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

Accuracy pecking order – How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays

Chaka Chaka[A]

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

Keywords

Accuracy;
AI;
AI detectors;
AI-generated and human-written content;
artificial intelligence;
English L1 and English L2;
false positive rates;
student essays;
true negative rates.

Correspondence

chakachaka8@gmail.com [A]

Article Info

Received 9 March 2024
Received in revised form 28 March 2024
Accepted 3 April 2024
Available online 10 April 2024

DOI: https://doi.org/10.37074/jalt.2024.7.1.33

Abstract

This study set out to evaluate the accuracy of 30 AI detectors in identifying generative artificial intelligence (GenAI)-generated and human-written content in university English L1 and English L2 student essays. 40 student essays were divided into four essay sets of English L1 and English L2 and two undergraduate modules: a second-year module and a third-year module. There are ten essays in each essay set. The 30 AI detectors comprised freely available detectors and non-premium versions of online AI detectors. Employing a critical studies approach to artificial intelligence, the study had three research questions. It focused on and calculated the accuracy, false positive rates (FPRs), and true negative rates (TNRs) of all 30 AI detectors for all essays in each of the four sets to determine the accuracy of each AI detector to identify the GenAI content of each essay. It also used confusion matrices to determine the specificity of best- and worst-performing AI detectors. Some of the results of this study are worth mentioning. Firstly, only two AI detectors, Copyleaks and Undetectable AI, managed to correctly detect all of the essay sets of the two English language categories (English L1 and English L2) as human written. As a result, these two AI detectors jointly shared the first spot in terms of the GenAI detection accuracy ranking. Secondly, nine of the 30 AI detectors completely misidentified all the essays in each of the four essay sets of the two language categories in both modules. Thus, they collectively shared the last spot. Thirdly, the remaining 19 AI detectors both correctly and incorrectly classified the four essay sets in varying degrees without any bias to any essay set of the two English language categories. Fourthly, none of the 30 AI detectors tended to have a bias toward a specific English language category in classifying the four essay sets. Lastly, the results of the current study suggest that the bulk of the currently available AI detectors, especially the currently available free-to-use AI detectors, are not fit for purpose.

Introduction

In academia, plagiarism and generative artificial intelligence (GenAI)-generated content are two different things. For instance, a student does not need a GenAI tool to plagiarise, but they need a GenAI tool to generate GenAI content. Notably, plagiarism predates the advent of GenAI content generation, especially as the latter is heralded by GenAI language models such as ChatGPT. As such, the possibility of plagiarism is always there with or without the use of GenAI tools, but GenAI-generated content is almost impossible to generate without using GenAI tools such as ChatGPT as its catalysts. With the launch of ChatGPT and the other related GenAI-powered chatbots, the quest for detecting GenAI-generated content in university student writing, in particular, has become unavoidable. What is even more pressing is the quest for differentiating between GenAI-generated and human-written content in student writing in higher education (HE). In the HE arena, universities and academics have always prided themselves in being the guardians and protectors of original and authentic academic writing in all disciplines. This guardianship and protectorship has often come under the banner of academic integrity (see Anthology White Paper, 2023; Blau et al., 2020; Gamage et al., 2020; Perkins, 2023; Sullivan et al., 2023; Uzun, 2023). It is no exaggeration to assert that academic integrity, guardianship and protectorship in HE almost borders on a frenzy due to, mainly, though not exclusively, pressure points brought by GenAI-powered chatbots like ChatGPT. In this frenzied scrambling, GenAI-generated content and plagiarism feature as proxies for academic dishonesty.

However, viewing academic integrity through the prism of its nemesis, like academic dishonesty that comprises GenAI-generated content and plagiarism, is simplistic and superficial. This conception of academic integrity has to do with the practice of text- or content-matching that chimes with plagiarism-detection software programmes in which plagiarism and GenAI-generated content, are deemed a twin threat to academic integrity (cf. Blau et al., 2020; Gamage et al., 2020; Ifelebuegu, 2023; Rudolph et al., 2023; Sobaih, 2024). As Gamage et al. (2020) contend, this view of academic integrity overlooks other elements of academic dishonesty or other violations of academic integrity (see Blau et al., 2020). In addition to GenAI-generated content and plagiarism, examples of elements of academic dishonesty or violations of academic integrity include fraudulence, falsification, fabrication, facilitation, cheating, ghost-writing (Blau et al., 2020), contract cheating, and collusion (Gamage et al., 2020). Of course, some of these elements or violations may overlap: fraudulence with falsification and fabrication, ghost-writing with contract cheating, and facilitation with collusion (cf. Blau et al., 2020; Gamage et al., 2020). Additionally, both cheating and fraudulence can be used as overarching terms for academic dishonesty. Therefore, reducing academic dishonesty to GenAI-generated content and plagiarism alone tends to obscure its other facets, such as the ones furnished here.

With the surge of GenAI-generated content and plagiarism being a threat to academic integrity in HE, several AI content detectors have been released, while existing traditional plagiarism detection tools have upgraded their offerings to include AI content detection features (see Anil et al., 2023; Chaka, 2023a, 2024; Bisi et al., 2023; Dergaa et al., 2023; Ladha et al., 2023; Uzun, 2023; Wiggers, 2023; Weber-Wulff et al., 2023). The cardinal function of AI content detectors is to do exactly what they are designed to do: detect GenAI-generated content in different types of academic and scholarly writing. To this effect, there have been studies that have tested the effectiveness or reliability of AI content detectors in detecting GenAI-generated content in academic writing, or in distinguishing between GenAI-generated and human-written content in academic writing. These studies have tested different types of AI content detectors that include single AI content detectors (see Habibzadeh, 2023; Perkins et al., 2024; Subramaniam, 2023), two AI content detectors (see Bisi et al., 2023; Desaire et al., 2023; Ibrahim, 2023), three AI content detectors (see Cingillioglu, 2023; Elali & Rachid, 2023; Gao et al., 2023; Homolak, 2023; Ladha et al., 2023; Wee & Reimer, 2023), four AI content detectors (Abani et al., 2023; Alexander et al., 2023; Anil et al., 2023), and multiple AI content detectors (Chaka, 2023a; Odri & Yoon, 2023; Santra & Majhi, 2023; Walters, 2023) (see Chaka, 2024).

