Vol.7 No.1 (2024)

**Journal of Applied Learning & Teaching**

**ISSN** : **2591-801X**

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

# Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review

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

## Keywords

Academic and scientific writing;
AI detection accuracy;
AI detection reliability;
AI detection tools;
AI-generated text;
higher education;
human-written text;
large language models (LLMs);
review study.

## Correspondence

chakachaka8@gmail.com[A]

## Article Info

Received 16 December 2023
Received in revised form 18 January 2024
Accepted 3 February 2024
Available online 7 February 2024

**DOI:** https://doi.org/10.37074/jalt.2024.7.1.14

## Abstract

The purpose of this study was to review 17 articles published between January 2023 and November 2023 that dealt with the performance of AI detectors in differentiating between AI-generated and human-written texts. Employing a slightly modified version of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol and an aggregated set of quality evaluation criteria adapted from A MeaSurement Tool to Assess systematic Reviews (AMSTAR) tool, the study was conducted from 1 October 2023 to 30 November 2023 and guided by six research questions. The study conducted its searches on eleven online databases, two Internet search engines, and one academic social networking site. The geolocation and authorship of the 17 reviewed articles were spread across twelve countries in both the Global North and the Global South. ChatGPT (in its two versions, GPT-3.5 and GPT-4) was the sole AI text generator used or was one of the AI text generators in instances where more than one AI text generator had been used. Crossplag was the top-performing AI detection tool, followed by Copyleaks. Duplichecker and Writer were the worst-performing AI detection tools in instances in which they had been used. One of the major aspects flagged by the main findings of the 17 reviewed articles is the inconsistency of the detection efficacy of all the tested AI detectors and all the tested anti-plagiarism detection tools. Both sets of detection tools were found to lack detection reliability. As a result, this study recommends utilising both contemporary AI detectors and traditional anti-plagiarism detection tools, together with human reviewers/raters, in an ongoing search for differentiating between AI-generated and human-written texts.

## Introduction

Since the launch of ChatGPT on 30 November 2022, much research, both academic and non-academic papers, and numerous preprints have been published on the multiple uses for which generative artificial intelligence (AI) chatbots or AI-powered large language models (LLMs) can be put to educational settings. These types of chatbots are also known as AI text generators. The multifarious affordances of these AI text generators are now well documented. Some of these affordances include educational content generation (Ifelebuegu et al., 2023; Kasneci et al., 2023; Perkins et al., 2023; cf. Chaka, 2023a; Rudolph et al., 2023), enhancing online assessments and supporting collaborative assessments (Gamage et al., 2023; Ifelebuegu, 2023a; Kasneci et al., 2023), language learning and personalised learning (Chaka, 2023a, 2023b; Hew et al., 2023; Ifelebuegu et al., 2023; Jeon & Lee, 2023), student learning (Abbas et al., 2023; Hew et al., 2023; Sullivan et al., 2023); essay writing (Chaka, 2023a; Yeadon et al., 2023); student/teaching assistants (Jeon & Lee, 2023; Kasneci et al., 2023; Kuhail et al., 2023; Nah et al., 2023); and conducting and publishing research (Kooli, 2023; van Dis et al., 2023). Equally, the various challenges and risks AI chatbots pose in academia have now been profusely documented as well. Among these are academic dishonesty and plagiarism (Chaka, 2023a, 2023c; Cotton et al., 2023; Ifelebuegu, 2023; Ifelebuegu et al., 2023; Kasneci et al., 2023; Kleebayoon & Wiwanitkit, 2023; Kooli, 2023; Perkins et al., 2023; Rudolph et al., 2023; Sullivan et al., 2023) and bias and unfairness (Dwivedi et al., 2023; Kasneci et al., 2023; Nah et al., 2023; Ray, 2023). To this effect, some review studies have been conducted on the use of the new AI chatbots in education in general (see Baidoo-Anu & Ansah, 2023; Dergaa et al., 2023; Ifelebuegu et al., 2023; Perera & Lankathilaka, 2023; Pinzolits, 2023; Sullivan et al., 2023; Thurzo et al., 2023; Yang et al., 2023). For example, Baidoo-Anu and Ansah's (2023) study investigated, among other things, the potential benefits of ChatGPT in education as reported in peer-reviewed journal articles and/or in preprints published between November 2022 and March 2023. In addition, Dergaa et al.'s (2023) study explored the possible benefits and threats of ChatGPT and other natural language processing technologies in academic writing and research publications as reported in peer-reviewed journal articles indexed in Scopus's quartile 1.

Importantly, instances of AI tools that can detect AI-generated content and that can distinguish this type of content from the one written by humans have been investigated by a number of scholars. Scholars who have done so include, among others, Abani et al. (2023), Alexander et al. (2023), Anil et al. (2023), Chaka (2023c), Elkhatat et al. (2023), Gao et al. (2023), Perkins et al. (2023), and Uzun (2023). These scholars have done so in varying degrees and by focusing on different types of AI detection tools. The detection tools explored include single detection tools (Habibzadeh, 2023; Perkins et al., 2023; Subramaniam, 2023); two detection tools (Bisi et al., 2023; Desaire et al., 2023; Ibrahim, 2023); three detection tools (Cingillioglu, 2023; Elali & Rachid, 2023; Gao et al., 2023; Homolak, 2023; Ladha et al., 2023; Wee & Reimer, 2023); four detection tools (Abani et al., 2023; Alexander et al., 2023; Anil et al., 2023); and multiple detection tools (Chaka, 2023c; Odri & Yoon, 2023; Santra & Majhi, 2023; Walters, 2023). But more scholarly papers published in this area are preprints, which, at the moment, tend to outnumber journal articles and book chapters. However, unlike the picture painted above, there are, if any, few review studies that have been published in this area (cf. Baidoo-Anu & Ansah, 2023; Dergaa et al., 2023; Ifelebuegu et al., 2023; Perera & Lankathilaka, 2023; Pinzolits, 2023; Sullivan et al., 2023; Thurzo et al., 2023; Yang et al., 2023). Rather, the bulk of scholarly papers published in this area are, again, preprints (see Aremu, 2023; Maddugo, 2023; Weber-Wulff et al., 2023) and, to some extent, conference proceedings (see Sarzaeim et al., 2023; Singh, 2023).

At the time of writing this paper, there was no published peer-reviewed review journal article on AI detection tools differentiating between AI-generated and human-written texts. Such review publications are essential for the purpose of framing a related work section to highlight and interrogate issues pertaining to specific AI detection tools that relevant review studies have explored. So, in the absence of such studies, the present paper will not have a related work section. This paper consists of the following sections: the purpose of the study, article characteristics and research questions, methods (search strategy, eligibility criteria and selection of peer-reviewed journal articles, quality evaluation, coding, and inter-rater reliability, data extraction and analysis), findings and discussion, and conclusion. All of these sections together constitute a review protocol (see Xiao & Watson, 2019).

## Purpose of the study, article characteristics, and research questions

AIn light of the points highlighted above, the purpose of this study was to review 17 articles published between January 2023 and November 2023 that focused on the performance of AI detection tools in differentiating between AI-generated and human-written texts. The focus of the study was on AI detection tools employed in the higher education (HE) sector during this period. This purpose was informed by the fact that the study wanted to establish which AI detection tools in the reviewed studies are reported to perform better in differentiating between AI-generated and human-written texts. Establishing which AI detection tools perform better and knowing whether their detection accuracy is reliable are some of the key factors confronting the HE sector since the release of ChatGPT and the proliferation of AI-powered chatbots that followed after its launch. The purpose of the study also had to do with the overall desire to contribute to review studies in this area of AI detection tools.

