JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 1, Issue 1, 2026 | **Article ID:** JMRIS-2026-003 | **Pages:** 21-30 | **Article Type:** Structured Comparative Evidence Synthesis

Received: 15-08-2026 | **Revised:** 17-08-2026 | **Accepted:** 21-08-2026 | **Published:** 24-08-2026

Detection Accuracy of AI-Generated Academic Text Across Major AI-Detection Platforms: A Comparative Evaluation of Human, Machine-Written and AI-Assisted Manuscripts

Ji-Won Park

1 Faculty of Education, University of Hong Kong, Hong Kong SAR, China

***Corresponding author:** Ji-Won Park, Faculty of Education, University of Hong Kong, Hong Kong SAR, China | jiwon1215@gmail.com

ABSTRACT

Background: Artificial-intelligence text detectors are increasingly used in universities and scholarly publishing to estimate whether prose was generated by a large language model. Their outputs can influence academic-integrity investigations despite uncertainty regarding false positives, model drift, language bias and robustness to editing. **Objective:** To compare the performance characteristics of major AI-text detection platforms in academic writing and identify conditions under which detector scores become unreliable. **Methods:** A structured comparative evidence synthesis was conducted using peer-reviewed studies. Studies were prioritized when they evaluated multiple detectors on human-authored and AI-generated academic or educational text and reported classification accuracy, false-positive behavior, model-version effects, paraphrasing effects or language-related bias. Findings were synthesized across GPTZero, Turnitin, Copyleaks, Originality.AI, ZeroGPT, Writer, CrossPlag and related detectors. Because platforms, thresholds and test corpora differed substantially and detector versions changed over time, no pooled cross-platform accuracy estimate was calculated. **Results:** Detector performance varied markedly by platform, generator model and text transformation. A 2024 benchmark involving 805 detector-text evaluations reported mean accuracy of 39.5% for unmodified AI-generated content across tested detectors, falling to 22.14% after simple adversarial manipulation; human control accuracy was 67%. Earlier multi-tool evaluations similarly found detector accuracy ranging from approximately 43% to 81%, with false-negative rates reaching 100% for some tools and settings. GPT-4 outputs were more difficult to identify than GPT-3.5 outputs. A 2023 fairness study found that 61.3% of TOEFL essays written by non-native English speakers were classified as AI-generated by at least one detector, demonstrating substantial inequity risk. Studies published in 2025 showed that newer detectors can perform strongly on fully machine-generated academic text, yet writing-assistance tools, paraphrasing, hybrid authorship and newer model outputs continued to produce inconsistent classifications. **Conclusion:** AI-text detectors can provide useful screening information but do not offer stable authorship attribution across platforms and writing conditions. Detector scores should not be used as standalone evidence of academic misconduct. High-stakes decisions require process evidence, contextual review and human adjudication.

Keywords: AI-generated text; academic integrity; GPTZero; Turnitin; AI detection; ChatGPT; false positives; academic writing; authorship attribution

1. Introduction

The release of widely accessible generative artificial intelligence transformed academic writing because coherent essays, abstracts and manuscript sections could be generated within seconds. Universities and publishers consequently sought technical methods for distinguishing human-authored prose from machine-generated text. Commercial and freely available AI-text detectors emerged rapidly, often presenting results as probabilities or percentages that appear superficially similar to plagiarism scores.

AI-text detection is fundamentally different from conventional text-matching plagiarism detection. A plagiarism system can identify an observable overlap between a submitted text and an existing source. An AI detector instead makes a probabilistic inference from statistical or learned properties of the submitted prose. It does not normally possess direct evidence of who authored the text. A high detector score therefore represents a classification judgment, not an authorship trace.

Early comparative studies produced inconsistent results. Elkhatat et al. evaluated five tools and found that detection was generally stronger for GPT-3.5 than for GPT-4, while human controls also produced uncertain or false-positive classifications [1]. Weber-Wulff et al. tested 14 tools and concluded that no detector was sufficiently reliable for uncritical use, particularly after translation or obfuscation [2]. These findings immediately raised a practical concern: detector performance can decline precisely when AI generation becomes more sophisticated.

