Khazar Journal of Humanities and Social Sciences

Volume 28 | Issue 4 Article 2

12-31-2025

AI Detection Tools: A Systematic Review of Empirical Evidence and Their Implications for Education

Halime Nuran CANER
Akdeniz University, Antalya, TURKİYE, nurancaner@akdeniz.edu.tr

ORIGINAL STUDY

AI Detection Tools: A Systematic Review of Empirical Evidence and Their Implications for Education

Halime Nuran CANER

Akdeniz University, Antalya, TURKİYE

ABSTRACT

The rapid advancement of generative artificial intelligence has significantly transformed academic writing practices, prompting institutions to implement tools designed to verify authorship and uphold academic integrity. Artificial intelligence detection systems have emerged as a prominent, albeit increasingly debated response to these challenges. This systematic review synthesizes empirical evidence to assess the reliability, fairness, and pedagogical implications of artificial intelligence text-detection tools in educational settings. Adhering to PRISMA 2020 standards, this review identified twenty-five peer-reviewed empirical studies, 18 of which were conducted directly within educational settings. This review synthesizes empirical studies published between 2022 and 2025, encompassing quantitative, qualitative, and mixed-methods designs across diverse disciplinary and linguistic contexts. The findings indicate that AI text detection tools are unsuitable for high-stakes academic integrity decisions in their current form. Furthermore, there is substantial variability and instability in detection accuracy across tools, genres, and linguistic backgrounds; a noticeable weakness in paraphrasing, translation, and other adversarial techniques; and systemic biases that disproportionately affect non-native English writers. Human judgment was also found to be inconsistent, reinforcing the difficulty in reliably distinguishing AI-generated text from human-authored text. Collectively, these results raise significant ethical, pedagogical and institutional concerns. This review underscores the need for integrity strategies that prioritize transparency, AI literacy, fairness-aware design, and process-based assessment rather than relying on detection-centered approaches. The findings suggest the necessity of hybrid approaches that combine watermarking and fairness-aware detection algorithms with process-oriented assessment, AI literacy initiatives, and cross-linguistic benchmarking, alongside interpretability-focused and longitudinal research on students' perceptions of AI detection.

Keywords: AI detection tools, Generative AI, Academic integrity, Systematic review, PRISMA, Authorship verification

Received xx xx xxxx; revised xx xx xxxx; accepted xx xx xxxx.
Available online 31 December 2025

E-mail address: nurancaner@akdeniz.edu.tr (H. N. CANER).

https://doi.org/10.5782/2223-2621.1580

Introduction

The advent of large language models (LLM) has significantly transformed academic writing, knowledge production, and assessment practices in higher education. As generative artificial intelligence (GenAI) tools facilitate the creation of fluent and contextually appropriate texts with remarkable ease, concerns regarding authorship, academic integrity, and the responsible use of technology have intensified (Baron, 2024; Kumar, 2024). Institutions globally, ranging from major research universities to national quality assurance bodies, are actively revising their academic integrity frameworks to address the pedagogical and ethical challenges introduced by GenAI.

The initial applications of artificial intelligence (AI) in education predominantly concentrated on adaptive tutoring and analytics-driven personalization, emphasizing efficiency and individualization rather than authorship verification or integrity enforcement. As generative AI (GenAI) tools have advanced in capability and accessibility, the focus has shifted from supporting learning processes to addressing authorship, accountability, and academic justice issues.

AI detection tools have emerged as a significant institutional response to these concerns. Broadly defined, AI detectors are algorithmic systems, typically stylometric classifiers or neural probability models, designed to estimate the likelihood that a given text was generated by an AI system. Commercial detectors such as *Turnitin*, *Copyleaks*, *GPTZero*, *Originality.ai*, and *ZeroGPT* rely on patterns of lexical distribution, syntactic regularity, and model-specific signals to infer authorship. However, these systems remain largely unclear to end users, offering probabilistic outputs without clearly articulating how such judgments are produced or how they should be interpreted in the assessment context. In this review, "AI use detection" refers to attempts to infer whether a given text was authored by a human or produced with generative AI, typically through stylometric or probabilistic classifiers trained on corpora of human- and AI-generated text.

