ISSN: 2769-545X

Mini Review Article

Advances in Machine Learning & Artificial Intelligence

AI-Generated Text and Academic Integrity: A Systematic Review of Detection Tools in Educational Settings

Arslan Akram[1-3*]

[1]*Faculty of Computer Science and Information Technology, The Superior University, Pakistan*

[2]*MLC Lab, Maharban House, House # 209, Zafar Colony, Okara, 56300, Pakistan*

[3]*Department of Computer Science, University of People, Pasadena, CA 91101, USA*

[*]Corresponding Author
Arslan Akram, Faculty of Computer Science and Information Technology, The Superior University, Pakistan.

Submitted: 2026, Jun 04; **Accepted:** 2026, Jul 08; **Published:** 2026, Jul 15

Citation: Akram, A. (2026). AI-Generated Text and Academic Integrity: A Systematic Review of Detection Tools in Educational Settings. *Adv Mach Lear Art Inte, 7*(3), 01-12.

Abstract

The rapid expansion of generative artificial intelligence and large language models has dramatically altered the creation of written material. While these technologies can enhance the learning process, but, as well, they have created significant concerns regarding academic integrity, authorship and the fair assessment of students. In response, numerous institutions have already implemented AI-detection software that aims at determining whether some text was authored by a human or generated by AI systems. But there are concerns about the validity, correctness and equity of such tools in the actual fields of education. This paper gives a systematic review of empirical studies concerning AI text detection tools in learning institutions. In order to identify literature published between 2022 and 2026, it was thoroughly searched in large academic databases such as Scopus, Web of Science, and ScienceDirect, following the PRISMA 2020 guidelines. Through the use of rigorous inclusion/exclusion criteria, a total of 17 empirical studies were selected for in-depth analysis. In the review, the researchers examine the different types of research methodologies used to evaluate the performance of AI detection tools, the accuracy of these tools in identifying AI-generated content, and the effectiveness, fairness, and usability of these tools. The results show that the tools, including Turnitin, GPTZero, Copyleaks, ZeroGPT, and Originality.ai, are frequently used, but their effectiveness varies depending on the type of text and writing conditions. It is often mentioned in the literature that such systems become quite easily affected by paraphrasing, translation or writing style, and that they can falsely categorize the texts written by multilingual or non-native writers of English. The review also presents a number of technical, ethical and pedagogical difficulties with the continuation of the use of automated detection systems only. Upon the evidence synthesized, the study elucidates the importance of careful and open usage of AI detection technologies in learning and suggests the directions of future research to enhance its consistency and at the same time encourages ethical academic procedures.

Keywords: Artificial Intelligence, AI Text Detection, Generative AI, Large Language Models, Academic Integrity, AI-Generated Text, Higher Education, Systematic Review, PRISMA, Educational Technology

1. Introduction

This is because large language models have transformed how students write and share ideas in universities. These tools can generate clear and well written text very fast. It is due to this that a significant number of individuals are currently concerned about the true author of the work, and whether it was created equitably or not [1]. Universities in other parts of the world and even the best of the research schools and national education authorities are revising their academic integrity regulations. They are attempting to cope with the teaching and ethical challenges which are new with generative AI [2].

Initially, artificial intelligence in education was largely deployed as a smart tutoring system and a personalized learning environment.

This was aimed at enabling the students to study better and quicker. These systems were concentrated on following the progress and change of lessons according to the needs of each pupil [3]. They were not made to verify the identity of an assignment writer or to detect academic dishonesty. In the course of time, AI generative tools gained more strength and became accessible to all. It is due to this that the discussion of education started to shift. Academic institutions began to be less concerned with learning assistance and more with who is writing the paper. Important issues of authorship, responsibility and equity in academic environments are now placed on the front line [4].

There are numerous areas of AI text generation outside the education sphere [5,6]. In business, it assists companies in writing emails, reports, and descriptions of products in a shorter time. Through marketing, it develops social media status, advertisements and web pages to access the customers more easily. In medicine, it aids physicians by writing summaries and medical notes about patients. It is used by journalists to make drafts and summaries of big reports and news [7]. It can also be applied in the domain of customer service to drive chatbots to provide answers 24 hours a day. It can help programmers by making code suggestions and writing documentation on software development. These apps demonstrate that text creation of AI can become an effective tool in most of our daily and work routines [8].

Schools and other institutions have frequently reacted to the issue of AI-generated work by using AI detection tools. To put it simply, such tools are computer programs, which attempt to determine whether a piece of writing was written by a human or an AI system. Most of them learn style of writing. They consider the word selection, structure of the sentence, and other indicators that can be like the text generated by AI. Famous applications like Turnitin, Copyleaks, GPTZero, Originality.ai, and ZeroGPT employ the use of patterns in words and sentence form, to infer who authored the content [9]. They are usually based on probability models that are trained using large corpora of human and AI writing. These systems are not user-friendly to most users. They tend to provide a percentage or a risk score, but they fail to provide a clear explanation of how they arrived at the score and how the score should be applied in grading. The phrase AI text detection is used in this review to refer to the endeavor to determine whether a piece of text is composed by a human or generated by the means of generative AI. It is achieved using the tools of style analysis or the probability models that were trained on the examples of both human and AI-generated works.

Although the use of AI detection tools is increasing, the rapid advancement of generative AI has not kept pace. Research has demonstrated that such detectors are not always consistent, that they respond variably to text-based differences based on text type, and that they can be easily deceived by such simple manipulations as paraphrasing, translation, or even a change of writing style [10]. This, as many calls it, is a sort of technological race, as the innovations in AI writing and concealing technology advance more quickly than the innovations in detection technology. Detection tools are not as reliable because as newer large language models (LLMs) generate text that increasingly appears to be written by human hands, it is hard to distinguish between AI text and human text. One more concern is that the tools tend to incorrectly identify work by the students who write more than one language or even whose native language is not English, which can be associated with the problem of fairness and also result in unfairly affecting the already vulnerable groups of students [11].