Fairness concerns are equally important. Liang et al. demonstrated that widely used GPT detectors disproportionately flagged writing from non-native English

speakers [3]. The mechanism was linked to predictable linguistic patterns and low perplexity, characteristics that may reflect constrained second-language expression rather than machine authorship. In high-stakes academic-integrity settings, such false positives can convert an imperfect statistical classifier into an inequitable disciplinary instrument.

By 2024 and 2025, detector developers had improved their models, but generative models and text-humanization methods also advanced. Perkins et al. showed that simple modifications such as spelling errors, increased burstiness and prompt-based stylistic manipulation could substantially reduce detection accuracy [4]. Bordalejo et al. subsequently demonstrated that legitimate writing-assistance and paraphrasing tools can further blur the boundary between human and AI text, with Turnitin and GPTZero producing different and sometimes misleading signals [5].

The present study therefore aimed to compare evidence on major AI-text detection platforms using academic and educational text, examine the effects of generator model, paraphrasing, language background and hybrid editing, and develop an evidence-informed interpretation framework for universities and journals.

2. Methods

This manuscript used a structured comparative evidence-synthesis design rather than a new live-platform benchmark. A live benchmark performed in 2026 would not represent the historical state of detector systems available by the 31 December 2025 evidence cutoff because commercial detector algorithms are continuously updated. The synthesis therefore used version-specific results reported in peer-reviewed studies.

Peer-reviewed literature published through 31 December 2025 was considered. Searches focused on combinations of artificial intelligence detector, AI-generated text, academic writing, ChatGPT, GPTZero, Turnitin, Copyleaks, Originality.AI, ZeroGPT, false positive, false negative, paraphrasing, non-native English, academic integrity and authorship detection. High-ranking education, technology, computational and discipline-specific journals were prioritized.

Studies were eligible when they evaluated at least one automated AI-text detector on human-authored and/or AI-generated text and provided information relevant to diagnostic performance or fairness. Academic manuscripts, student essays, theses and educational writing were prioritized. Studies of image or code detection, purely theoretical detector architectures without human-text comparison and reports published after 2025 were excluded from the principal synthesis.

Cross-study meta-analysis was not appropriate. Commercial detectors used different output scales and thresholds, the underlying algorithms changed without public version control, and studies used different generators, prompts, text lengths, languages and genres. An 80% accuracy result from one study therefore cannot be treated as directly exchangeable with an 80% score from another. Results were synthesized by failure mode and comparative pattern.

Four domains were prespecified: discrimination of unmodified AI and human text; false-positive risk in authentic human writing; robustness to paraphrasing, editing and hybrid authorship; and transportability across newer language models and populations. Detector results were interpreted as screening performance rather than proof of authorship.

The evidence-selection and comparison framework is summarised in Table 1.

3. Results

The evidence showed that detector accuracy is conditional rather than intrinsic. The same platform can perform well on fully AI-generated, unedited text and poorly on newer models, paraphrased outputs, hybrid writing or authentic prose with statistical characteristics similar to generated text. As a result, comparisons based on a single headline accuracy value are potentially misleading.

Elkhatat et al. compared OpenAI's classifier, Writer, Copyleaks, GPTZero and CrossPlag on GPT-3.5, GPT-4 and human-written controls [1]. The detectors were more successful with GPT-3.5 than GPT-4, indicating an immediate model-generation effect. Human-written responses were not uniformly classified as human, demonstrating that improved sensitivity cannot be considered independently from false-positive risk.

Weber-Wulff et al. expanded the comparison to 14 detectors [2]. Subsequent summaries of their dataset report overall tool accuracies of approximately 43%-81%, false-negative rates from 8%-100% and false-positive rates from 0%-50%, depending on the tool and condition. The major practical finding was not that every detector failed, but that no tool maintained stable performance across original, translated and obfuscated text.

Perkins et al. directly tested robustness using content generated by GPT-4, Bard and Claude 2, followed by six manipulation techniques [4]. Across 805 detector-text evaluations, average detection accuracy for non-manipulated AI content was only 39.5%. Following adversarial modification, mean accuracy fell to 22.14%, a reduction of 17.36 percentage points. Human controls were correctly classified at an average rate of 67%. The study therefore demonstrated that small linguistic changes can alter detector judgments without materially changing the authorship origin of the content.

