American Journal of IR 4.0 and Beyond (AJIRB)

ISSN: 2837-4738 (ONLINE)

VOLUME 4 ISSUE 1 (2025)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of IR 4.0 and Beyond(AJIRB)

Volume 4 Issue 1, Year 2025
ISSN: 2837-4738 (Online)
DOI: https://doi.org/10.54536/ajirb.v4i1.3795
https://journals.e-palli.com/home/index.php/ajirb

Accuracy and Reliability of AI-Generated Text Detection Tools: A Literature Review

Jezreel Edriene J. Gotoman[1], Harenz Lloyd T. Luna[1], John Carlo S. Sangria[1], Cereneo S. Santiago Jr.[1*], Danel Dave Barbuco[1]

Article Information

Received: September 15, 2024

Accepted: October 19, 2024

Published: February 18, 2025

Keywords

Artificial Intelligence, AI Detection, AI-Generated Text, AI Text Detector, AI Text Accuracy

ABSTRACT

Artificial intelligence has become a significant tool for completing a wide range of tasks, from simple to complex, though its use is subject to various considerations and preferences. This study explored one aspect of the varied usages of artificial intelligence, the AI-generated text (AIGT) detection. This study used a literature review wherein pertinent studies were gathered and selected to discover potential implications. Research objectives were defined to assess the accuracy and reliability of the AI text detectors and identify which AI detectors were evaluated. Three online databases were used to search for relevant literature, of which 34 articles were finalized. Results show that despite most detectors attaining accuracy above 50%, they are unreliable. Paid tools generally perform better than free ones, but there are concerns about bias against non-native English speakers. These tools also struggle with sophisticated AI content and tricks like paraphrasing, so using them carefully and relying on human judgment is important to avoid unfairly discrediting someone's work. AI-generated text detection technology still has a lot of room for improvement. Users should not rely completely on these tools but rather cooperate with those tools to better find the true writer of a text. Hence, authorities who use these AI detectors should only partially trust these tools, for they are imperfect and can still make mistakes in their judgment.

INTRODUCTION

Evaluating the usefulness, accuracy, and reliability of AI content identification systems in discriminating between AI-generated material and human-written content remains a less researched field (Chaka, 2023). AI detection is the process of determining whether information is created by AI or humans, often done using a detection tool that employs natural language processing and machine learning to find patterns usually present in artificially created content. AIGT detectors' primary reason for usage is to distinguish whether a text is genuinely made by a human author or generated through a machine such as a computer or a fellow AI. An AI Content Detector is a tool that employs complex algorithms and machine learning to evaluate whether the content was made by a real person or made artificially. It is capable of recognizing frequent patterns and irregularities in AI-generated writing (Needle, 2023). AI detection tools' method of detection does not guarantee that correct identification will happen every time since humans have diverse patterns of writing, and it is possible that their writing can match with the patterns that AI detection tools are looking for detection. Text generation is how an artificial intelligence system creates written text that mimics human language patterns and styles. Text generation has become more critical in various industries, including natural language processing, content production, customer support, and code help (Awan, 2023). Since it mimics human language patterns and styles, text generators need help understanding the meanings behind the words they generate. Human-made literature is valuable for any field of study or industry. If a machine generates literature, there is no guarantee that the words they generate tell facts and the truth.

AI detection tools are powered by cutting-edge technology but are only sometimes accurate. They might make mistakes, so trusting the tool is questionable (York, 2024). Just like humans, even AI detectors can make errors in their detection since they are also developed by humans, who can also make mistakes.

Artificial Intelligence has garnered well-deserved attention due to its wide range of uses in different fields, such as entertainment, medicine, and even education. ChatGPT will eventually be mentioned whenever AI is discussed due to its extreme popularity. ChatGPT is a generative AI tool that provides quality responses about a vast range of topics, which has a risk of being misused in different fields, including education (Akram, 2023). Research and scientific writing ethics strictly prohibit the use of AI (Guleria *et al.*, 2023). Recognizing this, educational institutions have started implementing AI-generated content detection tools to prevent students and the like from claiming the work of AI as their own. Unsurprisingly, students who plagiarize or cheat using AI will not admit their wrongdoing easily. It was found that AI users do not recognize themselves as the rightful owner of the AI-generated text and refuse to declare publicly that the content was authored by an AI, even going as far as claiming that it was written by a human ghostwriter (Draxler *et al.*, 2024). As one would realize, these AI-generated content detectors must not be blindly trusted as they are not foolproof like everything else. The difference in expertise of a language heavily influences the tools' detection rate, as the text produced by non-native English speakers was more prone to false positives

[1] Department of Information Technology, Cavite State Universtiy, Silang Campus, Biga 1 Silang, Cavite, 4112, Philippines
[*] Corresponding author's e-mail: cssantiago@cvsu.edu.ph

(Otterbacher, 2023). These facts have prompted the authors to conduct this literature review.

