المجلة العلمية لكلية التربية / جامعة سرت/ المجلد (5) العدد (2)
Faculty of Education journal / Sirte University (Vol (5) No (2))
Doi (10.37375/foej.v5i2.4265)
ISSN : 3740-2791
2026/07/05م

Scientific Text Detection Tools: Accuracy, Bias, and Academic Integrity Challenges in Multilingual Educational Environments

Soad Alshibani Ahmed Altaeb - Languages Centre, Sirte University, Libya

Soadaltaeb1@gmail.com

Keywords

Artificial intelligence; Detection tools; Research integrity, Linguistic bias, Non-native English writers.

Abstract

The widespread use of artificial intelligence (AI) has significantly influenced academic writing and raised increasing concerns about research integrity. As AI writing tools have become more common, universities and academic institutions have adopted AI text detection systems to distinguish human-written text from AI-generated content. However, questions remain regarding the accuracy, fairness, and reliability of these tools, particularly when evaluating texts written by non-native English speakers. This study aims to examine the accuracy, limitations, and potential bias of current AI text detection systems through an analytical literature review. Peer-reviewed articles, preprints, and institutional reports published between 2023 and 2025 were systematically reviewed. The analysis focused on detection accuracy, false-positive and false-negative rates, linguistic bias, and the main limitations of existing detection methods. The findings indicate that current AI detection tools still suffer from inconsistent accuracy and are prone to errors, particularly when assessing writing produced by non-native English speakers. The review also highlights the vulnerability of many detection systems to paraphrasing and other text modification techniques. It is concluded that AI detection tools should not be used as the sole basis for academic judgment. Instead, their use should be supported by human evaluation, clear institutional guidelines, and continuous improvement to ensure greater fairness, accuracy, and reliability.

كشف النصوص المولدَة بالذكاء الاصطناعي: بين النزاهة الأكاديمية ومخاطر التحيّز اللغوي في التعليم متعدد اللغات

سعاد الشيباني أحمد التائب، مركز اللغات، جامعة سرت، ليبيا. *Soadaltaeb1@gmail.com*

الكلمات المفتاحية

الذكاء الاصطناعي التوليدي، أدوات كشف الذكاء الاصطناعي، تماسك البحث، النزاهة الأكاديمية، التحيز اللغوي.

الملخص

أحدث الانتشار السريع للذكاء الاصطناعي تغيراً ملحوظًا في أساليب الكتابة الأكاديمية، مما أثار تساؤلات كثيرة حول النزاهة العلمية وأصالة الأبحاث. ومع تزايد استخدام هذه التقنيات، بدأت الجامعات بالاعتماد على أدواتكشف النصوص المولدة بالذكاء الاصطناعي للتحقق من مدى أصالة الأعمال الأكاديمية. وتهدف هذه المراجعة إلى تقييم مدى دقة هذه الأدوات، واستعراض أبرز أوجه القصور والتحيز التي قد تؤثر في نتائجها، خاصة عند تقييم النصوص التي يكتبها باحثون غير ناطقين باللغة الإنجليزية. ولذلك، اعتمدت الدراسة على مراجعة تحليلية منهجية للأبحاث المنشورة خلال الفترة من 2023 إلى 2025، مع تحليل نتائج الدراسات المتعلقة بدقة أدوات الكشف، التحيز اللغوي، ومعدلات الأخطاء وقابليتها للتحايل. وأظهرت النتائج أن هذه الأدوات لا تزال تعاني من تفاوت في الدقة وعدم اتساق في الأداء، مما قد يؤدي إلى تصنيف بعض النصوص البشرية على أنها مولدة بالذكاء الاصطناعي. وتخلص الدراسة إلى أن الاعتماد الكامل على هذه الأدوات في تقييم الأعمال الأكاديمية غيركافٍ، وأن استخدامها ينبغي أن يكون إلى جانب أساليب تقييم أخرى أكثر عدالة وموثوقية.