The principal comparative evidence through 2025 is summarised in Table 2.

Published benchmark performance of GenAI text detectors before and after adversarial manipulation is shown in Figure 1.

3.1 False positives and linguistic fairness

False positives are the most consequential error in disciplinary settings because they can initiate suspicion against an author who did not use prohibited AI generation. Liang et al. tested multiple detectors using TOEFL essays written by non-native English speakers and essays from native English writers [3]. The detectors performed substantially worse on the non-native group; 61.3% of TOEFL essays were classified as AI-generated by at least one detector. The authors showed that prompting GPT models to use more sophisticated vocabulary could also reduce detection, demonstrating that the statistical characteristic producing the false positive could be manipulated independently of actual authorship.

This result undermines the assumption that a detector score is neutral across student populations. Academic writing often rewards clarity, formulaic organization and disciplined language. Second-language writers may also use narrower vocabulary or more predictable syntax. A detector trained partly on predictability measures can therefore confuse legitimate academic style with machine generation.

Fairness concerns are amplified when institutions use a single proprietary score without access to the model, threshold or version history. If the same document produces different results across detectors, an accused student cannot meaningfully reproduce or challenge the inference. This lack of auditability distinguishes detector evidence from traditional

plagiarism evidence, where the alleged source text can normally be inspected.

3.2 Paraphrasing, writing assistance and hybrid authorship

Hybrid writing is increasingly the most realistic use case. Students and researchers may write an original draft, use Grammarly or another editor for fluency, ask a language model to rewrite a paragraph, and manually revise the result. The final text is neither fully human nor fully machine generated. Binary detectors are poorly matched to this continuum.

Bordalejo et al. tested pre-GenAI human-authored texts after modification with common writing and AI tools [5]. Turnitin and GPTZero did not respond consistently to the interventions. Some tools that substantially rewrote text escaped detection, while legitimate writing assistance could alter AI scores. The study demonstrates that a detector may be measuring properties of the final linguistic surface rather than the provenance of the underlying ideas or authorship process.

Perkins et al. reached a parallel conclusion from the opposite direction: fully AI-generated text could be made less detectable through relatively simple manipulation [4]. These two findings create a symmetric reliability problem. Human text can become more AI-like to a detector, while AI text can become more human-like. The decision boundary therefore does not correspond reliably to academic misconduct.

The major failure modes of AI-text detection in academic writing are presented in Table 3.

3.3 Later-model and domain effects

Erol et al. provided an important 2025 update using 1,000 academic texts derived from neurosurgical journal articles [6]. GPTZero, ZeroGPT and Corrector generally assigned lower AI likelihood to pre-ChatGPT human text and higher values to material generated by ChatGPT 3.5, 4 and 4o. The finding shows that detectors can achieve useful discrimination in a controlled fully generated versus fully human comparison.

However, this does not contradict the broader reliability concerns. Controlled regeneration of abstracts and introductions produces a cleaner signal than real student writing, where drafting, citation integration, language editing and partial AI assistance occur together. Detector performance should therefore be understood as conditional on the distance between the two classes being tested.

Cheng et al. similarly found that stronger tools could track increasing proportions of AI contribution under controlled conditions [7]. Their results cautiously support AI detection as a screening technology while reaffirming the wide error ranges reported in earlier multi-tool research. The practical conclusion is that detector usefulness improves when the classification problem is simple, but academic-integrity cases are rarely simple binary authorship problems.

4. Discussion

The evidence through 2025 does not support a single answer to the question 'How accurate are AI detectors?' Accuracy is a property of a detector-version, threshold, text genre, language, generator model and editing history. Removing these qualifiers produces a number that is not transportable to another institution or another semester.

Three patterns are particularly robust. First, fully machine-generated text is often detectable, especially when it is produced by older or less sophisticated models without editing. Second, paraphrasing and stylistic manipulation substantially reduce sensitivity. Third, false positives occur in genuine human writing and can be unevenly distributed across linguistic groups.