The creation and deployment of AI content detectors and classifier tools emphasize the rising relevance and necessity to distinguish between human-written and AI-generated material in various domains, including education and content marketing. No research has yet to thoroughly investigate the ability of these AI content detectors and classifiers to discriminate between human and AI-generated material (Elkhatat *et al.*, 2023). Hence, this study explores the capabilities of several newly released AI content detectors and classifier tools in distinguishing between human-written and AI-generated material. It signifies the concept of an AI-generated tool that can detect a word between an AI and human material.

MATERIALS AND METHODS

The Design

A systematic literature review on the topic of AIGT detectors used in academics was conducted utilizing the six-step principles for performing a systematic literature review (Durach *et al.*, 2017). Defining the research question was prioritized and accomplished first. Then, the attributes needed for the study were identified. The process is then continued by retrieving possibly relevant material and selecting appropriate literature. The next step is to synthesize the pertinent material from the literature, and the final stage is to summarize the findings of the review. Three research objectives determined to be useful were then identified for the study.

Research Question and Research Objectives

This literature review aims to answer the research question: To what extent can AI-produced text detection techniques distinguish between text generated by AI and text written by humans? Moreover, it seeks to accomplish the following objectives:

(1) To identify the different AI-Generated Text Detection Tools used;

(2) To investigate the accuracy of AI-Generated Text Detection Tools; and

(3) To investigate the reliability of AI-generated text Detection Tools.

Retrieving and Selecting Pertinent Literature

The systematic literature review made use of three credible and large-scale, comprehensive online databases to search for relevant literature that can be of use for the systematic literature review. To ensure quality and relevance, the selected articles are retrieved from reputable online databases for research, namely IEEE, Science Direct, and Google Scholar, containing relevant information and published studies covering topics in the field of computer science. The search strings or keywords employed in the search for pertinent literature, as well as the online database where those keywords were used, how many articles appeared on the search results, and how many of those articles were selected to use for the study (in parentheses), are summarized in the table below.

Table 1: Number of articles that appeared on the online databases using the different search keywords as of April 4, 2024

Keywords	IEEE	Science Direct	Google Scholar
Different AI-generated text detection tools	2 (0)	15,204 (6)	15,100 (32)
Different AI-generated text detectors	0	9,420 (7)	11,400 (35)
Testing AI-generated text detectors	1 (1)	7,760 (7)	10,300 (34)
Testing AI-generated text detection tools	3 (1)	13,798 (8)	13,300 (31)
Accuracy of AI-generated text detection tools	2 (1)	10,287 (10)	13,300 (35)
Reliability of AI-generated text detection tools	2 (1)	5,795 (1)	12,600 (33)

The total number of articles from the search results using the search strings and keywords across all three online databases is 138,274. Ten articles were retrieved from IEEE, 62,264 from ScienceDirect, and 76,000 from Google Scholar. These numbers do not reflect the number of unique articles retrieved from the search results, as the searches can return the same articles across different search strings.

The retrieved articles were assessed based on the criteria including their relevance and the recency of their publishment, resulting in 80 viable articles. The contents of the articles must be related to the objectives of the literature review. To finalize the selection of pertinent literature, the screening process involved reading the abstract, conclusions or results, methodology, and assessing each article, finalizing 34 articles.

Retrieving and Selecting Pertinent Literature

The following selected studies/literature were published from 2020 to 2024. A total of 34 chosen articles were found useful for this systematic literature review. Information from the following studies was extracted, and the findings and procedures of the studies were analyzed and summarized in Tables 2 - 4 below.