Most crucially, there is one study that has discovered that AI detectors tend to be biased against non-English language speakers (Liang et al., 2023; Mathewson, 2023; Shane, 2023; cf. Adamson, 2023; Gillham, 2024). This finding resonates, in a different but related scenario, with the view that some studies have established that currently available automatic speech recognition technologies poorly detect, if any, and discriminate against the English spoken by Black people, especially African American Language (AAL), thereby exposing their racial bias and demographic discrimination against this type of English (Martin & Wright, 2023). Linguistic and racial biases are but two of the instances of bias that GenAI models, and not just AI detection models, have to contend with in their everyday deployment. Other instances of bias GenAI models have to grapple with are cultural, ideological, political, temporal, and confirmation biases (see Ferrara, 2023). Thus, in addition to simply detecting GenAI-generated content, or distinguishing it from its human-written counterpart, these biases are some of the pressing challenges that these models have to wrestle with on an ongoing basis.

Against this background, the current study set out to:

- evaluate the accuracy of 30 AI detectors in differentiating between GenAI-generated and human-written content in university English L1 and English L2 student essays for two different undergraduate modules;
- establish whether these 30 AI detectors will classify these four sets of student essays differentially based on their English L1 and English L2 categories; and
- discover which language category within these four sets of student essays is assigned more false positives.

On this basis, the overarching purpose of this study is to contribute to the ongoing debate about the effectiveness (accuracy, precision, and reliability) of AI content detectors in distinguishing between GenAI-generated and human-written content in the essays produced by English L1 and English L2 students. The student essays in this study were written by English L1 and English L2 students who registered for a second-year undergraduate module and a third-year undergraduate module offered by an English department at a university in South Africa in 2018, 2020, and 2022.

Given the points highlighted above, this study seeks to answer the following research questions (RQs):

- RQ1: What is the accuracy of the 30 AI detectors in differentiating between GenAI-generated and human-written content in university English L1 and English L2 student essays for two different undergraduate modules?

- RQ2: Do these 30 AI detectors classify these four sets of student essays differentially based on their English L1 and English L2 categories or not?

- RQ3: Which language category within these four sets of student essays is assigned more false positives by these AI detectors?

Critical studies approach to AI

In a surreal world, AI, algorithms, and machine learning would be devoid of any bias: racial, demographic, gender, sexuality, disability, and training data bias (see Lindgren, 2023; also see AIContentfy team, 2023; Chaka, 2022; Ferrara, 2023; Wu et al., 2023). In real-world contexts, though, that is not the case. This rings true for AI detectors. Their efficacy is largely determined by, among other things, their training data, their algorithms, and their computing prowess (AIContentfy team, 2023). All of this, together with the types of bias mentioned and those stated earlier, leads to AI detectors having shortcomings and deficiencies. As such, they end up not being as effective and efficient as they are made out to be or as they often claim to be. This is where a critical studies approach to AI comes in. This approach draws on some of the ideas propounded by Chaka (2022), Couldry and Mejias (2019), Lindgren (2023), Mohamed et al. (2020), Ricaurte (2019), who adopt a critically driven approach to dealing with and studying technology, algorithms, data, and datafication. Importantly, it draws on Lindgren's (2023) notion of critical studies of AI.

In this paper, in particular, the critical studies approach to AI entails recognising that AI detectors are not 100% efficient and effective: they have limitations, deficiencies, and biases. This is so notwithstanding the accuracy percentage claims that these models may arrogate to themselves on their landing pages. This approach also acknowledges that AI detectors are constrained by contextual factors such as domains, algorithms, training data, performance, robustness, and adversarial testing. The latter refers to how well an AI detector performs when tested with an adversarial input like edited or paraphrased content (see Captain Words, 2024; Wu et al., 2023) or such as single spacing (Cai & Cui, 2023). This latter aspect highlights the fact that AI detectors can be tricked by manipulating or reworking input content (see Chaka, 2023a; Lee, 2023). This is one of the limitations AI detectors have, which is recognised by the critical studies approach to AI as framed here. Finally, this approach contends that the limitations and deficiencies of AI detectors should not be reduced to technologism alone: they are also a reflection of their designers, architecture, or otherwise.

Related literature

This related literature section is unconventional in that it selectively deals with a few studies that have a bearing on the current study. To this end, it wants to foreground a few points. First, save for Liang et al.'s (2023) study, there is a paucity of studies that have tested how currently available AI detectors tend to be biased against non-native English writers/students vis-à-vis native English writers/students. Secondly, as pointed out briefly earlier, since the release of ChatGPT and the other related GenAI-powered chatbots, several AI detectors have been designed and launched, which are intended to detect GenAI-generated content or distinguish between GenAI-generated and human-written content. In keeping with this attempt to detect GenAI-generated content, existing traditional plagiarism detection software programmes have been upgraded to accommodate AI detection tools in their offerings (see Anil et al., 2023; Bisi et al., 2023; Chaka, 2023a, 2024; Dergaa et al., 2023; Ladha et al., 2023; Uzun, 2023; Wiggers, 2023; Weber-Wulff et al., 2023). Again, as stated earlier, some studies have evaluated the effectiveness of single AI detectors (Habibzadeh, 2023; Subramaniam, 2023), two AI detectors (Desaire et al., 2023; Ibrahim, 2023), three AI detectors (Cingillioglu, 2023; Elali & Rachid, 2023; Wee & Reimer, 2023), four AI detectors (Alexander et al., 2023; Anil et al., 2023), and multiple AI detectors (Chaka, 2023a, 2024; Odri & Yoon, 2023; Walters, 2023).

In the midst of so many and varied studies that have been conducted in the aftermath of ChatGPT's launch, I will, in this section, briefly discuss a select few studies that have explored or tested the effectiveness of multiple AI detectors in detecting GenAI-generated content or distinguish between GenAI-generated from human-written content in given subject areas. Elsewhere, Chaka (2024) conducted a review of studies that tested the effectiveness of different AI detectors in distinguishing between GenAI-generated and human-written content in different subject areas. It is also worth mentioning that some of the studies that have investigated the effectiveness of multiple AI detectors, in this regard, are preprints like Webber-Wulff (2023) and Wu et al. (2023). Others are AI detectors' in-house studies such as AIContentfy Team, 2023; Captain Words, 2024). The first study that has some bearing on the present study is Liang et al.'s (2023) study. This study set out to evaluate the effectiveness of seven AI detectors in detecting GenAI-generated text in a dataset of 91 human-written Test of English as a Foreign Language (TOEFL) essays and in a dataset of 88 U.S. 8th-grade essays extracted from the Hewlett Foundations' Automated Student Assessment Prize (ASAP). The first dataset was sourced from a Chinese educational forum. The seven AI detectors employed to evaluate these two essay datasets were ZeroGPT, GPTZero, Crossplag, OpenAI, Sapling, Quillbot, and Originality. These detectors detected and classified the U.S. 8th-grade essay dataset almost accurately. Nonetheless, they misidentified more than half of the TOEFL essay dataset as generated by GenAI, with a mean false positive rate (FPR) of 61.22%. In addition, these AI detectors accorded the misidentified TOEFL essays a very low perplexity due to the limited linguistic variability of these essays, which was easily predictable. But, after ChatGPT was employed to improve the linguistic expressions of the TOEFL essays to those of a native English speaker, their misidentification by the said AI detectors decreased, with their mean FPR concomitantly decreasing to 11.77%, and their perplexity significantly improving as well.