There were twelve article characteristics investigated in each review article. These were as illustrated in Table 1. To this end, the study had the following research questions (RQs):

- RQ 1: What types of articles have the current review study identified, and what discipline do they belong to?
- RQ 2: What is the purpose of each article?

- RQ 3: What are the AI-generated and human-written texts tested?
- RQ 4: What is the number and what are the names of the AI detection tools used, and what are the best- and worst-performing AI detection tools reported?
- RQ 5: What are the detection accuracy rate and the detection accuracy reliability reported?
- RQ 6: What are the main findings and the key conclusions of the 17 reviewed articles?

## Method

There are different typologies of review studies. For instance, Grant and Booth (2009) identify fourteen different types of review studies, of which rapid reviews, scoping reviews, literature reviews, systematic reviews, meta-analyses, and integrative synthesis reviews are but a few examples (cf. Xiao & Watson's 2019 sixteen types of review studies). These review types differ mainly in terms of their foci, aims, search strategies, appraisals, analyses, and syntheses (Grant & Booth, 2009). Due to space constraints, I will briefly describe a literature review and a synthesis review as they constitute the current study. The present study is a review that comprises literature and synthesis review components. In its literature review component, the study focused on currently published peer-reviewed journal articles on AI detection tools differentiating between AI-generated and human-written texts in more than one field of study. Its searches were comprehensive but constrained by a given timeline, and its quality assessment was proscribed by the scarcity of published peer-reviewed journal articles on its focus area.

Additionally, the study employed a thematic analysis and a narrative synthesis. In its integrative synthesis outlook, the study integrated and compared peer-reviewed journal articles currently published in its focus area and selected all the relevant articles that were retrievable from the online search platforms on which it conducted its search strategies. Similarly, its analysis and synthesis were thematic and narrative, respectively. Importantly, the aim of an integrative synthesis is to broaden how a given phenomenon is understood (see Grant & Booth, 2009; cf. Chaka, 2022, 2023d; Snyder, 2019; Xiao & Watson, 2019). When two different types of reviews have been fused, as is the case in this study, such a product is referred to as a hybrid review (see Xiao & Watson, 2019; also see Bacon, 2017). This type of hybrid review entails summarising and synthesising findings from reviewed studies (Bacon, 2017; Grant & Booth, 2009; Snyder, 2019).

Even though this study is a hybrid review study as specified above, for transparency purposes, it followed a slightly modified version of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol in its review process, as spelt out below. Four key features of the PRISMA reporting protocol are comprehensiveness, systematicity, transparency, and rigour in the review process (e.g., literature searches, screening and identifying eligible articles (publications), data extraction and analysis, and summarising and synthesising findings) (see Chaka, 2022, 2023d; Ismail et al., 2023; Stracke et al., 2023; Yang et al., 2023).

### Search strategy

A literature search for potential peer-reviewed journal articles was carried out from 1 October 2023 to 30 November 2023. The search was conducted on Internet search engines, online databases, and one academic social networking site. These online search platforms were as follows: two Internet search engines (Google search and Microsoft Bing search), eleven online databases (Google Scholar, Semantic Scholar, Taylor & Francis Online, Wiley Online Library, ScienceDirect, Scopus, SpringerLink, IEEE Xplore Digital Library, ERIC, JSTOR, and BASE), and ResearchGate. Altogether, these constituted fourteen online platforms (see Figure 1; cf. Chaka, 2022, 2023d; Ismail et al., 2023; Stracke et al., 2023). All of these online platforms were easily accessible, while the others, such as EBSCO and Web of Science, had paywalls.

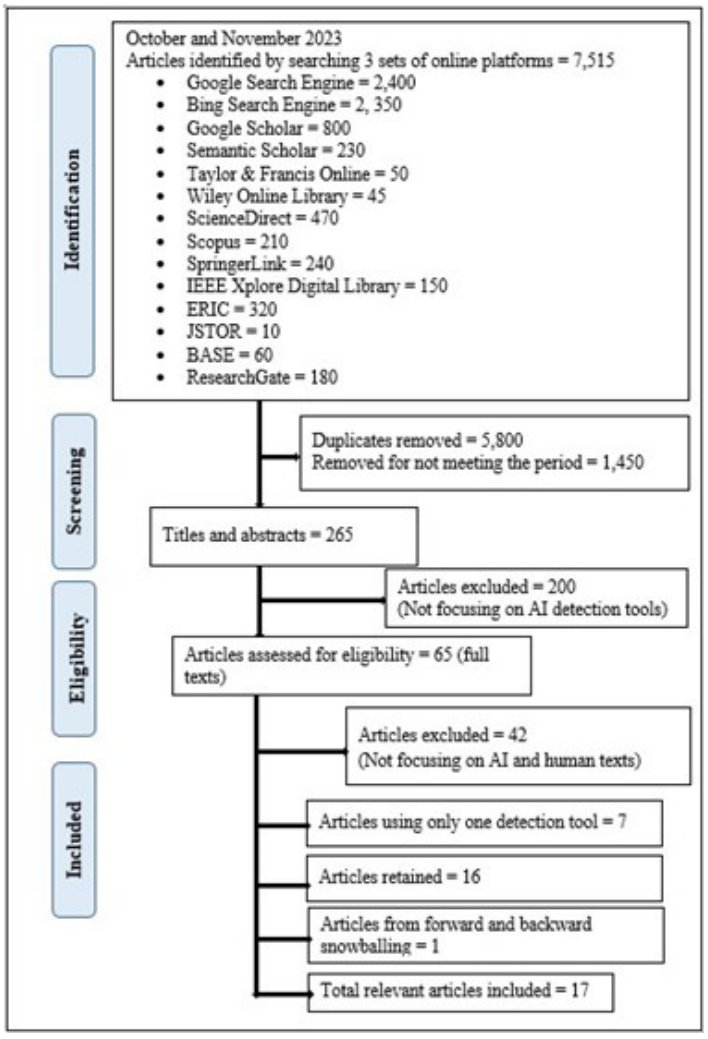


Figure 1: PRISMA flowchart for screening articles.

Search strings included keywords, phrases, and short clauses related to the focus area of the study: AI detection tools used in differentiating between AI-generated and human-written texts. Even though the application context of these AI detection tools was higher education, the search strings were left open-ended in order to source wide-ranging AI

detection tools. This was after the researcher had realised the scarcity of peer-reviewed journal articles focusing on this area at the time of conducting the present study. The search strings consisted of Boolean operators (AND or OR) (see Chaka, 2023d) and truncation symbols such as *, \\, or -, depending on the search platform. Moreover, the permutations of these search strings were used iteratively. Below are examples of the search strings that were employed:

- Published papers on AI-generated content detection tools in 2023 (Google, Bing, Google Scholar, and ResearchGate)
- Tools for detecting artificial intelligence-generated content (Taylor & Francis Online, Wiley Online Library, ScienceDirect, Scopus, and SpringerLink)
- Detecting AND AI texts OR human texts – 2023 (Google, Bing, Google Scholar, and ResearchGate)
- Differentiating between AI\-generated and human\-written text using AI detection tools (Semantic Scholar)
- Tools to detect \\\-written and human\\\-written text (Semantic Scholar)
- Differentiating between AI\-generated and human\-written text using AI detection tools (Wiley Online Library)
- Best AI tools to detect AI plagiarism; Plagia* detec*; Detect* artificial intellig* gener* cont*; Detect* artificial intellig* gener* text; Detect* tools artificial intellig* gener* text; Best artificial intellig* tools for detect* artificial intellig* gener* text; Artificial intellig* tools for detect* artificial intellig* gener* text; Artificial intellig* detect* tools (IEEE Xplore Digital Library).

