-based consensus methodologies to reduce false positives to safe margins (Hyatt et al., 2025), thus moving beyond the simple superficial intuition applied by many educators without mastering the system's key metrics (Alexander et al., 2023).

3.3 Normative and Institutional Repercussions on Educational Quality, Ethics, and Equity

The analysis of the evidence identifies that false accusations by AI detectors threaten educational equity and evaluative ethics, violating deontological principles such as the presumption of innocence (Angeles et al., 2024; Cingillioglu, 2023). This situation critically affects vulnerable populations, exposing them to algorithmic biases and unfair evaluations that damage their trajectories (Arvidsson, 2025; Chaka, 2024a; Fiedler & Döpke, 2025; Hyatt et al., 2025; Makhmutova et al., 2026). From an institutional perspective, excessive reliance on punitive automation erodes trust in quality assurance, damaging the reputation and the teacher-student bond (Hadra et al., 2026; Weber-Wulff et al., 2023). Furthermore, the high incidence of errors and the vulnerability to obfuscation techniques demonstrate their inefficiency in protecting integrity, making it necessary to reconsider their isolated use (Alexander et al., 2023; Asiri et al., 2025; Harwood et al., 2024; K. H. S. Ibrahim et al., 2025; Odri & Yoon, 2023; Poullet et al., 2025; Richards et al., 2024). Faced with this reliability crisis, the findings demand a regulatory shift toward hybrid pedagogical approaches, urging universities to redefine integrity and use detectors solely as initial formative alerts (Bordalejo et al., 2025; K. Ibrahim, 2023; Mizumoto, 2026; Poullet et al., 2024; Van Vlasselaer et al., 2026). To mitigate biases and ensure quality, it is proposed to combine expert human supervision with a higher threshold of suspicion (AL-Rawas et al., 2025, 2026; Cheng et al., 2025; Erol et al., 2025; Liu et al., 2024). Thus, the disruption generated by artificial intelligence forces a comprehensive adaptation of institutional regulations, promoting transparency, process evaluation, and the strict supervision of current scientific literature (Gosling et al., 2024; Kaya Ergen, 2025; Ozkara et al., 2025; Zhao et al., 2024).

4. Discussion

The purpose of this systematic review was to analyze the empirical evidence on the incidence of false positives in AI detectors, their causes, and their impact on academic integrity, contributing to a framework of critical evaluation against the hasty adoption of these technologies. When examining prevalence, the results reveal algorithmic variability where reliability depends on the tool and the writer's profile (Alexander et al., 2023; Asiri et al., 2025; Chaka, 2024a; Gosling et al., 2024; K. H. S. Ibrahim et al., 2025; Makhmutova et al., 2026; Mizumoto, 2026; Van Vlasselaer et al., 2026; Weber-Wulff et al., 2023). This prevalence acquires critical significance under Messick's Validity Theory (1990), where uncritical adoption introduces construct-irrelevant variance, penalizing intrinsic characteristics of formal styles and non-native speakers. These findings expand the literature, aligning with research pointing out that the application of automated detectors systematically violates educational equity through false and biased accusations (Liang et al., 2023; Za'rour & Altakhaine, 2026).

Regarding the causes, the corpus suggests that algorithms fail on genuine texts due to their technical dependence on reductionist predictive metrics, such as perplexity and burstiness, which are documented as incompatible with the nature of academic writing (AL-Rawas et al., 2025, 2026; Arvidsson, 2025; Bordalejo et al., 2025; Cheng et al., 2025; Cingillioglu, 2023; Erol et al., 2025; Fiedler & Döpke, 2025; Hadra et al., 2026; Hyatt et al., 2025; K. Ibrahim, 2023; Kaya Ergen, 2025; Liu et al., 2024; Odri & Yoon, 2023; Ozkara et al., 2025; Poullet et al., 2024, 2025; Richards et al., 2024; Zhao et al., 2024). Interpreted through Shannon's (1948), Information Theory, algorithmic classifiers appear to conflate the low-entropy signature of disciplined scientific prose with automated text generation. Beyond corroborating the technical boundaries identified by Verma et al. (2024) and Xu et al. (2024), this synthesis suggests that purely distributional metrics struggle to delineate legitimate computer-assisted editing from end-to-end synthetic generation, exhibiting diminished sensitivity when writers introduce stylistic alterations. Likewise, external studies (Jiang et al., 2024; Sun et al., 2025) indicate that current designs systematically underestimate human language variability.