Since the publication of Liang et al.'s (2023) study, there have been, in varying degrees, some comments about it (see Mathewson, 2023; Shane, 2023) and some reactions to it (see Adamson, 2023; Gillham, 2024). Among the reactions, Adamson's (2023) is the most interesting one as it shows how Liang et al.'s (2023) study seems to have ruffled up the veneer of AI detectors' effectiveness in detecting GenAI-generated text in student-written essays without being linguistically biased. To this effect, a Turnitin test was subsequently conducted to detect GenAI-generated text in three datasets of ASAP, ICNALE, and PELIC that comprised L1 English (ASAP = 2,481 and ICNALE = 400) and L2 English (ICNALE = 2,222 and PELIC = 4,000). The results of this test showed that for documents with a minimum 300-word threshold, the difference in the false positive rate (FPR) between L1 English essays and L2 English essays was fractional and, thus, was not statistically significant. This proved that the paper asserts that Turnitin, as an AI detector, did not evince any statistically significant bias against the two sets of English language essays. Moreover, the paper avers that even though each essay set's FPR was marginally higher than Turnitin's overall target of 0.01 (1%), none of the two essay sets' FPR was significantly different from this overall target. In contrast, the paper argues that for documents whose content was below the minimum 300-word threshold, there was a significant difference in the FPR between L1 English essays and L2 English essays. This difference was greater than Turnitin's 0.01 overall target. On this basis, the paper concludes that this finding confirms that AI detectors need longer essay samples for them to detect GenAI-generated content accurately and for them to be able to avoid producing a high rate of false positives (Adamson, 2023). An overall FPR target of 1% means that 10 human-produced student essays are likely to be misclassified as false positives in every 1,000 university essay scripts. This number is still concerning given those students who might be affected by this misclassification (see Anderson, 2023).

It is worth mentioning that Turnitin is not among the seven AI detectors tested by Liang et al. (2023). Despite this, there is no gainsaying that this resultant Turnitin test bears testimony to the ruffle that Liang et al.'s (2023) study has caused to the AI detection ecosystem, not only Turnitin but that of the other AI detectors as well. The other point to emphasise is that Liang et al.'s (2023) study has an element of a critical studies approach to AI. This element has to do with the way the study approached the seven AI detectors from a critical standpoint by highlighting their linguistic detection bias in dealing with native English speakers versus non-native English speakers in their written English. Moreover, this criticality element is related to the two adversarial prompts the study inputted into ChatGPT to write the two datasets differently with a view of tricking the seven AI detectors. It is when one applies this type of critical perspective which is grounded on relevant raw data to GenAI in general, and to AI detectors in particular, that one gets the owners and designers of AI detectors' attention as is the case with Adamson's (2023) paper. Without that criticality, nothing is likely to happen.

Among the studies that have evaluated multiple AI detectors in other subject areas than English is Odri and Yoon's (2023) study. This study had three objectives, which were to: evaluate 11 AI detectors' performance on a wholly GenAI-generated text, test AI detection-evading methods, and evaluate how effective these AI detection-evading methods were on previously tested AI detectors. It hypothesised that the 11 AI detectors to be tested were not all equally effective in identifying GenAI-generated text and that some of the evasion methods could render the GenAI-generated text almost undetectable. The GenAI text was generated from ChatGPT-4 and was tested on 11 AI detectors: Originality, ZeroGPT, Writer, Copyleaks, Crossplag, GPTZero, Sapling, Content at Scale, Corrector, Writefull, and Quill. The text was tested before applying AI detection evasion techniques and after applying them. The AI detection evasion techniques employed included: improving command messages (prompts) in ChatGPT, adding minor grammatical errors (e.g., a comma deletion), paraphrasing, and substituting Latin letters with their Cyrillic equivalents. The GenAI text was manipulated six times to produce its slightly modified versions using the aforesaid evasion techniques in ChatGPT. The study also tested a scientific text produced by a human (Sir John Charnley) in 1960 (Odri & Yoon, 2023). One plausible reason that can be extrapolated from the study about the use of this text is that it is freely available online. The other plausible reason is that the text predates the advent of GenAI models, particularly ChatGPT, by 62 years. Therefore, in 1960, there was no way any text could have been generated by GenAI models.

For the initial, unaltered GenAI text generated by ChatGPT, seven of the 11 AI detectors identified it as written mainly by humans. This is how these AI detectors fared in this text: GPTZero = human, Writer = 100% human, Quill = human, Content at Scale = 85% human, Copyleaks = 59.9% human, Corrector = 0.02% AI, and ZeroGPT = 25.8% AI. The more this text was slightly modified in sustained degrees (one modification after another as mentioned above), the more the 11 AI detectors misclassified it as human-written. Regarding the human-written text, only one of the 11 AI detectors (Originality) was able to correctly detect it as having 0% AI. It is important to mention that despite this correct detection, Originality is one of the four AI detectors that misidentified the final modified version of the GenAI-generated text as having 0% AI content (Odri & Yoon, 2023). Like Liang et al.'s (2023) study discussed above, the relevance of Odri and Yoon's (2023) study is that it has elements of a critical studies approach to AI. Its use of adversarial attacks in the form of prompt attacks is an example of an adversarial input that I earlier referred to as one of the contextual factors that degrades the efficacy of AI detectors (also see Anderson, 2023; Chaka, 2023a, 2024; Krishna et al., 2023; Sadasivan et al., 2023). From a critical perspective, prompt attacks expose the limitations and deficiencies of AI detectors.

Materials and methods

This study followed an exploratory research design, with the primary objective of exploring a given area, aspect, or phenomenon that has not been extensively researched. By its nature, exploratory research can tentatively analyse a new emerging topic, or suggest new ideas (Swedberg, 2020; see Makri & Neely, 2021). Testing the accuracy and effectiveness of AI detectors in identifying GenAI-generated and human-written content, or in distinguishing between these content types is still a relatively new area in many disciplines (see Chaka, 2023a, 2023b).