**Eligibility criteria and selection of peer-reviewed journal articles**

The eligibility criteria used to judge the suitability and relevance of the peer-reviewed journal articles for this study were based on the classical inclusion/exclusion format (see Chaka, 2022, 2023d; Ismail et al., 2023; Stracke et al., 2023). For example, the time-period inclusion criterion was peer-reviewed journal articles published between January 2023 and November 2023 (see Table 1). Eligible journal articles were determined through a search and screening process that was conducted on the fourteen aforementioned online search platforms during the specified coverage time frame. During this process, 7,515 articles were returned by the fourteen online search platforms (see Figure 1). Of these articles, 5,800 were duplicates and were removed while 1,450 did not meet the designated coverage time frame and were, also, accordingly, eliminated. The remaining articles (n = 265) were screened by reviewing their titles and abstracts. After this screening process, 200 articles were excluded as they did not focus on AI detection tools. A full-text review of the remaining 65 articles was conducted, after which 41 articles were eliminated due to their lack of focus on AI and human texts. Of the remaining 23 articles, 7 articles were excluded as they each used only one AI detection tool for distinguishing between AI-generated and human-written texts. This led to 16 articles being retained (see Figure 1).

Table 1: Inclusion/exclusion criteria.

| Criteria | Inclusion | Exclusion |
|---|---|---|
| Time period | Peer-reviewed journal articles published between January 2023 and November 2023 | Peer-reviewed journal articles not published between January 2023 and November 2023 |
| Types of articles | Articles published in peer-reviewed journals | Articles not published in peer-reviewed journals |
| Content and focus of articles | Peer-reviewed journal articles focusing on AI detection tools that can differentiate between AI-generated and human-written texts | Peer-reviewed journal articles not focusing on AI detection tools that can differentiate between AI-generated and human-written texts |
| Number of AI detection tools | More than one AI detection tool, one of which is meant to detect the currently available LLM chatbots* | Only one AI detection tool meant to detect the currently available LLM chatbots |
| Language of publication | Peer-reviewed journal articles published in English | Peer-reviewed journal articles not published in English |

* LLM chatbots released after ChatGPT or released to compete with it.

From the 16 retained articles, forward snowballing, and backward snowballing – also known as descendent and ancestry searches – were conducted to further identify suitable and eligible articles (see Chaka, 2022, 2023d; Wohlin et al., 2022). Forward snowballing entails searching and locating publications that cite the publications established during the search process; backward snowballing involves searching and locating publications listed in the reference lists of publications discovered during initial literature searches (see Chaka, 2022; Wohlin et al., 2022). The resultant dual snowballing search yielded one more relevant and eligible article. Overall, then, the total number of suitable and eligible articles for the present study was 17 (see Figure 1). The reviewing of the 17 articles was done manually.

**Quality evaluation, coding, and inter-rater reliability**

Evaluating and ensuring methodological quality is essential for review studies. This is so even when there is a scarcity of review studies in any given area of focus. There are quality assessment criteria recommended by scholars such as Kitchenham et al. (2009) and Shea et al. (2009). The present review study formulated and utilised an aggregated set of quality evaluation criteria adapted from A MeaSurement Tool to Assess systematic Reviews (AMSTAR) tool (Shea et al., 2009; Shea et al., 2017; also see Chaka, 2022; Li et al., 2022) and from the quality evaluation guidelines designed by Kitchenham et al. (2009) and Kitchenham and Brereton (2013). Based on these sixteen quality evaluation criteria, a checklist form was formulated (see Table 2). However, since this is not a systematic literature review, and as there was a dearth of peer-reviewed articles published in the focus area of this study, as mentioned earlier, the quality evaluation criteria used here are customised for this study, even though they have some universal applicability for review studies on AI detection tools. The application of the quality evaluation checklist was not rigid but flexible.

Concerning the reviewed articles, two raters (including the author of this article) independently evaluated each article using the checklist illustrated in Table 2. A "yes or "no" rating was allocated to each of the sixteen criteria for each

Table 2: Quality evaluation questions.

| Article quality evaluation questions |
|---|
| 1. Are the aims/purposes of the article clearly stated? |
| 2. Is the field of study (subject area) of the article mentioned? |
| 3. Are the genres of the AI and human texts tested provided? |
| 4. Is there a specific method/methodology specified? |
| 5. Is there more than one AI detection tool used? |
| 6. Are the names of the AI tools used to generate texts specified and are the protocol, procedure, prompts, re-prompting, revised prompts, etc., for generating AI texts mentioned? |
| 7. Is the nature of the humans used to generate texts specified mentioned? |
| 8. Is the justification for the choice of the AI tools to generate texts given? |
| 9. Are the names of the AI tools used to detect AI-generated and human-written texts specified? |
| 10. Is the justification for the choice of the AI tools used to detect AI-generated and human-written texts given? |
| 11. Are the AI-generated and human-written texts for detection by AI tools both sufficient and credible in terms of how they have been described? |
| 12. Is there transparency in the way the AI-generated and human-written texts were collected? |
| 13. Is there transparency in reporting the way in which the AI-generated and human-written texts were subjected to the AI detection tools used? |
| 14. Are the findings grounded on the data and credible? |
| 15. What contribution do the findings make to the existing knowledge or understanding? |
| 16. Are the conclusions based on the findings? |

article, with a "yes" rating allotted the number 1 (one) and a "no" rating assigned the number 0 (zero). The two raters' rating agreement scores were calculated following Cohen's kappa coefficient (κ) (see Cohen, 1960). Rating discrepancies between the two raters were resolved by discussing them and by reaching a consensus (Landis & Koch, 1977; Pérez et al., 2020). The inter-rater agreement was calculated using Landis and Koch's (1977) scoring and its related interpretation. The inter-rater agreement represents the extent of autonomy raters exhibit in scoring items by attempting to reach the same conclusion. Using Landis and Koch's (1977) κ values of <0 = poor, 0.00-0.20 = slight, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, and 0.81-1.00 = almost perfect, which are modifications of Cohen's (1960) original labels, the inter-rater agreement between the two raters was 0.82. As this joint agreement score falls within the 0.81-1.00 almost-perfect score range, it was deemed acceptable (also see Chaka, 2022, 2023d; McHugh, 2012).

### Data extraction and analysis

Based on the quality evaluation criteria, the coding procedure, and the inter-rater reliability described above, datasets were extracted from the peer-reviewed articles included in this study. These datasets were in the form of the twelve journal characteristics illustrated in Table 3. This table also served as an analytic scheme for thematic analysis that was conducted on the extracted datasets. Categories and themes that responded to the research questions for this study were developed from this analysis (see Chaka, 2022, 2023d).

Table 3: Twelve key journal characteristics investigated in each review study.

| | |
|---|---|
| Author(s) and year and month of publication | Number and names of AI detection tools |
| Country of origin | Best and worst performing AI detection tool(s) |
| Article type | Detection accuracy rate |
| Discipline/subject area | Detection accuracy reliability |
| Purpose | Main findings |
| AI and human texts tested | Key conclusions |

## Findings and discussion

The findings presented in this part of the paper are based on the datasets extracted from the 17 selected journal articles. They are presented according to the twelve journal characteristics and in line with the research questions (RQs) mentioned earlier. These findings are integrated with their discussion.

