Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools

Article *in* Journal of Applied Learning & Teaching · July 2023

DOI: 10.37074/jalt.2023.6.2.12

1 author:

Chaka Chaka
University of South Africa
85 PUBLICATIONS **929** CITATIONS

SEE PROFILE

Vol.6 No.2 (2023)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools

Chaka Chaka[A] [A] *Professor, University of South Africa, Pretoria, South Africa*

Keywords

ChatGPT;
Copyleaks AI Content Detector;
generative AI;
Giant Language model Test Room;
GPTZero,
higher education;
OpenAI Text Classifier;
Writer.com's AI Content Detector.

Correspondence

chakachaka8@gmail.com [A]

Article Info

Received 24 June 2023
Received in revised form 21 July 2023
Accepted 22 July 2023
Available online 24 July 2023

DOI: https://doi.org/10.37074/jalt.2023.6.2.12

Abstract

This paper set out to test the accuracy of five AI content tools, GPTZero, OpenAI Text Classifier, Writer.com's AI Content Detector, Copyleaks AI Content Detector, and Giant Language model Test Room, to detect AI-generated content in the responses generated by ChatGPT, YouChat, and Chatsonic. The responses were generated from these three AI chatbots using English prompts related to applied English language studies. Then, the ChatGPT-generated responses were Google-translated into German, French, Spanish, Southern Sotho, and isiZulu, and inputted into GPTZero for it to detect the AI-generated content in them. Additionally, the ChatGPT-generated responses Google-translated into German, French and Spanish were inputted into Copyleaks AI Content Detector for it to detect the AI-generated content in them. For the ChatGPT-, YouChat-, and Chatsonic-generated responses, Copyleaks AI Content Detector emerged as the top-most performing AI content detector among the five AI content detectors. It was followed by OpenAI's AI Text Classifier. Concerning the ChatGPT-generated responses that were Google-translated into five languages, GPTZero misidentified all of them as human-produced. For the ChatGPT-generated responses that were Google-translated into German, French and Spanish, Copyleaks AI Content Detector correctly identified three of the German-translated texts, five of the French-translated texts, and all the Spanish-translated texts as AI-generated. Thus, it is evident from this paper that all five AI content detectors seem not yet fully ready to accurately and convincingly detect AI-generated content from machine-generated texts in different contexts. This has dire consequences for AI-generated plagiarism in academic essay writing.

Introduction

The launch of ChatGPT, a generative artificial intelligence (AI) chatbot owned by OpenAI (OpenAI, 2022), on 30 November 2022, had a domino effect in cyberspace and in the real-life world. It not only rattled the AI world in which generative AI chatbots, which before ChatGPT were relatively unknown, suddenly emerged or announced their presence (Chaka, 2023a; Eliaçik, 2023a; Hetler, 2023; Kanran, 2023), but it also led to the emergence of AI content detection tools intended to detect and to differentiate between AI-generated and human-written texts. One such AI content detection tool, which was launched late in 2022 as a direct consequence of ChatGPT, is GPTZero. The first part of its name is directly linked to the last part of ChatGPT's name. Much more will be said about GPTZero below. In a manner almost resembling what happened after ChatGPT was released, similar AI detection tools emerged or announced their presence in the aftermath of GPTZero's launch. Examples of such tools are AI Text Classifier, Giant Language Model Test Room (GLTR), Writer.com's AI Content Detector, and Copyleaks AI Content Detector (Lim, 2023; Outlook Spotlight, 2023; Chrome, 2023; Copyleaks, 2023). Again, much more will be said about these tools below.

All these AI-powered content detection tools emerged when there were assertions that no current AI content tool could detect AI plagiarism in ChatGPT-generated responses (Chaka, 2023b; Cutcliffe, 2022; Heilweil, 2022). While these tools can be used to detect what Lim (2023) calls AI-assisted content in different text types in general, it is AI-assisted academic content that is the focus of this paper. This is more so as immediately after the release of ChatGPT, some schools and universities reacted by saying that they would ban it because of the temptation it had for students to use it in school- or university-level essays (Anders, 2023; Barnett, 2023; Caren, 2022; Ceres, 2023; Harris, 2022; Hern, 2022; Somoye, 2023; Stokel-Walker, 2022; Wingard, 2023). Of course, there were some academic and science journals that were said to have taken a stance to ban ChatGPT as well (Sample, 2023). So, at issue here is AI-assisted academic content that tends to characterise responses produced by generative AI chatbots such as ChatGPT and others similar to it. This type of plagiarism is a grave concern for schools and universities.

Literature review

With the advent of generative AI-powered large language model (LLM) chatbots, which was heralded by ChatGPT's release in November 2022, there has been a growing number of scholarly papers that focus on and explore these chatbots. Examples of such scholarly papers include, but are not limited to: Alser & Waisberg (2023), Chaka (2023a), Cotton et al. (2023), Ifelebuegu (2023), Popenici (2023), Rasul et al. (2023), Rudolph et al. (2023a, b), Sullivan et al. (2023), and Yeadon et al. (2023). Some of these papers are published papers, while others are preprints, a publication pattern that almost resembles that of papers published during the COVID-19 pandemic (Chaka, 2020). Among these two streams of scholarly papers, there are those that explore the risks posed by ChatGPT for academic integrity concerning student assessment (see Ifelebuegu, 2023; Khalil & Er, 2023; Perkins, 2023; Rudolph et al., 2023a, b; Sullivan et al., 2023; Ventayen, 2023; Yeadon et al., 2023). But the critical issue with regard to academic integrity for most educational institutions is detecting plagiarism and distinguishing AI-generated content from human-written content. This is more so for both student essay writing and scholarly writing. In addition to the AI content detection tools specified in the preceding section, other tools include OriginalityAI, Content At Scale, Kazan SEO, GPT-2 Output Detector (Outlook Spotlight, 2023), Crossplag AI Content Detector (Lim, 2023), Claude AI, AI Writing Check, GPT Radar, and CatchGPT (Wiggers, 2023). Additional tools are Corrector App AI Content Detector, Plagibot, CopyScape, Winston AI, Writefull GPT Detector, Turnitin (Uzun, 2023), SciSpace, Hive Moderation, Hello Simple AI (Awan, 2023), PlagiarismCheck, Check For AI, DetectGPT, Compilation, and Go Winston (Weber-Wulff et al., 2023).