Regarding regulatory repercussions, the data converge on quality issues, pointing out that punitive use fractures evaluative ethics (Angeles et al., 2024; Harwood et al., 2024). The findings are compared with perspectives on a policing surveillance paradigm that replaces pedagogical guidance and undermines trust (Deep et al., 2025; Giray, 2024). By contrasting these results with broader contexts (Kirsanov et al., 2026; Perkins et al., 2024), it can be argued that unfounded accusations lead to high levels of anxiety and force students to adopt defensive strategies. In agreement with previous literature (Lye

& Lim, 2024), it becomes evident that universities have favored ineffective technological control, neglecting their duty to guarantee fairness, thus making it necessary to remedy accusations and establish frameworks centered on formative assessments.

Despite the contributions of this synthesis, certain methodological limitations warrant consideration. First, regarding the review protocol, the retrieval executed on July 27, 2026, captures a cross-sectional temporal snapshot. While compliant with PRISMA 2020 reproducibility standards (Page et al., 2021), this single-date retrieval cannot reflect ongoing algorithmic recalibrations or updates across proprietary commercial detection software. Second, restricting the search to peer-reviewed, indexed databases ensured verifiable empirical benchmarks and enabled quality appraisal via the MMAT; however, this boundary circumscribed the evidence to formal academic publications, potentially excluding emerging exploratory trials hosted in preprint repositories.

Beyond procedural boundaries, the synthesized literature exhibits notable contextual constraints: a marked geographic concentration in the Global North (predominantly the United States and Western Europe), an almost exclusive focus on English-language academic corpora, an analytical reliance on a narrow set of commercial platforms, and a prevalence of cross-sectional designs with sample sizes under 500 textual artifacts (86.6%). Consequently, claims regarding institutional vulnerability must be contextualized within English-medium assessment environments rather than generalized as an established universal phenomenon across linguistically diverse educational systems (Chaka, 2024b; Ndayayisenga et al., 2026).

To address these gaps, future research could explore several coordinated avenues. On a computational level, empirical inquiries would benefit from exploring continuous auditing frameworks that integrate decentralized, sociolinguistically diverse macro-corpora, which may enable a more nuanced evaluation of Type I errors. In parallel, mixed-methods longitudinal studies are needed to track the sustained erosion of academic identity and the anxiety induced by false-positive classifications. Finally, expanding qualitative inquiry into how cognitive delegation reconfigures institutional integrity could prove instrumental in informing pedagogical scaffolding frameworks that aim to subordinate automated technical scrutiny to sustained human analysis.

5. Conclusions

The present systematic review indicates that the uncritical adoption of AI-generated text detectors introduces considerable risks to evaluative fairness and quality assurance mechanisms, primarily evidenced within English-medium higher education institutions where current research is situated. Empirical findings point to a sustained prevalence of sociolinguistic biases in these tools, which disproportionately penalize non-native writers and the formal writing styles inherent to scientific rigor.

Regarding underlying algorithmic mechanisms, the findings suggest that Type I errors largely stem from classifier reliance on simplified probabilistic indicators, particularly perplexity and burstiness. Because these metrics often align poorly with the standard uniformity of scholarly prose, automated tools frequently misidentify disciplined human academic writing—or the routine use of grammar-enhancement software—as synthetic output.

At the institutional and regulatory level, evidence indicates that the automated and punitive use of these detectors violates the fundamental principles of evaluative ethics and severely erodes the essential trust between faculty and students. Additionally, unfounded accusations of algorithmic plagiarism not only cause significant psychological damage characterized by anxiety over sanctions, but also force students to adopt defensive and evasive writing strategies, fracturing the authenticity of the teaching-learning process.