### Authors, countries of origin, article types, disciplines, and purposes

The 17 reviewed articles were produced by authors from twelve countries: India, the USA, Germany, Greece, France, South Africa, Australia, Hong Kong, Qatar, Croatia, Kuwait, and Malaysia. Three articles were written by authors from two countries: India and the USA. Two articles were produced by authors from France. The remaining articles (n = 9) were written by authors from nine different countries (see Table 4). At a geolocational vantage point, there is an infinitesimal difference between the number of articles contributed by countries deemed to represent the Global North and those by countries viewed to represent the Global North, notwithstanding a fractional edge the former block of countries have over the latter block in this review study. This geolocational and authorship distribution, which is often viewed as a proxy for the geopolitics and economy of knowledge production (see Chaka, 2023e; Müller, 2021; R'boul, 2022; also see Domínguez et al., 2023), seems not to resonate with the views and findings of Chaka (2023e), Müller (2021), and R'boul (2022), at least in the context of this study. While this does not invalidate or deny the views and findings of these scholars' studies, as their contexts and dynamics differ, the current study articulates one of the observations that emanates from it. Without denying the existence of the geopolitics of knowledge and of the geospatial entanglements of knowledge, this observation is instructive, though, since the study did not use *geopolitics and economy of knowledge production nor names of countries in its search strings*.

Table 4: Article numbers, types, authors, countries, texts tested, AI tools used, the best and worst performing AI tools.

| Worst performing tool | Turnitin | Open AI Classifier | DupliChecker | NOW | GLTR | Open AI Classifier | Open AI Classifier | NOW | Writer | iThenticate | NOW | Crossplag | Content at Scale | GPTZero & Quill | Ouriginal, iThenticate & DupliChecker | Sapling & Content Detector | GPT-2 Output Detector |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Best performing tool | OpenAI Classifier | Crossplag | Grammarly | NOW | Copyleaks | GPTZero | ZeroGPT | NOW | Crossplag | GPT-2 Output Detector | NOW | GPT-2 Output Detector | Writer | Originality | Content at Scale & Sapling | Copyleaks & Turnitin | Content at Scale |
| Number AI detection tools used | 4 | 4 | 4 | 2 | 5 | 3 | 2 | 3 | 5 | 3 | 3 | 2 | 3 | 11 | 8 | 16 | 3 |
| Texts tested | Two sets of 3 research papers (AI&H) | Four academic writing essays (AI&H) | 100 essays (AI) | 425 original & MT articles | Three sets of AI English texts (n = 21) | 150 essays (75 = AI; 75 = human) | 300 paragraphs (100 = human; 200 = AI) | A fabricated abstract | 15 paragraphs (AI) | 100 abstracts (50 = human; 50 AI) | 200 abstracts (172 = human; 28 = AI) | 240 essays (AI&H) | Four research articles (H&AI) | 8 texts (7 = AI; 1 = human) | 80 academic writing samples (AI) | 126 essays ( 42 = human; 84 = AI) | 16 essays (H, AI & MT) |
| Country | Germany | Greece | India | France | South Africa | Australia | Hong Kong | USA | Qatar | USA | Croatia | Kuwait | India | France | India | USA | Malaysia |
| Author(s) | Abani et al. (2023) | Alexander et al. (2023) | Anisi et al. (2023) | Bisi et al. (2023) | Chaka (2023) | Cingillioglu (2023) | Desaire (2023) | Elali & Rachid (2023) | Elkhatat et al. (2023) | Gao et al. (2023) | Homolak (2023) | Ibrahim (2023) | Ladha et al. (2023) | Odri & Yoon (2023) | Santra & Majhi (2023) | Walters (2023) | Wee & Reimer (2023) |
| Article number & type | Art. 1 (OP) | Art. 2 (RP) | Art. 3 (RP) | Art. 4 (RP) | Art. 5 (RP) | Art. 6 (RP) | Art. 7 (Report) | Art. 8 (OP) | Art. 9 (OA) | Art. 10 (BC) | Art. 11 (Comment.) | Art. 12 (RP) | Art. 13 (RP) | Art. 14 (RP) | Art. 15 (RP) | Art. 16 (RP) | Art. 17 (VP) |

**Notes**: NOW = no outright winner; AI&H = artificial intelligence-generated & human-written; MT = machine translated/translation; OP = opinion paper; RP = research paper; OA = original article; BC = brief communication; comment. = commentary; VP = viewpoint

**NB:** The early online versions of Bisi et al. (2023 and Odri and Yoon (2023) are the ones that were reviewed. Their latest versions were published in December after the study had been conducted. Except for their bibliographic details, the contents of these articles are the same in both versions.

The articles reviewed in this study were of different types: research papers (n = 11), opinion papers (n = 2), commentary (n = 1), report (n = 1), brief communication (n = 1), and viewpoint (n =1) (see Table 4). All of these articles were published or available online between 10 March 2023 and 15 November 2023, with three articles published in April and October, respectively (see Figure 2). Research articles, or original papers, predominated the other article types. This is an unexpected but not a surprising development since most AI-related scholarly papers, including scholarly papers on AI detection tools that can differentiate between AI-generated and human-written texts, were an instantaneous response to ChatGPT after its viral launch on 30 November 2022. This particular development tends to resemble, albeit for different

reasons and for dissimilar dynamics, the exponential growth in the number of scholarly papers and preprints that were published immediately after the outbreak of the COVID-19 pandemic. During this period, too, many commentaries, reports, and viewpoints were instantly published (see Chaka, 2020).

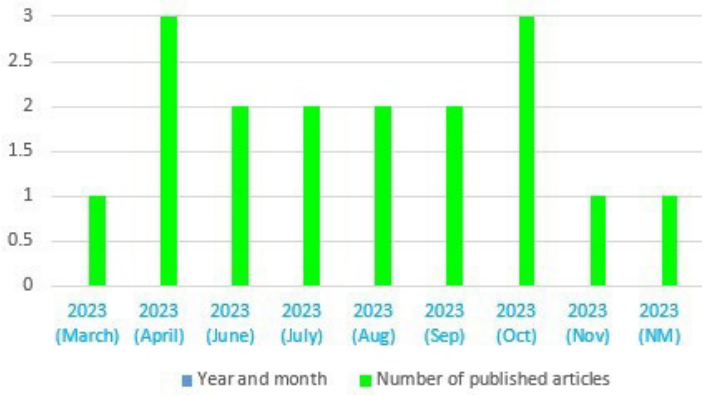


Figure 2: Number of published articles and years and months of publication.

The academic disciplines covered by the 17 reviewed articles are many and varied. Medical and biomedical sciences, together with hard sciences (e.g., chemistry), dominated, followed by English language studies (see Table 5). This observation should be seen against the backdrop of the disciplines in which ChatGPT and the other generative AI models seem to pose the biggest threat in terms of academic integrity. For medical and biomedical sciences and hard sciences, it is the integrity of scientific writing that generative AI models like ChatGPT threaten. Similarly, for English language studies, which have academic essay writing and composite studies as some of its flagship assessment methods, the emergence of generative AI models is not a transient fad: it is a big issue that goes to the heart of its existence. So, for all of these disciplines, testing or evaluating which detection tools can discriminate between AI-generated and human-written texts with the highest accuracy and the maximum reliability is a matter of life and death (see Kenwright, 2024; Uzun, 2023; cf. Lim et al., 2023).