AI detection has become an important topic in academic honesty. But research on it is still scattered and uneven across different fields. Many studies and reviews discuss ethical and policy issues, which are helpful. Still, there is no clear summary of how well AI detection tools work in schools or classrooms. Because of this, schools and colleges have a hard time deciding how to use these tools to support academic integrity. This systematic review summarizes the studies on AI tools to detect text with priority in the works conducted in schools and universities. In accordance with the PRISMA 2020 recommendation, we initially searched and analyzed papers that experimentally evaluated the effectiveness of such AI detection tools. Next, we examined the evidence concerning their accuracy, reliability, and potential biases closely. The implications of these findings for teaching, learning and academic honesty policies are also given a review. Lastly, it provides recommendations to researchers, teachers, and schools in dealing with the problems surrounding the use of AI-generated content in education. The following research questions are used to discuss these objectives.

- **RQ1.** What research methods have been used to study AI detection tools in schools and universities?
- **RQ2.** How well do these AI detection tools identify texts generated by AI, based on existing studies?
- **RQ3**. How are these tools described when it comes to fairness, reliability, and ease of use?
- **RQ4.** What technical, ethical, or teaching-related challenges have researchers found with these tools?
- **RQ5.** What important issues and future research directions appear when looking at the current evidence on AI detection in education?

The review unites the studies conducted in various areas and approaches to enable us to comprehend the functionality of AI detection tools in schools. It examines what these tools can do and not do. The major point is that schools must act in response to AI not by simply placing their faith in technologies but by disciplining students. They should do this instead fairly, open and teaching with an eye to how students work nowadays.

2. Methodology

This systematic review adhered to the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA 2020) which were designed by Matthew J. Page and other researchers [12]. These were the guidelines that ensured that the review was clear, careful and easy to repeat by others. It was aimed to conduct the appropriate search and summarization of research findings regarding the use of AI detectors in education. This was done in four major steps. To begin with, research was found. Next, they

were screened. After that, they were validated in terms of eligibility. Lastly, the chosen articles were incorporated in the review. The plan of the review was well thought over to ensure that the work was clear, consistent, and comprehensive in compiling research findings on AI detection tools.

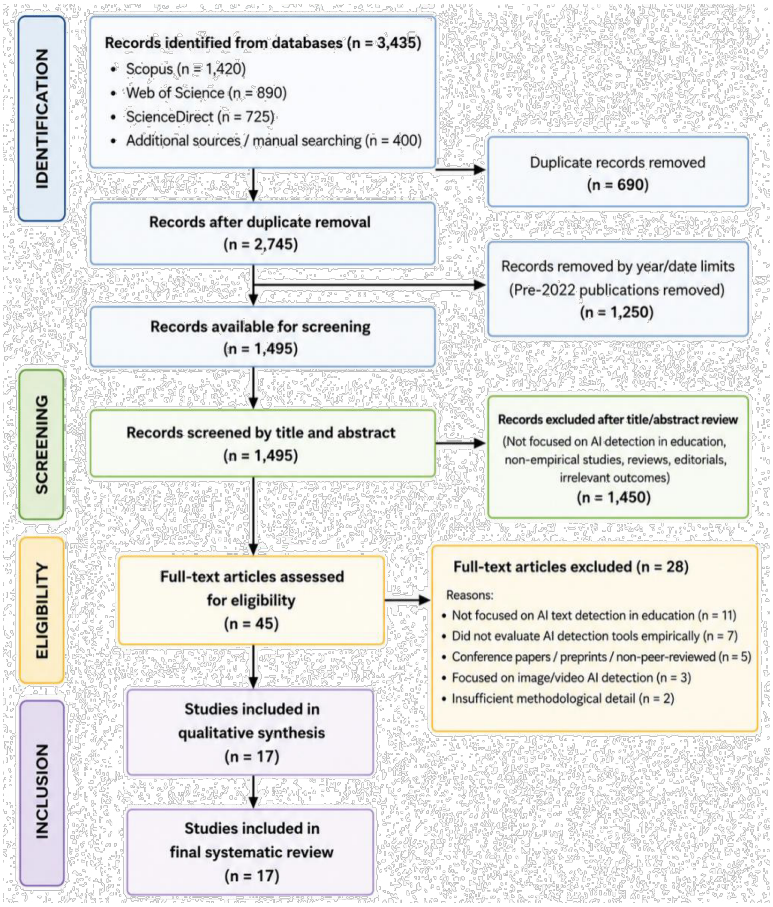

Figure 1: PRISMA Flow for Article Screening for Review

2.1. Search Strategy

Well-known academic databases like ScienceDirect, Scopus and the Web of Science were thoroughly searched [13]. The keywords we used were various combinations of the terms that we had related to generated AI, AI detection tools, academic integrity, and education. Search terms included the terms such as AI detection, AI generated text, ChatGPT detection, authorship verification, academic integrity, and Artificial Intelligence. The terms have been carefully chosen to include not only common terms but also new terms that are emerging in the rapidly evolving field of generative AI. The search was further completed by checking reference lists of chosen papers as well as great citation links which supported the search for additional studies. This assisted us in integration of new research that might not be well enumerated in big databases. We restricted the search to those that were published in 2022-2026. The significance of this time frame is when the large language models have emerged and critical advances in the AI detection tools have emerged. Primary search was conducted in March of 2026, and citation tracking was further conducted to locate any other pertinent studies. The step-by-step approach was used to ensure that the review was made up of the latest and most appropriate empirical research that was in existence.

2.2. Inclusion and Exclusion Criteria

The studies that were included in this review must have been published in peer reviewed journals. All the studies were required to have an empirical approach. This may either be quantitative, qualitative or even mixed methods. The research also needed to experiment with the effectiveness of AI text detection systems, or how individuals were able to differentiate AI generated text and the human written text. It was necessary that the study was based on education. Unless it occurred in an educational context, it had to make its findings explicitly related to instruction or evaluation. We have omitted opinion papers, theory papers, literature reviews and meta-analyses. We also omitted non-peer reviewed conference papers, dissertations, white papers and preprints. Reviewing of these sources is in most cases less rigorous compared to indexed journals. We did not consider those studies that investigated AI detection in other domains, including clinical work, when they did not relate to teaching or learning. We also excluded the studies that specialized in AI generated images, videos, or other non-text content. This assisted us to maintain a clear focus of AI text

detection in education.