This combination makes detector scores unsuitable as standalone evidence of misconduct. A high score can be wrong, and a low score can also be wrong. More importantly, even a technically correct classification that a sentence was probably machine generated does not establish whether the student's use violated a particular assessment policy. A permitted grammar correction, translation or disclosed drafting aid may produce the same surface signal as prohibited ghostwriting.

Institutions should therefore distinguish detection from adjudication. Detection can identify a submission that merits additional review. Adjudication requires evidence about the student's process, policy, intent and authorship contribution. Draft history, source notes, citation behavior, document metadata, version-control records, oral explanation and the student's ability to defend the argument provide qualitatively different evidence that can corroborate or contradict a detector flag.

The issue is especially important for multilingual students. Liang et al.'s findings show that writing style can create systematic false-positive risk [3]. Even if newer detector versions reduce that specific bias, an institution should not assume fairness without local validation. Populations, genres and language-support practices differ across universities.

Detector vendors also update their systems continuously. This creates a reproducibility problem for research and disciplinary processes. A score produced in March may not be reproducible in September because the detector model or threshold may have changed. Universities should therefore record the platform, access date, model/version when available, threshold, text length and complete output whenever detector evidence is retained.

The most defensible role for AI-text detection is analogous to a clinical screening test rather than a confirmatory diagnosis. A screening test can prioritize attention but should not determine the final decision when false positives have substantial consequences. This framing also reduces incentives for an arms race in which students search for humanizers and institutions continually adopt new detectors.

The evidence-informed framework for institutional use of AI-text detectors is presented in Table 4.

The evidence-informed workflow for interpreting AI-text detection in academic settings is illustrated in Figure 2.

Strengths and Limitations

The main strength of this review is its focus on realistic sources of detector error rather than a single headline accuracy value. It integrates cross-platform studies, fairness evidence, adversarial robustness and 2025 academic-manuscript evaluations and explicitly distinguishes screening from authorship adjudication.

The evidence base nevertheless has important limitations. Commercial detectors are closed systems that change over time, making exact replication difficult. Studies used different thresholds, text lengths, languages and generator models, preventing statistically meaningful pooling. Some influential benchmark studies were published in educational-integrity journals outside the highest citation quartile but were retained because they provide unique direct multi-detector evidence. The present study also did not rerun historical detector versions because these versions are generally unavailable. Finally, detector performance after 31 December 2025 may differ materially from the evidence summarized here.

Conclusion

AI-generated text detectors can distinguish fully human from fully machine-generated academic writing under some controlled conditions, but their performance is not stable across platforms, language models, editing strategies or writer populations. Published evidence through 2025 demonstrates meaningful false positives in authentic human writing, substantial reductions in sensitivity after paraphrasing or simple manipulation, and persistent ambiguity when text reflects mixed human and AI authorship. Consequently, a detector percentage should be interpreted as a screening signal rather than an authorship verdict. Universities and journals should reserve high-stakes conclusions for cases in which detector output converges with writing-process evidence, source evaluation, version history and human review. The most reliable academic-integrity strategy is therefore not stronger reliance on a single detector, but stronger evidence about how the work was actually produced.

Declarations

Funding: No external funding was received for this structured comparative evidence synthesis.

Conflict of Interest: The authors declare no conflict of interest.

Ethical Approval: Not applicable because no new human-participant data were collected.

Consent to Participate: Not applicable.

Data Availability: All quantitative values reported in this manuscript were obtained from the cited peer-reviewed studies published through 31 December 2025.