Table 2: Table synthesis about the different AI-generated text detection tools

Keywords	Google Scholar
Author	AI-Generated Text Detection Tools

Elkhatat *et al.* (2023)	OpenAI classifier, Writer, Copyleaks, GPTZero, and CrossPlag.
Weber-Wulff *et al.* (2023)	Check For AI, Compilation, Content at Scale, Crossplag, DetectGPT, Go, Winston, GPTZero, GPT-2 Output Detector Demo, OpenAI Text Classifier, PlagiarismCheck, Turnitin, Writeful GPT Detector, Writer, Zero GPT.
Chaka, C. (2023)	Copyleaks AI Content Detector, Giant Language Model Test Room (GLTR), GPTZero, OpenAI Text Classifier, and Writer.com's AI Content Detector.
Odri & Yoon (2023)	Content at scale, Copyleaks, Corrector, Crossplag, GPTZero, Originality, Sapling, Writefull et Quill, Writer, ZeroGPT.
Krishna *et al.* (2023)	GPTZero, DetectGPT, and OpenAI's text classifier.
Ippolito *et al.* (2020)	GPT-2.
Pu *et al.* (2023) [15]	Four online services powered by Transformer-based tools.
Ladha *et al.* (2023)	AI detection tools made by Content, CopyLeaks, and Writer.com.
Walters (2023)	Content at scale, ContentDetector.ai, Copyleaks, Crossplag, Grammica, GPT Radar, GPTZero, Ivy Panda, OpenAI, Originality.ai, Sapling, Scribbr, SEO.ai, TurnItIn, Writer, ZeroGPT.
Akram (2023)	GPTkit, GPTZero, Originality, Sapling, Writer, Zylalab.
Sadasivan *et al.* (2024)	OpenAI, GPT-2 detector models, DetectGPT.
Jawahar *et al.* (2020)	GPT-2, GROVER detector, RoBERTa detector.
Kumarage *et al.* (2023)	OpenAI-FT, DetectGPT.
Chakraborty *et al.* (2023)	GPTZero, DetectGPT, OpenAI.
Ghosal *et al.* (2023)	DetectGPT, GPTWatermark.
Gao *et al.* (2023)	GPT-2 Output Detector, Plagiarism Detector, iThenticate.
Bellini *et al.* (2024)	GPTZero, Originality, Writer ACD, and ZeroGPT.
Rashidi *et al.* (2023)	Non-API GPT-4-based AI-detector, RoBERTa-based detector model.
Chaka (2024)	Turnitin, OpenAI Classifier, DupliChecker, GLTR, Writer, iThenticate, Crossplag, Content At Scale, GPTZero, Quill, Ouriginal, Sapling, Content Detector, GPT-2 Output Detector, Copyleaks.
Orenstrakh *et al.* (2023)	AI Text Classifier, CheckForAI, CopyLeaks, GLTR, GPT-2 Detector, GPTKit, GPTZero, OriginalityAI.
Šigut (2023)	Compilatio, Turnitin, and GPT-2 Output Detector,
Wu & Flanagan (2023)	GPT-2 Output Detector Demo.
Perkins *et al.* (2023)	Turnitin.
Pan *et al.* (2024)	DetectGPT, Giant Language Model Test Room (GLTR), GPT-2 Detector, GPTZero, Sapling.
Ibrahim *et al.* (2023)	GPTZero, and OpenAI's Text Classifier.
Singh (2023)	GLTR, GPTZero, CopyLeaks, CrossPlag
Bhattacharjee & Liu (2023)	ChatGPT.
Singh *et al.* (2024)	ZeroGPT, Perplexity, Hugging Face, Writefull GPT Detector, Copyleaks, Writer AI Content Detector, Draft and Goal, Originality, ai, Content at Scale, OCR.
Perkins *et al.* (2024)	Turnitin, GPTZero, ZeroGPT, Copyleaks, Crossplag, GPT-2 Output Detector, GPTKit.
Desaire *et al.* (2023)	The researchers made a GPT -2 Output Detector and an AI detector model.
Uzun (2023)	GPT-2 Detector, AI Dungeon Detector, TELLER, FakeSpot, Sensity AI, Truepic, DALL-E Detector.

Table 3: Table synthesis about the accuracy of the AI-generated text detection tools

Author	**Accuracy**
Elkhatat *et al.* (2023)	AI identification accuracy is higher in detecting material created by GPT 3.5 compared to GPT 4 but shows inconsistency when applied to human-made control replies.
Weber-Wulff *et al.* (2023)	The present detection technologies have a strong bias toward identifying the output as human-written rather than detecting AI-generated material.
Chaka (2023)	All five AI content detectors must be fully capable of distinguishing AI-generated material from machine-generated text properly and convincingly.