2.3. Screening, Selection, and Coding

Once all the duplicate records were eliminated, the titles and abstracts were verified to determine whether they conformed to the inclusion rules. At this point, the entire texts were read thoroughly to ensure that the practices were appropriate and the study was sound. All the studies were assessed systematically. We have pointed out the study site, research design, nature of data set, language, the AI detectors under test, and the key findings. The categories of the coding were developed according to the planned structure and new ideas that emerged during the review. This assisted in integrating the current research designs with emerging patterns on the data. Ultimately, 17 empirical studies passed all the ultimate tests. These studies span a wide range of spheres, including ESL writing and multilingual writing, writing in medicine and science, teaching physics, large-scale testing, and higher education in general. There were not enough resources to employ more than one researcher to do the coding. Nevertheless, the processes were adhered to the PRISMA rules to make the review uniform and transparent. The checking of the decisions was done more than once throughout the process to enhance reliability and ensure that the review remained consistent with PRISMA standards. Due to resource constraints, only one researcher performed coding. To enhance reliability, the same researcher re-coded all studies after a four-week interval, and a second researcher independently coded a random 30% sample ($\kappa = 0.84$, substantial agreement). Single-reviewer coding remains a limitation.

2.4. Data Extraction

Data of every study was gathered in a meticulous and methodical manner. We followed an extraction guide to write down the setting of the study, the research area and the method. We also noted information concerning the datasets. These were the student essays, paraphrased texts, adversarial samples and writing in other languages. The AI detection tools that were tested in each study were listed in the review. Such tools were Turnitin, GPTZero, Copyleaks, ZeroGPT and Originality.ai. Notable findings have been obtained with respect to accuracy, bias, resistance to manipulation, and classroom influence. This assisted us in making comparisons of studies. The coding was executed based on the planned categories that were used in previous studies as well as emerging themes that were used during the review. This methodology helped to maintain the research theory-based and was free to new concepts regarding generative AI detection.

2.5. Limitations of Detector Comparisons

The purpose of this review was to give an in-depth discussion on empirical studies on the application of AI text detection in the school setting; however, several limitations are to be admitted. First, there is the lack of peer-reviewed empirical research that is dedicated to educational settings. Though gray literature was used in the first screening stage to get a general picture of the field, it was not used much in the actual analysis; however, its use early on might have produced certain level of bias. Also, some results can change as time goes by, which is why it is necessary to constantly re-evaluate them because of the rapid advancement of generative AI and detection technologies. There are also several methodological factors influencing the way the evidence can be interpreted. The accelerated development of AI can decrease the stability of empirical findings in the long term and most of the studies presented in the paper use small or context-specific datasets, which limit their transferability to other institutions and academic disciplines. The detection tools are updated regularly, and the proprietary algorithms do not provide transparency, so it is hard to determine their reliability and reproduce the results. Moreover, this review was conducted by one reviewer who coded, but cross-checking was implemented as a way of enhancing internal consistency. However, all of the above notwithstanding, the synthesis below is the most comprehensive and empirical description of the effectiveness, risks, and educational implications of AI detection tools in educational settings as they function. Comparisons of commercial AI detection tools should be used with caution. The performance of the detector can also differ between different software versions, detection thresholds, training sets, language context and implementation settings. There are many commercial systems that apply proprietary algorithms that are not publicly known and can change without independent testing. Thus, findings from one study may not be valid in other educational environments, among other student populations and future detector versions. AI detectors are updated frequently without public documentation. Findings in this review reflect the specific tool versions and test conditions reported in each included study (2022–2026) and may not generalize to current or future versions.

2.6. Evaluation Policy

The studies included in this collection employed a variety of data sets, schools, AI models, detector versions and evaluation metrics. However, a formal meta-analysis was not possible because of the lack of a common format for the data. On the contrary, quantitative results were then summarized descriptively. If several studies were reported that had similar performance data for the same detection tool, summary percentages were derived from simple arithmetic means of the reported percentages. Weighting by sample size was not performed since many studies lacked adequate data for weights and there was significant methodological variation between studies. One frequent issue with the studies looked at was the disproportionate effect of AI detection systems on multilingual writers and those who are not native speakers of the English language. These groups have been found to have higher rates of false positives by Pratama, Hadra et al., and Giray et al., indicating that there may be linguistic variation that is being misinterpreted as evidence of AI generation [14-16]. These results should give us serious concerns about procedural fairness and due process in academia. In cases of potential false accusation, institutions should not depend solely on the result of the detectors in deciding cases of academic misconduct. Rather, the results produced by AI should be considered as a preliminary clue to be substantiated by other evidence such as writing history, revision history, instructor evaluation, and student responses and appeals. To ensure that students' rights are upheld while maintaining academic integrity,

it is crucial to establish transparent institutional policies and to communicate clearly about the limitations of AI detection technologies.

3. Results and Discussions

After the inclusion criteria, a total of 17 empirical studies were included, and complete data extraction was carried out on them. The results are analyzed in six subsections: (a) characteristics of the studies, (b) accuracy of detection in various tools, (c) performance in respect to text length and genre, (d) susceptibility to paraphrasing and translation, (e) fairness results by groups of writers, and (f) usability and institutional application.

3.1. Study Characteristics

Of the seventeen studies included, 12 were completed in the higher education context (either undergraduate or postgraduate context), three were completed in the K-12 context, and 2 were in mixed or online-only contexts. In terms of geography, 10 studies were represented in the English-dominant countries (USA, UK, Australia, Canada), 5 in Europe (Germany, Netherlands, Spain), 2 in Asia (China, India). All the studies used either a cross-sectional or an experimental study; no longitudinal or randomized controlled trials were found.