Despite their rapid adoption, AI detection systems have not advanced at the same pace as Gen AI. Empirical research consistently highlights significant variability in detector accuracy, instability across various genres and disciplines, and marked susceptibility to basic evasion strategies such as paraphrasing, translation, or stylistic obfuscation (Elkhatat et al., 2023; Walters, 2023). This situation has been characterized as a technological arms race, wherein advancements in generative models and evasion tactics consistently surpass the incremental improvements in AI detection systems. As newer large language models (LLMs) produce increasingly human-like text, detection tools face challenges in distinguishing between AI-generated and human-authored writing, thereby diminishing their reliability over time. Of particular concern is the growing body of evidence indicating that AI detectors excessively misclassify work produced by multilingual and non-native English writers, raising significant issues regarding equity, procedural fairness, and the potential for harm to already marginalized student groups (Ibrahim, 2023; Jiang et al., 2024; Liang et al., 2023).

Significantly, these challenges extend beyond mere technical performance. Scholars often report uncertainty regarding the interpretation of detection scores, while students express anxiety about the possibility of being falsely accused in high-stakes assessment contexts (Lieberman, 2024). The vagueness of privately operated algorithms complicates transparency and appeals processes, resulting in a misalignment between institutional expectations for detection and the empirical realities of detector performance. Additionally, there is a growing pedagogical concern that detection-centered approaches may shift learning environments towards surveillance, deterrence, and compliance rather than fostering writing development, reflective practice, and AI literacy.

While AI detection has emerged as a central topic in discussions of academic integrity, the empirical research landscape remains fragmented and unevenly distributed across disciplines. Existing studies and conceptual reviews provide valuable insights into ethical and policy-related challenges (Bittle & El-Gayar, 2025; GarcíaLópez & TrujilloLiñán, 2025). However, a comprehensive synthesis that consolidates empirical findings on the performance of AI detection tools, specifically within educational contexts, is currently lacking. This gap constrains institutions' ability to make informed decisions regarding the integration of detection tools into integrity frameworks.

In response to this need, the current systematic review consolidates peer-reviewed empirical research on AI text-detection tools, with a specific emphasis on studies conducted in educational contexts. Adhering to the PRISMA 2020 guidelines, this review initially identified and evaluated empirical studies that assessed the performance of AI detection tools. Subsequently, it scrutinizes the evidence related to detection accuracy, reliability, and bias. Furthermore, it evaluates the implications of these findings for educational practice and academic integrity policy, offering recommendations for researchers, educators, and institutions as they navigate the challenges presented by the use of Gen AI in academic settings. These objectives are operationalized through the following research questions.

- **RQ1.** What empirical methodologies have been employed to examine AI detection tools within educational settings?
- **RQ2.** How effective are AI detection tools in discerning AI-generated texts, according to the studies reviewed?
- **RQ3.** How are these tools characterized in terms of their fairness, reliability, and usability?
- **RQ4.** What technical, ethical, and pedagogical limitations have been identified?
- **RQ5.** What critical issues and research priorities emerge from synthesizing the current evidence on AI detection in education?

This review synthesizes findings from various disciplines, methodological approaches, and linguistic contexts to establish an evidence-based framework for understanding the limitations and potential applications of AI detection technologies in educational settings. It contends that institutional responses to Gen AI should be rooted not in technological optimism or punitive measures, but in transparent, equitable, and pedagogically informed practices that align with the realities of contemporary academic work.

Methodology

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021) to ensure transparency, rigor, and replicability in identifying and synthesizing empirical studies on AI detection tools in the educational domain. The methodological process comprised four primary stages: identification, screening, eligibility assessment and inclusion. The review protocol was meticulously designed to ensure transparency, reproducibility, and comprehensiveness in synthesing empirical research on AI detection tools.

Search strategy

A thorough search was conducted across prominent academic databases—ScienceDirect, Scopus, and Web of Science—employing Boolean combinations of keywords pertinent to generative AI, AI detection tools, academic integrity, and education. The search terms

included phrases such as *AI detection*, *AI-generated text*, *ChatGPT detection*, *authorship verification*, *academic integrity*, and *artificial intelligence* (*AI*). These combinations were strategically designed to encompass both established terminologies and emerging descriptors in the rapidly evolving field of Gen AI.

To ensure comprehensive coverage, database searches were augmented with reference snowballing and targeted citation tracing. This approach facilitated the inclusion of recently published or marginally indexed empirical studies that may not yet have been fully integrated into major indexing systems. The search was confined to publications from 2022 to 2025, a period marked by the emergence of large language models and rapid advances in AI detection tools. The primary search was conducted in September 2025, with subsequent citation tracing to identify additional eligible studies. This multilayered strategy ensured that the review incorporated the most current and pertinent empirical work available.