Odri & Yoon (2023)	Paraphrasing or omitting commas reduces the AI detection's effectiveness while making the text seem human-made to all detection tools.
Krishna *et al.* (2023)	Paraphrasing is found to drastically reduce detection accuracy while keeping input semantics.
Ippolito *et al.* (2020)	The precision of human raters and classifiers varies significantly based on the decoding approach and length of produced sequences.
Pu *et al.* (2023)	The paired accuracy is calculated by combining real and synthetic articles so that the two pieces share the same information.
Ladha *et al.* (2023)	Since research papers follow the guidelines of scientific writing, an AI detector may find them to be robotic.
Walters (2023)	The precision of a detector is only little related to its paid or free status.
Akram (2023)	GPTkit has the lowest accuracy percentage of 55.29%, while Originality has the highest accuracy percentage of 97.09%.
Sadasivan *et al.* (2024)	The accuracy of AI-text detection tools drops significantly after using recursive paraphrasing and spoofing.
Jawahar *et al.* (2020)	Existing detectors exhibit poor cross-domain accuracy when dealing with articles from different publication formats.
Kumarage *et al.* (2023)	AIGT detectors' performance decreases significantly after implementing the evasive soft prompt framework.
Ghosal *et al.* (2023)	Paraphrasing can significantly reduce the accuracy of different types of AIGT detectors.
Gao *et al.* (2023)	The employed tool, GPT-2 Output Detector, has a higher chance of identifying text as AI-generated.
Bellini *et al.* (2024)	There was a significant disparity in the detecting capability of the used detectors, which made contradictory assessments.
Rashidi *et al.* (2023)	Most "real" abstracts were correctly predicted as non-AI-generated, a subset of abstracts (approximately 5%-10%) were erroneously assigned as AIGT with a high degree of confidence (ex. >80%).
Chaka (2024)	Traditional anti-plagiarism tools lack the ability to detect AIGT due to the differences in syntax and structure.
Orenstrakh *et al.* (2023)	Detectors are reasonably accurate but prone to errors when the generated data are paraphrased and need help with non-English languages, which lowers their accuracy.
Šigut (2023)	All detectors are more accurate with longer papers since they include more information.
Wu & Flanagan (2023)	Inserting spelling problems and processing them using QuillBot reduces the accuracy of detectors.
Perkins *et al.* (2023)	The AI tool has difficulty recognizing AI content when sophisticated prompting techniques are employed.
Pan *et al.* (2024)	The detectors experienced difficulties in discriminating between AI-generated code and human-made code.
Ibrahim *et al.* (2023)	OpenAI's Text Classifier incorrectly classified five percent of student contributions. GPTZero produces more false positives but fewer false negatives.
Perkins *et al.* (2023)	The employed detector recognized more than 90% of the submitted documents, including some AI-created data, but the overall detected material was just above 50%.
Singh (2023)	The employment of computational intelligence methodologies can contribute to the accuracy and competency of AI detectors.
Ma´jovsky *et al.* (2024)	There is insufficient scientific data to support claims of high detection tool accuracy.
Ormond & Eisgrau (2023)	It is noted that demand for such systems, however, is not measured by their fairness or accuracy.
Bhattacharjee & Liu (2023)	ChatGPT can detect AIGT in fewer than 50% of the samples and has a relatively high proportion of false negatives.
Singh *et al.* (2024)	After evaluating the numerous tools described, the study discovered that they need help to detect AI-generated text accurately.
Desaire *et al.* (2023)	Output Detector's paragraph-level performance of 85%-88% is not significantly distinctive from the 82% accuracy formerly reported on 100 scientific abstracts.
Uzun, L. (2023)	Online tools are only sometimes accurate and may need help recognizing information created by new or lesser-known AI models.

Table 4: Table synthesis about the reliability of AI-generated text detection tools