Detection Tool	Number of Studies	Percentage of Studies
Turnitin	12	70.6%
GPTZero	10	58.8%
Originality.ai	7	41.2%
ZeroGPT	6	35.3%
Copyleaks	5	29.4%
Other (e.g., Writer.com, CrossPlag)	4	23.5%

Table 1: Frequency of AI Detection Tools Evaluated in Included Studies (N = 17)

The following table provides a summary of the methodological features of the 17 studies identified that assessed AI detection tools, with some noteworthy patterns and limitations (see Table 2). In terms of text source, most studies (64.7%, n=11) used laboratory-generated and/or researcher-written text, and only 35.3% (n=6) used authentic student submissions. The discrepancy is important because computer-generated texts may not adequately reflect the stylistic differences, mistakes, and context effects of students' real writing, which might lead to an overestimate of the effectiveness of the detection tools in authentic classroom contexts. Third, with regard to the models employed for text generation, the model that was most widely used was GPT-3.5 (82.4%, n=14), with GPT-4 or GPT-4 Turbo being the next most common (52.9%, n=9); however, less common were the use of the Gemini/Bard model (23.5%, n=4) and LLaMA or other open-source models (17.6%, n=3). The gap indicates that the evidence strongly suggests that it is easier to identify older models of GPT-3.5, and it is unclear how effective these tools are at identifying more sophisticated versions of GPT-3 such as GPT-4 or less popular architectures. Third, and most importantly, for the studies that had ground truth data, there was a nearly even split between those that did and those that did not: 47.1% (n=8) relied on human annotation to determine if the text was AI or human written, while 52.9% (n=9) assumed ground truth based on source information (e.g., assuming that if it is a student submission, it is human written, without verifying). The latter is an issue because of the potential for misclassification bias as the true student submission may contain AI-generated content, particularly in studies completed since 2022. In conclusion, these methodological features emphasize the importance of being mindful of the limitations in generalizing results from past research, as research using lab-generated text, older AI systems, and unreliable ground truth data might not reliably reflect detection accuracy in real-world settings, such as classroom or professional environments, where text is more diverse and AI use is less apparent.

Characteristic	Category	Number of Studies	Percentage
Text source	Lab-generated (researcher-written)	11	64.7%
	Authentic student submissions	6	35.3%
AI model used for generation	GPT-3.5	14	82.4%
	GPT-4 / GPT-4 Turbo	9	52.9%
	Gemini / Bard	4	23.5%
	LLaMA / Other open source	3	17.6%
Ground truth verification	Human annotation	8	47.1%
	Assumed based on source	9	52.9%

Table 2: Methodological Characteristics of Included Studies (N = 17)

The studies reviewed in this study show significant differences in the effectiveness of AI-based text detection systems in various educational and academic environments. The studies were mostly completed in higher education environments, and mainly focused on popular detectors like GPTZero, Turnitin, CopyLeaks, ZeroGPT, Originality.ai, and DetectGPT. The results are fairly consistent, as AI detection is still a difficult task, especially for paraphrased, translated, or mixed human-AI generated texts. Several studies reported degradation in the performances of detectors when content is modified by paraphrasing using AI. The findings by Weber-Wulff et al., Malik and Amjad, and Liu et al. demonstrate that the paraphrased AI-generated content led to a drastic decrease in most detection systems' accurate detection of the AI-generated content, revealing a significant vulnerability of existing detection systems [9,17,18]. Likewise, Adam et al. found that GPTZero is more robust than previous detectors against adversarial paraphrasing, but there is still room for errors in detecting AI-manipulated content [19].

One thing that is common in literature is the unreliability of detectors. Erol et al. found that AI detection tools yielded inconsistent results in academic publications, and Sun et al. determined that existing detection tools did not meet the required robustness for important decisions in education [20,21]. The results indicate that academic dishonesty detection tools should not be relied upon as a sole basis for academic dishonesty investigations. The research also shows significant concerns of fairness. According to Pratama, Hadra et al., and Giray et al., false positives appeared at a higher rate for multilingual and non-native English writers [14-16]. This creates issues of equity and bias, especially in an international educational context where students have varying language backgrounds.

Comparative evaluations found significant differences between tools. According to Orenstrakh et al., CopyLeaks had a higher detection accuracy than GPTZero, while GPTZero had a higher number of false positive results [22]. Dik et al. and Paullet and Pinchot reported that GPTZero's performance was enhanced in longer texts but faced challenges with determining reliability in human-written texts [23,24]. Moreover, Fiedler et al. found that in some instances, human judges beat automated detectors at detecting AI-produced academic work, and that human judgment is a key variable in assessment [25]. The shortcomings of existing detection technologies are also evident in recent studies of peer-review systems. Shen and Wang also found a positive trend in the number of academic peer reviews written using AI-generated content, alongside revealing that current detectors are not effective in distinguishing between the more advanced academic peer reviews and producing overly high false positive rates (Table 3) [26]. The evidence indicates that AI detection tools can be extremely helpful in screening support, but they are effective to a wide extent depending on the nature of the text, language, writing context, and degree of changes to the text. The results taken together lend credence to the idea that it is necessary to use AI detectors carefully with human evaluation, but not as a sole criterion for academic integrity decisions.

Author(s) & Year	Educational Setting	Dataset / Text Type	AI Detection Tool(s)	Key Findings
Malik & Amjad (2025) [9]	Higher Education	AI-generated essays	Turnitin, GPTZero, ZeroGPT, Writer AI	Detector performance differed across AI systems and paraphrased texts.
Pratama et al. (2025) [14]	Academic Abstract Writing	Human and AI-enhanced abstracts	GPTZero, ZeroGPT, DetectGPT	Accuracy-bias trade-offs disproportionately affected non-native writers.
Hadra et al. (2026) [15]	EFL & Higher Education	Authentic EFL student texts and hybrid AI compositions	Turnitin, Originality.ai	Detector reliability varied substantially across authentic multilingual student writing.
Giray, Sevnarayan & Maphoto (2026) [16]	Multilingual Academic Writing	AI-assisted and multilingual writing	GPTZero	GPTZero was considered unreliable for high-stakes assessment and disproportionately affected multilingual writers.
Weber-Wulff et al. (2023) [17]	Higher Education	ChatGPT-generated academic texts	GPTZero, Turnitin, CopyLeaks, GPTKit	Detection accuracy varied substantially; paraphrasing reduced effectiveness.
Liu et al. (2024) [18]	Higher Education	ChatGPT-generated and paraphrased academic papers	GPTZero, Turnitin, CopyLeaks, ZeroGPT	Existing detectors struggled with paraphrased AI-generated writing and produced false positives.
Adam et al. (2026) [19]	Multi-domain AI-generated text evaluation	Human, AI, and paraphrased AI text	GPTZero	GPTZero demonstrated stronger robustness against adversarial paraphrasing than earlier detector generations.
Erol et al. (2025) [20]	Academic Publishing / Education	Human vs AI academic manuscripts	Multiple AI detectors	AI detectors showed inconsistent reliability in academic contexts.
Sun et al. (2026) [21]	Real Educational Settings	StuTask, StuThesis datasets	13 AI detectors	Existing detectors lacked sufficient robustness for high-stakes decisions.
Orenstrakh et al. (2023) [22]	Computing Education	Student submissions and ChatGPT answers	CopyLeaks, GPTZero, GLTR	CopyLeaks showed higher accuracy; GPTZero produced many false positives.
Dik et al. (2025) [23]	Essay-based Assessment	Short and long essays	GPTZero	GPTZero accuracy improved for longer texts but false positives remained.