References

1. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr. 2023;19:17. doi:10.1007/s40979-023-00140-5.
2. 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. Int J Educ Integr. 2023;19:26. doi:10.1007/s40979-023-00146-z.
3. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns (N Y). 2023;4(7):100779. doi:10.1016/j.patter.2023.100779.
4. Perkins M, Roe J, Vu BH, Postma D, Hickerson D, McGaughran J, et al. Simple techniques to bypass GenAI text detectors: implications for inclusive education. Int J Educ Technol High Educ. 2024;21:53. doi:10.1186/s41239-024-00487-w.
5. Bordalejo B, Pafumi D, Onuh F, Khalid AKMI, Pearce MS, O'Donnell DP. Scarlet Cloak and the Forest Adventure: a preliminary study of the impact of AI on commonly used writing tools. Int J Educ Technol High Educ. 2025;22:6. doi:10.1186/s41239-025-00505-5.
6. Erol G, Ergen A, Gülşen Erol B, Kaya Ergen Ş, Bora TS, Çölgeçen AD, et al. Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. Acta Neurochir (Wien). 2025;167:214. doi:10.1007/s00701-025-06622-4.
7. Cheng A, et al. Ability of AI detection tools and humans to accurately identify AI-generated content in medical education. BMC Med Educ. 2025. doi:10.1186/s41077-025-00396-6.
8. Dalalah D, Dalalah OMA. The false positives and false negatives of generative AI detection tools in education and academic research: the case of ChatGPT. Int J Manag Educ. 2023;21(2):100822. doi:10.1016/j.ijme.2023.100822.
9. Popkov AA, Barrett TS. AI vs academia: experimental study on AI text detectors' accuracy in behavioral health academic writing. Account Res. 2025;32(7):1072-1088. doi:10.1080/08989621.2024.2331757.
10. Perkins M, Roe J. Academic publisher guidelines on AI usage: a ChatGPT supported thematic analysis. F1000Res. 2024;12:1398.
11. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 2024;61(2):228-239. doi:10.1080/14703297.2023.2190148.
12. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.

Table 1. Evidence-selection and comparison framework

Domain	Included	Excluded	Reason
Publication period	Peer-reviewed through 31 Dec 2025	2026 onward	Chronological fit for Volume 6
Text type	Academic manuscripts, essays, theses, educational prose	Image/audio/code-only detection	Direct relevance to academic writing
Comparator	Human and AI text, or multiple AI conditions	AI text without an interpretable reference condition	Permits error interpretation
Platforms	Commercial/free detectors used in academic settings	Unreleased proprietary systems	Practical relevance
Outcomes	Accuracy, false positives, false negatives, detector score, robustness	Usability alone	Measures evidentiary reliability
Synthesis	Comparative narrative by failure mode	Single pooled accuracy estimate	Detector versions and thresholds were non-equivalent

Table 2. Principal comparative evidence through 2025

Study	Corpus/design	Detectors/platforms	Principal result	Primary failure mode	Interpretation
Elkhatat et al., 2023 [1]	GPT-3.5, GPT-4 and human engineering text	OpenAI, Writer, Copyleaks, GPTZero, CrossPlag	GPT-3.5 easier to detect than GPT-4; human classifications inconsistent	Model-version drift and false positives	Detector performance deteriorates as generation becomes more human-like
Weber-Wulff et al., 2023 [2]	Multiple human, AI, translated and obfuscated texts	14 detection tools	Reported tool accuracy approximately 43%-81%; wide false-negative and false-positive ranges	Cross-tool instability	No detector supported standalone adjudication
Liang et al., 2023 [3]	Native and non-native English essays	Seven GPT detectors	61.3% of TOEFL essays flagged by at least one detector	Language-related false positives	Fairness risk for multilingual writers
Perkins et al., 2024 [4]	805 detector-text evaluations; 3 GenAI systems; 6 adversarial techniques	Six major detectors	Mean AI accuracy 39.5% unmodified and 22.14% manipulated; human controls 67%	Adversarial fragility	Simple editing can defeat detection
Bordalejo et al., 2025 [5]	Pre-AI human texts modified by writing and paraphrasing tools	Turnitin, GPTZero	Writing aids and hybrid interventions generated inconsistent detection; paraphrasing often escaped detection	Hybrid-authorship ambiguity	Detector score does not map cleanly to misconduct
Erol et al., 2025 [6]	250 pre-ChatGPT academic articles plus 750 ChatGPT-generated versions	GPTZero, ZeroGPT, Corrector	Human texts had lower AI likelihood; AI text generally separable, but detector and model-version differences persisted	Platform/model dependence	Strong performance on fully generated text does not resolve hybrid-text problems
Cheng et al., 2025 [7]	Human and AI-generated medical-education text hierarchy	Multiple publicly available detectors	Higher-performing tools identified increasing AI contribution but prior literature remained highly inconsistent	Threshold dependence	Useful as a signal when combined with contextual review