Table 5: Articles, disciplines and purposes.

| Articles | Disciplines | Purposes |
|---|---|---|
| Art. 1 | Veterinary neurology | Exploring the possible advantages and limitations of ChatGPT in scientific writing related to veterinary neurology |
| Art. 2 | English as a Second Language (ESL) | Sharing information on the challenges confronted by English as a Second Language (ESL) lecturers in identifying AI-(ChatGPT-) generated texts. |
| Art. 3 | Ten/Miscellaneous fields of study: animal use, cosmetics and pharmaceutical industry, cosmology, engineering, environment, evolution of sports, finance, gender roles, medical technology, and chronic diseases. | Comparing overall similarity index (OSI) of four plagiarism detection tools and evaluating the factors affecting their effectiveness in detecting plagiarism. |
| Art. 4 | Orthopaedics and traumatology | Screening for AI-generated content in articles published in Orthopedics & Traumatology: Surgery & Research (OTSR) before and after the release of ChatGPT. |
| Art. 5 | Applied English language studies (AELS) | Testing 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, in detecting AI-generated content. |
| Art. 6 | Data-driven predictive policing | Discussing the method and strategies to maintain academic integrity in educational settings. |
| Art. 7 | Physical science (Chemistry) | Accurately detecting AI text when ChatGPT is told to write like a chemist. |
| Art. 8 | Orthopaedics and rheumatology | Determining how AI-generated chatbots can be used to fabricate research in the medical community; and testing the accuracy of free, online AI detectors. |
| Art. 9 | Chemical engineering/ Chemistry | Investigating the capabilities of various AI content detection tools in distinguishing between human and AI-authored content. |
| Art. 10 | Biomedical sciences | Determining if ChatGPT can write convincing medical research abstracts. |
| Art. 11 | Pharmacology | Examining the use of ChatGPT in academic writing and to evaluate the dependability of existing AI detectors. |
| Art. 12 | English as a Second Language (ESL) writing | Examining the potential of two AI-based classifiers to detect AI-assisted plagiarism in ESL composition classes. |
| Art. 13 | Scientific writing | Assessing the AI detection sites on a text generated wholly by the AI; testing the methods provided for evading AI detection. |
| Art. 14 | Orthopaedics and traumatology | Assisting publishers to identify AI-generated text in scientific research, academic work, and assignments as a means of regulating and promoting the ethical usage of AI in academia. |
| Art. 15 | Library and information science | Examining the capability and shortcomings of typical plagiarism detectors to identify machine-generated scholarly text. |
| Art. 16 | Social sciences, the natural sciences, and the humanities | Evaluating the accuracy of 16 AI text detectors in distinguishing between AI-generated and human-generated writing. |
| Art. 17 | Bioscience | Examining the detection limits of AI detection tools to identify human-written essays, ChatGPT-generated essays, and partially AI-aided essays by checking AI procedurally generated essays in English and three other languages (Malay, Mandarin, Japanese, and their AI-translated versions). |

Note. For full versions of some of these purposes, see Appendix A.

Moreover, the 17 articles had their specific purposes. While these purposes appear to be many and divergent, the convergence point is examining, evaluating, assessing, or testing the capabilities, potential, or accuracy and shortcomings or limits of AI detection tools in identifying AI-generated texts or in distinguishing between AI-generated and human-written texts in varying degrees (see Table 5). Two articles' (Art. 6 and Art. 10) purposes are generic. However, the purpose of Art. 6 is to preserve academic integrity by utilising AI detection tools in higher education. All of these purposes are about the detection of and the differentiation between AI-generated and human-written content as mediated largely by the AI detection tools utilised by the respective articles. Elsewhere, one of the purposes of Maddugoda's (2023) paper, which has some resonance with the convergence points of the purposes of the 17 articles, was to assess the efficacy of traditional anti-plagiarism software tools against some of the current AI detectors in identifying AI-generated content.

## The AI-generated and human-written texts tested

Twelve of the 17 articles utilised both AI-generated and human-written texts, with ChatGPT as a common text generator in all of them. Some of them had varying versions of ChatGPT-generated texts, such as original, fabricated, slightly modified, paraphrased or translated versions. Five articles employed only AI-generated texts. Of these, three articles employed ChatGPT as their preferred AI text generator (see Art. 3, Art. 8, and Art. 15). One article (Art. 9) used both ChatGPT-3.5 and ChatGPT-4, while another (Art. 5) utilised ChatGPT, YouChat, and Chatsonic. In their paper, Wu et al. (2023) provide large language model- (LLM) generated and human-written datasets that can be used as test datasets for detecting LLM-generated and human-written texts. Among these datasets are ChatGPT- or AI-generated datasets that can serve as the basis for comparing ChatGPT-generated text with human-written text. Likewise, Weber-Wulff et al.'s (2023) paper compared AI detection tools that could reliably distinguish between ChatGPT-generated texts and human-written texts. In these two studies, as is the case with the current study, ChatGPT-generated text serves as one of the pieces of AI-generated text.

## Number and names of the AI detection tools used

The number of AI detection tools employed by the 17 reviewed articles ranged from two to sixteen. Six articles (Art. 6, Art. 8, Art. 11, Art. 13, Art. 17, Art. 20) each used three AI tools, while three articles utilised two AI detection tools and four AI detection tools, respectively. Only two articles employed five AI tools. The remaining articles tested eight, eleven, and sixteen AI detection tools apiece. In this case, articles that employed three AI tools predominated. A paper that compared three AI tools is Singh's (2023), whereas Weber-Wulff et al.'s (2023) paper tested fourteen AI detectors.

**Best- and worst-performing AI detection tools reported, and the detection accuracy rate and the detection accuracy reliability reported**

Concerning the best-performing AI detection tools, OpenAI Text Classifier, Crossplag, Grammarly, Copyleaks, for Art. 1, Art. 2, Art. 3, and Art. 5, respectively, had a better detection accuracy than their counterparts. The same is the case for Originality and Crossplag, Content at Scale and Sapling, Copyleaks and Turnitin, and Content at Scale in Art. 14, Art 15, Art. 16, and Art. 17, correspondingly. With regard to Art. 6, GPTZero's detection accuracy was fractionally better than that of Copyleaks, while Crossplag had a marginal advantage over the other four detection tools in terms of detection accuracy in Art. 9. What is noteworthy is that in the case where eleven AI detection tools were tested, Originality and Crossplag did fairly better than the other nine tools. And, where sixteen AI detection tools were evaluated, Copyleaks and Turnitin had a higher detection accuracy than the other fourteen detectors. At a simple numerical level, Crossplag can be regarded as the best-performing AI detection tool as it topped or as it was one of the top-performing tools in at least three of the 17 reviewed articles (see Art. 2, Art. 9, and Art. 14). It is followed by Copyleaks that topped and co-topped in Art. 5 and Art. 16, respectively.

Concerning the other reviewed articles, the AI detection tools they tested either had a low detection accuracy (see Art. 4, Art. 5, Art. 7, Art. 8, and Art. 10), or displayed inconsistencies in their detection accuracy (see Art. 9, Art. 11, Art. 12, and Art. 13). Two AI detection tools that had tended to perform badly in the two instances (articles) in which had been used are Duplichecker (Art. 3 and Art. 15) and Writer (Art. 5 and Art. 9).