Inclusion and exclusion criteria

Studies were included in this review if they were published in peer-reviewed journals, employed empirical methodologies (quantitative, qualitative, or mixed-methods), and examined either the performance of AI text detection tools or the human ability to distinguish between AI-generated and human-authored writing. Studies were deemed eligible if they focused on educational settings or provided findings with clear relevance to educational practice. Conceptual or opinion-based publications, theoretical analyses, reviews and meta-analyses, and other non-empirical reports were excluded. Conference proceedings, dissertations, non-peer-reviewed white papers, and preprints that had not undergone peer review were excluded because of their less standardized peer-review processes compared with those of indexed journals. Furthermore, studies focusing on AI detection in non-educational domains (e.g., purely clinical applications without a teaching or assessment focus) were omitted to preserve the review's educational scope. Finally, studies addressing the detection of non-textual AI outputs, such as images or videos, were excluded to maintain a clear focus on text-based AI detection.

Screening, selection, coding, and reliability

Following the removal of duplicate entries, titles and abstracts were screened for relevance according to the inclusion criteria, followed by a comprehensive full-text review to evaluate methodological appropriateness and empirical rigor. Each study was systematically coded for context, methodological design, dataset type, language profile, detectors evaluated, and principal findings of the study. The coding categories were developed using a combined deductive–inductive approach to integrate established analytical frameworks with sensitivity to emerging patterns of data. Ultimately, 25 empirical studies satisfied the final inclusion criteria, encompassing a diverse array of educational and disciplinary contexts, including ESL and multilingual writing, medical and scientific writing, physics education, large-scale exam assessments, and broader higher education environments. Although the coding was conducted by a single researcher owing to resource limitations, the procedures were cross-verified against the PRISMA guidelines to enhance internal consistency. Decisions were revisited iteratively to improve internal consistency, drawing on the reliability procedures recommended in the PRISMA-aligned systematic review guidelines.

Data extraction

Data from each study were systematically coded using a structured extraction protocol that documented the study context, disciplinary focus, methodological design, and characteristics of the datasets involved, such as student essays, paraphrased or adversarial texts, and multilingual writing samples. The protocol also recorded the specific detection tools evaluated, including *Turnitin*, *GPTZero*, *Copyleaks*, *ZeroGPT*, and *Originality.ai*, as well as comparisons between human and automated classification performance. Key findings related to accuracy, bias, robustness, and pedagogical implications were extracted to enable a cross-study synthesis. Coding followed a combined inductive–deductive strategy, allowing emergent patterns to complement established analytical categories in the literature and ensuring that the review remained theoretically grounded while being sensitive to novel developments in the detection of Gen AI.

Limitations

While this review sought to offer a comprehensive examination of empirical research on AI text detection in educational contexts, several limitations must be acknowledged. First, the quantity of peer-reviewed empirical studies specifically addressing educational settings remains limited. Although gray literature was utilized during the initial stages of screening to map the landscape, it was largely excluded from the final analysis; nevertheless, its initial use may have introduced minor bias. Furthermore, given the rapid evolution of Gen AI and detection technologies, some findings may be time-sensitive, highlighting the need for ongoing reassessments.

Several methodological constraints influence the interpretation of this evidence. The rapid evolution of AI may diminish the permanence of empirical results, and many included studies rely on small or context-specific datasets, thereby limiting their generalizability across institutions and disciplines. Detection tools are frequently updated, and their privately operated algorithms restrict transparency, complicating efforts to evaluate their reliability and replicate findings. Furthermore, although single-reviewer coding was employed in this review, cross-checking procedures were implemented to enhance internal consistency. Despite these constraints, the synthesis presented here offers the most comprehensive and empirically grounded account currently available of the performance, risks, and pedagogical implications of AI detection tools in the educational context.

Results

The rapid proliferation of generative artificial intelligence in the educational sector has heightened concerns regarding authorship verification and academic integrity. The synthesis of twenty-five empirical, peer-reviewed studies in this review uncovers distinct and recurring patterns related to the performance, fairness, and educational implications of contemporary AI detection systems. Despite their growing institutional adoption, empirical evidence indicates that current tools remain technically unstable, susceptible to manipulation, and systematically biased, particularly in linguistically and disciplinarily diverse contexts.

PRISMA flow outcomes

The search process yielded 142 records, from which 109 titles and abstracts were screened after removing duplicates. A total of 76 studies were subjected to full-text review, of which 25 met the eligibility criteria. The excluded studies predominantly comprised conceptual or commentary-only works, non–peer-reviewed sources, and analyses not relevant to AI detection. The final corpus encompasses areas such as higher education, disciplinary writing, ESL/EAP contexts, scientific communication, and large-scale standardized assessment.