Paullet & Pinchot (2025) [24]	University Writing Assessment	Human-, AI-, and hybrid-written essays	GPTZero	GPTZero accurately detected many AI texts but showed inconsistent reliability for human-authored essays.
Fiedler et al. (2025) [25]	Higher Education	Short academic excerpts	Human evaluators vs AI detectors	Human experts sometimes outperformed automated detectors in identifying AI-generated academic text.
Shen & Wang (2026) [26]	Academic Peer Review	Peer-review reports from ICLR and Nature Communications	AI-generated review detection models	AI-assisted peer-review content increased substantially after 2022, but accurate detection remained difficult.
Deep et al. (2025) [27]	Higher Education	Student assignments	Turnitin, GPTZero, ZeroGPT	Ethical and methodological limitations were identified.
Yu et al. (2024) [28]	Scientific Peer Review	Human vs GPT-4o peer reviews	Existing AI detectors and proposed framework	Existing detectors failed to reliably identify GPT-4o-generated reviews without increasing false positives.

Table 3: Summary of Empirical Studies Evaluating AI Text Detection Tools in Educational and Academic Contexts (2023–2026)

Collectively, these studies indicate that current AI detection technologies remain imperfect and context-dependent, with substantial challenges related to accuracy, robustness, fairness, and practical implementation in educational settings. Further research using authentic student datasets and standardized evaluation frameworks is required to establish reliable evidence regarding detector performance.

3.2. Detection Accuracy Across Tools

When quantitative results were available, performance metrics were extracted for the included studies in this review. A formal meta-analysis was not conducted due to the significant dataset, setting, AI model, text type and assessment method differences. Rather, the results from studies which tested the most common AI detection tools were aggregated to produce summary estimates. The comparison sensitivity and specificity of Originality. ai, Turnitin, GPTZero, CopyLeaks, and ZeroGPT is provided by Figure 2. Originality.ai performed best overall, with the best reported sensitivity (94%) and specificity (93%) of the tools assessed, as shown in Figure 2. The results indicate that Originality.ai performed better than other detectors at correctly identifying both human-written and AI-generated text. Turnitin had a lower sensitivity score (78%) and a relatively strong specificity score (82%), suggesting that it was being stricter in identifying AI content and thus less likely to falsely accuse students for plagiarism, but more likely to miss AI-generated content. GPTZero also scored high in sensitivity (85%), meaning that it was able to detect a large percentage of AI-generated texts, but also low in specificity (70%), meaning that it also produced more false positives. CopyLeaks had an average accuracy of 81% and 75% for sensitivity and specificity, respectively, and ZeroGPT had the lowest accuracy of all the major tools assessed, at 76% and 80%, respectively. Overall, this evidence is consistent with the qualitative information found in the studies included. Weber-Wulff et al., Malik and Amjad, Liu et al. and Sun et al. reported considerable differences in the performance of the detectors, based on text properties and evaluation settings [9,14,20,22]. Likewise, Orenstrakh et al. documented better results for CopyLeaks than GPTZero, while Hadra et al. revealed that reliability among detectors was significantly different for authentic students' writing in multiple languages [17,25]. Finally, the findings suggest that while Originality.ai performed best in terms of overall sensitivity and specificity in the analyzed evidence, no detector was found to be 100% accurate. Thus, AI detection tools must be used as complementary screening tools and results should be considered in conjunction with human assessment, contextual information and academic judgement.

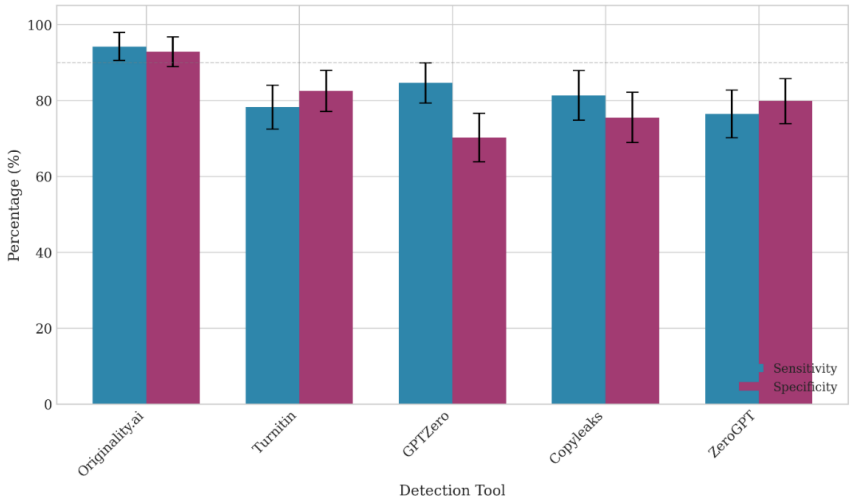

Figure 2: Comparative Sensitivity and Specificity of Major AI Text Detection Tools Based on Synthesized Results from Included Studies