Ultimately, the value and contribution of this research lie in highlighting the apparent unsustainability of university regulations anchored exclusively in technological solutionism. While not constituting a definitive solution, the analysis suggests that, to safeguard academic integrity, universities might consider integrating authentic assessment models, where technology functions solely as a guiding resource. Under this premise, the subordination of algorithmic analysis to teaching judgment emerges as a hypothetical pathway to preserve equity and evaluative justice in the face of artificial intelligence challenges.

Acknowledgments

The authors thank the Peninsula of Santa Elena State University for its institutional support, as well as those who contributed valuable insights.

Author contributions

Eliana Geomar Leon Abad and Wilson Alexander Zambrano Vélez were responsible for conceptualization. Wilson Alexander Zambrano Vélez and Gioryi Augusto Sornoza Zavala developed the methodology. Wilson Alexander Zambrano Vélez, Rudy Jonathan Parraga Solorzano, and Eduardo Henríquez Mendoza carried out the formal analysis.

Wilson Alexander Zambrano Vélez and Rudy Jonathan Parraga Solorzano performed the investigation. Rosa Elena Niola Sanmartín and Gioryi Augusto Sornoza Zavala were responsible for data curation and validation. Eliana Geomar Leon Abad supervised the study, and Wilson Alexander Zambrano Vélez managed the project administration and drafted the original manuscript. Eliana Geomar Leon Abad, Rudy Jonathan Parraga Solorzano, Rosa Elena Niola Sanmartín, and Eduardo Henríquez Mendoza reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by the Peninsula of Santa Elena State University, grant number 0041.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Redfame Publishing.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

References

- Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting ai-generated writing in second language education in higher education. *Teaching English With Technology*, 2023(2). <https://doi.org/10.56297/BUKA4060/XHLD5365>
- AL-Rawas, M., Abdul Qader, O. A. J., Lin, G. S. S., Beh, Y. H., Sabarudin, M. A., Ang, Y., ... & Noorani, T. Y. (2026). Human versus artificial intelligence: Investigating ability of young academics from research and non-research institutions to identify ChatGPT-generated dental research abstracts. *Scientific Reports*, 16(1), 12287. <https://doi.org/10.1038/s41598-026-42555-3>
- AL-Rawas, M., Qader, O. A. J. A., Othman, N. H., Ismail, N. H., Mamat, R., Halim, M. S., ... & Noorani, T. Y. (2025). Identification of dental related ChatGPT generated abstracts by senior and young academicians versus artificial intelligence detectors and a similarity detector. *Scientific Reports*, 15(1), 11275. <https://doi.org/10.1038/s41598-025-95387-y>
- Amnie, A. G. (2026). AI-resistant and AI-resilient assessment in higher education: A systematic review of validity-grounded strategies, institutional frameworks, and equity implications. *Assessment & Evaluation in Higher Education*, 1–23. <https://doi.org/10.1080/02602938.2026.2695377>
- Angeles, C. N., Samson, B. D., Mama, B. R. Z. I., Luriaga, R. L., Delizo, J. P. D., & Ching, M. R. D. (2024). Students' perception of the use of AI Detector System by faculty members in determining the originality of submitted academic requirements. *Proceedings of the 2024 8th International Conference on E-Commerce, E-Business, and E-Government*, 56–61. <https://doi.org/10.1145/3675585.3675587>
- Arvidsson, K. (2025). AI-generated language in introductory physics lab reports in the first year of ChatGPT. *Discover Artificial Intelligence*, 5(1), 408. <https://doi.org/10.1007/s44163-025-00742-7>