Overview of included empirical studies

Within the corpus, research has been conducted across various domains, including educational technology, applied linguistics, computer science, health sciences, and academic integrity. The methodologies employed encompassed quantitative benchmarking, adversarial testing, human–AI classification comparisons, and field-based evaluations of writing practices among students and educators. The tools frequently analyzed included *Turnitin*, *GPTZero*, *Copyleaks*, *Originality.ai*, and *ZeroGPT* (e.g., Abd-Elaal et al., 2022; Elkhatat et al., 2023; Walters, 2023).

Eighteen studies were conducted in educational settings, involving areas such as ESL writing (Ibrahim, 2023), medical education (Liu et al., 2024), physics education (Yeadon et al., 2023), and large-scale assessment validation (Jiang et al., 2024). These investigations provide significant insights into AI detector accuracy, robustness, linguistic bias, and the limitations inherent in human judgment.

Benchmarking studies frequently compare AI-generated outputs, such as those from GPT-3.5 and GPT-4, with human-authored texts, including first-year composition essays, laboratory reports, and research abstracts. These human-written texts were often collected prior to the public release of Gen AI tools.

Numerous studies have also developed adversarial test sets by employing techniques such as paraphrasing, synonym substitution, and automated text-obfuscation tools to assess the robustness of detectors. Table 1 provides an overview of the included empirical studies.

Thematic synthesis of findings

Variability and instability in detection accuracy

Empirical research has consistently demonstrated significant variability in accuracy across various tools, genres, languages and generative models. Benchmarking studies have revealed that detection systems frequently yield false positives and false negatives, with some performing only slightly better than random chance (Elkhatat et al., 2023; Walters, 2023). Comprehensive multilingual evaluations (Weber-Wulff et al., 2023) indicate that these detectors exhibit inconsistent behavior across different linguistic profiles, often failing when applied to multilingual or highly formal academic texts.

When analyzed within specific disciplinary contexts, these challenges become increasingly apparent. Research in the fields of medicine (Liu et al., 2024; Erol et al., 2025) and physics education (Yeadon et al., 2023) has demonstrated that detectors frequently misclassify technical, structured writing, underscoring the discrepancy between the detector training data and actual academic genres. Under controlled conditions, some tools initially appeared to be accurate; however, their performance significantly declined when applied to authentic student works.

Table 1. Overview of empirical peer-reviewed studies on AI detection.

Study	Context/Discipline	Method	Dataset	Detectors / Tools	Key Findings
Abd-Elaal et al. (2022)	Higher Education	Quant	Human/AI essays	Turnitin-style tools	Academics struggle to detect AI writing
Alharthi et al. (2025)	Computer Science	Quant	Mixed AI datasets	Hybrid neural networks	Improved accuracy through feature fusion
Baron (2024)	Educational Integrity	Mixed	Student essays	Turnitin, others	Detection scores unreliable/misinterpreted
Casal & Kessler (2023)	Applied Linguistics	Quant	Linguistics texts	Expert judgments	Experts perform poorly distinguishing AI text
Chaka (2023)	General	Quant	Multi-model outputs	5 detectors	Large accuracy variance
Cingillioglu (2023)	Higher Education	Quant	AI essays	Turnitin & others	Detectors struggle with ChatGPT writing
Dalalah & Dalalah (2023)	Higher Education	Quant	Human vs AI	Various detectors	High false-positive/negative rates
Elkhatat et al. (2023)	Higher Education	Quant	Short prose	Several detectors	Inconsistent accuracy; false positives
Elkhatat (2023)	Higher Education	Quant	ChatGPT responses	Multiple tools	Human + AI judgments unreliable
Erol et al. (2025)	Higher Education	Quant	Medical writing	Commercial detectors	Accuracy limitations in scientific prose
Fishchuk & Braun (2024)	Medical Writing	Quant	Adversarial texts	Black-box detectors	Easily bypassed by simple attacks
Fleckenstein et al. (2024)	Computational Linguistics	Quant	Student essays	Human judges	Teachers misclassify AI texts frequently
Gao et al. (2023)	K–12/HE	Quant	Scientific abstracts	Humans + detectors	Both groups misclassify frequently
Ibrahim (2023)	Medicine	Mixed	ESL essays	GPTZero, others	High false positives for ESL writers
Jiang et al. (2024)	Scholarly Communication	Quant	36,000+ essays	GPTZero, Turnitin	Bias against non-native English writers
Liu et al. (2024)	ESL Writing	Quant	Medical writing	Human vs AI detectors	Unreliable across academic prose
Liang et al. (2023)	Writing Assessment	Quant	L1/L2 English	Multiple detectors	Systemic bias against L2 writers
Lieberman (2024)	Medical Education	Qual	Faculty interviews	N/A	Confusion interpreting detection scores
Perkins (2023)	General	Qual	Academic staff	N/A	Faculty unsure how to use detectors
Perkins et al., 2024a)	Higher Education	Mixed	Student essays	Detectors + humans	Combined judgment still limited
Perkins et al., 2024b)	Higher Education	Quant	Paraphrased texts	Multiple detectors	Simple paraphrasing defeats detectors
Popkov & Barrett (2024)	Higher Education	Quant	Behavioral writing	Various detectors	Inconsistent accuracy
Pratama (2025)	Behavioral Health	Quant	Multilingual data	Various detectors	Strong accuracy–bias trade-offs
Walters (2023)	Computer Science	Quant	16 detectors	Wide corpus	Large variation; some near random
Weber-Wulff et al. (2023)	General	Quant	Multilingual academic texts	12+ detectors	Instability across languages, models
Yeadon et al. (2023)	Multilingual HE	Quant	Physics essays	Human + detectors	Short essays prone to false positives