Since most of these AI content detectors are new, not much research has been conducted to evaluate their efficacy, accuracy, and reliability in terms of distinguishing between the content generated by current AI-powered LLM chatbots and the content written by humans. So, this is a new and growing area that still needs a lot of research. Of the few scholarly papers focusing on this area, a lot of them are preprints. Five such preprints are Aremu (2023), Cai and Cui (2023), Guo et al. (2023), Ventayen (2023), and Weber-Wulff et al. (2023). Two of these papers, Aremu (2023 and Weber-Wulff et al. (2023), have some relevance to the current paper. These two papers are briefly reviewed by discussing only aspects of them that have some bearing on this paper.

Aremu's (2023) paper investigated the capability of six AI text detectors, Sapling AI, Crossplag AI Content Detector, OpenAI Text Classifier, ZeroGPT, GPTZero, and Content At Scale, to accurately identify different essay types written by humans and those generated by AI (ChatGPT). The essay types in question were argumentative, descriptive, expository, and narrative essays. Their sample numbers were as follows: argumentative = 13; descriptive = 17; expository = 11; and narrative = 11. These sample numbers were split almost equally between the two datasets: human-written and AI-generated essay types. The prompts for the four essay types were as follows, respectively: Gun control; A day at the beach; The benefits of regular exercise; and A journey towards self-discovery. The human-written essay samples were obtained from the Internet, and were pre-2022 (before the advent of ChatGPT), while the AI-generated essays were sourced from ChatGPT by using the same four prompts with their attendant enhancements. In the main, these AI detectors performed well in accurately recognising human-written essays. In contrast, they performed poorly in detecting ChatGPT-generated and enhanced essays. Crossplag and Content At Scale outperformed the other AI detectors by accurately identifying human-authored essays with consistency and reliability, while ZeroGPT and GPTZero outdid the other detectors in terms of identifying ChatGPT-generated essays. This indicates their being robust and resistant to content deception (Aremu, 2023).

Weber-Wulff et al.'s (2023) paper employed 14 AI detection tools to examine their accuracy and error types in distinguishing between human-written text and AI-generated (ChatGPT-generated) text. These tools consisted of 12 publicly available AI detection tools and two commercial plagiarism detection tools. They were: Check For AI; Compilatio; Content at Scale; Crossplag; DetectGPT; Go Winston; GPT Zero; GPT-2 Output Detector Demo; OpenAI Text Classifier; PlagiarismCheck; Turnitin; Writeful GPT Detector; Writer; and ZeroGPT. All of these tools were non-premium versions. The paper used 54 test cases that were divided into the following five categories of English-language files:

- human-written;
- human-written in a non-English language with a resultant AI/machine translation to English;
- AI-generated text;
- AI-generated text with resultant human manual edits; and
- AI-generated text with resultant AI/machine paraphrase.

The human-written test cases were produced by nine people (eight researchers and one collaborator), and represented diverse disciplines such as academic integrity, computer science, civil engineering, economics, history, linguistics, and literature. They were written in Bosnian, Czech, German, Latvian, Slovak, Spanish, and Swedish and were machine-translated into English using DeepL (3 cases) and Google Translate (6 cases). Different prompts were used to generate AI texts through ChatGPT. Two additional texts were generated from ChatGPT using fresh prompts. One set of them was manually edited by exchanging words with their synonyms or by re-ordering sentence parts. The other set was automatically rewritten by employing an AI-powered tool, Quillbot. In terms of detection accuracy across all text cases, Turnitin (ranked 1) and Compilatio (ranked 2) topped the other tools, while PlagiarismCheck (ranked 13) and Content at Scale (ranked 14) were the most poorly-performing tools. The paper concludes that its findings failed to confirm the accuracy claims made by the detection tools it used, as these tools are unsuitable for providing evidence of academic misconduct. It also concludes that these tools are amenable to gaming, especially through paraphrasing and machine translation (Weber-Wulff et al., 2023).

Research problem

With the rising number of generative AI-powered LLM chatbots, there are growing concerns about the risks these chatbots pose to academic integrity by academics and educational institutions. To address these concerns, a number of online AI content detection tools have been released following the launch of ChatGPT. All these tools make bold claims (mostly undercut by concomitant disclaimers) about their accuracy rate and their reliability in detecting AI-generated content (see Chaka, 2023a, b; Chrome, 2023; Copyleaks, 2023; Kirchner et al., 2023; Outlook Spotlight, 2023; Tech Desk, 2023; Tyrrell, 2023; Weber-Wulff et al., 2023; Wiggers, 2023). Amid this burgeoning number of AI content detection tools, there is a need to evaluate the accuracy and reliability of these tools to differentiate between AI-generated content and human-produced content. This is critical as their efficacy in doing so will help academics and educational institutions know when student content is human-written and when it is AI-generated. The distinction between the content generated by an AI tool and the one produced by a human, or what Uzun (2023) calls the author factor, becomes trickier to determine as manipulating any form of content tends to elude most of the currently available AI detection tools (see Aremu, 2023; Cai & Cui, 2023; Guo et al., 2023; Uzun, 2023; Ventayen, 2023; Weber-Wulff et al., 2023). Related to the author factor is the content factor, the validity and reliability of the content produced, both AI-generated and human-written content (see Uzun, 2023).

Against this background, the purpose of this paper is three-fold: to test the accuracy of five AI content detection tools to detect the content generated by three AI chatbots, ChatGPT, YouChat, and Chatsonic, in its original English version; to evaluate the accuracy of one of these five AI content detection tools to detect the German, French, Spanish, Southern Sotho, and isiZulu versions of this content as machine-translated by Google Translate; and to test the accuracy of one of these five AI content detection tools to detect the German, French, and Spanish versions of this content as machine-translated by Google Translate. The five AI content detection tools are: GPTZero, OpenAI Text Classifier, Writer.com's AI Content Detector, Copyleaks AI Content Detector, and Giant Language Model Test Room. Relatedly, the paper's research questions are:

- What is the accuracy of the five AI content detection tools (GPTZero, OpenAI Text Classifier, Writer.com's AI Content Detector, Copyleaks AI Content Detector, and Giant Language Model Test Room) in detecting the content generated by ChatGPT, YouChat, and Chatsonic, in its original English version?
- What is the accuracy of GPTZero in detecting the German, French, Spanish, Southern Sotho, and isiZulu versions of this content as machine-translated by Google Translate?
- What is the accuracy of Copyleaks AI Content Detector in detecting the German, French, and Spanish versions of this content as machine-translated by Google Translate?