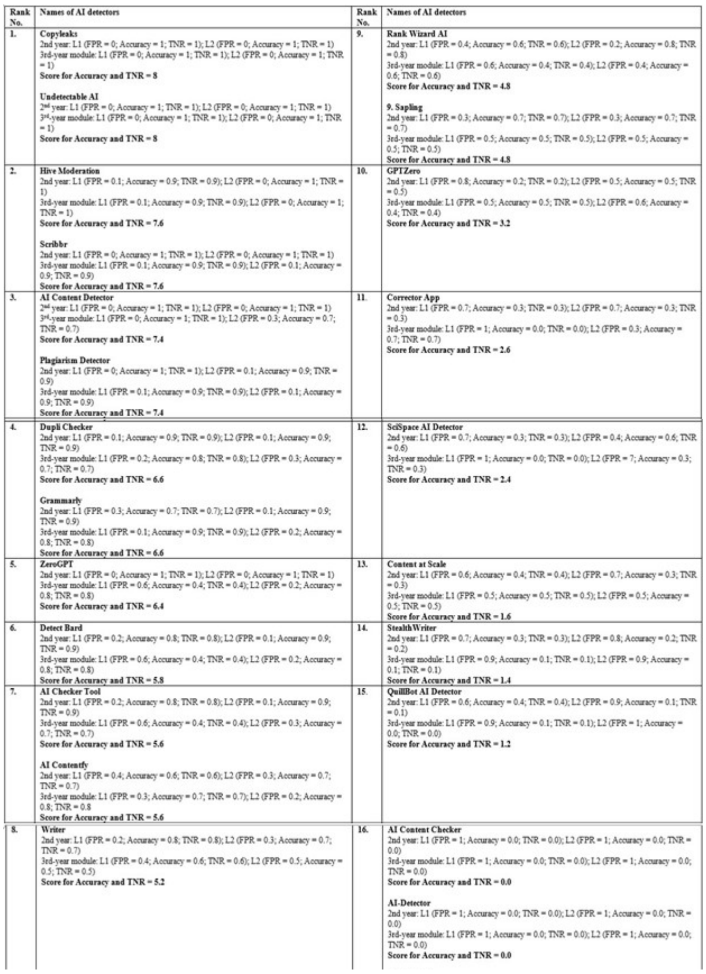
Data collection

The data collection process for this study comprised three stages. The first stage entailed selecting student (human) essay samples. These essays consisted of four datasets of university English L1 and English L2 student essays. They were selected from a pool of essays that had been submitted as assignment responses for two undergraduate modules offered by an English department at an open-distance and e-learning university in South Africa. The modules were second and third-year, major modules. Each dataset had ten essays. The two sets of essays for a second-year major module were submitted in 2018 (second semester), 2020 (first and second semesters), and 2022 (first and second semesters). The submission details of the ten essays in the English L1 essay set were as follows: 2018 first semester (n = 1), 2020 first semester (n = 4), 2020 second semester (n = 3), 2022 first semester (n = 1), and 2022 second semester (n = 1). The corresponding English L2 essay set for the second-year module consisted of the following essays in relation to their years and semesters of submission: 2020 first semester (n = 3), 2022 first semester (n = 1), and 2022 second semester (n = 6). Both sets of essays (English L1 and English L2) for a third-year, major module, each of which with ten essays, were submitted in the first semester of 2020.

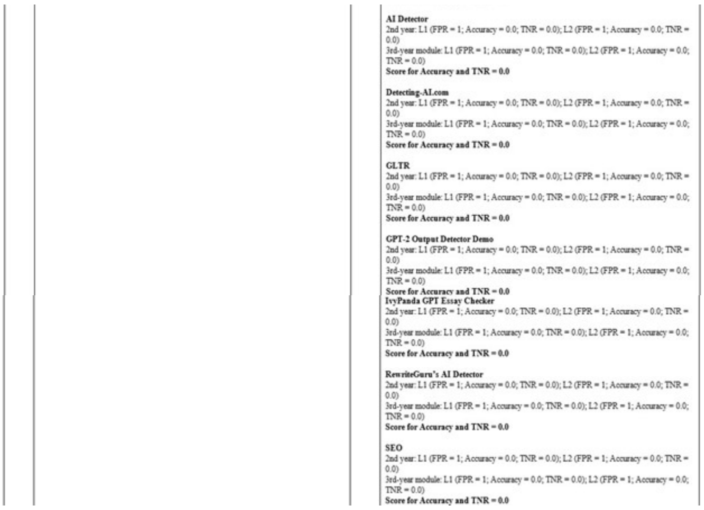
As is evident from the points presented above, the four datasets used in this study together had 40 essays. The essays were randomly selected from assignment scripts that served as either dummy or moderation scripts that are generally emailed to module team members by module primary lecturers. It is from this pool of essays that the current student essays were selected for this study. These essays were categorised as English L1 and English L2 based on whether the students who wrote them had identified English as their home language (English L1) or had identified a different language other than English as their home language (English L2) in their module registration information. All the selected essays for the four datasets were copied from their original PDF files and pasted into an MS Word file without changing anything. Thereafter, two MS Word files, English L1 and English L2 essay sets, were compiled for the two modules. The ten English L1 essays for the second-year module had a total word count of 4,465, with a mean word count of 446.5; their counterpart English L2 essays had a total word count of 4,322, with a mean word count of 432.2. The total word count of the ten English L1 essays for the third-year module was 4,504, with a mean word count of 450.4. Their corresponding English L2 essays had a total word count of 4,404, with 440.4 as their mean word count. The essay selection and compiling process took place between 18 December 2023 and 20 December 2023. Before the study was conducted, ethical clearance was secured, and the certificate number of this ethical clearance is Ref #: 2021_RPSC_050.

The second stage in the data collection process involved choosing free, publicly available online AI detectors. This process happened between 21 December 2023 and 28 December 2023. During which, many online AI detectors were identified. After trialling some of them, 30 AI detectors were chosen for use in this study (see Table 1). Then, from 02 January 2024 to 20 February 2024, the third stage occurred. Each essay from the four datasets was submitted to each of the 30 AI detectors for GenAI-generated content scanning. The test scores for each essay scan were copied and transferred to relevant tables, each of which was labelled English L1 and English L2 for each of the two modules, with each AI detector's name used as a caption for each table. However, to avoid having 30 individual tables, two tables were merged into one (see Table 1).

Table 1: Names of 30 AI detectors and their accuracy ranking.