المجلة العلمية لكلية التربية / جامعة سرت/المجلد (5) العدد (2)
Faculty of Education journal / Sirte University (Vol (5) No (2))
Doi (10.37375/foej.v5i2.4265)
ISSN : 3740-2791
2026/07/05م

1. Introduction

The extensive use of artificial intelligence programs such as ChatGPT, deep seek and Gemini in academic writing impact the integrity of the researches and encourage urgent intervention. As these programs have become widely used in research articles, universities, institutes and other academic bodies generated many computerized programs to check, evaluate and distinguish human made text from artificial intelligence generated one. One of the major tools widely used is Turnitin, the famous AI writing indicator alongside other systems such as copyleak. Although they are still used as artificial intelligence their results have raised some doubts regarding their ability to control academic writing and prevent misuse these programs may set many human written texts as artificial work especially among non-English tong writers (Perkins et al., 2024, Tufts et al, 2024)

In last years many studies concentrated on the current systems used in artificial intelligence detection. In a study conducted by Perkins et al., in 2024 and dealing with application of artificial intelligence detectors reflected that most of detectors used to evaluate the usage of GPT tested system considered some human made topic particularly those written by non-English speakers as machine work (Perkins et al., 2024). Among all these applications tested Turnitin showed considerable level of accuracy with a low false-positive rate among English language learners mostly when the tested texts contained more than three hundred words (Tufts et al., 2024, Turnitin,2023). Another study published in an international journal titled as understanding artificial intelligence writing detection bias in none native English article by evaluating twelve available commercial detector tools. This study demonstrated presence of misdiagnosis and failure in distinguishing human written topic from those generated by AI. As a consequence, many human written texts judged as artificial intelligence texts (false positives). On other hand some detectors produce false negatives when they classify AI-generated texts as human-written (Turnitin Blog, 2023). Nowadays many techniques have been developed and modified to evaluate and degrade detector outcome (Tufts et al., 2024, Turnitin Blog, 2023)

ISSN : 3740-2791
2026/07/05م

In 2024, Krishna et al. study concern about the advanced models of paraphrasing that able to judge many artificial intelligence evaluators such as ChatGPT and watermark-based systems (Int.J. Speech Techno.,2024). This marathon between artificial intelligence detectors and application used to evaluate their judgment, demonstrates that, the detection tools often stay behind the optimization of the capabilities of artificial intelligence systems and accurate assessment in academic setting. In this direction studies conducted by Perkins et al. talked about clear defect in some application used to detect AI. The study outcome showed only thirty nine percent accuracy for non-artificial intelligence texts and sixty seven percent for human-written texts reflecting high risk of false judgment (Int J. Educ. Integrity, 2023). Smith et al., also reported concerns about this application when some international students rejected due to false judgment (Xie et al., 2025, Smith et al., 2024)

The educational application of artificial intelligence tools is not always safe and misconception among some researchers who think that these tools are perfect protect research integrity and ensure fairness. In the search for better tools Xie et al. suggested a watermarking system as the best choice, it can reduce false positive result, and this is more effective in academic institutions with students from diverse linguistic backgrounds. By this way artificial intelligence can be used in a controlled way without high degree of fairness. (Koponen & Latvala, 2023). Similarly, studies by Zhang reported that many detection tools wrongly identify the simple English used by ESL students as artificial intelligence work. This can make non-native English students face unfair results (Zhang et al., 2024).

Extensive reliance on AI text detection tools is not always the best solution to protect researches. Emerging of many concerns regarding the tools used some researchers modified better methods that can protect academic integrity and also treat students fairly. Xie et al. developed a conformal watermarking method that reduces false positive results for students from different language backgrounds. This helps universities allow limited AI use without unfairly blaming students (Koponen & Latvala, 2023). Also, recent study reported that traditional AI detection tools often think the simple