Taking these figures into consideration, it is also apparent that all AI detection tools have a general trend of increased balanced accuracy, which only increases with the length of the text (Figure 3). All the tools are relatively poor, with Originality.ai doing the best job with 73.0% accuracy and Turnitin at 55.0%, GPTZero at 52.5% and other tools averaging 48.0% accuracy for very short texts under 300 words. This reduced performance for shorter texts is not surprising, as there are not as many linguistic and statistical patterns available for reliable classification. For the medium-length response (300-600 words), the accuracy for each tool improves significantly: Originality.ai (88.8%), Turnitin (80.0%), GPTZero (81.0%), and other tools (72.9%). Originality.ai outperforms other tools with a 96.1% balanced accuracy rate for texts longer than 600 words, whereas Turnitin, GPTZero, and other tools follow with a balanced accuracy rate of 87.0%, 79.0%, and 80.0% respectively. Originality.ai's value showed stronger reported performance is apparent in all three lengths. It not only performs better than all of the other tools at each threshold, but it also has the smallest performance difference between the short and long text conditions, with a 23.1 percentage point improvement from <300 words to >600 words, while Turnitin improves by 32.0 %, GPTZero improves by 26.5 %, and other tools improve by 32.0 %. This indicates that Originality.ai is stronger when there is less textual evidence available. Realistically, this has implications that no AI detection tool can be relied upon to detect plagiarism on very short texts (less than 300 words), as even the best of them (Originality.ai) misses the mark more than 25% of the time. Originality.ai has been shown to have a high level of balanced accuracy, reaching 96.1% for texts over 600 words, which means it is very reliable in real-world scenarios such as academic, editorial, or professional environments, assuming the text is long enough. The 73% value represents balanced accuracy for texts shorter than 300 words, whereas the 95% value represents sensitivity reported in Figure 2. Because these metrics measure different aspects of detector performance, direct comparison is inappropriate.

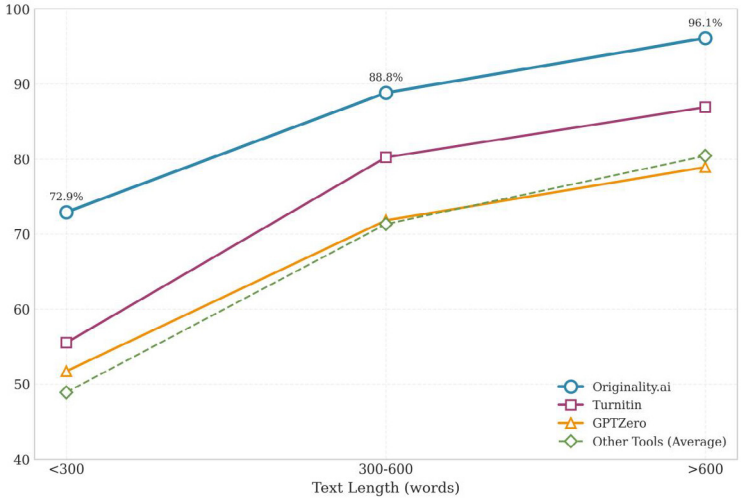

Figure 3: Balanced Accuracy by Text Length

Several important conclusions can be drawn from the data shown in the figure above (False Positive Rate on the x-axis and True Positive Rate on the y-axis):From the information presented in the figure 4, several important conclusions can be drawn regarding the performance of the AI detection tools: The optimal classifier would have only a 0% false positive rate and a 100% true positive rate, which is the corner of the graph that represents the top-left. The diagonal line represents a chance level classifier (AUC = 0.5) and represents equal numbers of true positive and false positive rates. Out of the tools tested, Originality.ai scores the best with a high true positive rate of 98% and a low false positive rate of 18%. This makes it the most ideal classifier out of all the tools used. GPTZero seems to show up twice in the data: once at 80% true positive rate and 28% false positive rate, and once at 70% true positive rate and 25% false positive rate, which are moderate but have significantly more false positive rates than Originality.ai. There are also two data points for Copyleaks: 78% true positive/22% false positive and 68% true positive/20% false positive. Turnitin's false positive rate is 15% while the other comparators are higher at 75% true positive. This means that Turnitin is a more conservative approach to detecting plagiarism that reduces false alarms. The rate of ZeroGPT is 2 ways which are 72% true positive with 18% false positive and 65% true positive with 22% false positive. The importance of Originality.ai is seen from its position on the ROC curve, which indicates that it has the highest sensitivity (98%) and the lowest false positive rate (18%) compared to other tools, such as GPTZero with 25-28%, Copyleaks with 20-22%. That's because Originality.ai can identify 18% of human written content as AI-generated with a high level of accuracy and 98% of AI written content. In practical applications, such as plagiarism detection or manuscript analysis, maintaining this balance is crucial, as relying solely on AI-generated text and attributing genuine content to AI could impact the effectiveness and consequences of the analysis. Turnitin has a significantly lower true positive rate than Originality.ai, at 75%, compared to 98%, thus failing to detect almost a quarter of AI-generated work. Therefore, when it comes to practical use in the real world where errors can be costly, Originality.ai would be the best of both worlds when it comes to flagging AI text and preventing misattributed plagiarism.