However, a word of caution is needed here. Notwithstanding the fact that some of the aforesaid AI detection tools did better than their counterparts as indicated above, they, nevertheless, fared badly in the other instances in which they were tested in some of the reviewed articles. For instance, the following AI detectors did badly in the reviewed articles indicated in parentheses: OpenAI Text Classifier (Art. 2, Art. 6 & Art. 7); Crossplag (Art. 12); Content at Scale (Art. 13); Sapling (Art. 16); and GPTZero (Art. 14). This points to some inconsistencies in these detection tools' accuracy when it comes to differentiating between AI-generated and human-created texts. Elkhatat et al. (2023) highlight this inconsistency bluntly when referring to the five AI detection tools they tested (see Art. 9) by opining that their performance was not completely reliable. This is because the AI detection tools they tested were inconsistent: they correctly identified some of the content of control responses (human-created texts) as having not been AI-generated while simultaneously displaying false positives and undecided classifications for the other portions of the same content. In fact, Wu et al. (2023) contend that none of the current state-of-the-art AI detection tools is infallible. In particular, the detection efficacy of AI detectors gets reduced by adversarial attacks, which are techniques or attempts to deliberately modify, fabricate, or manipulate text that goes beyond simple prompts (see Sayeed, 2023). For example, AI detectors are eluded by tampering with punctuation marks (e.g., removing a comma) in a text, and by applying synonym substitution, paraphrasing/rewording, and translating a text (Wu et al., 2023; also see Krishna et al., 2023). In addition, they can be tricked by instances of single spacing (Cai & Cui, 2023; also see Chaka, 2023c). Moreover, most of the current AI detectors do not perform well in multilingual texts due to their monolingual AI detection algorithms (see Chaka, 2023c; Wu et al., 2023).

So, if reliability is construed to refer to any AI detection tool's capability to consistently detect AI-generated text with 100% precision (with no false positives) and human-written text with 100% precision (with no false negatives) across all contexts of writing, then, all reviewed AI detectors in this study cannot be regarded as reliable as none of them met this reliability requirement. Most crucially, because of their varying degrees of inconsistency in their detection efficacy as pinpointed in the preceding paragraph, all of them were highly unreliable. This aspect, again, brings into sharp focus Wu et al.'s (2023) view that the currently available AI detectors are fallible. This view resonates with Chaka's (2023c) contention that most of the current AI detection tools are not yet fully ready to convincingly and accurately detect AI-generated content from machine-generated texts in different domains. Actually, Sayeed (2023) goes on to assert that detecting AI-generated text in a reliable way is increasingly becoming mathematically impossible for the current AI detection tools. Given the findings of the present review study, I am strongly persuaded to concur with this contention. While on this issue of AI detection unreliability and inaccuracies, it is worth mentioning that OpenAI, the company behind ChatGPT-3.5 and GPT-4, is reported to have quietly discontinued its own AI detection tool, OpenAI Text Classifier, due to its detection unreliability and inaccuracies. It is reportedly mulling over bringing a better version of its AI detection tool (see Dreibelbis, 2023) back to business.

**Main findings and key conclusions**

Some of the main findings of the reviewed articles touted the opportunities– potential solutions – offered by LLMs like ChatGPT, while flagging the challenges or threats posed by LLMs, especially in the area of academic and scientific writing. The opportunities relate to how such LLMs can benefit non-native English speakers in enhancing their academic and scientific writing (see Art. 2 and Art. 17). However, the catch is the plagiarism and the scientific dishonesty that LLMs encourage for academic and scientific writing (see Art. 2, Art. 13, Art. 15, and Art. 17). This set of main findings reflects how LLMs are double-edged or Janus-faced AI tools, at least for now. This is not a new observation, though. Well before the advent of ChatGTP, a paper by Sumakul et al. (2022) explored whether AI was a friend or a foe in English in foreign language (EFL) classrooms. After the release of ChatGPT, many papers have been published highlighting the benefits and challenges of ChatGPT in higher education. One such paper is Rasul et al.'s (2023). The other set of main findings concerns the inconsistencies of the AI detection tools tested in accurately and reliably distinguishing between AI-generated and human-written text. More than half of the reviewed articles reported on the inconsistencies of the AI detection tools they tested in their main findings (see Art. 1, Art 3, Art. 4, Art. 5, Art. 9, Art. 10, Art. 11, Art. 13,

Art 14, and Art. 17).

The detection inconsistencies of the AI detectors used in the reviewed articles have been dealt with and contextualised in the preceding section. Suffice it to say that one article (Art. 15) had as part of its main findings the fact that traditional anti-plagiarism tools (e.g., Turnitin, Grammarly, iThenticate) lack the ability to detect AI-generated text due to the differences in syntax and structure between machine-generated and human-written text. Dalalah and Dalalah (2023) take this shortcoming a step further by pointing out that discriminating between AI-generated text and simply copied text is rather difficult as AI detection algorithms are merely configured to detect whether a given text is machine-generated or not. A rider needs to be added to this point. AI detectors can only determine whether a text is AI-generated or not: they cannot establish the originality of a text even if it is copied. Doing so is the province of anti-plagiarism detection tools such as Turnitin, Grammarly, and iThenticate. The irony of anti-plagiarism detection tools, however, is that they do not necessarily detect plagiarism, but, rather, similarity indices. Added to this is the finding of Art. 15, which seems to loom large over them. Differentiating between and detecting plagiarised text and copied text, in addition to differentiating between and detecting AI-generated text and human-written text, is likely to become an even murkier minefield for AI detection tools as Microsoft's generative AI assistant, Microsoft 365 Copilot, is ready to be integrated into Microsoft 365 apps such as Word, Outlook, Teams, Excel, and PowerPoint. A similar generative AI assistant is likely to be integrated into the Google suite comprising Gmail, Docs, Slides, and Forms by Google (see Finnegan, 2023). While this generative AI integration might be beneficial for text predicting and for automating writing (e.g., drafting emails and creating slideshows), its downside is its potential to make up facts (hallucinate) and to spew inaccurate and false information (see Finnegan, 2023). All of this, then, adds another layer of AI-generated writing that AI detection tools will need to contend with in addition to simply differentiating between AI-generated and human-written texts.

Pertaining to the key conclusions, one set flagged the fact that the detection capability of most AI detection tools is largely confined to English (see Art. 4, Art. 5, Art. 9, Art. 12, Art. 17). The inability of some of the current AI detectors to function in texts written in other languages than English (including major European languages) is raised by, among others, Chaka (2023c) and Wu et al. (2023). For instance, Wu et al. (2023) argue that the main current AI detectors are designed to detect pieces of LLM-generated text meant for monolingual, and not multilingual, applications (also see Wang et al., 2023). Another key conclusion reported in this study is the need to use more than one AI detection tool, while another key conclusion is that AI detection tools need to be complemented by human reviewers. To add to these two points, in the unfolding environment of rapidly increasing AI text-generation tools and their attendant refinement, I think there is a need to employ a set of AI detection tools comprising traditional anti-plagiarism detection tools and AI detectors, on the one hand, and to enlist human reviewers/raters, on the other hand, for purposes of distinguishing between AI-generated text and human-written text.

Finally, the other key conclusions are about the need for more development and refinement of AI content detection tools, the necessity to provide digital literacy training for teachers/human raters, and the need for journals to review their existing evaluation policies and practices in the light of AI. All of this calls for doing things differently across all domains, especially in academia, in the era of LLMs like ChatGPT.