Rank No.	**Names of AI detectors**	**Rank No.**	**Names of AI detectors**
1.	**Copyleaks** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) **Score for Accuracy and TNR = 8** **Undetectable AI** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) **Score for Accuracy and TNR = 8**	**9.**	**Rank Wizard AI** 2nd year: L1 (FPR = 0.4; Accuracy = 0.6; TNR = 0.6); L2 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8) 3rd-year module: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.4; Accuracy = 0.6; TNR = 0.6) **Score for Accuracy and TNR = 4.8** **9. Sapling** 2nd year: L1 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) 3rd-year module: L1 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5); L2 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5) **Score for Accuracy and TNR = 4.8**
2.	**Hive Moderation** 2nd year: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0; Accuracy = 1; TNR = 1) **Score for Accuracy and TNR = 7.6** **Scribbr** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) **Score for Accuracy and TNR = 7.6**	**10.**	**GPTZero** 2nd year: L1 (FPR = 0.8; Accuracy = 0.2; TNR = 0.2); L2 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5) 3rd-year module: L1 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5); L2 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4) **Score for Accuracy and TNR = 3.2**
3.	**AI Content Detector** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) **Score for Accuracy and TNR = 7.4** **Plagiarism Detector** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) 3rd-year module: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) **Score for Accuracy and TNR = 7.4**	**11.**	**Corrector App** 2nd year: L1 (FPR = 0.7; Accuracy = 0.3; TNR = 0.3); L2 (FPR = 0.7; Accuracy = 0.3; TNR = 0.3) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) **Score for Accuracy and TNR = 2.6**
4.	**Dupli Checker** 2nd year: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) 3rd-year module: L1 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) **Score for Accuracy and TNR = 6.6** **Grammarly** 2nd year: L1 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) 3rd-year module: L1 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9); L2 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8) **Score for Accuracy and TNR = 6.6**	**12.**	**SciSpace AI Detector** 2nd year: L1 (FPR = 0.7; Accuracy = 0.3; TNR = 0.3); L2 (FPR = 0.4; Accuracy = 0.6; TNR = 0.6) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 7; Accuracy = 0.3; TNR = 0.3) **Score for Accuracy and TNR = 2.4**
5.	**ZeroGPT** 2nd year: L1 (FPR = 0; Accuracy = 1; TNR = 1); L2 (FPR = 0; Accuracy = 1; TNR = 1) 3rd-year module: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8) **Score for Accuracy and TNR = 6.4**	**13.**	**Content at Scale** 2nd year: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.7; Accuracy = 0.3; TNR = 0.3) 3rd-year module: L1 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5); L2 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5) **Score for Accuracy and TNR = 1.6**
6.	**Detect Bard** 2nd year: L1 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) 3rd-year module: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8) **Score for Accuracy and TNR = 5.8**	**14.**	**StealthWriter** 2nd year: L1 (FPR = 0.7; Accuracy = 0.3; TNR = 0.3); L2 (FPR = 0.8; Accuracy = 0.2; TNR = 0.2) 3rd-year module: L1 (FPR = 0.9; Accuracy = 0.1; TNR = 0.1); L2 (FPR = 0.9; Accuracy = 0.1; TNR = 0.1) **Score for Accuracy and TNR = 1.4**
7.	**AI Checker Tool** 2nd year: L1 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8); L2 (FPR = 0.1; Accuracy = 0.9; TNR = 0.9) 3rd-year module: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) **Score for Accuracy and TNR = 5.6** **AI Contentfy** 2nd year: L1 (FPR = 0.4; Accuracy = 0.6; TNR = 0.6); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) 3rd-year module: L1 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7); L2 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8 **Score for Accuracy and TNR = 5.6**	**15.**	**QuillBot AI Detector** 2nd year: L1 (FPR = 0.6; Accuracy = 0.4; TNR = 0.4); L2 (FPR = 0.9; Accuracy = 0.1; TNR = 0.1) 3rd-year module: L1 (FPR = 0.9; Accuracy = 0.1; TNR = 0.1); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 1.2**
8.	**Writer** 2nd year: L1 (FPR = 0.2; Accuracy = 0.8; TNR = 0.8); L2 (FPR = 0.3; Accuracy = 0.7; TNR = 0.7) 3rd-year module: L1 (FPR = 0.4; Accuracy = 0.6; TNR = 0.6); L2 (FPR = 0.5; Accuracy = 0.5; TNR = 0.5) **Score for Accuracy and TNR = 5.2**	**16.**	**AI Content Checker** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **AI-Detector** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0**



		AI Detector 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **Detecting-AI.com** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **GLTR** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **GPT-2 Output Detector Demo** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0**
		IvyPanda GPT Essay Checker 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **RewriteGuru's AI Detector** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0** **SEO** 2nd year: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) 3rd-year module: L1 (FPR = 1; Accuracy = 0.0; TNR = 0.0); L2 (FPR = 1; Accuracy = 0.0; TNR = 0.0) **Score for Accuracy and TNR = 0.0**