Upon examining the reviewed literature, it becomes evident that instability in detection accuracy is a widespread and recurring issue that consequently compromises the reliability of procedures intended to maintain academic integrity.

Vulnerability to paraphrasing, translation, and evasion techniques

The second prominent theme pertains to the ease with which AI-generated text can evade detection. Research utilizing paraphrasing, synonym substitution, or other forms of text obfuscation consistently demonstrates significant reductions in detection accuracy (Fishchuk & Braun, 2024; Perkins et al., 2024). Even minor lexical or syntactic alterations, including translation cycles (L1→L2→L1), render many detectors ineffective.

These vulnerabilities collectively highlight a significant "moving target" issue, wherein generative AI technologies advance at a considerably faster rate than detection systems, resulting in detectors consistently lagging behind the former. The synthesized evidence regarding susceptibility to paraphrasing, translation, and evasion techniques suggests that current detection systems fail to reliably identify AI-generated text following realistic student revisions, thereby questioning their capability in real-world educational settings.

Fairness and bias in misclassification patterns

Empirical evidence consistently indicates that AI detectors excessively identify multilingual and non-native English writers as AI-generated. Large-scale studies have revealed elevated false-positive rates for L2 English writers (Ibrahim, 2023; Jiang et al., 2024; Liang et al., 2023), while multilingual benchmarks have corroborated the tendency of detectors to misconstrue lower lexical diversity or predictable syntactic structures as "AI-like" (Weber-Wulff et al., 2023).

Bias was also evident in discipline-specific writing. Genres such as engineering, medicine, and physics—characterized by formulaicity and a highly technical register—were frequently misclassified as AI-generated (Giray, 2024; Popkov & Barrett, 2024; Yeadon et al., 2023). The synthesis of findings on misclassification patterns raises substantive concerns about equity, especially for multilingual learners and writers in highly structured academic disciplines, who appear to be unreasonably at risk of incorrect flagging by detection systems.

Limitations of human judgment

Numerous studies have investigated the ability of educators, linguists, and subject specialists to differentiate between AI-generated and human-authored writing. Across the corpus, human accuracy ranged from 40% to 55%, which is comparable to random chance (Casal & Kessler, 2023; Fleckenstein et al., 2024). Confidence ratings were not reliable indicators of correctness, and experts did not outperform students. Hybrid human–AI decision systems have yielded only modest improvements, with issues of misclassification and interpretability persisting (Perkins et al., 2024a).

The synthesized evidence regarding the limitations of human judgment indicates that human evaluators do not consistently address the deficiencies of the current detection tools. This inconsistency complicates and, in certain instances, undermines the reliability of approaches that combine human and detector evaluations.

Pedagogical, institutional, and ethical implications

Educators have expressed confusion regarding the interpretation of detection scores and discomfort with the lack of transparency in commercial systems (Lieberman, 2024; Perkins, 2023). Students, particularly multilingual writers, have reported experiencing anxiety due to the potential for false accusations. Settings that focus on detection risk shift the emphasis

of assessment from learning to surveillance, thereby undermining trust and discouraging the development of authentic writing skills (Baron, 2024).

The ethical implications of this are equally concerning. False positives have significant effects on student well-being, due process, and institutional credibility. Collectively, the intersection of bias, opacity, and technical instability indicates that current detection systems risk affecting structural inequity and compromising academic justice. Furthermore, some studies suggest that enforcement focused on detection may incentivize students to intentionally simplify or degrade their writing style to appear 'less AI-like,' potentially hindering the development of advanced academic literacy.

