Review

Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education

Promethi Das Deep *, William D. Edgington, Nitu Ghosh, Md. Shiblur Rahaman *

Posted Date: 30 July 2025

doi: 10.20944/preprints202507.2233.v1

Keywords: generative AI; AI detection tools; higher education; ethical implications

 doi:10.20944/preprints202507.2233.v1

Review

Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education

Promethi Das Deep [1,*], William D. Edgington [2], Nitu Ghosh [3] and Md. Shiblur Rahaman [4,5,*]

[1] Department of Educational Leadership, College of Education, Sam Houston State University, Huntsville, TX 77341-2119, USA

[2] School of Teaching and Learning, Sam Houston State University, SHSU Box 2119, Huntsville, TX 77341, USA

[3] Department of English, College of Humanities and Social Sciences, Sam Houston State University, Huntsville, TX 77341-2146, USA

[4] Department of Public Health, College of Health Sciences, Sam Houston State University, Huntsville, TX 77341-2177, USA

[5] Department of Environmental Science and Disaster Management, Noakhali Science and Technology University, Noakhali 3814, Bangladesh

* Correspondence: pxd033@shsu.edu (P.D.D.); mxr291@shsu.edu or shiblu@nstu.edu.bd (M.S.R.)

Abstract

The rapid rise of generative AI tools such as ChatGPT has prompted significant shifts in how higher education institutions approach academic integrity. Many universities have implemented AI detection tools like Turnitin AI, GPTZero, Copyleaks, and ZeroGPT to identify AI-generated content in student work. This qualitative evidence synthesis draws on peer-reviewed journal articles published between 2021 and 2024 to evaluate the effectiveness, limitations, and ethical implications of AI detection tools in academic settings. While AI detectors offer scalable solutions, they frequently produce false positives and lack transparency, especially for multilingual or non-native English speakers. Ethical concerns surrounding surveillance, consent, and fairness are central to the discussion. The review also highlights gaps in institutional policies, inconsistent enforcement, and limited faculty training. It calls for a shift away from punitive approaches toward AI-integrated pedagogies that emphasize ethical use, student support, and inclusive assessment design. Emerging innovations such as watermarking and hybrid detection systems are discussed, though implementation challenges persist. Overall, the findings suggest that while AI detection tools play a role in preserving academic standards, institutions must adopt balanced, transparent, and student-centered strategies that align with evolving digital realities and uphold academic integrity without compromising rights or equity.

Keywords: generative AI; AI detection tools; higher education; ethical implications

1. Introduction

In recent years, the integration of generative artificial intelligence (GenAI) tools into higher education has prompted educators and administrators to reevaluate traditional approaches to learning and assessment [1–3]. This wave of technological advancement differs from previous digital tools in its ability to autonomously produce high-quality, human-like academic outputs [4–6]. As these tools become more embedded in educational settings, institutions face growing pressure to develop effective responses that ensure both innovation and academic integrity [1,7,8].The discussion now centers not on whether AI belongs in education, but on how to implement it responsibly through clear policies, inclusive pedagogical strategies, and ethically grounded practices [2–4].

 doi:10.20944/preprints202507.2233.v1

1.1. Generative AI in Higher Education: Evolution, Impact, and Institutional Response

Artificial intelligence (AI), which once served mainly to automate basic tasks or process data, has rapidly evolved into a set of systems that can replicate advanced human functions like language, reasoning, and composition [9]. This transformation became especially noticeable in higher education with the release of large language models (LLMs) like ChatGPT [10]. These generative models can produce coherent, grammatically correct, and contextually appropriate essays, responses, and academic texts, often indistinguishable from those written by students [10]. Generative tools such as ChatGPT are AI co-authors that allow individuals to generate complete essays with limited original writing. Although they can be used to boost productivity, they pose severe threats to academic integrity, authorship, and student learning, especially in assessment environments where originality and writing competencies are crucial [11].

The use of AI in education goes beyond ChatGPT. Earlier tools like Grammarly and Quillbot focused on grammar correction, style enhancement, and paraphrasing [1]. However, ChatGPT can now independently generate large amounts of academically comparable text and that has prompted institutions to rethink their teaching and policy approaches [1]. As the potentially unauthorized use of AI as an academic tool grew, institutions began adopting AI detection systems. GPTZero, Turnitin's AI writing indicator, and OpenAI's classifier are designed to identify texts likely generated by machines by analyzing linguistic features and token-level entropy [10]. These tools were quickly integrated into academic integrity protocols, especially in universities where writing-heavy assessments are common [10].

1.2. Assessing the Effectiveness of AI Detection Tools and Ethical Issues in Teaching and Practice

While detection programs offer a technological fix, intense debates exist about their effectiveness and soundness [10]. A recent study evaluated 14 AI text detection tools and found their accuracy inconsistent, especially when analyzing AI-generated content that has been manually edited or paraphrased [12]. The authors concluded these tools are generally unreliable and perform poorly with obfuscated text. Adopting open-source AI models poses enforcement challenges, as they can be operated on local machines or even third-party servers, making them difficult to monitor [12]. Essay mills might use such tools to evade detection, particularly from jurisdictions lacking extensive legal cooperation [13]. Contract cheating refers to situations where third parties improperly assist or complete a student's work, often in exchange for money [13]. In the United States, 17 states, including California, Maryland, Florida, Massachusetts, and Pennsylvania, have enacted laws that make contract cheating services, like essay mills, illegal. Contract cheating laws, as they exist today, are ill-equipped to control general-purpose AI but are already in use in certain nations. However, existing contract cheating laws are insufficient to address the challenges posed by general-purpose AI tools, such as ChatGPT. The paper recommends revisiting these laws to differentiate between misuse and legitimate use, proposing measures like watermarking, safe harbors, and institutional collaborations to support responsible AI use [13]. Non-native English students are most at risk for such mistakes because of the similarities in syntax and structure with AI-written language. These flaws weaken detection tools' credibility and call into question their admissibility as evidence in academic conduct proceedings [14].

Beyond technical correctness issues, ethical concerns regarding computer detection are central to the argument [15]The unilateral deployment of AI detection software could breach students' due process rights, especially in ESL situations where language usage could be similar to AI-produced writing [16]The issue here is fairness and transparency, since a lack of proficiency in English may increase the risk for false positives. Schools should have explicit, inclusive policies to notify all, particularly ESL, students regarding their use of detection strategies to enforce academic integrity ethically and fairly [17]. Heavy dependence on AI detection tools may lead to a surveillance-oriented approach to education, ultimately eroding trust between instructors and students [18]. They argue that such tools are inherently flawed and risk unjustly penalizing students due to false positives. Rather than relying on detection technologies, the authors recommend establishing thoughtful

 doi:10.20944/preprints202507.2233.v1

assessment strategies that encourage meaningful student engagement and guide learners in using AI as a constructive, problem-solving resource [18].

Broader use of AI co-writing tools, which enable students to author complete essays while contributing minimally, puts at risk perceived notions of authorship and intellectual integrity [11]. Teachers, she contends, should embed such technologies in creative ways to uphold, not compromise, intellectual standards [11]. Students tend to be bewildered or inconsistently informed about the acceptable usage of such tools. Inconsistently practiced policies in uncertain learning environments tend to foster doubts about fairness and have an unequal effect on already vulnerable student populations [11]. Such environments punish exploration and deter intellectual curiosity, both of which are bedrocks of higher education. As campuses contend with these tensions, academics debate the implementation of a more integrated and pedagogic response to AI [19]. Instead of solely punitive enforcement, some argue for integrating AI ethics in curriculum planning, providing guidelines for clear usage, and creating assessments that accommodate collaboration with AI while maintaining academic integrity [20,21].

1.3. Identifying Gaps in Existing Research and Rationale for Study

Despite the growing adoption of AI-detection tools in higher education, their accuracy and reliability remain questionable [22]. Commonly used detectors such as OpenAI's classifier, Copyleaks, GPTZero, and others often produced inconsistent results, particularly when evaluating advanced outputs from ChatGPT-4. Notably, some tools yielded false positives by misclassifying human-written content as AI-generated [22]. These limitations raise significant concerns about fairness and due process in academic integrity proceedings. The authors emphasize the need for continued refinement of detection technologies and caution against relying on these tools as the sole basis for determining authorship in educational contexts [22].