As pointed out above, there is currently a paucity of research that has been conducted in the area of study highlighted above. Thus, the current paper attempts to make a contribution to this area of study.

Reviewing of the five AI content detectors

GPTZero

GPTZero is an AI content detection tool built by a senior computer science student at Princeton University shortly after the release of ChatGPT. As its name indicates, it is intended to detect whether a text generated by ChatGPT is AI-generated or human-written (Chaka, 2023a; Ofgang, 2023; Tech Desk, 2023; Tyrrell, 2023). Of course, in this sense, it has a wider application beyond the ChatGPT-generated text to text generated by other generative AI tools, including ordinary human-written responses or essays that have nothing to do with AI generation. Therefore, it can also be referred to as an AI content detection app.

How, then, does it detect whether a text is AI-generated or human-produced? It does so by identifying two measures: perplexity and burstiness. Perplexity measures a text's randomness. The understanding here is that a human-written text displays randomness or chaoticness and, thus, is likely to perplex or be unfamiliar to a language model such as GPTZero. The higher the perplexity of the text, the higher the likelihood that it is human-written. The converse is true: the lower the text's perplexity, the lower the likelihood that it is human-written. This lower perplexity index signals that a text is AI-generated. Burstiness measures the complexity of sentences or how highly varied sentence usage is in a text. The belief here is that humans are prone to varying the types and the length of their sentences when they write, while machines are not. So, burstiness relates to sentence variability or sentence bursting (Chaka, 2023a; Ofgang, 2023). Most importantly, GPTZero sometimes highlights or flags an AI-generated text in yellow in any given sample and allocates perplexity and burstiness scores to text samples. Higher scores for both measures indicate that a text is human-generated, while lower scores for both measures signal that a text is AI-generated. One of the drawbacks of this tool is that it sometimes misclassifies or misrecognises portions of a text as either AI-generated or human-generated, even in instances where that is not the case (Tyrrell, 2023). So, it is not 100% per cent accurate (Chaka, 2023b).

OpenAI AI Text Classifier

OpenAI AI Text Classifier is an AI detector owned by OpenAI, which also owns ChatGPT. It was released at the beginning of 2023 after the launch of ChatGPT in November 2022. Its main function is to differentiate between AI-generated and human-written text (Eliaçik, 2023b; Ismail, 2023; Tyrrell, 2023). In one of its blogs, its mother tech company asserts that it has "trained a classifier to distinguish between text written by a human and text written by AIs from a variety of providers" (Kirchner et al., 2023, par. 1). It also makes some disclaimers that it is not feasible to fully reliably detect every AI-generated text and that its classifier is not yet fully reliable. It, then, points out that when it tested its classifier in one use case, it had a 26% true positives rate (it correctly identified 26% of AI-generated text) and a 9% false positives rate (it misidentified 9% of human-produced text as AI-generated). According to OpenAI, some of the limitations its text classifier has are as follows:

- Unreliability on shorter texts having fewer than 1,000 characters;
- Only the first 5,000 characters are displayed in the free version;
- Sometimes, the classifier misidentifies longer texts and wrongly labels human-produced text as AI-generated;
- The classifier currently works better on English texts and has a high degree of unreliability on texts written in other languages;
- Unreliability to identify predictable text, especially identifying whether the first 1,000 prime numbers are AI-written or not;
- Edited AI-generated text can evade the classifier; and
- Poor detection of text fine-tuned outside the original training data (Eliaçik, 2023b; Ismail, 2023; Kirchner et al., 2023; Tyrrell, 2023).

Writer.com's AI Content Detector

This AI content detector tool, which is owned by Writer.com, is touted as reliable (Outlook Spotlight, 2023). Unlike most of its peers, it is a no-sign-up or a no-create-an-account tool for usage. It evaluates a text and identifies (by calculating) how much of it is likely AI-generated through percentage scores. It has a 1,500-character limit per text/prompt. Text can be added to this detector by pasting or writing it or by providing a URL of the intended text. The AI tool does not have a 100% accuracy rate, and sometimes, it can be tricked by certain texts (Help Center, 2023; see Lim, 2023). It can also be used for editing and generating text, and its parent company, Writer.com, has offerings such as products (e.g., Grammarly alternative, ChatGPT alternative, and Jasper alternative) and resources (e.g., Inclusive language and AI content generator) (Help Center, 2023; Outlook Spotlight, 2023).

Copyleaks AI Content Detector

Copyleaks AI Content Detector is a free-to-use AI tool that can determine whether a text is generated by AI chatbots like ChatGPT and many others or whether a text is written by a human. According to Copyleaks, this tool has, among others, the following differentiating features:

- A 99.12% detection accuracy rate
- In-depth, detailed analysis
- Detecting GPT-J, GPT-3, GPT-3.5, ChatGPT, GPT-4, and other related AI language models

- Detecting AI content written in multiple languages such as English, Spanish, Polish, Italian, and a few other languages, with more other languages being currently considered

- Verifying the authenticity of social media posts, online news articles, online reviews, etc. (Chrome, 2023; Copyleaks, 2023).

Giant Language Model Test Room

Giant Language Model Test Room (GLTR) is an online tool that employs an algorithm capable of detecting any content related to AI-generated text produced by AI chatbots. It executes a forensic inspection of language model elements on texts to establish whether they are AI- or human-generated. It is supported by a database of predicted words, in which such predicted words are highlighted in green, yellow, and red. The more predicted words a text has, the more likely that it is AI-generated than human-generated. It can also analyse a text for its realness. Its major drawback is that it works better on GPT-2 texts than on GPT-3 texts produced by bots such as ChatGPT (Lim, 2023; Outlook Spotlight, 2023).

All of the five AI content detectors reviewed above were employed in this paper in their free-to-use or non-premium versions. As pointed out earlier, Weber-Wulff et al.'s (2023) paper also evaluated the efficacy of fourteen AI detection tools in their non-premium versions.

Methodology

The paper used an exploratory study design. One key aspect of this study design, which resonates with the present paper, is exploring a topic or an area that has not been studied before (Chaka, 2023a; Elman et al., 2020; Singh, 2021). Evaluating the efficacy, accuracy and reliability of AI content detection tools in differentiating between AI-generated content and human-written content is still a less researched area.