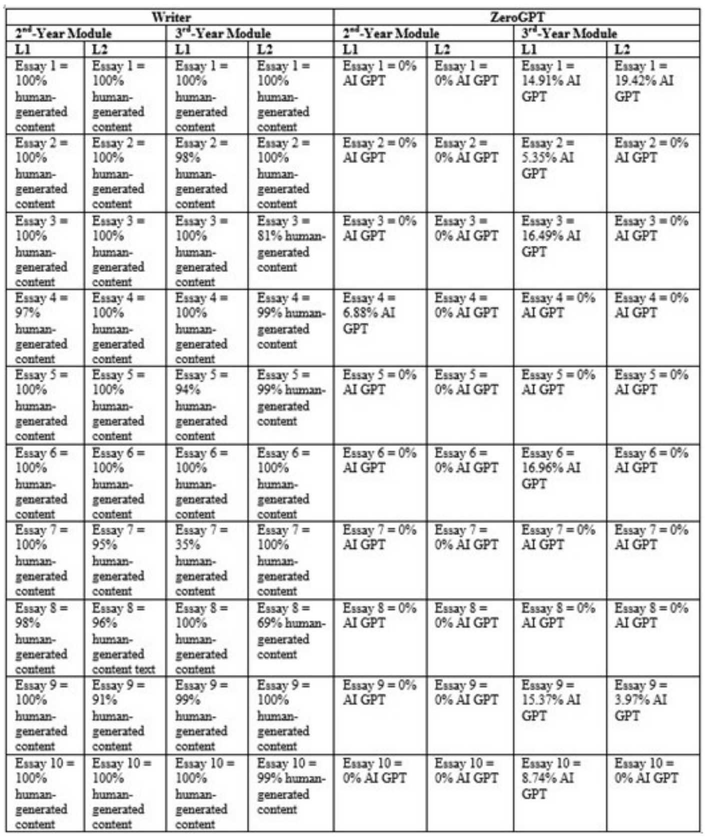
Data analysis

After the scan results for each of the relevantly labelled tables had been captured under the English L1 and English L2 categories for each of the two modules, the GenAI and human content probability scores (as percentages) and their accompanying statements as yielded by each AI detector, were entered in an MS Word file. The GenAI and human content probability scores for each set of English L1 and English L2 essays were calculated and summed. The sum for each set was averaged to get the mean score. This procedure was done for all essay datasets whose AI detector scans yielded GenAI and human content probability scores. For those essay datasets whose AI detector scans yielded only statements, those statements were captured accordingly in a tabular form. The mean scores of all the scan scores for all AI detectors were compared in each language category. Additionally, false positives (human-written essays misclassified as GenAI-generated) and true negatives (correctly detected human-written essays) for each AI detector were calculated with a view to getting false positive rates (FPRs) and true negative rates (TNRs) within each AI detector and between all AI detectors. The accuracy, specificity, and negative predictive value (NPV) of AI detectors whose test results were a direct opposite of each other were measured using confusion matrices (see Captain Words, 2024; Colquhoun, 2014; Gillham, 2024; Weber-Wulff et al., 2023; Wu et al., 2023) and compared with those of its counterparts.

Results

The GenAI test scores that were yielded by scanning each of the 30 AI detectors were compiled in a table (see Table 2). These test results were captured in the manner in which each AI detector displayed them without any modification. An example of such results is shown in Table 2. The exception is the phrasing about the colour red and the colour purple provided for GLTR AI test results. But even for this AI detector, this phrasing was formulated in keeping with how this AI detector itself explains its colour-coded scan scores. Where each AI detector's scan scores made it possible, the GenAI and human content probability scores for each set of English L1 and English L2 essays, together with their respective means, were calculated (see Tables 2 and 3). As is evident from Table 2, various GenAI and human content probability scores, expressed in percentages and percentage points, have been displayed as generated by Writer's and ZeroGPT's scan scores (raw data) for each of the ten essays for each of the two sets of essays for English L1 and English L2. These two AI detectors are used here for illustrative purposes since the scan scores of each of the 30 AI detectors cannot be displayed for lack of space. For example, Writer detected eight essays and seven essays for L1 and L2, respectively, under the 2nd-year module, as having 100% human-generated content. For the 3rd-year module, Writer classified six essays and five essays for L1 and L2, apiece, as containing 100% human-generated content. In contrast, under the 2nd-year module, ZeroGPT classified nine essays and none as containing 0% AI GPT content for L1 and L2 respectively. It, then, identified four essays for L1 and eight essays for L2 under the 3rd-year module, as having 0% AI GPT content.

Table 2: An example of how scan/test results were captured.