Author	Reliability
Elkhatat *et al.* (2023)	Inconsistencies when applied to human-written text.
Weber-Wulff *et al.* (2023)	The present detection technologies are neither accurate nor dependable.
Chaka (2023)	Text alteration lowers the effectiveness of existing detection tools.
Odri & Yoon (2023)	Most detectors are ineffective in detecting text created by generative AI, and even human writing can be detected as AI-generated.
Krishna *et al.* (2023)	Paraphrasing reduces detection accuracy significantly across all language models, detectors, and tasks, regardless of diversity control codes.
Ippolito *et al.* (2020)	The length of the excerpt improves the performance of the detectors.
Pu *et al.* (2023)	The model can generate genuine news stories that people only recognize as synthetic 52% of the time.
Ladha *et al.* (2023)	Human reviewers should not be excluded when reviewing materials, even if AI detection techniques are present.
Walters (2023)	AI text detectors provide both qualitative and quantitative estimates if the document was AI-created.
Akram (2023)	Among the text detection tools, Originality is the most reliable.
Sadasivan *et al.* (2024)	The application of AI-text detection tools is unreliable.
Jawahar *et al.* (2020)	The detectors' results will vary even with minuscule changes in the text input.
Kumarage *et al.* (2023)	Experiments proved that AIGT detectors are unreliable when evasive soft prompts are used to generate the AI text.
Chakraborty *et al.* (2023)	AIGT detectors cannot distinguish human-written text from text generated by new Large Language Models.
Ghosal *et al.* (2023)	AI text detectors are unreliable due to their negative bias towards English written by those of foreign origin.
Gao *et al.* (2023)	AI detectors and humans could recognize a fraction of the abstracts created by ChatGPT, although neither were perfect discriminators.
Bellini *et al.* (2024)	Writer ACD classified text as primarily human-written, while Originality usually detected ChatGPT-4 samples as AI-generated.
Rashidi *et al.* (2023)	The AI text detector misidentified up to 8% of known actual abstracts as AI-generated material. This shows the existing limits of these detection technologies.
Chaka (2024)	Inconsistency in the detection efficacy of the tested AI detectors and anti-plagiarism detection techniques was found to have low detection reliability.
Orenstrakh *et al.* (2023)	LLM-generated text detectors produce more accurate findings for human-made input than ChatGPT-created input, yet they are still unreliable for use in the academic setting.
Šigut (2023)	The tools were unreliable in providing proper responses. As a result, their decisions should be approached with caution.
Wu & Flanagan (2023)	GPT-2 Output Detector Demo showed substantial performance fluctuation, notably in the context of edited writings.
Pan *et al.* (2024)	Existing detectors do a poor job of discriminating AI-generated code against human-made code.
Ibrahim *et al.* (2023)	AI-text classifiers are unreliable in detecting ChatGPT's usage in schoolwork because of their proclivity in labeling human-made replies as AI-generated.
Perkins *et al.* (2023)	Employing strategies for designing AI prompts successfully avoids AI detection technologies, indicating that AI detection software has to be improved.
Singh, A. (2023)	The tools are limited in their ability to identify every form of machine-generated text, and some carry the risk of false identification.
Ma´jovsky *et al.* (2024)	While current detection tools promise excellent accuracy rates, several research suggests they are unreliable.
Ormond & Eisgrau (2023)	No presently available detection technology is sufficiently reliable to exclusively base potentially life- and career-altering decisions.

Bhattacharjee & Liu (2023)	GPT-4 labels everything as 'AI-generated' even when its predictions are incorrect, so their projections are extremely unreliable.
Singh *et al.* (2024)	For certain categories of prompts and subjects, the most efficient web technologies for recognizing authored material can only achieve a 90% success rate.
Perkins *et al.* (2024)	Detectors have a possibility of wrong allegations as well as unreported instances, thus making it unreliable.
Desaire *et al.* (2023)	At the paragraph level, 94% of approximately 1200 cases were properly categorized, while 99.5% were correctly allocated at the document level.
Uzun (2023)	AI-generated material can be purposefully created to resemble human writing styles, rendering detection tools unreliable.

RESULTS AND DISCUSSION

AI-Generated Text Detection Tools

Across multiple different articles, previous studies have studied and tested the same AIGT detection tools, such as Turnitin, CopyLeaks, Crossplag, GPTZero, and GPT-2 Output Detector, along with other less-known AIGT detection tools. Most of the AIGT detection tools evaluated in the studies are publicly available online for free, although some require payment or a subscription to be usable. The names of the identified AI-generated text detection tools listed in the table synthesis per article are based on the name or alias written in that particular article. This does not discount the possibility that multiple different articles pertain to the same AIGT detection tool, only under various names and aliases.