Data collection process

The data collection process for this paper consisted of two stages. In the first stage, the content was generated using ChatGPT, YouChat, and Chatsonic. This content was generated by inputting three sets of English prompts into these three AI chatbots, with each set of prompts for each AI chatbot. The prompts were queried to the three AI chatbots on two different dates. ChatGPT's prompts were queried on 31 January 2023, while the prompts for YouChat, and Chatsonic were inputted on 07 March 2023. This time-lapse was occasioned by the fact that I only became aware of the last two AI chatbots in March 2023 (see Chaka, 2023a). The prompts for these three AI chatbots were based on some of the aspects of applied English language studies (AELS). The latter is one of the areas of my research interests. These prompts are indicated below.

ChatGPT's prompts

- What are decolonial applied English language studies?
- What is critical southern decoloniality?
- Who are the authorities on decolonial linguistics?
- Who are the leading scholars of critical southern decoloniality?
- What is translanguaging?
- What is the difference between translanguaging, multilanguaging, and languaging?

YouChat's prompts

- What are decolonial applied English language studies?
- What is critical southern decoloniality?
- What are Chaka's (2020) views of translanguaging?
- Who are the authorities on decolonial linguistics?
- What is translanguaging?
- What are the latest theories for translanguaging, multilanguaging, and languaging?
- What is the difference between translanguaging, multilanguaging, and languaging?

Chatsonic's prompts

- What is decolonial applied linguistics?
- What are decolonial applied English language studies?
- What is critical southern decoloniality?
- What are Chaka's (2020) views of translanguaging?
- Who are the authorities on decolonial linguistics?
- What is translanguaging?
- What is the difference between translanguaging, multilanguaging, and languaging?

In the second stage, the responses generated from the three AI chatbots were inputted into the five AI content detectors mentioned earlier in three phases from 30 March 2023 to 02 April 2023. During the first phase, the English-only responses were fed into the five AI content detectors. In the second phase, the ChatGPT-generated responses were machine-translated into five languages using Google Translate and inputted into GPTZero. The five languages were German, French, Spanish, Southern Sotho, and isiZulu. The reason for choosing GPTZero for the translated responses is that it was the only AI detector that recognised all five languages at the time of conducting the study. The Southern Sotho that Google Translate uses is the Lesotho orthography of the Sesotho language and not the South African Sesotho orthography. In respect of isiZulu, Google Translate refers to it as Zulu. Henceforth, the paper uses (isi)Zulu to indicate that Zulu has an isi- prefix. During the third phase, the Google-translated German, French, and Spanish responses were fed into Copyleaks AI Content Detector. Currently, Copyleaks AI Content Detector does not support Southern Sotho and (isi)Zulu.

The three sets of AI-generated English responses and the ChatGPT-generated English responses that were Google-translated into the five languages specified above constituted the datasets for this study. After they had been generated and translated, all these datasets were copied and transferred to their respective MS Word files in their original forms. They were not tampered with or manipulated, except that the ChatGPT-generated English responses were Google-translated into the five specified languages. So, they were inputted into the five AI detection tools in their original generated and translated versions.

Results

Detection of the ChatGPT-, YouChat-, and Chatsonic-generated English responses by five AI content detectors

All the ChatGPT-, YouChat-, and Chatsonic-generated responses were subjected to the five AI content detectors, GPTZero, OpenAI AI Text Classifier, Writer.com's AI Content Detector, Copyleaks AI Content Detector, and GLTR for them to detect AI-generated content from these three sets of responses. Concerning ChatGPT-generated responses, all five AI content detection tools yielded their detection results, as illustrated in Table 1. For example, GPTZero correctly classified four texts as AI-generated, while it was indecisive about two texts. Its lowest and highest perplexity scores were 38 and 90. The same kind of classification pattern was yielded by OpenAI AI Text Classifier. Writer.com's AI Content Detector classified five texts inaccurately as human-generated, while its classification of one text was accurate. Its lowest and highest percentages for human-generated content were 1% and 99%. In contrast, Copyleaks AI Content Detector classified five texts accurately, but classified one text inaccurately. Its lowest and highest probability percentages for AI-generated texts were 94% and 99.8%, while its probability percentage for human-generated text was 19.5%. For GLTR, it correctly classified one text as machine-generated, but misclassified five texts. The idea of classified is a proxy for predicted as these tools attempt to predict whether a given text response is AI- or human-generated more than just classifying a given text.

Table 1: Detection of the ChatGPT-, YouChat-, and Chatsonic-generated English responses by five AI content detectors.

Name of AI Content Detector	ChatGPT-Generated Responses[1]	YouChat-Generated Responses	Chatsonic-Generated Responses	Correct detection vs. Incorrect Detection	Indecisiveness	Ranking All the Five AI Content Detectors
Copyleaks AI Content Detector	5 texts' classification (prediction) highly accurate; 1 text's classification inaccurate **NB:** Lowest and highest probability percentages for AI-generated texts: 94% and 99.8% **NB:** Probability percentage for human-generated text: 19.5%	5 texts' classification (prediction) highly accurate; 2 texts' classification inaccurate **NB:** Lowest and highest probability percentages for AI-generated texts: 99.6% and 99.9% (3 texts) **NB:** Lowest and highest probability percentages for human-generated texts: 19.4% and 19.8%	5 texts' classification (prediction) highly accurate; 3 texts' classification inaccurate **NB:** Lowest and highest probability percentages for AI-generated texts: 92.3% and 99.9% (3 texts) **NB:** Lowest and highest probability percentages for human-generated texts: 20% and 98.6%	15 vs. 6	0	1
OpenAI AI Text Classifier	4 texts' classification (prediction) accurate; 2 texts' classification indecisive	3 texts' classification (prediction) accurate; 2 texts' classification indecisive; 2 texts' classification inaccurate	2 texts' classification (prediction) almost accurate; 5 texts' classification indecisive; 1 text's classification inaccurate	9 vs. 3	9	2
GPTZero	4 texts' classification (prediction) accurate; 2 texts' classification indecisive **NB:** Lowest and highest perplexity scores: 38 and 90	1 text's classification (prediction) accurate; 6 texts' classification inaccurate **NB:** Lowest and highest perplexity scores: 40 and 167	3 texts' classification (prediction) accurate; 1 text's classification indecisive; 4 texts' classification inaccurate **NB:** Lowest and highest perplexity scores: 32 and 118	8 vs. 10	3	3
Writer.com's AI Content Detector	5 texts' classification (prediction) inaccurate; 1 text's classification very nearly accurate **NB:** Lowest and highest percentages for human-generated content: 1% and 99%	1 text's classification (prediction) 100% accurate; 3 texts' classification almost accurate; 3 texts' classification inaccurate (one of which was 100% inaccurate) **NB:** Lowest and highest percentages for human-generated content: 0% and 100%	2 texts' classification (prediction) highly and almost accurate; 6 texts' classification inaccurate (two of which were 100% inaccurate) **NB:** Lowest and highest percentages for human-generated content: 2% and 100%	7 vs. 14	0	4
GLTR	1 text machine-generated; 5 texts misclassified	All the 7 texts misclassified	6 texts misclassified; 2 indecisive	1 vs. 18	2	5