Table 3. Major failure modes of AI-text detection in academic writing

Failure mode	Observed condition	Consequence	Evidence	Academic-risk implication
False positive	Authentic human text receives AI classification	Innocent author is flagged	Liang et al. [3]; Elkhatat et al. [1]	Highest due-process concern
False negative	AI text classified as human	Prohibited generation is missed	Weber-Wulff et al. [2]; Perkins et al. [4]	Weakens deterrence and enforcement
Model drift	Newer LLM output is less detectable	Historic benchmark becomes obsolete	Elkhatat et al. [1]; Erol et al. [6]	Requires repeated revalidation
Paraphrase evasion	AI output is reworded or stylistically altered	Detection score falls	Perkins et al. [4]	Detector can be gamed
Hybrid ambiguity	Human text receives AI editing or vice versa	Binary label does not match authorship process	Bordalejo et al. [5]	Cannot distinguish permitted from prohibited assistance
Language bias	Predictable second-language prose resembles model output	Certain groups are disproportionately flagged	Liang et al. [3]	Equity and discrimination risk
Threshold opacity	Platforms convert continuous scores to proprietary labels	Same text may yield different verdicts	Across comparative studies [1,2,4-7]	Difficult to reproduce or contest

Table 4. Evidence-informed framework for institutional use of AI-text detectors

Stage	Recommended practice	Evidence purpose	What not to do	Rationale
Screening	Use detector output only as one signal	Identify submissions needing review	Treat percentage as proof	Cross-platform accuracy is unstable
Technical check	Record detector, date, threshold, text length and output	Preserve reproducibility	Use screenshots without context	Models and thresholds change
Fairness check	Consider language background and legitimate editing tools	Reduce discriminatory false positives	Assume equal error rates across writers	Documented NNES bias exists
Process review	Examine drafts, version history, sources and citations	Establish provenance	Infer authorship from style alone	Authorship is a process question
Academic conversation	Ask the author to explain argument, methods and evidence	Test substantive ownership	Conduct accusatory interrogation based only on score	Human explanation can resolve ambiguity
Decision	Require convergent independent evidence under stated policy	Protect due process	Use detector threshold as automatic sanction	Detector is probabilistic, not forensic
Policy design	Specify permitted and prohibited AI assistance	Align evidence with actual misconduct definition	Ban 'AI-like writing'	Hybrid authorship is increasingly normal