Three of the studies have not specified or mentioned which AIGT detection tools were used in their study, although they still achieved results for the objectives of investigating the accuracy and reliability of the AIGT detection tools (Pu *et al.*, 2023; Májovský *et al.*, 2023; Ormond Eisgrau, 2023). One group of researchers has also made their own AI detector model to test for its accuracy, although the name for this novel detector was yet to be determined as of April 04, 2024 (Desaire *et al.*, 2023) Another study also used AI detectors based on language models, which were unnamed (Rashidi *et al.*, 2023). A total of 13 articles included the GPT-2 detectors among the AIGT detectors they studied, making it the most evaluated AIGT detector among the selected pertinent literature.

One study used ChatGPT, which is more commonly used as a generative AI tool, as a tool used to detect AI-generated content, although the results indicate that this approach is not accurate all the time (Bhattacharjee & Liu , 2023). This shows that generative AI tools also have the potential to be used against themselves as AIGT tools. Among all the identified AIGT detection tools listed in the table synthesis, the highest recorded accuracy reached 99% for the yet unnamed AI detector model created by a group of researchers (Desaire, 2023). Although these tools have made significant technological advancements, their performance is not yet flawless, and detection may not always be accurate. It is important for authorities to carefully interpret the results to prevent the possibility of misidentifying the work of individuals who have completed their tasks diligently and honestly.

Accuracy of the AI-Generated Text Detection Tools

Each study has used its own preferred criteria for A.I. detection accuracy. The researchers used the criteria – the type of error and the accuracy – to gauge the classification of these tools. More often than not, the methods for categorizing are assessed based on these characteristics: recall, precision, and accuracy. The study's authors also investigated to check for potential errors, as different types of errors have different meanings depending on educational background (Weber-Wulff *et al.*, 2023). According to what is seen in the table of synthesis previously mentioned, the results show that the AI identification methods were more accurate in detecting material created by GPT 3.5 than GPT 4 (Elkhatat *et al.*, 2023). The researchers also did an error analysis since various errors have distinct implications in the educational environment.

TurnItIn was performed with exceptional results in the analysis, using papers written by humans. Generally, the detector's accuracy is moderately related to its cost, such as free or paid. The most accurate detectors require the users to register and pay to utilize their complete features, while the rest, Sapling, GPT Radar, and GPTZero, are middling in terms of accuracy (Walters, 2023). AIGT detectors' performance declines dramatically after using the evasive soft prompt framework. It is determined that the accuracy of the paid AIGT detection tools averages 87%, while those that do not need a subscription are at 77%.

A modest link exists between detector accuracy and paid or free status (Walters, 2023). This means that there is an unfairness among the different AIGT detection tools since their accuracies differ, further worsened by the fact that their accuracy depends on whether the tool is paid or free. So, it would be unfair to people who lack the resources to use the more accurate AIGT detection tools because they usually require payment.

In one study, a confusion matrix was used to ascertain the accuracy regarding the origins of submitted written works, estimating the rate of correct and incorrect detections created by the AIGT detection tool (Ibrahim, 2023). The Output Detector's paragraph-level performance (85%-88%) is not significantly different from the 82% accuracy previously reported on 100 scientific abstracts, and this may be representative of the device's ability to accurately detect a single paragraph of scientific academic text (Desaire, 2023).

Reliability of AI-Generated Text Detectors

One of the most important aspects to be considered when using AIGT detectors is their reliability. Based on the studies contained in the table 4, it is generally proven that AIGT detectors are unreliable. This is not to say that the detectors fail at their purpose almost every single time, but it has not reached a level in which faith can be put in the results provided by these detectors. In one of the studies, the detectors performed well with content generated by GPT 3.5 over those generated by GPT 4 (Elkhatat *et al.*, 2023). Through this, the authors have inferred that the growth rate of AI content generators exceeds that of AI content detectors. This disparity between the performance of the opposite sides will only continue to increase over time.

As AIGT detectors become more prevalent, the academic authorities and perpetrators become more aware of what the opposite side is doing to win against the other. To avoid being caught by these detectors, a method that plagiarists and cheaters use when copying or stealing someone else's work, which was already being used on human-made content way before the rise of AI usage, is paraphrasing. This method has proven extremely effective to the point that, in some instances, the detection accuracy drops by more than 60%, demonstrating that the detectors are not advanced enough to see through these 'tricks' and be deemed reliable (Krishna *et al.*, 2023).