Although an ever-growing volume of literature about AI in education has been generated, much of the research on detection tool research to this point is of limited scope and lacking in institutional breadth [2]. A recent study demonstrated this in detailing one such qualitative study comprised of a mere six ESL lecturers, highlighting narrow empirical breadth and ongoing needs for broader policy-focused research to inform institutional responses to student work produced via AI [2]. Correspondingly, researchers emphasize the necessity to depart from reactive institution-level responses in urging large-scale, theory-informing studies to investigate in what ways AI-enabled misconduct is framed through institutional policies, behavioral patterns among students, and international education environments [7]. Collectively, such investigations identify an ongoing need to conduct additional comprehensive, comparative research to guide ethically informed and practically effective frameworks in academic integrity [7]. A contemporary investigation into the reliability of AI-detection technologies highlighted notable constraints, such as the exclusive use of ten free tools and a limited sample of AI-generated and paraphrased texts. The researchers also pointed out that commercial versions may yield greater sensitivity and proposed that future studies incorporate subscription-based software to enhance detection accuracy [23]. Building on prior critiques, a recent scholarly analysis underscored that much of the existing research is based on artificial test scenarios rather than authentic educational contexts. The authors contend that this approach masks the complexity of detecting AI use in real-world classroom settings, where such behavior is often subtle and multifaceted. These research and policy shortcomings carry critical implications for students' engagement with and understanding of AI technologies in educational environments [24]. While instructors and students express mixed attitudes toward AI in education, the literature focuses predominantly on faculty responses. Students' real-world experiences, especially their struggles with unclear institutional guidelines on AI usage, remain underexplored and insufficiently documented [18]. Students may struggle to discern when using AI authoring tools is appropriate, indicating a lack of clarity around ethical usage at different stages of the learning process [11]. ESL instructors often judged grammatically accurate and sophisticated writing as AI-

 doi:10.20944/preprints202507.2233.v1

generated, revealing a bias against second-language writers and highlighting how current detection systems may unintentionally penalize linguistic diversity [2].

Filling the existing literature gap, this research examines how universities embrace AI-based text-detection tools. The research is founded upon peer-reviewed scholarly publications that have emerged recently and is organized along three principal research topics.

1. Institutional Implementation of AI Detection Tools in Academic Integrity Frameworks and Assessment Systems
2. Evaluating the Effectiveness and Limitations of Current AI Detection Tools in Distinguishing AI-Generated Content from Student Work
3. The Ethical, Procedural, and Pedagogical Concerns that AI Detection Systems Raise for Students and Faculty in Higher Education

Through this inquiry, the study aims to contribute to a more critical and comprehensive understanding of AI's role in academic integrity, one that recognizes both the technological promise, and the human complexity embedded in higher education systems.

2. Materials and Methods

This research employs the narrative literature review approach. Narrative reviews are distinct from systematic reviews in accommodating diverse studies and viewpoints without demanding stringent inclusion or exclusion criteria [25]. In comparison with seeking extensive or exhaustive coverage, narrative reviews seek to yield significant synthesis through subjective critique and interpretation, guided by researchers' contextual, organizational, or historical perspectives [25]. Narrative reviews, rooted in interpretivist and subjectivist paradigms, are suited for investigating the sophisticated features of novel educational practices. The approach enables the refinement of research questions and scope adjustment through the iterative review process, allowing the synthesis to be reactive to the complexity and diversity of the literature [26]. To this effect, the current review is based on peer-reviewed publications accessed through ERIC, EBSCOhost, and JSTOR, with attention to literature that discusses the pedagogical and ethical impacts of ChatGPT in tertiary education.

Table 1 presents the Boolean search plan employed to find peer-reviewed literature reporting on generative AI tools like ChatGPT and their roles in postsecondary environments. The operators and keywords were selective in capturing varied threads, such as accuracy for detecting AI, scholarly integrity, moral issues, perceptions among academics and students, and teaching reactions. This organized plan secured an extensive yet targeted analysis of peer-reviewed research considering both technological and pedagogic facets of GPTs within university environments.

Table 1. Boolean Search Strategy: Keywords, Operators, and Rationale.

Keyword/Concept	Boolean Operators	Purpose/Focus
AI Detection Tools	"AI detection tools" OR "AI writing detectors" OR "AI content detection"	To identify literature discussing technologies designed to detect AI-generated content in educational contexts.
Higher Education Context	"Higher education" OR "university" OR "tertiary education" OR "college"	To narrow the scope to post-secondary education environments where such tools are being evaluated and implemented.

 doi:10.20944/preprints202507.2233.v1

Academic Integrity	"Academic integrity" AND "plagiarism" OR "misconduct" OR "academic honesty"	To capture studies exploring how AI detection tools intersect with integrity frameworks and institutional misconduct policies.
Generative AI Tools	"Generative AI" OR "ChatGPT" OR "large language models" OR "GPT-3" OR "GPT-4"	To focus on content related to the technologies most relevant to recent shifts in academic writing and integrity concerns.
Specific Detection Tools	"Turnitin" OR "GPTZero" OR "ZeroGPT" OR "Copyleaks" OR "CrossPlag"	To retrieve empirical studies that tested or reviewed commonly used detection platforms in educational contexts.
Detection Accuracy & Effectiveness	"accuracy" OR "detection performance" AND "false positives" OR "false negatives"	To locate studies that evaluated the technical reliability of AI detectors, particularly in distinguishing human vs. machine-generated texts.
Student and Faculty Perceptions	"Student perception" OR "faculty perception" AND "AI detection" OR "AI surveillance"	To understand user attitudes, concerns, and ethical views toward institutional use of AI detection technologies.
Ethical Implications	"ethics" OR "ethical implications" AND "privacy" OR "fairness" OR "consent" OR "bias" OR "equity"	To uncover scholarship discussing AI detection's moral and social dimensions, especially around data use, consent, and equity in enforcement.
Assessment and Curriculum Redesign	"Assessment design" OR "curriculum redesign" AND "AI-generated writing" OR "ChatGPT"	To identify pedagogical responses to AI, including innovative assessments that reduce reliance on detection and emphasize human-centered learning.
AI Literacy and Responsible Use	"AI literacy" OR "responsible AI use" AND "teaching practice" OR "student support"	To gather conceptual and practice-oriented papers promoting student engagement with AI tools in ethical and educationally beneficial ways.

Table 2 shows the selection criteria for including and excluding studies for this narrative review. The requirements are designed to concentrate on current, peer-reviewed, English-language articles published between 2021 and 2024 that discuss AI detection tools such as GPTZero and Turnitin AI in post-secondary contexts. Priority was given to having a transparent methodology, relevance for stakeholders, and consideration of ethics to maintain rigor and soundness concerning the study's research questions.

 doi:10.20944/preprints202507.2233.v1

Table 2. Inclusion and Exclusion Criteria with Rationale.

Criteria	Inclusion	Exclusion	Rationale
Publication Date	Peer-reviewed articles published between 2021 and 2024.	Articles published before 2021.	Ensures the review reflects the post-ChatGPT era, significantly accelerating the deployment and institutional scrutiny of AI detection tools in higher education. The 2021–2024 window aligns with the rapid technological and policy transformation period in academia.
Language	Studies published in English.	Non-English publications.	English was selected as the review language to maintain linguistic consistency, avoid translation inaccuracies, and reflect the dominance of English in scholarly discourse, particularly in AI and higher education research domains.
Peer-Review Status	Only peer-reviewed journal articles and conference proceedings included.	Excludes preprints, white papers, blogs, media articles, and non-peer-reviewed sources.	Peer-reviewed sources ensure methodological rigor and quality, according to the SANRA (Scale for the Assessment of Narrative Review Articles) criteria for evaluating the included literature.
Focus on Higher Education	Articles focused on university-level or tertiary education settings, including institutional, student, or faculty perspectives.	Research on K–12 education, vocational training, or non-academic sectors.	Ensures alignment with the study's scope—AI detection tools are used within university-level academic integrity systems, not other educational levels or industries.
AI Detection Systems	Articles discussing AI writing detection tools (e.g., Turnitin AI, GPTZero, Copyleaks, ZeroGPT, CrossPlag, OpenAI classifiers).	Articles discussing AI in education generally and which do not reference detection tools.	The tool-specific review assesses detection systems' technical, ethical, and procedural roles in enforcing academic integrity.