## Conclusion

This study set out to review 17 articles published between January 2023 and November 2023 that dealt with the performance of AI detectors in differentiating between AI-generated and human-written texts. It was guided by six research questions (RQs). Authors from twelve countries wrote the reviewed articles. Viewed within its context, the geolocational and authorship dispersion of these articles tend not to dovetail with the geopolitics and economy of knowledge production as advanced by scholars such as Chaka (2023e), Müller (2021), and R'boul (2022). While the reviewed articles were of diverse types, the predominant article types were research papers, a finding that suggests that within less than a year after the release of ChatGPT, there were already studies conducted on AI detection tools that could distinguish between AI-generated and human-written texts. Among the academic disciplines explored, medical and biomedical sciences, together with hard sciences, dominated. They were followed by English language studies.

Even though the purposes of the 17 articles were many and varied, they converged in terms of examining, evaluating, assessing, or testing the capabilities, potential, or accuracy and shortcomings or limits of AI detection tools in identifying AI-generated texts or in differentiating between AI-generated and human-written texts in different contexts. The types of texts evaluated by these articles were AI-generated and human-written texts or AI-generated texts. In these two sets of texts (the former and latter text sets), ChatGPT (in its two versions, GPT-3.5 and GPT-4) was the sole AI text generator used or was one of the AI text generators in instances where more than one AI text generator had been used. The lowest number of AI detection tools was two, whereas the highest number of AI detection tools was sixteen. The names of the AI detectors used are displayed in Table 4.

In relation to the best-performing AI detection tools, Crossplag topped the other AI detectors in the three articles (Art. 2, Art. 9, and Art. 14) in which it had been tested. Copyleaks did so in two articles (Art. 5 and Art. 16). This finding should be seen in its context – the context of the 17 reviewed articles in this study as different AI detection tools tend to be prone to inconsistencies in the different contexts in which they are tested. Regarding the worst-performing AI detection tools, both Duplichecker and Writer fared badly in the articles in which they had been tested. However, the same caveat provided for the best-performing AI detectors above applies to them as well.

Lastly, one major aspect flagged by the main findings of the 17 reviewed articles is the inconsistency of the detection efficacy of all the tested AI detectors and all the tested anti-plagiarism detection tools. To this end, both sets of AI detection tools lacked detection reliability. Owing to this AI detection deficiency and the AI detection unreliability, the current study recommends employing both contemporary AI detectors and traditional anti-plagiarism detection tools, together with human reviewers/raters, in the pursuit of differentiating between AI-generated and human-written texts.

## References


Abani, S., Volk, H. A., De Decker, S., Fenn, J., Rusbridge, C., Charalambous, M., … Nessler J. N. (2023). ChatGPT and scientific papers in veterinary neurology; Is the genie out of the bottle? *Frontiers in Veterinary Science, 10*(1272755), 1-7. https://doi.org/10.3389/fvets.2023.1272755

Abbas, N., Ali, I., Manzoor, R., Hussain, T., & Hussaini, M. H. A. (2023). Role of artificial intelligence tools in enhancing students' educational performance at higher levels. *Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 3*(5), 36-49. https://doi.org/10.55529/jaimlnn.35.36.49

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology, 23*(20), 25-43. https://doi.org/10.56297/BUKA4060/XHLD5365

Anil, A., Saravanan, A., Singh, S., Shamim, M. A., Tiwari, K., Lal, H., …Sah, R. (2023). Are paid tools worth the cost? A prospective cross-over study to find the right tool for plagiarism detection. *Heliyon, 9*(9), e19194, 1-11. https://doi.org/10.1016/j.heliyon.2023.e19194

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&type=2

Bacon, C. K. (2017). Multilanguage, multipurpose: A literature review, synthesis, and framework for critical literacies in English language teaching. *Journal of Literacy Research, 49*(3), 424-453. https://doi.org /10.1177/1086296X17718324

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52-62.

Bisi, T., Risser, A., Clavert, P., Migaud, H., & Dartus, J. (2023). What is the rate of text generated by artificial intelligence over a year of publication in orthopedics and traumatology: Surgery and research? Analysis of 425 articles before versus after the launch of ChatGPT in November 2022. *Orthopaedics and Traumatology: Surgery and Research, 109*(8), 103694. https://doi.org/10.1016/j.otsr.2023.103694

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

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*. https://doi.org/10.21203/rs.3.rs-61482/v1

Chaka, C. (2022). Is Education 4.0 a sufficient innovative, and disruptive educational trend to promote sustainable open education for higher education institutions? A review of literature trends. *Frontiers in Education, 7*(824976), 1-13. https://doi.org/10.3389/feduc.2022.824976

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*, 1150499, 1-10. https://doi.org/10.3389/feduc.2023.1150499

Chaka, C. (2023c). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching, 6*(2), 94-104. https://doi.org/10.37074/jalt.2023.6.2.12

Chaka, C. (2023d). Fourth industrial revolution—A review of applications, prospects, and challenges for artificial intelligence, robotics and blockchain in higher education. *Research and Practice in Technology Enhanced Learning, 18*(2), 1-39. https://doi.org/10.58459/rptel.2023.18002

Chaka, C. (2023e). The geopolitics of knowledge production in applied English language studies: Transknowledging and a two-eyed critical southern decoloniality. *Journal of Contemporary Issues in Education, 18*(1), 3-20. https://doi.org/10.20355/jcie29507

Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *The International Journal of Information and Learning Technology, 40*(3), 259-268. https://doi.org/10.1108/IJILT-03-2023-0043

Cohen, J. (1960). A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

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

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. The *International Journal of Management Education, 21*(100822),

1-13. https://doi.org/10.1016/j.ijme.2023.100822

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport, 40*(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623

Desaire, H. A., Chua, A. E., Isom, M., Jarosova, 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), 1-2. https://doi.org/10.1016/j.xcrp.2023.101426

Domínguez, G. E., Ramírez-March, A., & Montenegro, M. (2023). The geopolitics of knowledge production, or how to inhabit a contradiction: Introduction to the special issue on the narrative productions methodology. *Qualitative Research in Psychology, 20*(40), 525-541. https://doi.org/10.1080/14780887.2023.2255104

Dreibelbis, E. (2023). *OpenAI quietly shuts down AI text-detection tool over inaccuracies*. https://www.pcmag.com/news/openai-quietly-shuts-down-ai-text-detection-tool-over-inaccuracies

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., .... Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 1-63. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns, 4*, 1-4. https://doi.org/10.1016/j.patter.2023.100706

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity, 19*(17), 1-16. https://doi.org/10.1007/s40979-023-00140-5

Finnegan, M. (2023). *M365 Copilot, Microsoft's generative AI tool, explained.* https://www.computerworld.com/article/3700709/m365-copilot-microsofts-generative-ai-tool-explained.html

Gamage, K. A. A., Dehideniya, S. C. P., Xu, Z., & Tang, X. (2023). ChatGPT and higher education assessments: More opportunities than concerns? *Journal of Applied Learning and Teaching, 6*(2), 358-369. https://doi.org/10.37074/jalt.2023.6.2.32

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. *Npj Digital Medicine, 6*(75). 1-5. https://doi.org/10.1038/s41746-023-00819-6

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*, 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Habibzadeh, F. (2023). GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study. *Journal of Korean Medical Sciences, 38*(38), e319. https://doi.org/10.3346/jkms.2023.38.e319

Hew, K. F., Huang, W., Du, J., & Jia, C. (2023). Using chatbots to support student goal setting and social presence in fully online activities: Learner engagement and perceptions. *Journal of Computing in Higher Education, 35*(1), 40-68. https://doi.org/10.1007/s12528-022-09338-x