Pertaining to YouChat-generated responses, GPTZero classified one text correctly as AI-generated, while it misclassified the six other texts as human-written. Its lowest and highest perplexity scores for all these texts were 40 and 167. OpenAI AI Text Classifier classified three texts accurately as AI-generated but misclassified two texts. It was indecisive about two texts. Writer.com's AI Content Detector correctly classified four texts as AI-generated, but misclassified three texts as human-generated. Its lowest and highest percentages for human-generated content for these texts were 0% and 100%. For its part, Copyleaks AI Content Detector correctly classified five texts as AI-generated and incorrectly classified the other two texts as human-generated. Its lowest and highest probability percentages for AI generated texts were 99.6% and 99.9%, respectively, with three texts having a 99.9% tie. Its lowest and highest probability percentages for human-generated texts were 19.4% and 19.8%. In contrast to the other four AI detectors, GLTR misidentified all seven texts as human-written.

With reference to Chatsonic-generated responses, GPTZero correctly classified three texts as AI-generated, but misclassified four texts as human-generated. It was indecisive about one text. It recorded the lowest and highest perplexity scores for these eight texts as 32 and 118. OpenAI AI Text Classifier identified two texts correctly as AI-generated but detected one text incorrectly. It was indecisive about five more texts. Writer.com's AI Content Detector correctly identified two texts as AI-generated but misidentified six texts as human-written. Its lowest and highest percentages for human-generated content were 2% and 100%. In this regard, Copyleaks AI Content Detector correctly classified five texts as AI-generated but misclassified three texts as human-written. Its lowest and highest probability percentages for AI-generated texts were 92.3% and 99.9%, with three texts having a 99.9% tie. However, its lowest and highest probability percentages for human-generated texts were 20% and 98.6%, respectively. Again, in contrast to the

other four AI detectors, GLTR misclassified six texts, while it was indecisive about two texts.

When the five AI content detectors were judged together in terms of their overall correct identification of the three sets of responses generated by the three AI tools, they rank as follows: Copyleaks AI Content Detector (1); OpenAI AI Text Classifier (2); GPTZero (3); and Writer.com's AI Content Detector (4), and GLTR (5) (see Table 1).

Detection by GPTZero of the ChatGPT-generated responses Google-translated into German, French, Spanish, Southern Sotho, and (isi)Zulu

In this section, what is at issue is not the accuracy and correctness of the Google-translated texts for all five languages but rather GPTZero's ability to classify them correctly as machine-generated texts. All the ChatGPT-generated responses were subjected to GPTZero for it to detect if they were AI-generated or human-written (see Table 2). GPTZero incorrectly classified all the translated texts in all five languages as human-written. Its high and lowest perplexity scores for the texts translated into each of these languages were as follows: 110 and 2,478 (German); 57 and 221 (French); 108 and 361 (Spanish); 602 and 1,715 (Southern Sotho); and 651 and 938 ((isi)Zulu).

Table 2: Detection by GPTZero of the ChatGPT-generated responses Google-translated into German, French, Spanish, Southern Sotho, and (isi)Zulu.

Name of AI Content Detector	Google-German-Translated ChatGPT-Generated Responses	Google-French-Translated ChatGPT-Generated Responses	Google-Spanish-Translated ChatGPT-Generated Responses	Google-Southern Sotho-Translated ChatGPT-Generated Responses	Google-(isi)Zulu-Translated ChatGPT-Generated Responses
GPTZero	All 6 the texts' classification (prediction) inaccurate **NB:** All high perplexity scores, with the lowest and highest scores being 110 and 2,478.	All 6 the texts' classification (prediction)inaccurate **NB:** Lowest and highest perplexity scores being 57 and 221.	All 6 the texts' classification (prediction) inaccurate **NB:** Lowest and highest perplexity scores being 108 and 361.	All 6 the texts' prediction inaccurate **NB:** All high perplexity scores, with the lowest and highest scores being 602 and 1,715.	All 6 the texts' classification (prediction) inaccurate **NB:** All high perplexity scores, with the lowest and highest scores being 651 and 938.

Detection by Copyleaks AI Content Detector of the ChatGPT responses Google-translated into German, French, and Spanish

Similarly, here, all the ChatGPT-generated responses, which were Google-translated into the three languages as mentioned above, were subjected to Copyleaks AI Content Detector for it to detect whether they were AI-generated or not (see Table 3). This AI content detector correctly classified three German-translated texts as AI-generated but misclassified three texts as human-written. Its lowest and highest probability percentages for AI-generated texts were 83.7% and 99.9%, while its lowest and highest probability percentages for human-generated texts were 14.8% and 53.4%. It, then, correctly identified five French-translated texts as AI-generated but misidentified one text as human-written. Here, its lowest and highest probability percentages for AI-generated texts were 94.9% and 99.9%, with three texts having a tie at 99.9%. Its probability percentage for the human-generated text was 6.9%. Lastly, Copyleaks AI Content Detector correctly classified all the Spanish-translated texts as AI-generated. Its lowest and highest probability percentages for these texts were 99. 5% and 99.9%, with two texts and four texts tied at 99.5% and 99.9%, respectively.

Table 3: Detection by Copyleaks AI Content Detector of the ChatGPT responses Google-translated into German, French, and Spanish.