 doi:10.20944/preprints202507.2233.v1

Thematic Relevance	Studies that explore: (a) institutional policy integration, (b) effectiveness and accuracy of detection tools, or (c) ethical and pedagogical impacts.	Studies lacking focus on institutional policy, accuracy, or ethical/pedagogical implications.	These themes correspond directly to the three research questions (RQ1–RQ3) guiding the review and structure the analysis across policy, performance, and pedagogy/ethics.
Methodological Clarity	Research with explicit methodological frameworks, such as case studies, comparative evaluations, theoretical models, or structured reviews.	Articles lacking any straightforward methodological approach or lacking an evaluative structure.	This ensures replicability, transparency, and analytical depth, especially where empirical claims about tool performance or institutional practice are made.
Stakeholder Involvement	Studies involving university stakeholders: students, faculty, administrators, policy makers, or institutional frameworks.	Articles focusing on AI tool developers, industry marketing, or technical benchmarking without a user perspective.	It focuses on the academic community, which is most impacted by AI detection, ensuring that findings remain grounded in educational practice, student experience, and faculty decision-making.
Ethical and Social Dimensions	Articles addressing fairness, surveillance, student rights, linguistic bias, transparency, or AI misuse/misidentification.	Papers focused purely on technical algorithms or architecture without discussing social or ethical implications.	Ethical risk is a significant pillar of the review. Including studies that engage with due process, trust, and academic freedom ensures a holistic assessment of the tools' impact, beyond raw accuracy.

Table 3 overviews prominent studies on generative AI and scholarly integrity in postsecondary schools. These consist of various geographical locations, focal groups (e.g., students, educators, policy makers), and methodologies, varying from empirical investigations to conceptual examinations. The table identifies each study's aim, research design, and principal findings, providing a comparative understanding of how tools such as ChatGPT and discovery tools are transforming scholarly practices, understandings, and integrity models within international postsecondary environments.

 doi:10.20944/preprints202507.2233.v1

Table 3. Overview of Key Studies on Generative AI and Academic Integrity in Higher Education.

No	Reference (In-Text)	Study Location	Target Group	Research Objective	Research Approach	Principal Outcomes
1	[1]	Universities in Pakistan	University students	To examine the causes and consequences of ChatGPT usage among university students	Quantitative	Academic workload and time pressure increase ChatGPT use; reward-sensitive students use it less; use is linked to procrastination, memory loss, and lower CGPA.
2	[4]	Saudi Arabia	Higher education students	To explore students' awareness, use, impact perception, and ethical considerations of Generative AI (GenAI) tools in academic research	Quantitative survey	Students reported high awareness and positive experiences with GenAI, optimism about its future, and strong ethical concerns. However, training/support was limited.
3	[2]	Cyprus	ESL lecturers in higher education	Evaluate the effectiveness of AI detectors and human judgment in ESL assessment.	Qualitative	The paper reports that AI detectors (e.g, Turnitin, GPTZero, Crossplag) were generally more accurate than the ESL lecturers, particularly in identifying fully AI-generated essays
4	[19]	Global	Higher education faculty and students	Critically analyze generative AI detection tools' effectiveness, vulnerabilities, and ethical implications in academic assessments.	Narrative analysis	AI detectors are unreliable, biased, and misaligned with educational goals. The study advocates for replacing detection tools with robust, AI-inclusive assessment frameworks.

 doi:10.20944/preprints202507.2233.v1

5	[7]	Global	Commentary	Conceptual analysis using the situational crime prevention (SCP) framework, supplemented by academic misconduct literature.	Theoretical analysis	AI-misconduct can be tackled using opportunity-reduction strategies from SCP, including redesigning assessments, increasing detection risks, reducing temptations, and clarifying institutional policies.
6	[3]	Global	Students and Educators in Higher Education	To explore the benefits and challenges of AI chatbots (e.g., ChatGPT) in education	Narrative Review	AI chatbots can support adaptive, real-time, and personalized learning and teaching, but raise concerns about ethics, academic integrity, misinformation, and accessibility
7	[27]	Global	Students and Educators in Higher Education	To evaluate the accuracy and ethical concerns surrounding AI-content detection tools	Narrative Review	Limitations include false positives, a lack of transparency, and fairness concerns; Turnitin AI, GPTZero, and Copyleaks were analyzed.
8	[22]	UAE	University Students	To explore students' perceptions of using ChatGPT in academic writing, particularly regarding fairness, bias, and trust	Quantitative	Most students acknowledged ChatGPT's usefulness, but raised concerns over fairness, accuracy, and trustworthiness.
9	[13]	Global	Legal scholars, policymakers, and	Engaged in rigorous legal and policy analysis, reviewing	Conceptual Commentary	LLM providers may unintentionally fall under contract cheating laws; legal frameworks are unclear and need reform to

 doi:10.20944/preprints202507.2233.v1

			education regulators.	statutes, AI services, and advertising practices		protect legitimate AI use while targeting essay mills.
10	[28]	Not Applicable — the study is conceptual	Assessment professionals, educators, and students (Both K–12 and higher education)	Explore how LLMs and generative AI affect educational assessment	Conceptual	AI improves test creation and scoring efficiency but introduces concerns around fairness, validity, and security.
11	[9]	Not location-based; global focus	General readers, scholars, and business leaders	To provide a historical overview, current applications, and future outlook of AI in business and society	Narrative review/editorial	The article outlines the historical development of AI and its current applications in business contexts like HR, marketing, and decision-making. It also discusses future challenges related to ethics, regulation, and societal impact.
12	[14]	Kuwait	University ESL students	Test the effectiveness of AI detectors for identifying AI-generated plagiarism.	Empirical descriptive comparative study	Crossplag showed greater sensitivity to machine-generated texts, as reflected by a higher concentration of scores
13	[29]	India	Academic integrity experts	To assess how well free AI-detection tools can identify AI-generated and AI-paraphrased academic texts	Quantitative	Detection capability varied across tools; a few accurately identified original and paraphrased AI-generated content, while others failed to detect AI-generated text reliably.
14	[30]	Poland	AI-generate	To evaluate and compare	Quantitative	TF-IDF and Bag of Words with simple classifiers

 doi:10.20944/preprints202507.2233.v1

			d vs. student-written essays	different numerical text representation methods for classifying AI-generated essays		effectively detected AI-generated essays, while pretrained models performed poorly.
15	[31]	Not location-specific	Higher education (implicitly)	To evaluate how paraphrasing affects the performance of AI-text detectors and to test retrieval-based defense mechanisms	Experimental	Paraphrasing weakens detection tools, but retrieval-based methods offer a more reliable way to identify AI-generated content
16	[5]	Global	Scholars & publishers	To assess the ethical, practical, and academic implications of using ChatGPT in scholarly publishing.	Conceptual review	Identified benefits (efficiency, editing) and ethical concerns (bias, authorship, plagiarism)
17	[8]	Ecuador	Higher Education (EFL Students)	To explore Ecuadorian EFL students' perceptions of generative AI (e.g., ChatGPT) in academic writing, particularly regarding academic dishonesty, usage, and institutional responses.	Quantitative	Students partially understood AI misuse, overestimated detection tools, and favored ethical guidance.

 doi:10.20944/preprints202507.2233.v1

18	[32]	Malaysia & Singapore	Higher education students	To assess the performance and limitations of AIGC detectors in identifying AI-generated code in education contexts	Empirical (Quantitative)	AIGC detectors struggled to distinguish between AI- and human-written code. Most tools frequently misclassified code, indicating a need for improved detection models tailored to programming content. Sapling and GLTR showed relatively better performance with specific code variations.
19	[17]	Vietnam	Higher Education Institutions (HEIs), academic staff, and students	Examine academic integrity issues from student use of AI/LLMs like ChatGPT	Literature review	LLMs like ChatGPT produce undetectable, fluent content. Misconduct occurs when their use isn't disclosed.
20	[6]	UK	Higher education students	Compare the performance of ChatGPT-generated essays with human essays analyzing Old English poetry, and evaluate detection accuracy.	Mixed-methods approach	AI essays scored similarly to student essays but lacked depth and cultural insight. Human markers identified AI essays around 79% of the time; AI detectors like Quillbot were even more accurate, at around 95%. The study urges reassessing how academic integrity is upheld as AI tools advance.
21	[18]	USA	Higher Education (Students &	To find out if people can tell the difference between AI-generated and	Quantitative	Instructors struggled to detect AI essays, experience didn't help, and both groups had mixed views on AI use.

 doi:10.20944/preprints202507.2233.v1

			instructors)	student-written essays		
22	[12]	Cross-country	Higher education academic integrity context	Evaluate detection tools for AI-generated text	Quantitative	Tools are mostly inaccurate; Turnitin was the most reliable
23	[11]	Singapore	Higher Education (TESOL educators and second/foreign language learners)	To explore how AI writing tools affect academic integrity, authorship, teaching, and assessment in language education.	Conceptual analysis	AI challenges authorship & integrity. Risk of misuse but learning potential. Teachers should integrate AI ethically.
24	[10]	USA	Higher Education	To detect AI-generated and AI-revised content in Letters of Recommendation (LORs) and Statements of Intent (SOIs)	Quantitative;	Domain-specific AI detectors (e.g., BERT, DistilBERT) achieved near-perfect accuracy in distinguishing human vs AI content in admissions materials; general detectors showed poor cross-domain performance.