Homolak, J. (2023). Exploring the adoption of ChatGPT in academic publishing: Insights and lessons for scientific writing. *Croatian Medical Journal, 64,* 205-207. https://doi.org/10.3325/cmj.2023.64.205

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), 1-28. https://doi.org/10.1186/s40468-023-00260-2

Ifelebuegu, A. (2023). Rethinking online assessment strategies: Authenticity versus AI chatbot intervention. *Journal of Applied Learning and Teaching, 6*(2), 385-392. https://doi.org/10.37074/jalt.2023.6.2.2

Ifelebuegu, A. O., Kulume, P., & Cherukut, P. (2023). Chatbots and AI in Education (AIEd) tools: The good, the bad, and the ugly. *Journal of Applied Learning & Teaching, 6*(2), 332-345. https://doi.org/10.37074/jalt.2023.6.2.29

Ismail, F., Tan, E., Rudolph, J., Crawford, J., & Tan, S. (2023). Artificial intelligence in higher education. A protocol paper for a systematic literature review. *Journal of Applied Learning & Teaching, 6*(2), 56-63. https://doi.org/10.37074/jalt.2023.6.2.34

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28,* 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kenwright, B. (2024). Is it the end of undergraduate dissertations?: Exploring the advantages and challenges of generative AI models in education. In S. Hai-Jew (Ed.), *Generative AI in teaching and learning* (pp. 46-65). IGI Global. https://doi.org/10.4018/979-8-3693-0074-9.ch003

Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology, 55*, 2049-2075. http://dx.doi.org/10.1016/j.infsof.2013.07.010

Kitchenham, B., Brereton, P. O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology, 51*, 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Kleebayoon, A., & Wiwanitkit, V. (2023). Artificial intelligence, chatbots, plagiarism and basic honesty: Comment. *Cellular and Molecular Bioengineering, 16*(2), 173-174. https://doi.org/10.1007/s12195-023-00759-x

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*(7), 5614. https://doi.org/10.3390/su15075614

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). *Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense*. https://www.semanticscholar.org/reader/1c13af186d1e177b85ef1ec3fc7b8d33ec314cfd

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-74.

Li, L., Asemota, I., Liu, B., Gomez-Valencia, J., Lin, L., Arif, A. W., … Usman, M. S. (2022). AMSTAR 2 appraisal of systematic reviews and meta-analyses in the field of heart failure from high-impact journals. *BioMed Central (BMC), 11*(147), 1-8. https://doi.org/10.1186/s13643-022-02029-9

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education, 21*(2), 100790, 1-13. https://doi.org/10.1016/j.ijme.2023.100790

Maddugoda, C. (2023). *A comprehensive review: Detection techniques for human-generated and AI-generated texts.* https://www.researchgate.net/publication/374542625_A_Comprehensive_Review_Detection_Techniques_for_Human-Generated_and_AI-Generated_Texts

McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Müller, M. (2021). Worlding geography: From linguistic privilege to decolonial anywhere. *Progress in Human Geography, 45*(6), 1440-1466. https://doi.org/10.1177/0309132520979356

Nah, F. F-H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277-304. https://doi.org/10.1080/15228053.2023.2233814

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

Perera, P., & Lankathilaka, M. (2023). AI in higher education: A literature review of ChatGPT and guidelines for responsible implementation. *International Journal of Research and Innovation in Social Science (IJRISS), vii*(vi), 306-314. https://doi.org/10.47772/IJRISS

Pérez, J., Díaz, J., Garcia-Martin, J., & Tabuenca, B. (2020). Systematic literature reviews in software engineering—Enhancement of the study selection process using Cohen's Kappa statistic. *The Journal of Systems & Software, 168*, 110657, 1-12. https://doi.org/10.1016/j.jss.2020.110657

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, 1-25. https://doi.org/10.1007/s10805-023-09492-6

Pinzolits, R. F. J. (2023). AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities, 4*, 37-50. https://doi.org/10.53880/2744-2373.2023.4.37

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

R'boul, H. (2022). Epistemological plurality in intercultural communication knowledge. *Journal of Multicultural Discourses, 17*(2), 173-188. https://doi.org/10.1080/17447143.2022.2069784

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3,* 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Santra, P. P., & Majhi, D. (2023). Scholarly communication and machine-generated text: Is it finally AI vs AI in plagiarism detection? *Journal of Information and Knowledge, 60*(3), 175-183. https://doi.org/10.17821/srels/2023/v60i3/171028

Sarzaeim, P., Doshi, A. M., & Mahmoud, Q. H. (2023). *A*

*framework for detecting AI-generated text in research publications*. https://proceedings.icatsconf.org/conf/index.php/ICAT/article/download/36/21/117

Sayeed, A. M-U. (2023). *Reliably detecting AI-generated text is mathematically impossible*. https://www.linkedin.com/pulse/reliably-detecting-ai-generated-text-mathematically-sayeed

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., … Boers, M. (2009). AMSTAR is a reliable and valid measurement to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62*, 1013-1020. https://doi.org/10.1016/j.jclinepi.2008.10.009

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., … Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal, 358*, 1-9. http://dx.doi.org/10.1136/bmj.j4008

Singh, A. (2023). *A comparison study on AI language detector*. https://doi.org/10.1109/CCWC57344.2023.10099219.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104,* 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stracke, C. M., Chounta, I-A., Holmes, W., Tlili, A., & Bozkurt, A. (2023). A standardised PRISMA-based protocol for systematic reviews of the scientific literature on Artificial Intelligence and Education (AI&ED). *Journal of Applied Learning and Teaching, 6*(2), 64-70. https://doi.org/10.37074/jalt.2023.6.2.38

Subramaniam, R. (2023). Identifying text classification failures in multilingual AI-generated content. *International Journal of Artificial Intelligence and Applications (IJAIA), 14*(5), 57-63. https://doi.org/10.5121/ijaia.2023.14505

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://doi.org/10.37074/jalt.2023.6.1.17

Sumakul, D. T. Y. G., Hamied, F. A., & Sukyadi, D. (2022). Artificial intelligence in EFL classrooms: Friend or foe? *LEARN Journal: Language Education and Acquisition Research Network, 15*(1), 232-256.

Thurzo, A., Strunga, M., Urban, R., Surovková, J., & Afrashtehfar, K. I. (2023). Impact of artificial intelligence on dental education: A review and guide for curriculum. *Education Sciences, 13*(150), 1-15. https://doi.org/10.3390/educsci13020150

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

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224-226.

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*(20220158), 1-24.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., … Nakov, P. (2023). *M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection*. https://arxiv.org/pdf/2305.14902.pdf

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

Wee, H. B., & Reimer, J. D. (2023). Non-English academics face inequality via AI-generated essays and countermeasure tools. *BioScience, 73,* 476-478. https://doi.org/10.1093/biosci/biad034

Wohlin, C., Kalinowski, M., Felizardo, K. R., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology, 147*(106908), 1-12. https://doi.org/10.1016/j.infsof.2022.106908

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). *A survey on LLM-generated text detection: Necessity, methods, and future directions*. https://arxiv.org/pdf/2310.14724.pdf

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93-112. https://doi.org/10.1177/0739456X17723971

Yang, J., Chen, Y. L., Por, L. Y., & Ku, C. S. (2023). A systematic literature review of information security in chatbots. *Applied Sciences, 13*(11), 6355. https://doi.org/10.3390/app13116355

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.