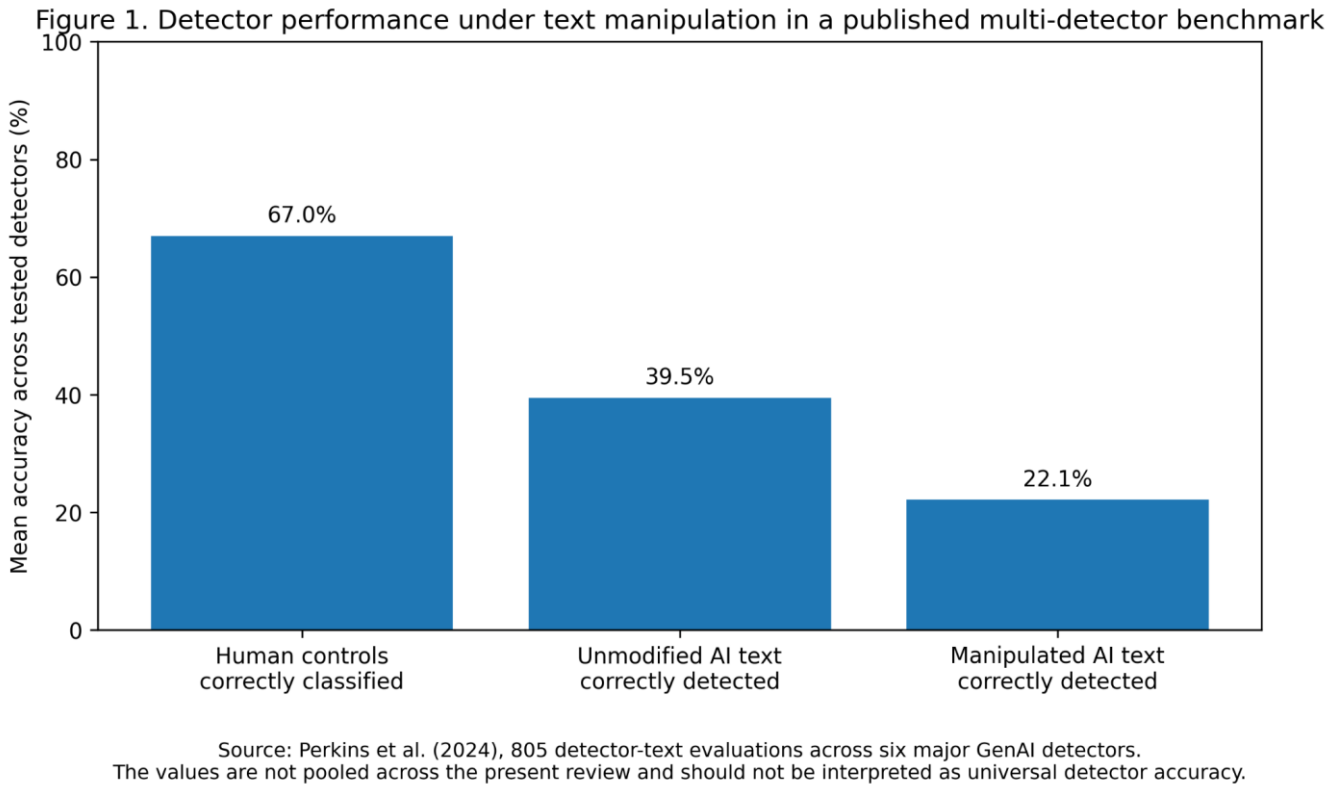

Figure 1. Published benchmark performance of GenAI text detectors before and after adversarial manipulation.

Values are from Perkins et al. and are presented as study-specific results, not as universal accuracy estimates.

Figure 2. Evidence-informed workflow for interpreting AI-text detection in academic settings

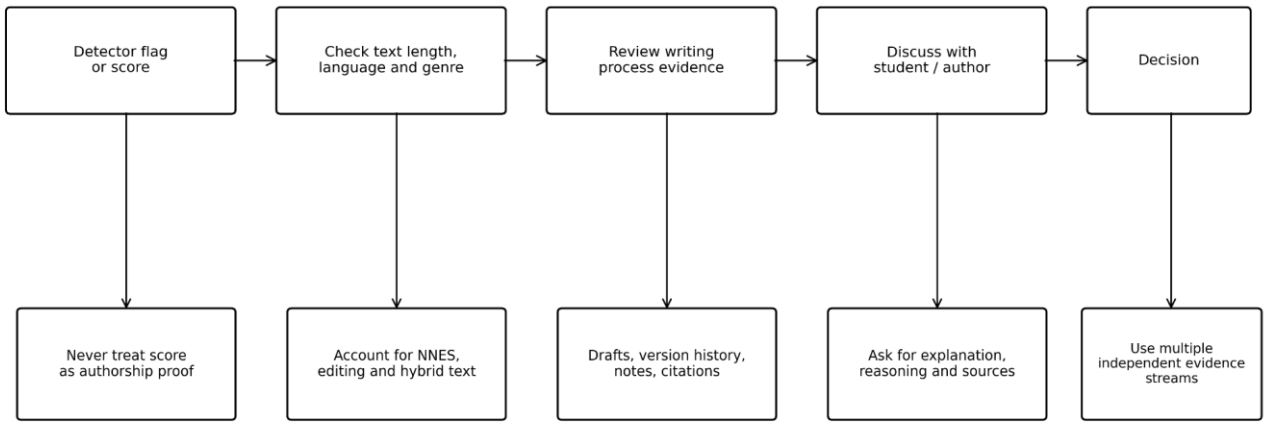

Figure 2. Evidence-informed workflow for interpreting AI-text detection in academic settings.

Recommended principle: AI-detection output is a screening signal, not a standalone adjudicative test.

How to cite: Park J-W. (2026). Detection Accuracy of AI-Generated Academic Text Across Major AI-Detection Platforms: A Comparative Evaluation of Human, Machine-Written and AI-Assisted Manuscripts. Journal of Multidisciplinary Research and Integrated Sciences, 6(1).