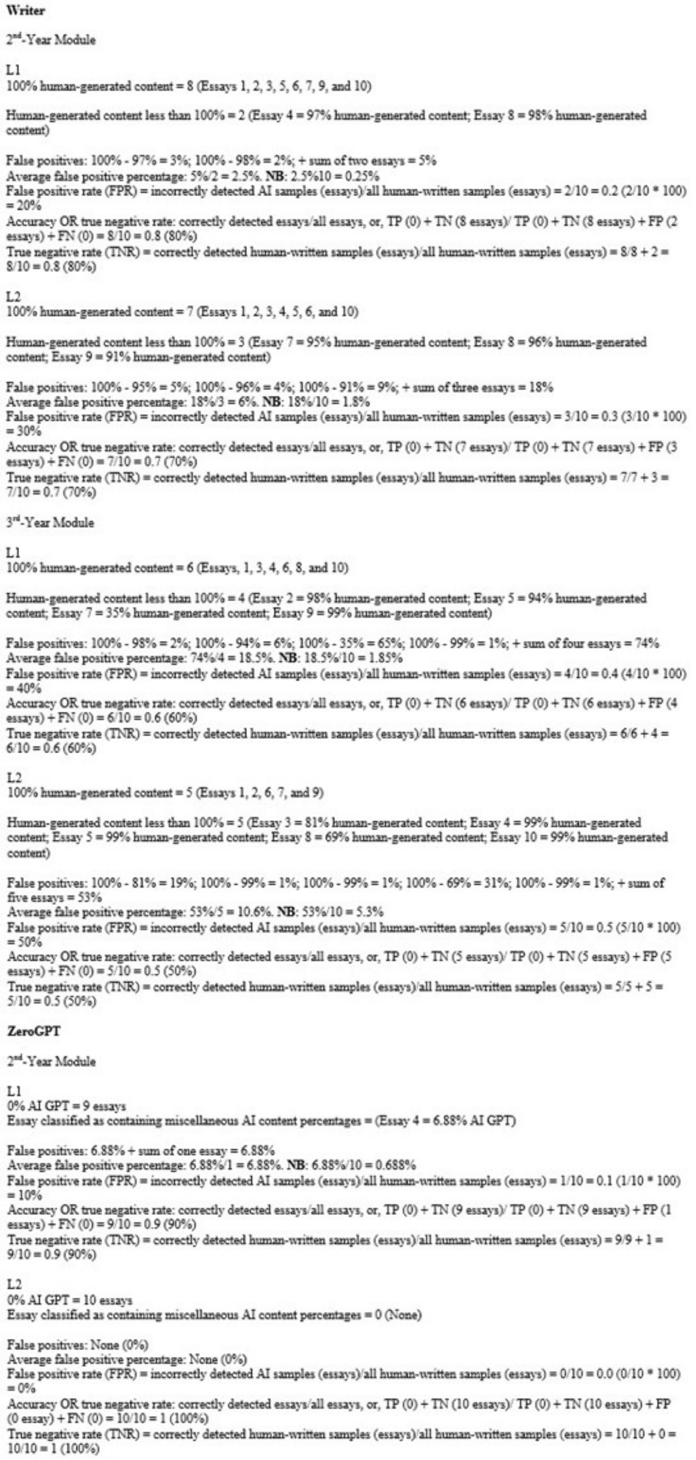


Writer				ZeroGPT			
2^{nd}-Year Module		3^{rd}-Year Module		2^{nd}-Year Module		3^{rd}-Year Module	
L1	L2	L1	L2	L1	L2	L1	L2
Essay 1 = 100% human-generated content	Essay 1 = 100% human-generated content	Essay 1 = 100% human-generated content	Essay 1 = 100% human-generated content	Essay 1 = 0% AI GPT	Essay 1 = 0% AI GPT	Essay 1 = 14.91% AI GPT	Essay 1 = 19.42% AI GPT
Essay 2 = 100% human-generated content	Essay 2 = 100% human-generated content	Essay 2 = 98% human-generated content	Essay 2 = 100% human-generated content	Essay 2 = 0% AI GPT	Essay 2 = 0% AI GPT	Essay 2 = 5.35% AI GPT	Essay 2 = 0% AI GPT
Essay 3 = 100% human-generated content	Essay 3 = 100% human-generated content	Essay 3 = 100% human-generated content	Essay 3 = 81% human-generated content	Essay 3 = 0% AI GPT	Essay 3 = 0% AI GPT	Essay 3 = 16.49% AI GPT	Essay 3 = 0% AI GPT
Essay 4 = 97% human-generated content	Essay 4 = 100% human-generated content	Essay 4 = 100% human-generated content	Essay 4 = 99% human-generated content	Essay 4 = 6.88% AI GPT	Essay 4 = 0% AI GPT	Essay 4 = 0% AI GPT	Essay 4 = 0% AI GPT
Essay 5 = 100% human-generated content	Essay 5 = 100% human-generated content	Essay 5 = 94% human-generated content	Essay 5 = 99% human-generated content	Essay 5 = 0% AI GPT	Essay 5 = 0% AI GPT	Essay 5 = 0% AI GPT	Essay 5 = 0% AI GPT
Essay 6 = 100% human-generated content	Essay 6 = 100% human-generated content	Essay 6 = 100% human-generated content	Essay 6 = 100% human-generated content	Essay 6 = 0% AI GPT	Essay 6 = 0% AI GPT	Essay 6 = 16.96% AI GPT	Essay 6 = 0% AI GPT
Essay 7 = 100% human-generated content	Essay 7 = 95% human-generated content	Essay 7 = 35% human-generated content	Essay 7 = 100% human-generated content	Essay 7 = 0% AI GPT	Essay 7 = 0% AI GPT	Essay 7 = 0% AI GPT	Essay 7 = 0% AI GPT
Essay 8 = 98% human-generated content	Essay 8 = 96% human-generated content text	Essay 8 = 100% human-generated content	Essay 8 = 69% human-generated content	Essay 8 = 0% AI GPT	Essay 8 = 0% AI GPT	Essay 8 = 0% AI GPT	Essay 8 = 0% AI GPT
Essay 9 = 100% human-generated content	Essay 9 = 91% human-generated content	Essay 9 = 99% human-generated content	Essay 9 = 100% human-generated content	Essay 9 = 0% AI GPT	Essay 9 = 0% AI GPT	Essay 9 = 15.37% AI GPT	Essay 9 = 3.97% AI GPT
Essay 10 = 100% human-generated content	Essay 10 = 100% human-generated content	Essay 10 = 100% human-generated content	Essay 10 = 99% human-generated content	Essay 10 = 0% AI GPT	Essay 10 = 0% AI GPT	Essay 10 = 8.74% AI GPT	Essay 10 = 0% AI GPT

In terms of false positives, Writer had two false positives and three false positives for the 2nd-year module's L1 and L2 essay sets, respectively. The first set collectively had 5% AI content, with an average false positive percentage of 2.5% AI content, while the second set contained 18% AI content, with an average false positive percentage of 6% AI content. With regard to the 3rd-year module, the L1 essay set consisted of four false positives that contained an overall AI content of 74%. Collectively, they had an average false positive percentage of 18.5% AI content. Its counterpart L2 essay set had four false positives, whose aggregate AI content was 53%. Its average false positive percentage was 10.6% AI content.

For ZeroGPT, the 2nd-year module's L1 and L2 essay sets had one false positive and no false positive, respectively. The first set contained 6.88% AI content, which was also its average false positive percentage. The second set had 0% AI content and 0% AI content as its average false positive percentage. ZeroGPT's 3rd-year module's L1 and L2 essay sets had six false positives and two false positives each. The first set had an aggregate AI content of 77.82%, with 12.97% as its average false positive percentage for its AI content. By contrast, the second essay set contained an overall AI content of 23.39%, with 11.695% being its average false positive percentage for its AI content (see Table 3).

Table 3: How the AI and human content probability scores and means were calculated.



Writer

2nd-Year Module

L1
100% human-generated content = 8 (Essays 1, 2, 3, 5, 6, 7, 9, and 10)

Human-generated content less than 100% = 2 (Essay 4 = 97% human-generated content; Essay 8 = 98% human-generated content)

False positives: 100% - 97% = 3%; 100% - 98% = 2%; + sum of two essays = 5%
Average false positive percentage: 5%/2 = 2.5%. **NB**: 2.5%10 = 0.25%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 2/10 = 0.2 (2/10 * 100) = 20%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (8 essays)/ TP (0) + TN (8 essays) + FP (2 essays) + FN (0) = 8/10 = 0.8 (80%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 8/8 + 2 = 8/10 = 0.8 (80%)

L2
100% human-generated content = 7 (Essays 1, 2, 3, 4, 5, 6, and 10)

Human-generated content less than 100% = 3 (Essay 7 = 95% human-generated content; Essay 8 = 96% human-generated content; Essay 9 = 91% human-generated content)

False positives: 100% - 95% = 5%; 100% - 96% = 4%; 100% - 91% = 9%; + sum of three essays = 18%
Average false positive percentage: 18%/3 = 6%. **NB**: 18%/10 = 1.8%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 3/10 = 0.3 (3/10 * 100) = 30%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (7 essays)/ TP (0) + TN (7 essays) + FP (3 essays) + FN (0) = 7/10 = 0.7 (70%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 7/7 + 3 = 7/10 = 0.7 (70%)

3rd-Year Module

L1
100% human-generated content = 6 (Essays, 1, 3, 4, 6, 8, and 10)

Human-generated content less than 100% = 4 (Essay 2 = 98% human-generated content; Essay 5 = 94% human-generated content; Essay 7 = 35% human-generated content; Essay 9 = 99% human-generated content)

False positives: 100% - 98% = 2%; 100% - 94% = 6%; 100% - 35% = 65%; 100% - 99% = 1%; + sum of four essays = 74%
Average false positive percentage: 74%/4 = 18.5%. **NB**: 18.5%/10 = 1.85%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 4/10 = 0.4 (4/10 * 100) = 40%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (6 essays)/ TP (0) + TN (6 essays) + FP (4 essays) + FN (0) = 6/10 = 0.6 (60%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 6/6 + 4 = 6/10 = 0.6 (60%)

L2
100% human-generated content = 5 (Essays 1, 2, 6, 7, and 9)

Human-generated content less than 100% = 5 (Essay 3 = 81% human-generated content; Essay 4 = 99% human-generated content; Essay 5 = 99% human-generated content; Essay 8 = 69% human-generated content; Essay 10 = 99% human-generated content)