2.1. Review Process

The selection process began by checking the relevance of titles, abstracts, and conclusions according to the review criteria. Duplicates were removed, and the remaining articles were screened in full to verify whether they fit within the study's scope. In the end, 24 studies were included. Narrative aspects of the appraisal followed the SANRA (Scale for the Assessment of Narrative Review Articles) criteria, a six-item tool developed for evaluating non-systematic narrative overviews [33]. SANRA covers essential areas such as the importance and purpose of the evaluation, the definition of literature searches, referencing, scientific reasoning, and the provision of relevant explanations. Each item is rated from low (0) to high (2), with a total possible score of 12 points. The original validation of the tool showed satisfactory internal consistency (Cronbach's α = 0.68) and

 doi:10.20944/preprints202507.2233.v1

inter-rater reliability (ICC = 0.77)[33]. To ensure transparency in documenting the selection and screening process, we used the PRISMA-ScR flow diagram PRISMA-ScR (Figure 1) [34].

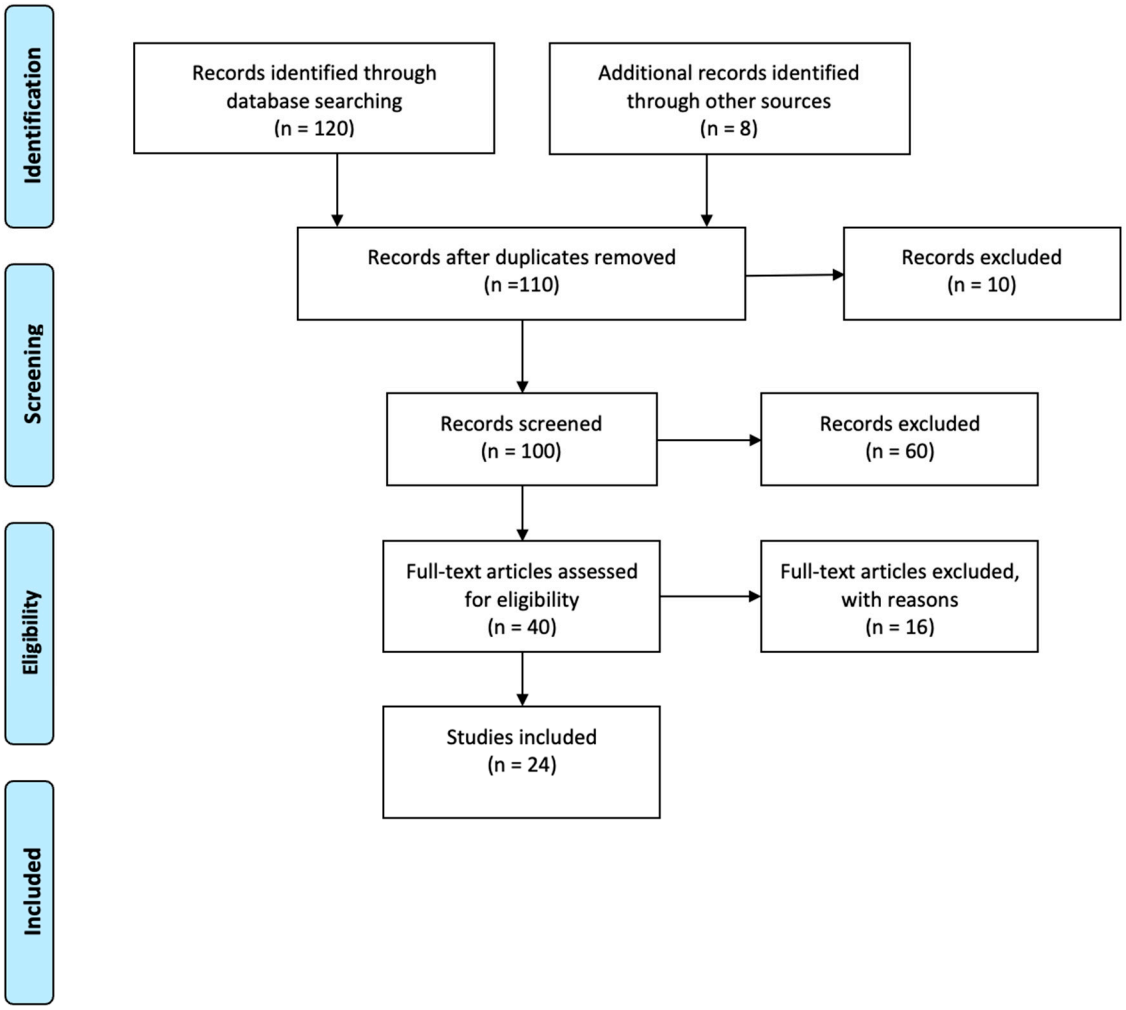

Figure 1. PRISMA Flow Diagram of Study Selection Process.

3. Results

3.1. Screening Summary

The initial database search yielded 128 records. After removing duplicates and conducting a preliminary screening based on titles and abstracts, a subset of articles was assessed in full. Applying the predefined inclusion and exclusion criteria led to the synthesis' final selection of 24 studies. The detailed screening and selection process is visualized in Figure 1.

3.2. Study Characteristics

The 24 papers included varied research designs and methodological approaches. These comprised narrative literature reviews, empirical studies using qualitative and quantitative methods, and some that employed mixed-methods design. The studies represent a broad range of geographical settings and institutional contexts, offering global perspectives on the use and impact of AI detection tools in higher education. Each paper was systematically reviewed to extract key details such as the country of origin, research objectives, methodological design, and observed outcomes. A comprehensive overview is provided in Table 3.

 doi:10.20944/preprints202507.2233.v1

3.3. Thematic Analysis

A qualitative thematic analysis was undertaken to identify recurring concepts and critical issues from the reviewed literature. The study revealed three dominant thematic domains: (1) institutional implementation of AI detection technologies and evolving policy frameworks; (2) technical performance, limitations, and variability in detection accuracy; and (3) ethical, social, and pedagogical concerns related to surveillance, fairness, and student autonomy. These themes were inductively derived and refined to align with this review's three guiding research questions. Table 4 summarizes the thematic categories, illustrating their relevance to the broader discourse on academic integrity and the responsible integration of AI technologies in educational assessment.

Table 4. Overview of Key Themes in AI Detection Tools in Higher Education.

Central Theme (Research Question)	Co-Themes	Structural Interpretation
1. Institutional Policy and Implementation	a. Lack of AI-Specific Academic Integrity Policies	Many institutions have yet to update academic integrity policies to reflect generative AI, creating confusion among faculty and students.
	b. Policy Gaps and Faculty Uncertainty	Faculty members often rely on personal judgment in AI-related cases due to vague or missing guidelines, which can lead to inconsistency.
	c. Student Confusion and Inconsistent Enforcement	Students face unclear expectations, with institutional responses varying by instructor or department, increasing the risk of unfair accusations.
2. Detection Accuracy and System Performance	a. False Positives and False Negatives	AI detectors often misclassify human writing as AI-generated and vice versa, undermining trust in academic decisions.
	b. Tool Performance with Evolving LLMs	Many detection tools are not trained to detect the most recent language models (e.g., GPT-4), limiting their effectiveness.
	c. Gaps in Validation Across Student Populations	Detection tools often perform poorly with diverse writing styles, especially those from multilingual or non-traditional students.
3. Ethical and Pedagogical Implications	a. Transparency and Due Process Issues	Students are rarely informed how their work is analyzed, raising concerns about consent, privacy, and fairness in enforcement.
	b. Equity and Bias Concerns	Linguistic bias leads to disproportionate scrutiny of non-native English speakers, creating equity concerns.

 doi:10.20944/preprints202507.2233.v1

	c. Need for Inclusive Assessment Design	Redesigning assessments (e.g., drafts, oral exams) can reduce reliance on detection tools and promote authentic student work.