The analysis of pedagogical, institutional, and ethical dimensions indicates that the limited benefits of detection-centered integrity systems, as currently implemented, are surpassed by their significant pedagogical and ethical risks.

Overall synthesis of empirical evidence

Across the twenty-five studies synthesized, four overarching conclusions emerged.

1. The accuracy of detection is inconsistent and contingent upon various factors, including the specific tool, model, language, and genre employed
2. Detection systems are highly vulnerable to evasion, as even basic paraphrasing or translation can significantly undermine their reliability
3. Systemic biases excessively affect multilingual writers and authors who use formulaic disciplinary prose.
4. Human judgment does not offer a dependable alternative, and combined human–AI methodologies remain inadequate.

In summary, empirical evidence indicates that current AI text detection tools lack the necessary robustness, fairness, and interpretability to make high-stakes decisions regarding academic integrity. Consequently, institutions should incorporate detection within broader pedagogy-centered integrity strategies that emphasize transparency, process-based assessment, and equitable practice.

Discussion

The findings of this systematic review indicate that current AI detection tools encounter significant technological, ethical, and pedagogical challenges that undermine their appropriateness for high-stakes academic integrity decision making. An analysis of empirical studies reveals a consistent pattern: commercially available detectors such as GPTZero, Copyleaks, Turnitin's AI module, Originality.ai, and ZeroGPT demonstrate unstable accuracy, systematic linguistic bias, low robustness to adversarial strategies, and limited interpretability (Elkhatat et al., 2023; Walters, 2023; Weber-Wulff et al., 2023). Collectively, these deficiencies highlight that current systems are not yet equipped to provide reliable or equitable authorship judgments in educational settings.

A primary technological challenge lies in the evidenced instability of detection accuracy across various genres, linguistic varieties, and disciplinary writing contexts. Research consistently highlights significant variability in false-positive and false-negative rates, sensitivity to surface-level stylistic features, and rapid performance degradation when texts are subjected to even minimal paraphrasing, translation, or synonymous word insertion (Fishchuk & Braun, 2024; Perkins et al., 2024b). This vulnerability is particularly evident in medical, scientific, and ESL/EAP writing, where deviations from algorithmic training

norms result in systematic misclassifications (Liu et al., 2024; Erol et al., 2025; Yeadon et al., 2023). Consequently, the technological landscape is volatile, with detectors frequently failing to generalize authentic writings, thereby limiting their educational value and effectiveness.

Empirical evidence indicates a constant, ethically significant bias against multilingual writers. In several large-scale assessments and benchmarking studies, non-native English writers were disproportionately identified as AI-generated (Ibrahim, 2023; Jiang et al., 2024; Liang et al., 2023; Weber-Wulff et al., 2023). Characteristics of L2 academic writing, such as reduced lexical diversity, less idiomatic phrasing, and more predictable syntactic structures, are often misinterpreted as "LLM-like" patterns. These systematic disparities raise significant equity concerns, particularly given the potential severity of academic misconduct allegations against affected students. The intersection of multilingualism, disciplinary genre conventions, and detector instability suggests that these systems inadvertently preserve the linguistic hierarchies embedded in their training data (Popkov & Barrett, 2024).

Empirical research indicates that human judgment does not serve as a dependable corrective measure. Educators, linguists, and other subject matter experts consistently perform at chance levels when attempting to differentiate between AI-generated and human-authored texts (Casal & Kessler, 2023; Fleckenstein et al., 2024). There was a weak correlation between confidence ratings and accuracy, and evaluators often overestimated their ability to identify AI-generated content. Even hybrid human–AI decision-making frameworks yield only marginal improvements (Perkins et al., 2024a). These findings challenge the assumption that professional intuition can effectively compensate for the limitations of AI algorithms.

The pedagogical and institutional implications of these limitations are extensive. Educators have reported experiencing confusion and uncertainty regarding the interpretation of AI detection scores, while opaque proprietary classification processes impede transparency and procedural fairness (Lieberman, 2024; Perkins, 2023). Students, particularly multilingual learners, express increased anxiety about false accusations and inconsistent enforcement. Reliance on detection-centered policies risks transforming assessment cultures towards surveillance and risk avoidance, undermining opportunities for writing development, reflective practice, and AI literacy. Furthermore, the ease with which detectors can be bypassed through paraphrasing or translation incentivizes evasive behaviors that further detract from authentic learning experiences.