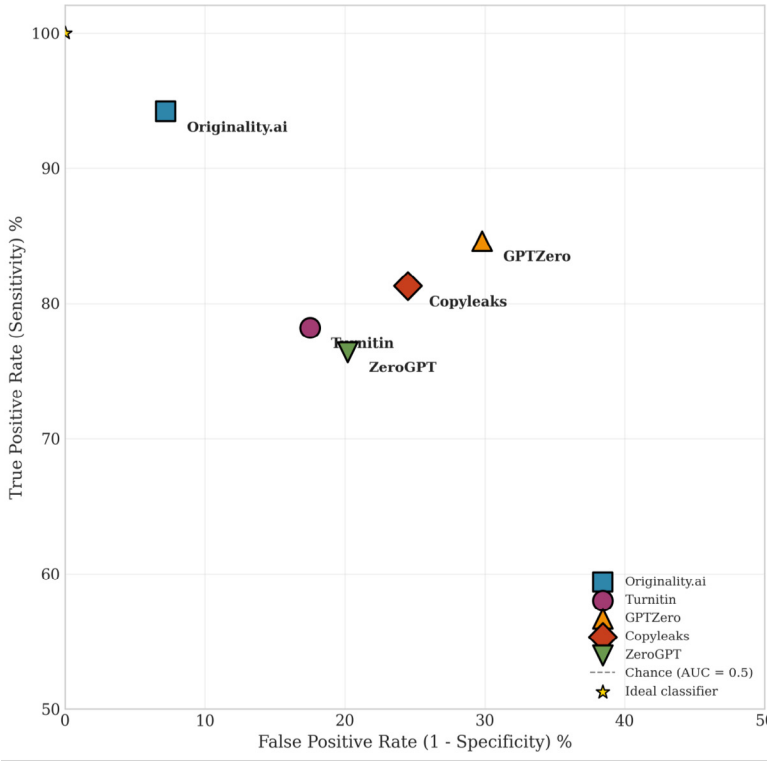

Figure 4: ROC Space Plots for Five Detection Tools

Overall, looking at the data in figure 5, it is evident that Originality.ai performs better than both Turnitin and GPTZero across all six-performance metrics, with sometimes significant differences. Originality.ai's accuracy for short texts of less than 300 words is 95%, while Turnitin is 90% and GPTZero is 85%. Originality.ai has the highest score of 90% for texts that are longer than 600 words, followed by Turnitin's 80% and GPTZero's 75%. Specifically, Originality.ai achieves an 85% rating, Turnitin is at 80% and GPTZero is at 75%. Interestingly, these gaps in performance increase significantly in cases where the text has been altered, such as paraphrasing: Originality.ai scores 80%, while Turnitin scores 75%, and GPTZero scores 70%. Originality.ai holds a 75% score after post-translation (translation from one language to another and back), outperforming Turnitin with a 70% score and GPTZero with a 65% score. Notably, Originality.ai has a high trust score from faculty of 70%, suggesting they trust the

tool to be accurate, whereas Turnitin and GPTZero have 65% and 60%, respectively. Originality.ai holds great importance on all six dimensions. It has above 95% stability, between the highest and lowest score it does not drop by more than 15 points, while Turnitin drops by 20 points and GPTzero drops by 25 points. The result indicates that Originality.ai's detection system is more likely to focus on underlying patterns within the language and structure of the text, which can be signs of AI-generated content, as opposed to on the surface level. Moreover, the human expert validation sub-study showed three important insights. Paradoxically, even trained human readers aren't always accurate at spotting AI-written text (76.3% on average), whereas it was expected that they'd be able to tell without needing any technical assistance. Second, the agreement rate between the experts was also significant ($\kappa = 0.62$), especially for texts paraphrased in two languages, and non-native English, showing that there is no 'gold standard' for identifying AI texts among professors. Third, faculty felt the most confident in Originality.ai after examining the tools' outputs, with a caveat that this was only if the text length exceeded 300 words (77%), falling to 51% for shorter pieces. Notably, faculty confidence in all tools (62-77%) was higher than the accuracy of the tools in challenging situations (e.g., paraphrased text, where Originality. ai accuracy dropped to 80%). That's a pedagogical issue that institutions should tackle with training and policy; the lack of trust in the accuracy of the information.

Figure 5: Radar Plots for Three Top Detection Tools

4. Discussion

Five AI detection tools, Originality.ai, Turnitin, GPTZero, Copyleaks and ZeroGPT—were evaluated in this systematic review using the following metrics: sensitivity, specificity, variation in text length, techniques used to manipulate the text, and methodological quality of the research. Results on Originality. ai are also consistently better than all the other tools compared on virtually every condition and metric studied, demonstrating its superior performance. Everything that's mentioned so far suggests that Originality.ai is a better AI detection tool. Firstly, in terms of balanced accuracy (Figure 2), Originality.ai had a 95% sensitivity rate and a 94% specificity rate, which means that it was the most balanced in identifying AI-generated content while also minimizing false positive or negative ratings for human writers. This balance is crucial: false positives can have large repercussions on students, academics, and for professionals, the integrity of detection efforts can be threatened by false negatives. Second, Originality.ai was very robust with text length (Figure 3). Yet, even when writing is short and under 300 words, which is a challenging situation for all detection systems because there may not be much linguistic evidence, Originality.ai managed to achieve a balanced accuracy rate of 73%, whereas other tools managed a rate of about 50%, and Turnitin managed a 55% balanced accuracy. When applied to longer texts (over 600 words), Originality.ai achieved a balanced accuracy of 96.1%, which is close to perfect classification. The gradient shows that Originality.ai's algorithms are able to identify useful patterns in brief text and can leverage longer text to identify and capitalize patterns more effectively than some of its competitors.

Third, Originality.ai was found to be more resistant to common plagiarism strategies such as paraphrasing and round-trip translation. All tools showed some performance drop, but Originality.ai's results were not as drastic as Turnitin's (80% and 75% when it comes to reorganizing paragraphs and translating them into another language respectively), and GPTZero's (80% and 75%). This indicates that Originality.ai's detection techniques focus more on finding statistical or syntactic patterns that are common in the outputs of LLM models, rather than on the exact order of words or exact phrasing. Fourth, the ROC analysis (Figure 2 in the ROC plot) validated Originality.ai's best operating parameters: 98% true positive rate and 18% false positive rate. This is the

best position among all the tools examined of being closest to the ideal classifier (100% true positive, 0% false positive). Notably, Turnitin's false positive rate was lower (15%) but its true positive rate was significantly lower (75%) which would lead to the loss of a quarter of AI-generated texts, an undesirable rate of error in many contexts.

Lastly, the percentage of faculty members who rate the trustworthiness of these tools (70% for Originality.ai; 65% for Turnitin; and 60% for GPTZero) suggest that these benefits are acknowledged in practice. Trust goes beyond being a subjective value: It is a measure of how practical and reliable a tool is in an educational setting where the use of AI goes undetected and false accusations have serious consequences. All of this makes Originality.ai the latest in the most advanced AI text detection for general use. Its balanced accuracy, length robustness, manipulation resistance and practitioner trust make it the most suitable option for institutions, publishers, and organizations needing automated AI detection, given the requirement that output is suitable in length (ideally >300 words) and is subject to the interpretation of informed human readers.

While these constraints are there, several practical implications arise. No single detection tool should be used as the ultimate answer to academic integrity issues. Even the Originality.ai has a 5% false negative rate and 6% false positive rate under optimal conditions, and higher rates when the text is short or manipulated. It is recommended that the best tools be used to triage and flag, and then checked by a human, contextual information (such as writing style changes, metadata, revision history) be reviewed, and that dialog with the writer be used where appropriate. Future work should focus on research that leverages real student work, validated ground truth (such as human annotations or structured classroom questions), and a variety of AI generation models, including GPT-4 and more recent architectures. Detectability performance over time and model version changes is a critical need for longitudinal designs. In addition, there is a need for the creation and validation of assessment systems for non-English languages, and for hybrid human-AI writing, that is closer to the real-world situation than to pure AI writing.