4. Discussion

The rapid rise of generative AI has prompted significant shifts in how higher education addresses academic integrity [35]. This discussion synthesizes key findings to evaluate how AI detection tools impact policy, accuracy, ethics, and pedagogy in higher education.

4.1. Institutional Implementation of AI Detection Tools in Academic Integrity Frameworks and Assessment Systems

As AI-generated content becomes more prevalent, institutions are revising academic integrity policies and assessment practices [16]. This section explores how higher education systems implement AI detection tools, adjust pedagogical strategies, and address related ethical and procedural challenges.

4.1.1. Policy Integration and Ethical Framing in the Age of AI

The rise of generative artificial intelligence (GenAI) tools like ChatGPT has prompted a critical shift in the academic integrity frameworks of higher education institutions [35]. One significant development is the explicit incorporation of AI use into institutional academic policies. These updates aim to clarify what constitutes academic misconduct when AI is involved, drawing a line between legitimate support and dishonest automation [17]. Faculty in ESL contexts often feel uncertain about addressing suspected AI-generated student work, highlighting the lack of comprehensive, AI-specific institutional guidelines [2]. Educators rely on subjective judgments without clear policies, sometimes misinterpreting linguistic sophistication as evidence of AI authorship. This underscores the urgent need to formally integrate AI into academic misconduct protocols [2].

Recent study recommends using situational crime prevention (SCP) models to combat AI-linked academic misconduct [7]. Their framework encourages proactive measures like raising the effort necessary to cheat, raising the perceived risk of discovery, and creating ethical choices by institutional culture [7]. They draw upon the established SCP model, which contains 25 techniques grouped within five main principles: raising effort, raising risk, reducing rewards, reducing provocations, and taking away excuses. Translated into educational environments such as supervised tests, randomized oral defenses, tracking of plagiarism, and declarations of authenticity, institutions can diminish opportunity for misconduct. This represents a strategic move away from discipline-based reaction toward organized deterrence, attaching prevention of AI misuse with time-tested crime reduction methods [7].

Student attitudes are instrumental in determining the success of policies aimed at preserving academic integrity. Although most students know that misuse of AI would be damaging to their learning, 87% showed an inadequate comprehension of the ramifications, including diluted understanding and surface-level scholarly success [8]. Students cited pressure from academia, lack of confidence, and dependence on AI among the most prevalent motivations for dishonest use. These observations led to the conclusion that policies should concentrate on more than merely addressing punishment, and priority should be given to AI literacy and appropriate usage, assisting students in utilizing these tools in accordance with established scholarly guidelines [8].

4.1.2. Technological Adoption and the Effectiveness of AI Detection Tools

AI detection tools are increasingly being implemented in higher education as part of a broader strategy to uphold academic standards in the face of advancing GenAI-generated content [12]. However, as multiple studies have shown, their reliability and integration into academic workflows

 doi:10.20944/preprints202507.2233.v1

remain inconsistent [12,32]. AI detectors in an ESL context and found notable discrepancies between the judgments made by humans regarding the origin of the content versus what machine detectors indicated. The authors noted that all the detection tools produced reports indicating the probability that the content had been generated by AI [2]and highlighted their probabilistic and often inconsistent nature. ESL instructors often relied on superficial linguistic cues, mistakenly equating fluency with artificiality. At the same time, AI detection tools showed inconsistent results and lacked transparency in their classification criteria, complicating their role in making academic decisions [36]

A comprehensive empirical study compared AI-generated and human-authored essays submitted to Oxford University, focusing on close readings of Old English poetry [6]. The findings revealed that AI essays achieved an average score of 60.46, closely trailing human submissions, which averaged 63.57 a difference not statistically significant ($p = 0.10$) [6]. Human essays demonstrated deeper engagement with cultural context and critical analysis, while AI responses were praised for their structure and clarity. Notably, the AI detection tool Quillbot achieved a 95.59% accuracy rate, outperforming other systems like GPTZero and ZeroGPT [6]. However, the study raised concerns about the fairness of detection algorithms, highlighting the risk that non-native English speakers could be disproportionately penalized due to linguistic patterns that diverge from native norms [6].

Although AI chatbots have innovative possibilities for education, they pose legitimate concerns for academic integrity, more so with the limitations of currently available detection tools [3]. The authors acknowledge that content produced by AI tends to slip past detection tools currently in place, thus calling for more stringent and ethical policies on AI integration for institutions [3]. They suggest that education systems incorporate proactive measures like oral exams, invigilation, and test designs which demand diagrams or graphical answers to deter misuse [37]. While ChatGPT enhances ESL learners' writing, engagement, and self-directed learning, its integration also poses risks such as plagiarism and over-reliance [3,27]. They advocate for structured implementation, AI literacy training, and ethical guidelines to ensure responsible use. A balanced approach combining AI support with human instruction is essential to maintain academic integrity and promote critical thinking [3,27].

4.1.3. Transformations in Assessment Design and Pedagogical Strategy

Increased uses of AI in higher education are leading to a reevaluation of core assessment approaches [4,10,32]. Traditional forms, including untimed take-home essays or standardized problem sets, grow increasingly vulnerable to exploitation by AI products of the sort of ChatGPT. This exploit concerns about scholarly integrity and the legitimacy of assignments submitted by students, especially where detection tools are not yet proven highly reliable or limited to a specific domain [10,32]. Educators thus find themselves forced to look at alternative methods of assessment that value critical thinking, originality, and process more highly than highly polished products [11]. AI-generated essays in humanities disciplines often meet acceptable grading criteria, mimicking the surface-level features of human writing [6]. While lacking critical analysis, these essays are well-organized, grammatically polished, and difficult to differentiate from human-written content without detailed forensic examination. This raises important questions about the validity of traditional assessments and highlights the need for a pedagogical shift toward formats that require personal engagement, creativity, and deeper contextual understanding [38].

Many institutions are moving toward more authentic and formative assessment practices. These include in-class written tasks, oral presentations, staged writing processes with feedback loops, and portfolio-based evaluations [38]. Emerging evidence points to highlight a need to redesign test forms, especially in ESL settings, to more effectively identify and prevent AI manipulation by looking for trends in students' writing progression [2]. But for this transition, more is required than structural reform. There is also a need for faculty members to be upskilled to know about the abilities of AI and how it can be used in pedagogy. In their investigation, ESL teachers were only 33% to 66% correct in identifying whether texts were composed by students or produced by AI [2], highlighting the dilemma teachers experience in terms of being able to make credible distinctions. This suggests the

 doi:10.20944/preprints202507.2233.v1

necessity of training teachers to identify AI-assisted writing and develop tests that minimize possibilities for illicit application of generative technologies [36]. Students were more apprehensive about their writing development being inhibited by AI rather than being disciplined for misconduct [8]. Institutions are advised by the authors to give higher priority to educational approaches which emphasize the process of education more than delivering a final, polished output [8]. This involves including the application of ethically used AI and transparency at classroom level and institutional policy for facilitating the promotion of academic integrity as a collective responsibility rather than detection being used merely as a deterrent [39].

4.2. Evaluating the Effectiveness and Limitations of Current AI Detection Tools in Distinguishing AI-Generated Content from Student Work

As AI-generated content becomes more common, institutions rely on detection tools to uphold academic integrity [40] This section examines their effectiveness, limitations, and associated ethical and legal concerns.

4.2.1. Effectiveness and Shortcomings of Current AI Detection Tools

Rapidly integrating large language models (LLMs) like ChatGPT into academic environments requires reliable detection tools to distinguish between AI-generated and human-written content [32]. While these tools show promise, their accuracy remains inconsistent, especially as language models continue to evolve. A recent evaluation tested a series of AI content detection programs OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag by looking at whether they could accurately distinguish content written by ChatGPT-3.5, ChatGPT-4, and human writers [22]. In their test, they found that while the programs were quite good at identifying content written by GPT-3.5, they were not very good at identifying content written by GPT-4. Furthermore, the programs were frequently in error about human-written content being written by AI, creating a large problem of false positives that could negatively impact students [22]. As the authors note, these discrepancies illustrate the weaknesses of currently available detection technology and must be approached cautiously in decision-making about academic integrity [41].