A more concerning problem found in several of these studies and literature is the prevalence of bias among the content detectors against those lacking English language skills. The fluency in the English language influences whether a human-written text will be wrongfully classified as artificially generated due to the limited range the detectors have over human language distribution (Ghosal *et al.*, 2023). This issue raises concerns about fairness and equal assessment for non-native English speakers. Upon reaching these observations, perseverance must be put into the effort to classify better whether the content is human-written or AI-generated. Until then, caution must be observed with the usage of these detectors, and ultimately, trust the judgment before risking the possibility of handing a wrongful verdict upon the credibility of someone's work.

CONCLUSION

This literature review has compiled the several issues and shortcomings that AI-generated content detectors have in their current state. Based on the results, it can be inferred that the AIGT detection tools are not accurate and reliable to use. The extent of AI detectors' accuracy is more effective on older generative AI models, wherein they start to lose their accuracy and reliability in their detection once newer generative AI models are used. As a result, individuals should not solely rely on the capabilities of AIGT detectors. These tools are designed to support individuals who use them, but they should not bear the burden of discovering who created a piece of text, it is still the users' responsibility to determine the truth.

The need for a better method of discerning AI-generated from human-made content continues to increase, and so does the improvement of AI content generation in presenting itself as genuine and human-like. Even so, the technology and academic sector must not cease its efforts to tackle this problem as it is essential to determine whether something is rightfully attributed to someone's work or made with the assistance of AI.

Moving forward, it is critical to prioritize continuous improvement and validation of these systems. Transparency, multidisciplinary collaboration, and regulatory frameworks are crucial. By assuring accuracy, reducing biases, and adhering to ethical norms, the potential advantages of AI-driven text detection can be maximized while limiting hazards. This collaborative endeavor will create trust, increase inclusion, and drive innovation in the field of AI technology. It should be taken into consideration that AI text detectors can be incorrect, so it must be used with caution and preferably alongside human reviewers. The repercussions of the results provided by the detectors must be kept in mind as they might cause irreversible damage to the authors' reputation as well as the institutions.

REFERENCES

Akram, A. (2023). *An empirical study of AI-generated text detection tools*. https://doi.org/10.48550/arXiv.2310.01423

Awan, A. A. (2023, May 24). What is text generation? *DataCamp*. https://www.datacamp.com/blog/what-is-text-generation

Bellini, V., Semeraro, F., Montomoli, J., Cascella, M., & Bignami, E. (2024). Between human and AI: Assessing the reliability of AI text detection tools. *Current Medical Research and Opinion, 40*(3), 353–358. https://doi.org/10.1080/03007995.2024.2310086

Bhattacharjee, A., & Liu, H. (2023). Fighting fire with fire: Can ChatGPT detect AI-generated text? *ACM SIGKDD Explorations Newsletter, 25*(2), 14-21. https://doi.org/10.1145/3655103.3655106

Chakraborty, M., Islam Tonmoy, S. M. T., Zaman, S. M. M., Sharma, K., Barman, N. R., Gupta, C., Gautam, S., Kumar, T., Jain, V., Chadha, A., Sheth, A. P., & Das, A. (2023). Counter Turing test CT²: AI-generated text detection is not as easy as you may think—Introducing AI detectability index. In *Proceedings of EMNLP 2023 Main*. https://doi.org/10.48550/arXiv.2310.05030

Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching, 7*(1). https://doi.org/10.37074/jalt.2024.7.1.14

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

Desaire, H., Chua, A. E., Isom, M., Jarošová, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science, 4*(6), 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction, 31*(2), Article 25, 1-40. https://doi.org/10.1145/3637875

Durach, C. F., Kembro, J., & Wieland, A. (2017). A new paradigm for systematic literature reviews in supply chain management. *Journal of Supply Chain Management, 53*(4), 67–85. https://doi.org/10.1111/jscm.12145

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

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Nature Digital Medicine, 6,* Article 19. https://doi.org/10.1038/s41746-023-00819-6

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities & impossibilities of AI-generated text detection: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2310.15264

Guleria, A., Krishan, K., Sharma, V., & Kanchan, T. (2023). ChatGPT: Ethical concerns and challenges in academics and research. *The Journal of Infection in Developing Countries, 17*(9), 1292-1299 https://doi.org/10.3855/jidc.18738

Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: 'The Terminator versus the machines'. *Language Testing in Asia, 13,* 46. https://doi.org/10.1186/s40468-023-00260-2

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808–1822). https://doi.org/10.18653/v1/2020.acl-main.163

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. S. (2020). Automatic detection of machine-generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics* (COLING). https://doi.org/10.48550/arXiv.2011.01314

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (NIPS '23) (pp. 27469-27500).