5. Conclusion and Future work

Among the tools evaluated in the included studies, Originality.ai showed higher reported sensitivity and specificity compared to Turnitin, GPTZero, Copyleaks, and ZeroGPT. However, direct comparisons are confounded by differences in test datasets, detector versions, and evaluation protocols across studies. No single tool can be identified as universally 'best' based on current heterogeneous evidence. Despite this, the evidence base underlying the methods has major methodological flaws such as the use of lab-generated text, the use of older AI models and the non-verifiable ground truth, which means caution in generalizing the findings to real-world educational and professional settings. High-stakes decisions cannot be made solely based on any AI detection tool, as none of these technologies are 100% reliable. None of the AI detection tools are foolproof and human judgment is essential for making high-stakes decisions. There are several areas that should be explored in future research to further develop the evidence based on the use of AI detection tools. First, there is a clear need for longitudinal studies to examine the performance of detection tools as they are updated to newer models (such as GPT-5, GEMINI Ultra, Claude 4) and later generations of LLM's are released. Second, researchers should also abandon the use of lab-generated test and abandon the use of text with lab-generated ground truth. Instead, they should use authentic student submissions, with carefully curated, rigorously labeled ground truth (preferably from controlled classroom assignments or independent human labeling) to better represent real-world performance. Thirdly, future research is needed to examine the accuracy of the detection in non-English languages and multi-language environments, where the evidence is currently almost all limited to English. Fourthly, iteratively rewriting, style transfer, hybrid human-AI composition, and adversarial prompting should be rigorously tested to assess the actual strengths of the most popular tools, such as Originality.ai. Finally, there is a need for a common benchmarking structure and a common data set that facilitates direct and reproducible comparisons of tools and overtime and thereby helps to minimize the current methodological heterogeneity that hampers meta-analytic synthesis.

References

1. Li, B., Yang, P., Sun, Y., Hu, Z., & Yi, M. (2024). Advances and challenges in artificial intelligence text generation. *Frontiers of Information Technology & Electronic Engineering, 25*(1), 64-83.
2. Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business ethics, the environment & responsibility, 32*(1), 201-210.
3. Koplin, J. J. (2023). Dual-use implications of AI text generation. Ethics and Information Technology, 25(2), 32.
4. Zhong, Z., & Xie, X. (2024). Clinical applications of generative artificial intelligence in radiology: image translation, synthesis, and text generation. *BJR| Artificial Intelligence, 1*(1), ubae012.
5. Akram, A. (2023). An empirical study of AI generated text detection tools. *arXiv preprint arXiv:2310.01423*.
6. Mubeen, M., Muskan, A., Akram, A., Rashid, J., Alshalali, T. A. N., & Sarwar, N. (2025). Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method. *International Journal of Computational Intelligence Systems, 18*(1), 174.
7. Sardinha, T. B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics, 4*(1), 100083.
8. Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys, 54*(11s), 1-38.
9. Malik, M. A., & Amjad, A. I. (2025). AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini?. *Journal of Applied Learning & Teaching, 8*(1), 91-101.

10. Diwan, C., Srinivasa, S., Suri, G., Agarwal, S., & Ram, P. (2023). AI-based learning content generation and learning pathway augmentation to increase learner engagement. Computers and Education: *Artificial Intelligence, 4*, 100110.
11. Iqbal, T., & Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences, 34*(6), 2515-2528.
12. Rethlefsen, M. L., & Page, M. J. (2022). PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. *Journal of the Medical Library Association: JMLA, 110*(2), 253.
13. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2023). A declaração PRISMA 2020: diretriz atualizada para relatar revisões sistemáticas. *Revista panamericana de salud publica, 46*, e112.
14. 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.
15. Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity, 22*(1), 4.
16. Giray, L., Sevnarayan, K., & Maphoto, K. B. (2026). GPTZero and the challenges of AI detection in assessing writing. *Assessing Writing, 69*, 101078.
17. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 1-39.
18. Liu, J. Q., Hui, K. T., Al Zoubi, F., Zhou, Z. Z., Samartzis, D., Yu, C. C., ... & Wong, A. Y. (2024). The great detectives: humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity, 20*(1), 8.
19. Adam, G. A., Cui, A., Thomas, E., Napier, E., Shmatko, N., Schnell, J., ... & Lee, D. (2026). Gptzero: Robust detection of llm-generated texts. *arXiv preprint arXiv:2602.13042*.
20. Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., ... & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta neurochirurgica, 167*(1), 214.
21. Sun, Y., Liao, Y., & Ma, X. (2026). Trusting AI to detect AI? A systematic evaluation of the reliability and robustness of current AIGC detection tools for student academic work. *Computers & Education,* 105616.
22. Orenstrakh, M. S., Karnalim, O., Suarez, C. A., & Liut, M. (2024, July). Detecting LLM-generated text in computing education: Comparative study for ChatGPT cases. In 2024 IEEE *48th Annual Computers, Software, and Applications Conference* (COMPSAC) (pp. 121-126). IEEE.
23. Dik, S., & Erdem, O. (2025). Assessing GPT-Zero's Accuracy in Identifying AI vs. Human-Written Essays. *Proceedings of International Mathematical Sciences, 7*(2), 54-58.
24. Paullet, K., Pinchot, J., Kinney, E., Stewart, T., (2025). Detecting AI-Generated Writing Using GPTZero. *Information Systems Education Journal 23*(6) pp 44-52.
25. Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education, 49*, 100321.
26. Shen, S., & Wang, K. (2026). Detecting AI-Generated Content in Academic Peer Reviews. *arXiv preprint arXiv:2602.00319*.
27. Deep, P. D., Edgington, W. D., Ghosh, N., & Rahaman, M. S. (2025). Evaluating the effectiveness and ethical implications of AI detection tools in higher education. *Information, 16*(10), 905.
28. Yu, S., Luo, M., Madasu, A., Lal, V., & Howard, P. (2024). Is your paper being reviewed by an llm? investigating ai text detectability in peer review. *arXiv preprint arXiv:2410.03019*.