The performance of two RoBERTa-based detectors GPT-2 Output Detector Demo and Crossplag was evaluated using a dataset of 240 essays. Both tools showed reasonable success in distinguishing between human- and AI-generated texts, although their accuracy declined with longer or more complex sample [15] Inconsistencies were partly a result of the fact that both detectors were trained on GPT-2 output and were hence less adept at identifying content created by more advanced models like GPT-3.5 or GPT-4[14]. The authors concluded that the ever-growing sophistication of large language models is problematic for currently existing detection tools, which tend to lag behind AI advancements [14].

The accuracy of ten widely available AI-detection tools was evaluated against ChatGPT-3.5-generated content and its paraphrased versions produced by QuillBot, Grammarly, and ChatGPT itself [23]. Only five of these tools were accurate in error-free detection of original AI content. Interestingly, 100% accuracy rates were shown by Sapling and Undetectable AI even after the content had been paraphrased by the other tools. Rates of detection accuracy among the other tools evaluated varied significantly, ranging from 0% to 100%. Such variations are of concern for educational detection setups because false detection would unfairly disadvantage students [23].

In summary, while AI detection tools have shown promise, they remain unreliable for high-stakes academic decisions [30]. Many current detectors rely heavily on outdated linguistic heuristics and were trained on earlier-generation models, making them less effective at identifying advanced AI-generated content. These limitations underscore the urgent need for more adaptive and context-aware detection systems that can keep pace with evolving generative AI technologies, ensuring fair and accurate outcomes in academic settings [42].

 doi:10.20944/preprints202507.2233.v1

4.2.2. Emerging Innovations and Methodological Approaches in AI Detection

The limitations of traditional AI detection tools have prompted researchers to develop more sophisticated techniques; many rooted in machine learning and natural language processing (NLP) [31]. The vulnerability of current detectors to paraphrased text. Using a powerful paraphrasing model (DIPPER), they showed that the detection accuracy of advanced systems like DetectGPT plummeted from 70.3% to just 4.6% when outputs were paraphrased measured at a fixed 1% false positive rate [31]. This underscores how surface-level linguistic features are easily manipulated by large language models (LLMs). Even transformer-based detectors such as DistilBERT experience a notable decline in performance when LLMs are prompted or fine-tuned to mimic a student's writing style [10]. These stylistic adaptations increase the likelihood of false negatives, suggesting that detection systems relying solely on syntactic or stylistic markers may be inadequate for identifying AI-generated content that closely resembles human writing [10].

In response, researchers have begun exploring white-box detection methods, which involve embedding watermarking techniques directly into AI-generated text. For instance, inference-time watermarking embeds imperceptible yet verifiable patterns during the generation process, enabling systems to trace authorship [30]. These white-box methods offer full access to a language model's internal mechanisms, supporting real-time behavior tracking. However, their effectiveness depends heavily on whether institutions or commercial platforms adopt standardized watermarking protocols which currently remains inconsistent. Watermarking as a promising NLP strategy in the fight against AI-assisted plagiarism, especially in educational contexts [14].

In a related development evaluated several numerical text representation methods including Bag of Words, TF-IDF, and fastText embeddings combined with multiple classifiers to enhance detection accuracy. These programs transform text into numerical data by counting the occurrence of every word in a document without consideration for grammar or word order [30]. This then allows the program to identify patterns that are more consistent with AI-generated text. Among these, TF-IDF paired with logistic regression achieved the highest overall accuracy of 99.82%. Additionally, methods like Bag of Words and fastText consistently produced accuracies above 96.50%. Meanwhile, sentence embeddings such as MiniLM and distilBERT yielded slightly lower accuracies (93.78%–96.63%), suggesting further room for improvement [30]. Notably, the XGBoost classifier delivered the most robust performance overall, with a minimum accuracy of 96.24% across different input features. These findings emphasize the importance of dataset-specific representation strategies and tailored preprocessing techniques in optimizing AI text classification systems [30].

4.2.3. Ethical, Legal, and Educational Implications

Using AI detection tools in higher education raises technical challenges and significant ethical, legal, and educational concerns. A central issue is the risk of false accusations when these tools incorrectly flag human-written work as AI-generated [40] False positives are particularly common when students use formal or refined writing styles, which detection tools may mistakenly flag as AI-generated. This raises fairness concerns and can undermine trust in institutions that rely on such tools to enforce academic integrity [41].

Many commercial AI detection systems are still experimental, with developers cautioning against relying on them for high-stakes academic decisions [14]. For instance, Turnitin advises against making integrity decisions based solely on AI detection scores if the likelihood of AI authorship is under 40%. These cautionary statements underscore the delicate balance universities must strike between embracing innovation and ensuring fairness [15].

From a legal standpoint, general-purpose AI tools like ChatGPT may inadvertently fall within the scope of contract cheating laws, such as England's Skills and Post-16 Education Act [13]. Although AI providers do not market themselves as essay mills, the paper highlights the legal ambiguity around their use in academic settings, particularly where the generated content could be considered part of an assignment that a student is required to complete personally [13]. The authors argue this legal uncertainty raises questions of platform liability, especially as students increasingly

 doi:10.20944/preprints202507.2233.v1

"cognitively offload" academic work to AI (p. 283). Rather than criminalizing tool providers or students, the authors recommend that educational institutions focus on regulation, accountability, and clearer policy boundaries to manage AI's role in learning [13].

In educational terms, the overreliance on AI detection tools risks cultivating a culture of surveillance that prioritizes policing over teaching [14]. Recent findings advocate for a human-centered approach that emphasizes educator AI literacy and human oversight [28]. By promoting collaboration between educators and AI systems, institutions can foster academic integrity while adapting to the realities of digital learning. Their framework positions detection tools as supportive aids in evaluating student work, rather than as sole decision-makers. Addressing the challenges posed by AI in education requires a coordinated response from policymakers, educators, and developers, as the limitations of current detection tools, legal ambiguities, and implementation gaps highlight the need for transparent technologies, informed policy, and educator support [13,14,22,28].

4.3. Ethical, Procedural, and Pedagogical Concerns Raised by AI Detection Tools in Higher Education

Adopting AI detection tools in higher education introduces a complex set of ethical, procedural, and pedagogical challenges for students and faculty [43]. These concerns span fairness, student rights, institutional consistency, and the evolving role of assessments in AI-enhanced learning environments. The existing detection tools have been described as lacking both precision and dependability [43].

4.3.1. Ethical Ambiguities and Student Vulnerabilities

Deploying AI detection systems in higher education presents significant ethical challenges, particularly around student rights, fairness, and systemic bias [12]. One key concern is the occurrence of false positives, where student-authored work is mistakenly flagged as AI-generated. This problem is intensified by the probabilistic nature of AI detection, which lacks definitive proof and often operates without clear transparency in its decision-making process [12]. These systems rely on surface-level linguistic markers that can unfairly penalize students for writing in a clear, structured, or stylistically sophisticated way. This is especially problematic for non-native English speakers, whose writing may appear more formulaic or mimic AI-generated content, leading to unjust penalties [19].

Furthermore, these systems are typically used without direct, informed consent of students, thus creating grave concerns about privacy and autonomy. These kinds of tools are most often launched without transparency, which negates students' right to data privacy and informed participation [19]. Wider ethical considerations surrounding the deployment of AI detection in testing, especially concerning authorship, integrity, and student agency [11]. Universities frequently integrate external AI tools into their assessment platforms without fully disclosing the data-sharing practices, leaving students unaware of how their work is being analyzed or stored [19]. Many commercial detection tools are opaque, with some even retaining student data indefinitely or using it to train proprietary models without students' knowledge or consent [12].