- Asiri, S. N., Alhazmi, A. S., & Mehta, V. (2025). *Assessing the Applicability of Artificial Intelligence (AI) in Fabrication and Plagiarism of Scientific Literature in the Field of Orthodontics*. 46(3).
- Bittle, K., & El-Gayar, O. (2025). Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information*, 16(4), 296. <https://doi.org/10.3390/info16040296>
- Bordalejo, B., Pafumi, D., Onuh, F., Khalid, A. K. M. I., Pearce, M. S., & O'Donnell, D. P. (2025). "Scarlet Cloak and the Forest Adventure": A preliminary study of the impact of AI on commonly used writing tools. *International Journal of Educational Technology in Higher Education*, 22(1), 6. <https://doi.org/10.1186/s41239-025-00505-5>
- Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., ... & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 16890. <https://doi.org/10.1136/bmj.l6890>
- Caner, H. N. (2025). AI Detection Tools: A Systematic Review of Empirical Evidence and Their Implications for Education. *Khazar Journal of Humanities and Social Sciences*, 28(4), 234–246. <https://doi.org/10.5782/2223-2621.1580>
- Chaka, C. (2024a). 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 & Teaching*, 7(1), 127–139. <https://doi.org/10.37074/jalt.2024.7.1.33>
- Chaka, C. (2024b). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching*, 7(1). <https://doi.org/10.37074/jalt.2024.7.1.14>
- Cheng, A., Lin, Y., Reedy, G., Joseph, C., Wirkowski, S., Mallette, V., Nagesh, V., Krieser, D., & Calhoun, A. (2025). Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. *Advances in Simulation*, 10(1), 66. <https://doi.org/10.1186/s41077-025-00396-6>
- Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *The International Journal of Information and Learning Technology*, 40(3), 259–268. <https://doi.org/10.1108/IJILT-03-2023-0043>
- COPE Council. (2019). *Discussion document: Predatory Publishing*. Committee on Publication Ethics. <https://doi.org/10.24318/cope.2019.3.6>
- Deep, P. D., Edgington, W. D., Ghosh, N., & Rahaman, Md. S. (2025). Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education. *Information*, 16(10), 905. <https://doi.org/10.3390/info16100905>
- Delina, L. L., & Del Rosario Rondilla, A. (2026). Struggling to learn or learning to struggle? Student identity, anxiety, and academic integrity in AI-assisted writing. *Learning, Media and Technology*, 1–17. <https://doi.org/10.1080/17439884.2026.2692945>
- Doğan, A. R., & Doğan, A. İ. (2026). From the Turing Test to AI detectors: An epistemological mismatch in scholarly publishing. *Brazilian Journal of Anesthesiology (English Edition)*, 76(2), 844740. <https://doi.org/10.1016/j.bjane.2026.844740>
- Duah, J. E., & McGivern, P. (2024). How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. *The International Journal of Information and Learning Technology*, 41(2), 180–193. <https://doi.org/10.1108/IJILT-11-2023-0213>
- Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., ... & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica*, 167(1), 214. <https://doi.org/10.1007/s00701-025-06622-4>
- Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, 49, 100321. <https://doi.org/10.1016/j.iree.2025.100321>
- Giray, L. (2024). The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism. *The Serials Librarian*, 85(5–6), 181–189. <https://doi.org/10.1080/0361526X.2024.2433256>
- Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond Policing: AI Writing Detection Tools, Trust, Academic Integrity, and Their Implications for College Writing. *Internet Reference Services Quarterly*, 29(1), 83–116. <https://doi.org/10.1080/10875301.2024.2437174>
- Gosling, S. D., Ybarra, K., & Angulo, S. K. (2024). A widely used Generative-AI detector yields zero false positives. *Aloma: Revista de Psicologia, Ciències de l'Educació i de l'Esport*, 42(2), 31–43. <https://doi.org/10.51698/aloma.2024.42.2.31-43>