Name of AI Content Detector	Google-German-Translated ChatGPT-Generated Responses	Google-French-Translated ChatGPT-Generated Responses	Google-Spanish-Translated ChatGPT-Generated Responses
Copyleaks AI Content Detector	2 texts' classification (prediction) highly accurate; 1 text's classification almost accurate; 3 texts' classification inaccurate **NB:** Lowest and highest probability percentages for AI generated texts: 83.7% and 99.9% (2 texts) **NB:** Lowest and highest probability percentages for human-generated texts: 14.8% and 53.4%	5 texts' classification (prediction) highly accurate; 1 text's classification inaccurate **NB:** Lowest and highest probability percentages for AI generated texts: 94.9% and 99.9% (3 texts) **NB:** Probability percentage for human-generated text: 6.9%	All the 6 texts' classification highly accurate. **NB:** Lowest and highest probability percentages for AI generated texts: 99. 5% (2 texts) and 99.9% (4 texts)

Discussion

For ChatGPT-generated responses, Copyleaks AI Content Detector had more correct classifications of the texts than the other four AI detectors. It was followed by GPTZero and OpenAI AI Text Classifier, which were joint second. In terms of misclassifications of texts (incorrect classifications of texts), GLTR topped the other four AI detectors with eighteen misclassifications; it was followed by Writer.com's AI Content Detector with fourteen misclassifications. AI detectors with the joint-most indecisive texts were GPTZero and OpenAI AI Text Classifier. GPTZero had a perplexity score of 90 for one of its AI-generated texts, which is a high score given that AI-generated texts are supposed to have a lower perplexity index compared to human-written texts (Chaka, 2023a; Ofgang, 2023; Tech Desk, 2023; Tyrrell, 2023). This, in a way, highlights an element of shakiness related to equating a high perplexity with human-only-written texts in an instance where machines, too, can generate texts with a high perplexity index (Heel, 2023; Lim, 2023; Wiggers, 2023). Writer.com's AI Content Detector recorded 1% and 99% as its lowest and highest percentages for human-generated content for two texts apiece. This means it identified the first text as 99% AI-generated, while it recognised the second text to be 1% AI-generated. These are two extremely contrasting detection rates for these texts when considering that all the texts in this set were ChatGPT-generated. For Copyleaks AI Content Detector, its probability percentage for a human-generated text was 19.4%, which is a bit high for texts that were exclusively machine-generated.

In relation to YouChat-generated responses, again, Copyleaks AI Content Detector had more correct identifications of the texts in this set than the other four AI detectors. It was followed by OpenAI AI Text Classifier. The other AI detectors had more misclassifications of texts than the correct classifications, with GLTR racking up the most misclassified texts. Only one AI detector (OpenAI AI Text Classifier) had two undecided texts. The highest perplexity score that GPTZero had for this set of texts is 167, which is a very high score for texts that were machine-generated. The concern

raised above about a high perplexity as an indicator of human-produced texts, applies here, too (Tech Desk, 2023; Iyer, 2023). In this context, Writer.com's AI Content Detector had 0% and 100% as its lowest and highest percentages for human-generated content: 0% and 100%. As is the case with the previous instance, these are two diametrically opposed detection rates for texts that were ChatGPT-generated. Copyleaks AI Content Detector recorded 19.4% and 19.8% as its lowest and highest probability percentages for human-generated texts. Again, these are high probability percentages for machine-generated texts.

Concerning Chatsonic-generated responses, the same trend as the one characterised above applies with minor variations. For example, Copyleaks AI Content Detector still had the most correct identifications of the texts in this set, but with GPTZero following it. Both Writer.com's AI Content Detector and GLTR had the most joint misidentified texts, followed by GPTZero. OpenAI AI Text Classifier had the most undecided texts. GPTZero had the highest perplexity score of 167, which, again, is a high score for texts that were machine-generated. Writer.com's AI Content Detector recorded the contrasting lowest and highest percentages of 2% and 100% for human-generated content. Copyleaks AI Content Detector's highest probability percentage of 98.6% for human-generated text was the highest ever for these machine-generated texts.

Overall, of the five AI content detectors tested across the three sets of texts, OpenAI AI Text Classifier and GLTR appeared to be most consistent in their detection rates if the indecisiveness of texts and the misclassification of texts are, respectively, used factors. In contrast, Copyleaks AI Content Detector tended to be the most consistent of the five AI content detectors if the correct identification of texts is used as a factor. Moreover, Copyleaks AI Content Detector trumped all the other four AI content detectors for the most correctly classified texts. GLTR had the highest text misclassification rate and could classify only one text correctly. As such, it ranked last among the five AI content detectors. In an instance in which seven AI text detectors, which included OpenAI AI Text Classifier, GPTZero and Copyleaks, were tested to detect AI-generated content created by Claude (a generative AI tool similar to ChatGPT), GPTZero was the top consistent performer. It was followed by ChatGPT and OpenAI AI Text Classifier. The writing samples were based on prompts related to different writing genres (Wiggers, 2023). However, in the current paper, GPTZero was the third-best performing AI content detector (see Table 1). As mentioned earlier, Aremu's (2023) paper that tested the detection capabilities of six AI detection tools found both ZeroGPT and GPTZero to have a higher level of deception robustness and resistance than the other four AI detection tools.

The fact that the five AI content detectors recognised some of the AI-generated texts inputted to them as human-written points to their propensity to false negativity. In their paper, Weber-Wulff et al. (2023) also found that fourteen of the AI detection tools they evaluated were prone to false negatives: they mistook AI-generated texts for human-produced texts. They call this tendency "misattributing" AI-generated texts to humans. This is what the present paper has referred to as misclassification and misidentification.

As mentioned earlier, as regards the ChatGPT responses that were Google-translated into the five aforementioned languages, GPTZero misidentified all of them as human-produced. One major reason it misidentified all these translated texts is the higher perplexity scores it assigned to them. This is particularly the case with the German, Southern Sotho and (isi)Zulu texts, whose highest perplexity scores were 2,478, 1,715 and 938, respectively. This, more than what has been said earlier, highlights the shakiness and, at times, the unreliability of a higher perplexity as an indicator of human-only-written texts. This is more so given that machine-translated texts can have inordinately higher perplexity scores, such as the ones for the five languages in this paper.

In contrast, Copyleaks AI Content Detector correctly identified three of the German-translated texts, five of the French-translated texts, and all the Spanish-translated texts as AI-generated. Again, here, Copyleaks AI Content Detector outperformed GPTZero, an outcome that contrasts with that of Wiggers' testing (2023) of seven AI content detectors in which GPTZero was a top performer.