Ethical concerns around equity also emerge in this context. Students with access to premium AI tools that can rephrase or "humanize" content can easily bypass detection systems, putting less-resourced students at a disadvantage [11]. This disparity reinforces existing educational inequities and encourages unethical behavior by students attempting to avoid unfair punishment. The ethical implications of AI detection systems cannot be divorced from broader surveillance, fairness, and justice issues within academia [10]. Addressing these concerns requires a rights-based approach prioritizing student dignity and institutional transparency over punitive enforcement [40]. Ultimately, a more thoughtful and equitable framework is needed to balance the benefits of AI detection with the need to uphold students' rights and foster fairness in academic settings [40].

 doi:10.20944/preprints202507.2233.v1

4.3.2. Procedural Inconsistencies and Institutional Risks

Procedural challenges surrounding the implementation and reliance on AI detection tools highlight current academic integrity enforcement weaknesses. A primary concern is the accuracy of detection systems [10]. Many commercial AI detectors exhibit significant inconsistency, with both high false positive and false negative rates. For instance, paraphrasing tools such as QuillBot can easily mask AI-generated content, allowing it to evade detection [29]. At the same time, genuine human writing especially when aided by advanced grammar tools may be wrongly flagged due to stylistic features that resemble AI output [12,22].

The lack of standardized procedures for flagged submissions compounds these technical shortcomings [2]. Though teachers themselves often report their capacity to identify AI-based work, empirical research shows the contrary. Furthermore, institutions are inconsistent in interpreting AI-assisted work under academic integrity policies [2]. While some universities allow limited AI use with proper disclosure, others consider any AI involvement a violation. These discrepancies confuse students and faculty, leading to uneven enforcement and potential legal risks for institutions that fail to maintain fair and consistent procedures [17]. The absence of clear guidelines, appeal mechanisms, and evidence standards further undermines procedural integrity. Students often have little recourse when wrongly accused, especially in cases where detection tools provide no transparent explanation or verifiable output [12,19]. Without proper auditing, accountability, and due process, AI detection risks becoming a blunt tool that undermines academic standards rather than safeguarding them [12,19].

The need for adequate institutional support, training, and ethical guidance, as many students and faculty lack the necessary knowledge to use AI tools appropriately, exacerbating misinterpretations during AI content reviews [4]. Domain-specific detection tools may offer better accuracy, but their effectiveness is contingent on access to large training datasets. That is something most institutions cannot afford, which raises equity and feasibility concerns in implementing fair AI-detection systems [10].

4.3.3. Pedagogical Consequences and the Future of Assessment

Introducing AI detection tools in higher education has significant pedagogical consequences, especially in reshaping teaching, learning, and assessment practices. One of the primary concerns is that these systems shift the focus from trust and engagement to surveillance and suspicion [35]. Rather than promoting ethical academic behavior, AI detection tools often encourage strategic avoidance tactics, such as paraphrasing services or altering sentence structures to bypass detection [12,31]. This undermines the goal of meaningful learning, as students are more focused on “passing” algorithmic checks than developing deeper critical thinking and reflective writing skills [31]. Over-reliance on these tools detracts from students’ ability to hone their critical and reflective skills [1] .Traditional essay-based assessments, already susceptible to plagiarism, become even more vulnerable to manipulation with AI tools [40]As generative AI tools like ChatGPT become increasingly capable of producing highly realistic and polished text, traditional essay-based assessments, already prone to issues like plagiarism, face even greater challenges [10]. In higher education settings, especially in contexts such as college admissions, distinguishing between genuine student writing and AI-assisted content has become increasingly difficult [10]. This blurring of authorship undermines the reliability of essay-based evaluations as indicators of students’ true abilities, raising serious concerns about fairness and integrity in academic assessment [10].

Recent study advocates for transitioning to more robust and authentic assessment formats, such as oral defenses, project-based tasks, and practical demonstrations, which are less susceptible to AI manipulation and better support academic integrity [19]. Similarly, another study highlights the vulnerability of traditional essay-based assessments, showing that even experienced instructors struggle to distinguish between AI- and student-written work, thus underscoring the need for rethinking how learning is evaluated [18]. Furthermore, the student voice is often absent in institutional decisions regarding AI detection policies. Most policies are created without involving

 doi:10.20944/preprints202507.2233.v1

students, leading to a top-down approach that erodes trust and overlooks the potential benefits of AI when used ethically [19]. When integrated thoughtfully, AI tools can support multilingual learners in language development. Teaching students to use these tools for tasks such as brainstorming or editing while maintaining originality can enhance digital literacy and promote inclusivity [11].

Ultimately, AI detection tools should complement not replace sound pedagogical strategies. Faculty must be supported not only in learning how to use detection tools effectively but also in redesigning curricula to reflect the realities of an AI-integrated educational landscape [19]. includes promoting ethical AI use, fostering students' AI literacy, and reinforcing the value of originality in academic work [21]. By integrating these elements, higher education can uphold its core mission of advancing meaningful learning while safeguarding academic integrity [11].

5. Limitations of Current Research

While significant progress has been made in reviewing AI detection tools in higher education, the current body of research remains limited in several key areas, both methodologically and contextually. First, there is a noticeable lack of large-scale, empirical, and longitudinal studies that assess the real-world implementation of detection tools across various geographic and policy contexts. Most existing research relies on small sample sizes, controlled experiments, or anecdotal case studies. These narrow study designs hinder findings' generalizability and external validity, limiting their applicability to diverse academic settings.

Second, much of the technical evaluation of AI detection tools takes place in controlled lab environments that fail to capture the complexity and unpredictability of how students interact with AI tools in actual academic settings. This disconnect means that the real-world accuracy of these tools, particularly in the face of paraphrasing, code-switching, and blending writing styles, remains unclear. As a result, their practical effectiveness remains uncertain.

Third, existing research tends to focus heavily on the perspectives of faculty, administrators, and technology developers, often neglecting the voices and experiences of students. This oversight limits our understanding of how AI detection tools impact student trust, anxiety, and behavior. Additionally, there is insufficient attention to equity issues, such as how linguistic, racial, and socioeconomic factors might contribute to higher false-positive rates and exacerbate institutional biases.

Fourth, interdisciplinary approaches are sorely lacking in the research on AI detection tools. Ethical, legal, and psychological perspectives remain underdeveloped, leaving crucial questions about privacy, due process, consent, and the mental health impact of detection tools inadequately addressed.

Finally, the rapid pace of generative AI advancements far exceeds the methodological adaptability of current research. Tools evaluated just a few months ago may already be outdated yet are still used as benchmarks for shaping academic policies. This lag between technological innovation and research methodologies creates a fragile foundation for long-term policy and pedagogical strategies.

Thus, despite valuable insights, current research presents an incomplete and uneven understanding of AI detection tools in higher education. A more robust, inclusive, and future-ready research agenda is urgently needed to address these gaps and ensure that AI detection tools are effectively integrated into academic practice.

6. Suggestions for Future Research

To better understand the role and impact of AI detection tools in higher education, future research needs to move beyond surface-level evaluations and tackle the deeper, more complex issues at play. A more comprehensive and equitable approach should be built on conceptual insight and real-world evidence.

 doi:10.20944/preprints202507.2233.v1

First, large-scale, cross-institutional studies are urgently needed. Research should examine how AI detection tools are implemented across different regions, cultures, languages, and policy environments. This would help identify common patterns and challenges while highlighting unique contextual factors that shape institutional practices.

Second, mixed-method research designs are crucial. While quantitative data can reveal how accurate detection tools are, qualitative input from student interviews, faculty discussions, and classroom observations offers valuable insight into how these tools affect people emotionally and pedagogically. These perspectives are often missing from technical evaluations but are vital to understanding the full impact of detection technologies.

Third, longitudinal studies are necessary to explore how AI and detection tools use evolves over time. Such research can uncover how students and institutions adapt, how writing practices shift, and what long-term positive or negative consequences might emerge from continued exposure to these systems.

Fourth, interdisciplinary collaboration should be a priority. By bringing together experts from fields such as computer science, education, ethics, psychology, and law, researchers can develop well-rounded frameworks that address technical effectiveness and issues of privacy, fairness, and student rights.

Fifth, there's a need to rethink assessment strategies. Future research should explore alternative assignment designs like scaffolded writing, oral presentations, and project-based learning less vulnerable to AI misuse and more supportive of genuine learning.

Finally, we should also look at co-use models, where AI is seen not just as a threat but as a tool for learning. By studying how transparent and ethical AI use can empower students especially those who are English language learners or neurodivergent we can design more inclusive and supportive academic environments.

In summary, future research must safeguard academic integrity, embrace inclusion, foster innovation, and promote critical AI literacy.