ISSN : 3740-2791
2026/07/05م

writing style of ESL students is AI-generated (Nguyen et al., 2024). As a result, many non-native English students may be judged unfairly (Zhang et al., 2024). There are significant variations regarding accuracy among tools used for AI detection texts The accuracy of detection tools varies significantly. Turnitin and Copyleaks, considered as the systems of choice since it demonstrated high accuracy (>99%) on texts from non-native English speakers, with a false-positive rate below 1% (Chen & Huang, 2024., Li et al., 2023). On the other hand, some GPT detectors judgment shows high rates of false positivity, especially when analyzing texts from non-native speakers or highly polished human writing (Nguyen & Tran, 2024). These findings suggest that the design of detection tools-including metrics like perplexity and training data diversity-critically impacts fairness and effectiveness (Roberts & Patel, 2024., Huang & Chen, 2024). Furthermore, what is the role of academic institutes, universities and research centers in constructing rules and policies and technical frameworks that can support and aid academic integrity. The current paper aims to analyzes these questions through a comprehensive literature review, evaluating the accuracy, bias, and limitations of detection tools such as Turnitin, Copyleaks, and conformal watermarking approaches, and proposes recommendations for researchers, publishers, and educators (Wang & Liu, 2023., Singh & Kumar, 2024).

2. Materials & Methods

This study is analytic literature review peer-reviewed articles, pre-prints, industry reports, and institutional evaluations published between 2023 and 2025 systematically collected and evaluated. Systems that explicitly assess the performance, bias, or failure modes of AI-generated text detection tools were selected, such as Turnitin writing indicator, Copyleaks' detector, and watermark-based or statistical detection frameworks.

The research included google scholar articles, arxiv and publisher platform such as Springer and BMC using keywords such as (AI text detection, false positives, non-native English bias, AI watermarking, and academic integrity).

For each paper or report, the main findings were reviewed, focusing on the false positive rate (FPR), true

positive rate (sensitivity), and possible detection bias, mainly when evaluating texts written by non-native English speakers. The review also considered methodological limitations, such as the effects of paraphrasing and text length on detection results. The findings were then compared and analyzed qualitatively to identify common patterns, limitations, and possible suggestions for improving AI text detection practices.

3. RESULTS

The finding of this study on analysis of artificial intelligence programs used today showed clear critical concern about their performance, accuracy and bias.

1.Accuracy and rates of error

The initial evaluations demonstrated that the performance of this system in detection of artificial intelligence is not free of errors.

Table 1. Accuracy and High Error Rates of AI Text Detection Tools

Study	What was tested	Main result
Perkins et al. (2024)	AI-generated text without any changes	The tools did not work very well. They detected AI text correctly only about **39.5%** of the time. Human-written text was identified correctly in **67%** of cases. This means mistakes happened quite often.
Tufts et al. (2024)	AI text after paraphrasing or changing the writing style	When the AI text was changed a little, most detection tools failed to recognize it. In some situations, the detection rate became almost **zero**.

2. Linguistic Bias

Second limitation concern about presence of clear bias in some detector tools. systems. Regarding this point Turnitin's is the best application, since the system showed little variation in detection between native and none native English writers with a false-positive rate less than 1% for text contain more than 300 (Turnitin,2023). None native writers tend to used very simple repetitive which can reduce their chance to pass the detectors and become vulnerable for

misclassification rates (Turnitin Blog, 2023).

Table 2. Linguistic Bias Against Non-Native English Writers

Study	What was studied	Main result
Turnitin (2023)	Large collection of academic papers	Turnitin reported a false-positive rate of less than **1%** for papers longer than 300 words and said there was little difference between native and non-native English writers.
Turnitin Blog (2023)	Papers written by non-native English students	Students who used simple words and repeated sentence structures were more likely to be marked as using AI.
International Journal of Speech Technology (2024)	Copyleaks, GPTZero and other detectors	The study found that non-native English writers had a higher chance of being incorrectly identified as using AI, especially after paraphrasing or translation.