- Gritsai, G., Voznyuk, A., Grabovoy, A., & Chekhovich, Y. (2024). *Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2410.14677>
- Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). *PRISMA2020*: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, 18(2), e1230. <https://doi.org/10.1002/cl2.1230>
- 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(1), 4. <https://doi.org/10.1007/s40979-026-00213-1>
- Harwood, H., Roulin, N., & Iqbal, M. Z. (2024). “Anything you can do, I can do”: Examining the use of ChatGPT in situational judgement tests for professional program admission. *Journal of Vocational Behavior*, 154, 104013. <https://doi.org/10.1016/j.jvb.2024.104013>
- Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., ... & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4), 285–291. <https://doi.org/10.3233/EFI-180221>
- Hyatt, J.-P. K., Bienenstock, E. J., Firetto, C. M., Woods, E. R., & Comus, R. C. (2025). Using aggregated AI detector outcomes to eliminate false positives in STEM-student writing. *Advances in Physiology Education*, 49(2), 486–495. <https://doi.org/10.1152/advan.00235.2024>
- Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: “The Terminator Versus the Machines.” *Language Testing in Asia*, 13(1), 46. <https://doi.org/10.1186/s40468-023-00260-2>
- Ibrahim, K. H. S., Al Otaibi, D., & Sibai, F. N. (2025). The Robustness of AI-Classifiers in the Face of AI-Assisted Plagiarism: The Case of Turnitin AI Content Detector. *International Journal of Computer-Assisted Language Learning and Teaching*, 15(1), 1–27. <https://doi.org/10.4018/IJCALLT.375428>
- Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, 217, 105070. <https://doi.org/10.1016/j.compedu.2024.105070>
- Kaya Ergen, S. (2025). Comparative Detector Analysis for the Identification of Academic Articles Synthesized by Artificial Intelligence in the Field of Ophthalmology. *Beyoglu Eye Journal*, 10(3), 175–180. <https://doi.org/10.14744/bej.2025.70973>
- Kirsanov, O., Kushwah, L., & Selvaretnam, G. (2026). Beyond Detection: How Students Use—and Hide—AI in Online Assessments and What Authentic Tasks Can Do About It. *Journal of Academic Ethics*, 24(1), 14. <https://doi.org/10.1007/s10805-025-09691-3>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Liu, J. Q. J., Hui, K. T. K., Al Zoubi, F., Zhou, Z. Z. X., Samartzis, D., Yu, C. C. H., Chang, J. R., & Wong, A. Y. L. (2024). The great detectives: Humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, 20(1), 8. <https://doi.org/10.1007/s40979-024-00155-6>
- Lye, C. Y., & Lim, L. (2024). Generative Artificial Intelligence in Tertiary Education: Assessment Redesign Principles and Considerations. *Education Sciences*, 14(6), 569. <https://doi.org/10.3390/educsci14060569>
- Makhmutova, A., Sharimbayev, B., Amirzhanov, A., & Shalkarbay-uly, A. (2026). Testing the Limits: Evaluating AI Detectors’ Accuracy and the Impact of Obfuscation Techniques on AI-Generated Text. *Journal of Advances in Information Technology*, 17(3), 438–449. <https://doi.org/10.12720/jait.17.3.438-449>
- Messick, S. (1990). VALIDITY OF TEST INTERPRETATION AND USE. *ETS Research Report Series*, 1990(1), 1487–1495. <https://doi.org/10.1002/j.2333-8504.1990.tb01343.x>
- Mizumoto, A. (2026). Corpus-informed AI-likeness screening for L2 writing: A proof-of-concept field test and two flagged cases. *Applied Corpus Linguistics*, 6(3), 100223. <https://doi.org/10.1016/j.acorp.2026.100223>
- Ndacyayisenga, J. C. Y., Kidega, C., & Aciro Can, L. (2026). The impact of generative artificial intelligence tools on assessment equity for non-native English-speaking medical students: A systematic review. *BMC Medical Education*, 26(1), 989. <https://doi.org/10.1186/s12909-026-09303-7>
- Nikolić, D., & Basta Nikolić, M. (2026). Designing AI-resilient assessment in higher education: A four-pillar conceptual framework. *Frontiers in Artificial Intelligence*, 9, 1841682. <https://doi.org/10.3389/frai.2026.1841682>
- Odri, G.-A., & Yoon, D. J. Y. (2023). Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research*, 109(8), 103706.