Collectively, the evidence suggests that AI text detection tools, in their current iteration, are unsuitable for making high-stakes academic integrity decisions. Their instability, bias, and lack of transparency are at odds with the fundamental principles of due process, proportionality, transparency, and academic justice. Instead, the extant literature advocates comprehensive institutional strategies that prioritize pedagogical integrity over technology enforcement. These strategies encompass the integration of AI literacy into curricula, promotion of process-based assessment frameworks, establishment of clear and student-centered AI use policies, and adoption of fairness-aware research and development practices for future detection models.

In conclusion, empirical evidence suggests that AI detectors may serve as valuable low-stakes contextual indicators; however, they should not be regarded as conclusive proof of misconduct. Ensuring sustainable academic integrity in the era of Gen AI necessitates a shift from detection-focused responses to strategies that integrate ethical oversight, transparency, instructional redesign, and pedagogical support. Only through such comprehensive frameworks can institutions effectively navigate the challenges and opportunities presented by Gen AI while upholding fairness, inclusivity, and meaningful learning.

These findings indicate that AI detection scores should not be regarded as self-evident or definitive values. The opacity of probability values and 'AI-likelihood' percentages often lead to misinterpretation by both faculty and students. Consequently, institutions require explanatory frameworks that highlight uncertainties, contextual evidence, and transparent reasoning. Future technical efforts should focus on producing interpretable outputs that clarify the rationale for flagging a text, rather than depending solely on singular numerical scores.

From a pedagogical standpoint, the evidence underscores the significance of implementing trust-building interventions, engaging in meta-discussions concerning artificial intelligence, and employing process-oriented assessments, such as drafts, reflections, and in-class writing. These strategies can alleviate anxiety related to false accusations, particularly among multilingual writers, and reframe Gen AI as a subject of critical inquiry rather than merely a source of risk.

Conclusion

This systematic review synthesizes empirical evidence from twenty-five peer-reviewed studies that investigated the accuracy, fairness, and pedagogical implications of AI text-detection tools in educational settings. Across various disciplinary domains, linguistic backgrounds, and methodological designs, the findings converge on a significant and consequential insight: current AI detectors lack the robustness, transparency, and equity necessary to support high-stakes academic integrity decisions. The accuracy of these tools varies significantly across different contexts and tools, fails with minimal paraphrasing or translation, and remains particularly unreliable for multilingual writers and authors of highly structured disciplinary prose. These patterns reflect not isolated technical flaws but systemic limitations inherent to detector architecture, training data, and the rapid pace of technological change.

In addition to these challenges, there are significant human and institutional constraints. Educators and domain experts, similar to automated systems, often demonstrate low accuracy in determining authorship and frequently express uncertainty in interpreting opaque or probabilistic detection outputs. This convergence of technical instability, inequitable misclassification, and limited interpretability highlights the risks associated with treating detection scores as definitive evidence of research misconduct. As the reviewed studies consistently illustrate, such reliance threatens due process, overly impacts marginalized students, and undermines trust in academic integrity procedures.

In the context of GenAI, maintaining academic integrity necessitates a significant shift from surveillance-oriented or punitive measures to comprehensive and pedagogically informed strategies. Educational institutions should emphasize assessment designs that highlight the writing process, iterative development, reflection, and explicit AI literacy. Policies must prioritize fairness, transparency, and student support, positioning detection tools—if employed at all—as low-stakes contextual indicators rather than definitive determinants of authorship. Moreover, when detection tools are employed, their function should be limited to serving as low-stakes contextual indicators that encourage dialogue and further investigation rather than being used as conclusive evidence in high-stakes determinations of misconduct.

Future research should focus on advancing explainable and fairness-aware detection methods, developing multilingual and discipline-specific benchmark datasets, and exploring alternative technical solutions, such as watermarking or provenance-based approaches. Equally important are longitudinal and ethnographic studies that investigate how detection

practices affect students' trust, perceptions of fairness, and engagement with writing and AI technologies. Ultimately, maintaining academic integrity in an AI-pervasive landscape will rely less on outpacing technological innovation and more on cultivating assessment ecologies that promote equity, meaningful learning and institutional accountability.

Responsible Use Statement for Artificial Intelligence Tools: During the preparation of this work, the author used artificial intelligence (AI) to improve the linguistic clarity and scholarly tone of the manuscript. After using Paperpal, the author reviewed and edited the content as needed and takes full responsibility for the final manuscript. All research design, data analysis, and substantive interpretations remain the original work of the author.