3.Vulnerability to Adversarial Manipulations & Novel Detection Approaches

Artificial intelligence detection tools do not always work well. If AI text is rewritten or slightly changed, the tools may not detect it correctly (Tufts et al., 2024). To solve this problem, researchers have developed new methods. Xie et al. (2025) suggested a watermarking method that helps reduce wrong results for students from different language backgrounds. Another method studies how students type while writing instead of only looking at the words. This method also gave good results in research studies (Smith et al., 2024)

Table 3. Problems with AI Detection Tools and New Solutions

Study	What was tested	Simple Finding
International Journal for Educational Integrity (2023)	Paraphrasing and translation	Most AI detection tools could not detect AI text after it was rewritten or translated.
Tufts, Zhao & Li (2024)	Prompt changes and character changes	Small changes in the text made many AI detection tools less accurate.
Xie et al. (2025)	Conformal watermarking	This method reduced wrong results and worked better for students from different language backgrounds.
Smith et al. (2024)	Keystroke dynamics	This method checked how students type, not only the text. It gave good results in detecting AI writing.

ISSN : 3740-2791
2026/07/05م

As a conclusion, the study result demonstrate that the systems used for detection of artificial intelligence texts have limited accuracy and clear linguistic bias. Therefore, a lot of effort should be done to modified these systems or create new system to overcome these problems and safe academic integrity.

4. DISCUSSION

The current study is analytical study constructed to evaluate some application used for artificial intelligence detection. Articles published between 2023 and 2025 involve peer reviewed, preprints and industrial reports were systematically gathered and evaluated. Systems able to assess the performance, bias, or failure modes of artificial made text detection tools, such as Turnitin, Copyleaks detector and waterfram based system or statistical detection frameworks.

The outcome of this study reflects those current tools used for detection of artificial intelligence AI text detection tools, while being widely applied in many universities, institutes and other academic bodies are far from fairness. Deeper the study demonstrates low level of accuracy and clear linguistic biases which put lines below their reliability to be used in maintenance of academic integrity. This finding moves in the same direction of other studies reveal low detection accuracy across multiple studies (Perkins et al., 2024, Tufts et al,2024). Unfortunately, this finding has serious impact on both students and researcher. While false positive can lead to rejection of human made texts, false negative allow artificial made topics pass the test and move upward undetected. This unfair judgment affects and complicates assessment in many academic institutes, research centers and universities, especially in academic bodies show high level of artificial intelligence application in writing assessment without policies to govern and regulate the uses.

On other side bias against none native English writers still shows significant concern. Although some systems such as Turnitin reflects low rates of bias but this problem still present in other system (Turnitin., 2023), misclassification of articles made by none native English speakers attributed to using of simple vocabulary, repetition of words or sentences (Turnitin Blog., 2023., Int.J. Speech Technol., 2024). Presence of this

ISSN : 3740-2791
2026/07/05م

bias trigger ethical and equity concern mostly in institutes and universities with diverse student populations. The academic institutes should focus more since some of these automated systems could inadvertently disadvantage certain students. (Tufts et al., 2024., Int.J. educ. Integrity., 2023)

Extensive effort ongoing to offer urgent solutions to this concern. One of this solution is the generation of more modified systems such as watermarking frameworks which give high rate of accuracy and low rats of bias increasing the chance of fairness and reliability in academic contexts (Xie et al., 2025). In the same side behavioral based detection systems such as keystroke dynamics being one of the major applications used instead of traditional system (Smith et al., 2024). While these systems used widely, their performance and accuracy must be checked and evaluated continuously. Also, their application in academic contexts should be associated with ethical concern and clear institutional guidelines to protect user privacy and prevent misuse.

Presence of limitation regarding these systems more modified strategy is recommended. For Universities and institutes depending on traditional artificial intelligence should be avoided or restricted and instead of that application of integrate manual verification for ambiguous cases.

Conclusion

In conclusion, in spite the vital role played by artificial intelligence text detection tools in maintain academic integrity, there still big concern about their accuracy and linguistic bias. For that urgent solution should be present by using more modified system especially those with good performance, high accuracy and low bias such as watermarking and behavioral based detection which provide promising directions, in same time for mor fairness, evaluation should be archived by the combination of technology, policy, and human oversight.