Kumarage, T., Sheth, P., Moraffah, R., Garland, J., & Liu, H. (2023). How reliable are AI-generated text detectors? An assessment framework using evasive soft prompts. In *Proceedings of the EMNLP 2023*. https://doi.org/10.48550/arXiv.2310.05095

Ladha, N., Yadav, K., & Rathore, P. (2023). AI-generated content detectors: Boon or bane for scientific writing? *Indian Journal of Science and Technology, 16*(39), 3435–3439. https://doi.org/10.17485/ijst/v16i39.1632

Majovský, M., Černý, M., Netuka, D., & Mikolov, T. (2024). Perfect detection of computer-generated text faces fundamental challenges. *Cell Reports Physical Science, 5*(1), 101769. https://doi.org/10.1016/j.xcrp.2023.101769

Needle, F. (2023, November 9). AI detection: How to pinpoint AI-generated text and imagery [+ detection tools]. *HubSpot*. https://blog.hubspot.com/marketing/ai-detection

Odri, G.-A., & Yoon, D. J. (2023). Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research, 109*(8), Article 103706. https://doi.org/10.1016/j.otsr.2023.103706

Orenstrakh, M. S., Karnalim, O., Suarez, C. A., & Liut, M. (2023). Detecting LLM-generated text in computing education: A comparative study for ChatGPT cases. *arXiv*. https://doi.org/10.48550/arXiv.2307.07411

Ormond, J., & Eisgrau, A. (2024). Can we ensure that systems for detecting generative AI are accurate and fair? *ACM.org*. https://www.acm.org/media-center/2023/october/systems-detecting-generative-ai

Otterbacher, J. (2023). Why technical solutions for detecting AI-generated content in research and education are insufficient. *Patterns, 4*(7), 100796. https://doi.org/10.1016/j.patter.2023.100796

Pan, W. H., Chok, M. J., Wong, J. L. S., Shin, Y. X., Poon, Y. S., Yang, Z., Chong, C. Y., Lo, D., & Lim, M. K. (2024). Assessing AI detectors in identifying AI-generated code: Implications for education. In *ICSE-SEET '24: Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training* (pp. 1–11). https://doi.org/10.1145/3639474.3640068

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgment and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(89), 89-113. https://doi.org/10.1007/s10805-023-09492-6

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). GenAI detection tools, adversarial techniques and implications for inclusivity in higher education. *arXiv*.

https://doi.org/10.48550/arXiv.2403.19148

Pu, J., Sarwar, Z., Abdullah, S. M., Rehman, A., Kim, Y., Bhattacharya, P., Javed, M., & Viswanath, B. (2023). Deepfake text detection: Limitations and opportunities. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy* (SP) (pp. 1613-1630). https://doi.org/10.1109/SP46215.2023.10179387

Rashidi, H. H., Fennell, B. D., Albahra, S., Hu, B., & Gorbett, T. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts are misidentified as AI creations by an AI text detection tool. *Journal of Pathology Informatics, 14,* 100342. https://doi.org/10.1016/j.jpi.2023.100342

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. https://doi.org/10.48550/arXiv.2303.11156

Šigut, P. (2023). *Evaluation of machine-generated text detectors* (Undergraduate thesis, Masaryk University). https://is.muni.cz/th/f5y2v/Bachelors_thesis.pdf

Singh, A. (2023). A comparison study on AI language detector. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference* (CCWC) (pp. 0489-0493). https://doi.org/10.1109/CCWC57344.2023.10099219

Singh, P., Singh, A. P., Rathi, S., & Vasesi, S. (2023). Unmasking the source: Identifying human vs ChatGPT-generated text through machine learning. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 622-625). https://doi.org/10.1109/ICACCTech61146.2023.00106

Uzun, L. (2023). ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. *Language and Education Technology, 3*(1), 45-54.

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

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., *et al.* (2023). Testing of detection tools for AI-generated text. *International Journal of Educational Integrity, 19,* 26. https://doi.org/10.1007/s40979-023-00146-z

Wu, H., & Flanagan, T. (2023). The limits of AI content detectors. *Journal of Student Research, 12*(3). https://doi.org/10.47611/jsrhs.v12i3.5064

York, A. (2024, March 20). 10 best AI detection tools & checkers in 2024. *ClickUp*. https://clickup.com/blog/ai-detection-tools/