Even though in the current study no texts were deliberately manipulated through editing, paraphrasing, or effecting an extra space (a space bar) between words, there are studies that have discovered that text manipulation reduces the detection efficacy of AI detection tools. For instance, Cai and Cui (2023) found that effecting a mere single space, what they call an extra space, results in text detection evasion. If this is the case, this points to one of the inherent problems with current AI detection tools: their lack of reliability and consistency in accurately differentiating between AI-generated texts and human-produced texts. Part of this problem might have to do with the algorithmic configuration of these AI detection tools, which assumes that distributional gaps exist between AI-generated and human-written content. Once these distributional gaps are destabilised by, for example, intentionally adding single space characters before commas in AI-generated content, these tools tend to misrecognise the content output (see Cai & Cui, 2023). In addition, Guo et al. (2023) point out that by removing indicating words such as Nope, My take is, and Hmm, from human-written content and by removing I regret to hear that, I'm an AI assistant, and Here are steps to follow, from an AI-generated content, most AI detection tools are likely to be tricked in their detection capability. These inherent algorithmic shortcomings are likely to be there in the premium versions of these AI detection tools as well.

Implications and limitations

With the release of ChatGPT, as a generative AI chatbot, for public use, a lot of AI content detectors were instantly launched, even though others could have been there before the launch of ChatGPT. Some of these AI content detectors, such as GPTZero and OpenAI AI Text Classifier, were specifically intended to detect AI-generated content from ChatGPT (Eliaçik, 2023a, b; Ismael, 2023; Iyer, 2023: Kirchner et al., 2023; Lim, 2023; Ofgang, 2033). Nevertheless, with their instant launch, these AI content detectors seem

not yet fully ready to accurately and convincingly detect AI-generated content from machine-generated texts in different contexts. Most of them seem to be beset by the algorithmic shortcomings mentioned above. This is one of the implications emanating from the five AI content detectors tested in this paper. Additionally, the five AI content detectors were not able to distinguish, in clear-cut terms, between AI-generated texts and human-produced texts. All they could do was to make estimates in percentages (e.g., Writer.com's AI Content Detector and Copyleaks AI Content Detector) or in probabilistic terms such as likely (e.g., GPTZero) and probability (e.g., Copyleaks AI Content Detector), or in combined percentages and probabilistic terms (e.g., Copyleaks AI Content Detector). Others, such as GLTR, used estimating histograms. Educational institutions, academic staff, and students are impatiently waiting for an AI content detector that will precisely, accurately, and correctly detect AI-generated and human-written texts every time they apply them to student writing and to academic essay writing. They are not interested in percentage and probabilistic estimates.

A major limitation of this paper is that it used free-to-use AI content detectors or the non-premium versions of some of these AI content detectors. In fact, when the data for this study were collected, all five AI content detectors had only free versions that were available to the public. Nonetheless, most of them now do have premium or paid-for versions. The tricky thing about the premium versions of these AI detection tools is that one has to have a paid-for subscription with them for one to be able to access and use them. This becomes almost impossible if a researcher wants to evaluate more of them at the same time. But it seems implausible that the premiums versions of these AI detection tools are free of the two algorithmic shortcomings mentioned above. Mostly, what their premiums versions boast of are differentiators such as increased word/character counts and uploading multiple full-text files as part of premium benefits. In the main, these differentiators are, at best, mechanical, and, at worst, not game-changing. One of the things needed to help improve the AI detection efficacy of AI tools is improved super-intuitive machine learning algorithms that can detect sophisticated and subtle stylometric patterns built into language use (see Uzun, 2023).

References

Alser, M., & Waisberg, E. (2023). Concerns with the usage of ChatGPT in academia and medicine: A viewpoint. *American Journal of Medicine Open, 9*(100036), 1-2. https://doi.org/10.1016/j.ajmo.2023.100036

Aremu, T. (2023). *Unlocking Pandora's box: Unveiling the elusive realm of AI text detection*. https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4470719_code5947956.pdf?abstractid=4470719&mirid=1

Anders, B. A. (2023). Why ChatGPT is such a big deal for education. *C2C Digital Magazine*. https://scalar.usc.edu/works/c2c-digital-magazine-fall-2022---winter-2023/why-chatgpt-is-bigdeal-education

Awan, A. A. (2023). Top 10 tools for detecting ChatGPT, GPT-4, Bard, and Claude. *KDnuggets*. https://www.kdnuggets.com/2023/05/top-10-tools-detecting-chatgpt-gpt4-bard-llms.html

Barnett, S. (2023). ChatGPT is making universities rethink content. *Wired*. https://www.wired.com/story/chatgpt-college-university-content/#intcid=_wired-amp-bottom-recirc_8fe776e7-19f4-4523-9ba8-1b1d23637669_text2vec1

Cai, S., & Cui, W. (2023). *Evade ChatGPT detectors via a single space*. Arxiv. https://arxiv.org/pdf/2307.02599.pdf

Caren, C. (2023). AI writing: The challenge and opportunity in front of education now. *Turnitin*. https://www.turnitin.com/blog/ai-writing-the-challenge-andopportunity-in-front-of-education-now

Ceres, P. (2023). ChatGPT is coming for classrooms. Don't panic. *Wired*. https://www.wired.com/story/chatgpt-is-coming-for-classrooms-dont-panic/#intcid=_wired-bottom-recirc_9c0d2ac5-941b-45c7-b9ac-7ac221fc2e33_wired-content-attribution-evergreen

Chaka, C. (2020). Higher education institutions and the use of online instruction and online tools and resources during the COVID-19 outbreak – An online review of selected U.S. and SA's universities. *Research Square,* 1-46. Preprint. https://doi.org 10.21203/rs.3.rs-61482/v1

Chaka, C. (2023a). Generative AI chatbots - ChatGPT versus YouChat versus Chatsonic: Use cases of selected areas of applied English language studies. *International Journal of Learning, Teaching and Educational Research, 22*(6), 1-19. https://doi.org/10.26803/ijlter.22.6.1

Chaka, C. (2023b). Stylised-facts view of fourth industrial revolution technologies impacting digital learning and workplace environments: ChatGPT and critical reflections. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1150499.