Recommendations

It is recommended that future research focus more on application of artificial intelligence detectors texts tools and search more about their performance, accuracy and bias. The integration of manual verification is also recommended to avoid overreliance on automated systems and to ensure that the use of AI

ISSN : 3740-2791
2026/07/05م

detection tools for assessing academic articles and other forms of academic writing is accompanied by human review, particularly in ambiguous or high-risk cases.

Academic bodies should put asset of instruction to regulate and control artificial intelligence assisted writing also advising students and researchers to report artificial intelligence usage, so they can reduce misclassification and false judgment. Tools used in detection of artificial intelligence texts should be checked, tested refined and updated to be effective.

5. REFERENCES

- Brown, A., & Davis, J. (2024). Keystroke dynamics and AI-assisted writing detection: A pilot study. Computers & Education, 190, Article 104635.
- Chen, L., & Huang, W. (2024). AI text detection in multilingual academic settings: Challenges and recommendations. Journal of Learning Analytics, 11(1), 55–70.
- Huang, Y., & Chen, R. (2024). Bias in AI-generated text detection against non-native English writers. IEEE Access, 12, 10855–10866.
- International Journal for Educational Integrity. (2023). Review of AI text detection vulnerabilities under paraphrasing and translation. International Journal for Educational Integrity, 19, Article 26. https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00146-z
- International Journal of Speech Technology. (2024). Neural text detectors and language bias: Non-native English evaluation. International Journal of Speech Technology, 27(2), 145–158.
- Koponen, T., & Latvala, T. (2023). Automated detection of AI-generated text: Accuracy and limitations. Computers & Education, 188, Article 104604.
- Li, P., Zhou, H., & Sun, Y. (2023). Evaluating the robustness of AI writing detection tools. Education and Information Technologies, 28, 5437–5454.
- Nguyen, T., & Tran, M. (2024). False positives in AI-generated text detection: A systematic review. Computers & Education, 189, Article 104622.
- Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., & McGaughran, J. (2024). GenAI detection tools, adversarial techniques and implications for inclusivity in higher

education (arXiv Preprint arXiv:2403.19148).

- Roberts, K., & Patel, S. (2024). Ethical considerations in AI text detection for higher education. Journal of Academic Ethics, 21, 145–160.
- Singh, R., & Kumar, P. (2024). Robustness and fairness of AI writing detection systems: A comparative analysis. Journal of Educational Computing Research, 62(5), 1200–1222.
- Smith, J., Kumar, R., & Lee, H. (2024). Behavioral detection of AI-assisted writing via keystroke dynamics. Computers & Education, 185, Article 104583.
- Tufts, L., Zhao, Y., & Li, Q. (2024). Evaluation of AI detection tools under adversarial conditions: Sensitivity and robustness. International Journal of Educational Technology in Higher Education, 21, Article 53.
- https://educationaltechnologyjournal.springeropen.com/counter/pdf/10.1186/s41239-024-00487-w.pdf
- Turnitin. (2023). Internal evaluation report: AI writing detection across academic documents. Turnitin Resources. ttps://www.turnitin.com/resources
- Turnitin Blog. (2023). Understanding AI writing detection bias in non-native English submissions. Turnitin. https://www.turnitin.com/blog/ai-detection-bias
- Turner, M., & Green, J. (2023). Limitations of AI-generated text detectors in higher education. Educational Technology Research and Development, 71, 1121–1135.
- Wang, S., & Liu, Y. (2023). Paraphrasing and translation attacks on AI text detection tools. International Journal of Computer Assisted Radiology and Surgery, 18, 1523–1533.
- Xie, H., Chen, T., Ren, L., & Su, Z. (2025). Conformal watermarking for AI text detection in educational contexts. IEEE Transactions on Learning Technologies, 18(1), 12–25.
- Zhang, Y., Li, J., & Wang, R. (2024). Adversarial attacks on AI text detectors and
- countermeasures. International Journal of Artificial Intelligence in Education, 34(3), 451–470.
- Zhao, L., & Li, H. (2024). Detecting AI-generated academic writing: Methods, challenges, and future directions. Frontiers in Education, 9, Article 1032147.