References

Abd-Elaal, E., Gamage, S. H. P. W., & Mills, J. E. (2022). Assisting academics to identify computer-generated writing. *European Journal of Engineering Education,* 47(5), 725–745. https://doi.org/10.1080/03043797.2022.2046709.

Alharthi, R., Ojo, S., Nathaniel, T.I., Abdel Samee, N., Umer, M., Jamjoom, M.M., Alsubai, S., & Khan, J. (2025). Responsible detection and mitigation of AI-generated text using hybrid neural networks and feature fusion: Toward trustworthy content management in the era of large language models. *International Journal of Computational Intelligence Systems,* 18(1), 274, 1–16. https://doi.org/10.1007/s44196-025-01025-w.

Baron, P. (2024). Are AI detection and plagiarism similarity scores worthwhile in the age of ChatGPT and other generative AI? *Scholarship of Teaching and Learning in the South*, 8(2), 151–179. https://doi.org/10.36615/sotls.v8i2.411.

Bittle, K., & El-Gayar, O. (2025). Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information*, 16(4), 296. 1–15. https://doi.org/10.3390/info16040296.

Casal, J. E. & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing? *Research Methods in Applied Linguistics*, 2(3), 1–12. https://doi.org/10.1016/j.rmal.2023.100068.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and ChatSonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching,* 6(2). https://doi.org/10.37074/jalt.2023.6.2.12.

Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *International Journal of Information and Learning Technology*, 40(3). 259–268. https://doi.org/10.1108/IJILT-03-2023-0043.

Dalalah, D. & Dalalah, O. M. A. (2023). 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). 1–13. https://doi.org/10.1016/j.ijme.2023.100822.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AIgenerated text. *International Journal for Educational Integrity*, 19(17). 1–16. https://doi.org/10.1007/s40979-023-00140-5.

Elkhatat, A. M. (2023). Evaluating the authenticity of ChatGPT responses. *International Journal for Educational Integrity*, 19(15), 1–23. https://doi.org/10.1007/s40979-023-00137-0.

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, 214. 1–12. https://doi.org/10.1007/s00701-025-06622-4.

Fishchuk, V., & Braun, D. (2024). Robustness of generative AI detection: Adversarial attacks on blackbox neural text detectors. *International Journal of Speech Technology,* 27(4), 861–874. https://doi.org/10.1007/s10772-024-10144-2.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence,* 6, 100209. 1–9. https://doi.org/10.1016/j.caeai.2024.100209.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1), 75. 1–5. https://doi.org/10.1038/s41746-023-00819-6.

GarcíaLópez, I. M., & TrujilloLiñán, L. (2025). Ethical and regulatory challenges of generative AI in education: A systematic review. *Frontiers in Education,* 10, 1565938. 1–13. https://doi.org/10.3389/feduc.2025.1565938.

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.

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). 1–28. https://doi.org/10.1186/s40468-023-00260-2.

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. 1–14. https://doi.org/10.1016/j.compedu.2024.105070.

Kumar, R. (2024). Beyond reproach: Navigating usage, detection, and future pathways of AI in education. *Brock Education Journal*, 33(3), 22–29. https://doi.org/10.26522/brocked.v33i3.1173.

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, Article 8. https://doi.org/10.1007/s40979-024-00141-8.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns,* 4(7), 1–4. https://doi.org/10.1016/j.patter.2023.100779.

Lieberman, G. (2024). The use and detection of AI-based tools in higher education. *Journal of Instructional Research,* 13, 70–80. https://doi.org/10.9743/jir.2024.13.4.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International journal of surgery*, 88, 105906. https://doi.org/10.1016/j.ijsu.2021.105906.

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice,* 20(2). https://doi.org/10.53761/1.20.02.07.

Perkins, M. Roe, J., Postma, D., McGaughran, J. & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics* 22(1). 89–113. https://doi.org/10.1007/s10805-023-09492-6.

Perkins, M., Roe, J., Vu, B.H., Postma, D., Hickerson, D., McGaughran, 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(53), 1–25. https://doi.org/10.1186/s41239-024-00487-w.

Popkov, A. A., & Barrett, T. S. (2024). AI vs academia: Experimental study on AI text detectors' accuracy in behavioral health academic writing. *Accountability in Research: Policies & Quality Assurance*, 32(7), 1072–1088. https://doi.org/10.1080/08989621.2024.2331757.

Pratama A. 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.

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science,* 7(1), 1–24. https://doi.org/10.1515/opis-2022-0158.

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(26), 1–39. https://doi.org/10.1007/s40979-023-00146-z.

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education,* 58(3), 1–13. https://doi.org/10.1088/1361-6552/acc5cf.