Chrome. (2023). *AI content detector – Copyleaks*. https://chrome.google.com/webstore/detail/ai-content-detector-copyl/gplcmncpklkdjiccbknjjkoidpgkcakd

Copyleaks. (2023). *ChatGPT and AI content detection*. https://copyleaks.com/blog/chatgpt-and-ai-content-detection

Cotton, D. R. E., Cotton, P. A., & Shipway, L. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 60,* 1-13. https://doi.org/10.1080/14703297.2023.2190148

Cutcliffe, J. (2022). ChatGPT, chatbots and artificial intelligence in education. *Ditch That Textbook*. https://ditchthattextbook.com/ai/

Ifelebuegu, A. (2023). Rethinking online assessment strategies: Authenticity versus AI chatbot intervention. *Journal of Applied Learning and Teaching, 6*(2). Advance online publication. https://journals.sfu.ca/jalt/index.php/

jalt/article/view/875

Eliaçık, E. (2023a). Best ChatGPT alternatives: What awaits us beyond ChatGPT? *DataConomy*. https://dataconomy.com/2023/02/beyond-chatgpt-alternatives-best-jasper/

Eliaçık, E. (2023b). *AI Text Classifier: OpenAI's ChatGPT detector indicates AI-generated text*. https://dataconomy.com/2023/02/openai-ai-text-classifier-chatgpt-detector/

Gerring, J., Mahomey, J., & Elman, C. (2020). Introduction. In C. Elman, J. Gerring & J. Mahoney (Eds.), *The production of knowledge: Enhancing progress in social science* (pp. 1-14). Cambridge University Press.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*. Arxiv. https://arxiv.org/pdf/2301.07597.pdf

Harris, M. (2022). ChatGPT—The game-changing app every teacher should know about. *Learners Edge*. https://www.learnersedge.com/blog/chatgpt-the-game-changing-app-everyteacher-should-know-about

Heilweil, R. (2022). AI is finally good at stuff, and that's a problem. *Vox*. https://www.vox.com/recode/2022/12/7/23498694/ai-artificial-intelligence-chat-gpt-openai

Help Center. (2023). *AI content detector*. https://support.writer.com/article/205-ai-content-detector

Hern, A. (2022, December 4). AI bot ChatGPT stuns academics with essay-writing skills and usability. *The Guardian*. https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stuns-academics-with-essay-writing-skills-and-usability

Hetler, A. (2023). Bard vs. ChatGPT: What's the difference? *TechTarget*. https://www.techtarget.com/whatis/feature/Bard-vs-ChatGPT-Whats-thedifference

Ismail, M. (2023). OpenAI releases AI classifier to check for Chat GPT written articles, still needs work. *TechNave*. https://technave.com/gadget/OpenAI-releases-AI-classifier-to-check-for-Chat-GPT-written-articles-still-needs-work-33313.html

Iyer, A. (2023). After GPTZero, Turnitin is developing tool to identify AI-generated text. *Analyticsindiamag*. https://analyticsindiamag.com/after-gptzero-turnitin-is-developing-a-tool-to-identify-ai-generated-text/

Kamran, S. (2023). *Microsoft Bing vs Google Bard: ChatGPT cloning battle has begun*. https://techcommunity.microsoft.com/t5/itops-talk/microsoft-bing-vs-googlebard-chatgpt-cloning-battle-has-begun/m-p/3737979

Khalil, M., & Er, E. (2023). *Will ChatGPT get you caught? Rethinking of plagiarism detection*. Arxiv. https://arxiv.org/pdf/2302.04335.pdf

Kirchner, J. H., Ahmad, L., Aaronson, S., & Leike, J. (2023). New AI classifier for indicating AI-written text. *OpenAI*. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

Lim, H. (2023). 5 content detection tools to tell if content is written by ChatGPT. *HongKiat*. https://www.hongkiat.com/blog/chatgpt-content-detection-tools/

Ofgang, E. (2023). What is GPTZero? The ChatGPT detection tool explained by its creator. *TechLearning*. https://www.techlearning.com/news/what-is-gptzero-the-chatgptdetection-tool-explained

Outlook Spotlight. (2023). Best AI content detection tools: Free ChatGPT output detector. *Outlook India*. https://www.outlookindia.com/outlook-spotlight/best-ai-content-detection-tools-free-chatgpt-output-detector-news-256773

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice, 20*(2), 1-26. https://doi.org/10.53761/1.20.02.07

Popenici, S. (2023). The critique of AI as a foundation for judicious use in higher education. *Journal of Applied Learning and Teaching, 6*(2). Advance online publication. https://journals.sfu.ca/jalt/index.php/jalt/article/view/883

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching, 6*(1), 41-56. https://journals.sfu.ca/jalt/index.php/jalt/article/view/787

Rudolph, J., Tan, S., & Tan, S. (2023a). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://journals.sfu.ca/jalt/index.php/jalt/article/view/689

Rudolph, J., Tan, S., & Tan, S. (2023b). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://journals.sfu.ca/jalt/index.php/jalt/article/view/771

Sample, I. (2023). Science journals ban listing of ChatGPT as co-author on papers. *The Guardian*. https://www.theguardian.com/science/2023/jan/26/science-journals-ban-listing-of-chatgpt-as-co-author-on-papers

Singh, A. (2021). *An introduction to experimental and exploratory research*. https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3789360_code3438088.pdf?abstractid=3789360&mirid=1

Somoye, F. L. (2023). *Can universities detect ChatGPT?* https://www.pcguide.com/apps/can-universities-detect-chat-gpt/

Stokel-Walker, C. (2022). AI bot CHATGPT writes smart essays — Should professors worry? *Nature*. https://doi.

org/10.1038/d41586-022-04397-7.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching, 6*(1), 31-40. https://journals.sfu.ca/jalt/index.php/jalt/article/view/731

Tech Desk. (2023). *GPTZero to help teachers deal with ChatGPT-generated student essays.* https://indianexpress.com/article/technology/gptzero-app-helpsteachers-catch-chatgpt-8378062/

Tyrrell, J. (2023). *ChatGPT screening: OpenAI text classifier versus GPTZero app*. https://techhq.com/2023/02/chatgpt-screening-openai-text-classifier-versusgptzero-app/

Ventayen, R. J. M. (2023). OpenAI ChatGPT generated results: Similarity index of artificial intelligence-based contents. *Advances in Intelligent Systems and Computing,* 1-6. http://dx.doi.org/10.2139/ssrn.4332664

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O. Šigut, P., & Waddington, L. (2023). *Testing of detection tools for AI-generated text*. arXiv https://doi.org/10.48550/arXiv.2306.15666

Wiggers, K. (2023). Most sites claiming to catch AI-written text fail spectacularly. *TechCrunch.* https://techcrunch.com/2023/02/16/most-sites-claiming-to-catch-ai-written-text-fail-spectacularly/

Wingard, J. (2023). ChatGPT: A threat to higher education? *LinkedIn*. https://www.linkedin.com/pulse/chatgpt-threat-higher-education-dr-jason-wingard

Yeadon, W., Inyang, O-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(035027), 1-13.