7. Conclusion

The rapid expansion of generative AI technologies in higher education has significantly shifted how institutions approach and uphold academic integrity. This review synthesizes current literature on adopting AI detection tools, evaluating their role within institutional policies, their technical reliability, and the ethical and pedagogical challenges they present. The findings suggest that integrating AI detection tools within universities is inconsistent and underdeveloped. While some institutions have updated their academic integrity policies to address AI-assisted writing, many still lack clear guidelines or enforcement procedures. This lack of standardization leads to varying interpretations by faculty and administrators, leaving students confused about the acceptable use of AI. These inconsistencies highlight the need for coherent policies and faculty training to align institutional practices with the rapidly evolving digital landscape.

From a technical perspective, the performance of AI detection tools remains limited. Although some tools have shown promising results in controlled environments, their effectiveness significantly decreases in real-world academic settings. One primary concern is the frequent occurrence of false positives, especially with work produced by non-native English speakers, raising doubts about the reliability of these tools in academic misconduct cases. Moreover, the rapid advancement of generative AI continues to outpace the development of detection systems, making many tools quickly outdated. Ethically and pedagogically, the use of AI detection tools presents additional challenges. These technologies are often implemented without sufficient transparency, informed consent, or safeguards, potentially violating student rights and exacerbating educational inequities. Relying too heavily on detection tools may also foster a culture of surveillance rather than engagement, eroding trust between students and institutions.

Overall, these findings highlight the need for a more comprehensive approach. Future strategies should focus on developing inclusive policies, promoting critical AI literacy, and rethinking

 doi:10.20944/preprints202507.2233.v1

assessment methods to ensure academic integrity while providing a more equitable and meaningful learning experience for all students.

Author Contributions: "Conceptualization, P.D.D. and M.S.R.; methodology, P.D.D.; software, P.D.D.; validation, P.D.D., M.S.R. and W.D.E.; formal analysis, P.D.D.; investigation, P.D.D.; resources, P.D.D. and M.S.R.; data curation, M.S.R. and W.D.E.; writing—original draft preparation, P.D.D.; writing—review and editing, N.G., M.S.R. and W.D.E.; visualization, P.D.D.; supervision, M.S.R. and W.D.E.; project administration, P.D.D; funding acquisition, M.S.R. All authors have read and agreed to the published version of the manuscript."

Funding: "This research received no external funding"

Institutional Review Board Statement: "Not applicable"

Informed Consent Statement: "Not applicable."

Data Availability Statement: All data in the manuscript.

Conflicts of Interest: "The authors declare no conflicts of interest."

References

1. Abbas M, Jam FA, Khan TI. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education. **2024**;21.
2. Alexander K, Savvidou C, Alexander C. Who wrote this essay? Detecting ai-generated writing in second language education in higher education. Teaching English with Technology. **2023**;23:25–43.
3. Davar NF, Dewan MAA, Zhang X. AI Chatbots in Education: Challenges and Opportunities. Information (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI); **2025**.
4. Al-Zahrani AM. The impact of generative AI tools on researchers and research: Implications for academia in higher education. Innovations in Education and Teaching International. **2024**;61:1029–43.
5. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. <scp>ChatGPT</scp> and a new academic reality: <scp>Artificial Intelligence-written</scp> research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol. **2023**;74:570–81.
6. Revell T, Yeadon W, Cahilly-Bretzin G, Clarke I, Manning G, Jones J, et al. ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. International Journal for Educational Integrity [Internet]. **2024**;20:18. Available from: https://edintegrity.biomedcentral.com/articles/10.1007/s40979-024-00161-8
7. Birks D, Clare J. Linking artificial intelligence facilitated academic misconduct to existing prevention frameworks. International Journal for Educational Integrity. **2023**;19.
8. Nelson AS, Santamaría P V., Javens JS, Ricaurte M. Students' Perceptions of Generative Artificial Intelligence (GenAI) Use in Academic Writing in English as a Foreign Language †. Educ Sci (Basel). **2025**;15.
9. Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. Calif Manage Rev. **2019;**61:5–14.
10. Zhao Y, Borelli A, Martinez F, Xue H, Weiss GM. Admissions in the age of AI: detecting AI-generated application materials in higher education. Sci Rep. **2024**;14:26411.
11. Yeo MA. Academic integrity in the age of Artificial Intelligence (AI) authoring apps. TESOL Journal. 2023;14.
12. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text. International Journal for Educational Integrity. **2023**;19.
13. Gaumann N, Veale M. AI providers as criminal essay mills? Large language models meet contract cheating law. Information and Communications Technology Law. **2024**;
14. Ibrahim K. Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines." Language Testing in Asia. **2023**;13:46.
15. Ibrahim K. Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines." Language Testing in Asia. **2023**;13.
16. Perkins M. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice. **2023**;20.

 doi:10.20944/preprints202507.2233.v1

17. Perkins M. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice. **2023**;20.
18. Waltzer T, Pilegard C, Heyman GD. Can you spot the bot? Identifying AI-generated writing in college essays. International Journal for Educational Integrity. **2024**;20.
19. Ardito CG. Generative AI detection in higher education assessments. New Directions for Teaching and Learning. **2024**;
20. Hao J, Von Davier AA, Davier V, College B, Harris DJ. Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI [Internet]. **2024**. Available from: https://openai.com/dall-e-3
21. Yeo MA. Academic integrity in the age of Artificial Intelligence (AI) authoring apps. TESOL Journal**. 2023**;14.
22. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. **2023**;19.
23. Kar SK, Bansal T, Modi S, Singh A. How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. Indian J Psychol Med. **2024**;
24. Waltzer T, Pilegard C, Heyman GD. Can you spot the bot? Identifying AI-generated writing in college essays. International Journal for Educational Integrity. **2024**;20.
25. Sukhera J. Narrative Reviews: Flexible, Rigorous, and Practical. J Grad Med Educ**. 2022**;14:414–7.
26. Sukhera J. Narrative Reviews: Flexible, Rigorous, and Practical. J Grad Med Educ. **2022**;14:414–7.
27. Deep P Das, Martirosyan N, Ghosh N, Rahaman MdS. ChatGPT in ESL Higher Education: Enhancing Writing, Engagement, and Learning Outcomes. Information. **2025**;16:316.
28. Hao J, von Davier AA, Yaneva V, Lottridge S, von Davier M, Harris DJ. Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. Educational Measurement: Issues and Practice. **2024**;43:16–29.
29. Kar SK, Bansal T, Modi S, Singh A. How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. Indian J Psychol Med. **2025**;47:275–8.
30. Krawczyk N, Probierz B, Kozak J. Towards AI-Generated Essay Classification Using Numerical Text Representation. Applied Sciences. **2024**;14:9795.
31. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. Adv Neural Inf Process Syst. Neural information processing systems foundation; **2023**.
32. Pan WH, Chok MJ, Wong JLS, Shin YX, Poon YS, Yang Z, et al. Assessing AI Detectors in Identifying AI-Generated Code: Implications for Education. Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training. New York, NY, USA: ACM; **2024**. p. 1–11.
33. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. **2019**;4.
34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. American College of Physicians; **2018**. p. 467–73.
35. Ardito CG. Generative AI detection in higher education assessments. New Directions for Teaching and Learning. 2024;
36. Alexander K, Savvidou C, Alexander C. Who wrote this essay? Detecting ai-generated writing in second language education in higher education. Teaching English with Technology. 2023;23:25–43.
37. Davar NF, Dewan MAA, Zhang X. AI Chatbots in Education: Challenges and Opportunities. Information (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI); **2025**.
38. Revell T, Yeadon W, Cahilly-Bretzin G, Clarke I, Manning G, Jones J, et al. ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. International Journal for Educational Integrity. **2024**;20.
39. Nelson AS, Santamaría P V., Javens JS, Ricaurte M. Students' Perceptions of Generative Artificial Intelligence (GenAI) Use in Academic Writing in English as a Foreign Language †. Educ Sci (Basel). **2025**;15.
40. Zhao Y, Borelli A, Martinez F, Xue H, Weiss GM. Admissions in the age of AI: detecting AI-generated application materials in higher education. Sci Rep. **2024;**14.
41. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. **2023**;19.

 doi:10.20944/preprints202507.2233.v1

42. Krawczyk N, Probierz B, Kozak J. Towards AI-Generated Essay Classification Using Numerical Text Representation. Applied Sciences (Switzerland). **2024**;14.
43. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text. International Journal for Educational Integrity. **2023**;19.