<https://doi.org/10.1016/j.otsr.2023.103706>

- Ozkara, B. B., Boutet, A., Comstock, B. A., Van Goethem, J., Huisman, T. A. G. M., Ross, J. S., ... & Castillo, M. (2025). Artificial Intelligence–Generated Editorials in Radiology: Can Expert Editors Detect Them? *American Journal of Neuroradiology*, 46(3), 559–566. <https://doi.org/10.3174/ajnr.A8505>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: Una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, 74(9), 790–799. <https://doi.org/10.1016/j.recesp.2021.06.016>
- Paullet, K., Pinchot, J., Kinney, E., & Stewart, T. (2024). Can GPTZero detect if students are using artificial intelligence to create assignments? *Issues In Information Systems*, 25(3), 165–174. https://doi.org/10.48009/3_iis_2024_113
- Paullet, K., Pinchot, J., Kinney, E., & Stewart, T. (2025). Precision check: A critical look at the reliability of AI detection tools. *Issues In Information Systems*, 26(3), 401–412. https://doi.org/10.48009/3_iis_2025_2025_132
- Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughan, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1), 53. <https://doi.org/10.1186/s41239-024-00487-w>
- Richards, M., Waugh, K., Slaymaker, M., Petre, M., Woodthorpe, J., & Gooch, D. (2024). Bob or Bot: Exploring ChatGPT's Answers to University Computer Science Assessment. *ACM Transactions on Computing Education*, 24(1), 1–32. <https://doi.org/10.1145/3633287>
- Rossi, E. (2023). Introducción a las revisiones sistemáticas y metaanálisis. *Acta Gastroenterológica Latinoamericana*, 53(1), 7–14. <https://doi.org/10.52787/agl.v53i1.291>
- Rudakova, L. V. (2025). Quality assurance in modern higher education. *Ekonomika i Upravljenje: Problemy, Resheniya*, 77(160), 209–218. <https://doi.org/10.36871/ek.up.p.r.2025.07.07.022>
- Schaaff, K., Schlippe, T., & Mindner, L. (2024). Classification of human- and AI-generated texts for different languages and domains. *International Journal of Speech Technology*, 27(4), 935–956. <https://doi.org/10.1007/s10772-024-10143-3>
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Sozon, M., Sia, B. C., Pok, W. F., & Alkharabsheh, O. H. M. (2025). Academic integrity violations in higher education: A systematic literature review from 2013–2023. *Journal of Applied Research in Higher Education*, 17(5), 1454–1468. <https://doi.org/10.1108/JARHE-12-2023-0559>
- Stone, B. W. (2025). Generative AI in Higher Education: Uncertain Students, Ambiguous Use Cases, and Mercenary Perspectives. *Teaching of Psychology*, 52(3), 347–356. <https://doi.org/10.1177/00986283241305398>
- Sun, L., Yang, Y., & Song, Y. (2025). AI-Generated Essay Identification: A Machine Learning Approach and Fairness Analysis. 2025 *IEEE Frontiers in Education Conference (FIE)*, 1–5. <https://doi.org/10.1109/FIE63693.2025.11328327>
- Van Vlasselaer, M., Van Droogenbroeck, F., & Spruyt, B. (2026). Who wrote this? Evaluating the reliability of AI detection tools in higher education. *International Journal for Educational Integrity*, 22(1), 16. <https://doi.org/10.1007/s40979-026-00226-w>
- Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2024). Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1702–1717. <https://doi.org/10.18653/v1/2024.naacl-long.95>
- 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), 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Xu, Z., Xu, R., & Sheng, V. S. (2024). ChatGPT-Generated Code Assignment Detection Using Perplexity of Large Language Models (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23688–23689. <https://doi.org/10.1609/aaai.v38i21.30527>
- Za'rour, R., & Altakhaineh, A. R. M. (2026). False Positive Risk in AI Detection of L2 Academic Writing: A Case Study. *Discourses on Culture*, 25(1), 37–78. <https://doi.org/10.2478/doc-2026-0003>
- Zhao, Y., Borelli, A., Martinez, F., Xue, H., & Weiss, G. M. (2024). Admissions in the age of AI: Detecting AI-generated application materials in higher education. *Scientific Reports*, 14(1), 26411. <https://doi.org/10.1038/s41598-024-77847-z>