False positives: 100% - 81% = 19%; 100% - 99% = 1%; 100% - 99% = 1%; 100% - 69% = 31%; 100% - 99% = 1%; + sum of five essays = 53%
Average false positive percentage: 53%/5 = 10.6%. **NB**: 53%/10 = 5.3%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 5/10 = 0.5 (5/10 * 100) = 50%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (5 essays)/ TP (0) + TN (5 essays) + FP (5 essays) + FN (0) = 5/10 = 0.5 (50%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 5/5 + 5 = 5/10 = 0.5 (50%)

ZeroGPT

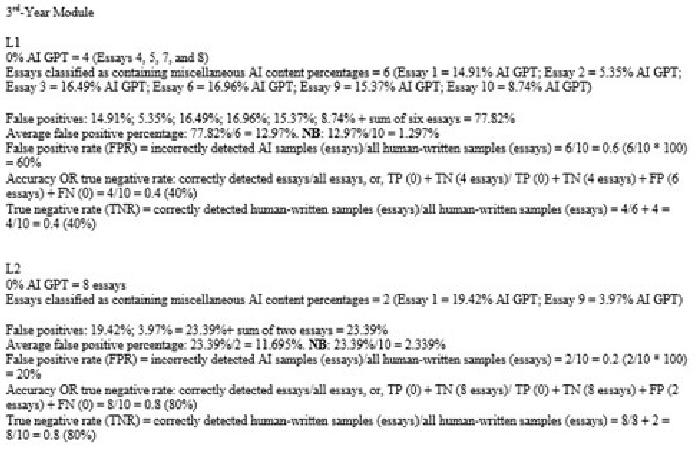
2nd-Year Module

L1
0% AI GPT = 9 essays
Essay classified as containing miscellaneous AI content percentages = (Essay 4 = 6.88% AI GPT)

False positives: 6.88% + sum of one essay = 6.88%
Average false positive percentage: 6.88%/1 = 6.88%. **NB**: 6.88%/10 = 0.688%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 1/10 = 0.1 (1/10 * 100) = 10%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (9 essays)/ TP (0) + TN (9 essays) + FP (1 essays) + FN (0) = 9/10 = 0.9 (90%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 9/9 + 1 = 9/10 = 0.9 (90%)

L2
0% AI GPT = 10 essays
Essay classified as containing miscellaneous AI content percentages = 0 (None)

False positives: None (0%)
Average false positive percentage: None (0%)
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 0/10 = 0.0 (0/10 * 100) = 0%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (10 essays)/ TP (0) + TN (10 essays) + FP (0 essay) + FN (0) = 10/10 = 1 (100%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 10/10 + 0 = 10/10 = 1 (100%)



3rd-Year Module

L1
0% AI GPT = 4 (Essays 4, 5, 7, and 8)
Essays classified as containing miscellaneous AI content percentages = 6 (Essay 1 = 14.91% AI GPT; Essay 2 = 5.35% AI GPT; Essay 3 = 16.49% AI GPT; Essay 6 = 16.96% AI GPT; Essay 9 = 15.37% AI GPT; Essay 10 = 8.74% AI GPT)

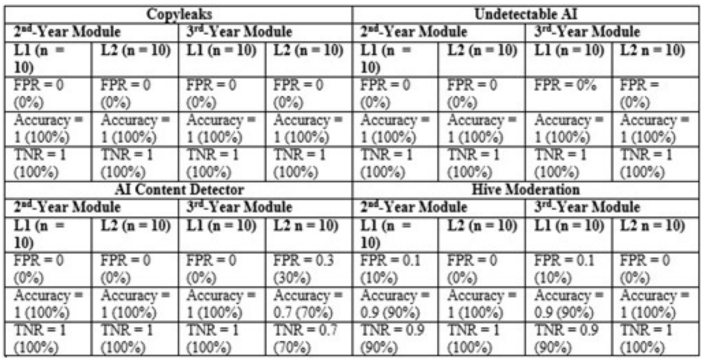
False positives: 14.91%; 5.35%; 16.49%; 16.96%; 15.37%; 8.74% + sum of six essays = 77.82%
Average false positive percentage: 77.82%/6 = 12.97%. **NB**: 12.97%/10 = 1.297%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 6/10 = 0.6 (6/10 * 100) = 60%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (4 essays)/ TP (0) + TN (4 essays) + FP (6 essays) + FN (0) = 4/10 = 0.4 (40%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 4/6 + 4 = 4/10 = 0.4 (40%)

L2
0% AI GPT = 8 essays
Essays classified as containing miscellaneous AI content percentages = 2 (Essay 1 = 19.42% AI GPT; Essay 9 = 3.97% AI GPT)

False positives: 19.42%; 3.97% = 23.39%+ sum of two essays = 23.39%
Average false positive percentage: 23.39%/2 = 11.695%. **NB**: 23.39%/10 = 2.339%
False positive rate (FPR) = incorrectly detected AI samples (essays)/all human-written samples (essays) = 2/10 = 0.2 (2/10 * 100) = 20%
Accuracy OR true negative rate: correctly detected essays/all essays, or, TP (0) + TN (8 essays)/ TP (0) + TN (8 essays) + FP (2 essays) + FN (0) = 8/10 = 0.8 (80%)
True negative rate (TNR) = correctly detected human-written samples (essays)/all human-written samples (essays) = 8/8 + 2 = 8/10 = 0.8 (80%)

Since the raw false positives and their corresponding average false positive percentages as discussed above are not a reliable measure of the accuracy of AI detectors, false positive rates (FPRs), true negative rates (TNRs), and the accuracy of the scan scores of the 30 AI detectors for the four sets of essays were calculated (see Captain Words, 2024; Colquhoun, 2014; Gillham, 2024; Weber-Wulff et al., 2023; Wu et al., 2023; also see Table 3). In particular, the FPRs, the TNRs, the accuracy, and the specificity of the AI detectors whose scan scores were direct opposites of each other, were chosen and calculated for comparative analysis. Included in the 30 AI detectors are the AI detectors that correctly classified all ten essays in each of the four essay sets (two sets for English L1 and two sets for English L2), which were tested by the 30 AI detectors. They also encompassed the AI detectors that completely misclassified all ten essays in each of these four essay sets. In this context, two AI detectors, Copyleaks and Undetectable AI, correctly classified all ten essays in each of the four essay sets (see Table 4). Contrariwise, nine AI detectors completely misclassified all ten essays in each of these four essay sets. These nine AI detectors were AI Content Checker, AI-Detector, AI Detector, Detecting-AI.com, GLTR, GPT-2 Output Detector Demo, IvyPanda GPT Essay Checker, RewriteGuru's AI Detector, and SEO (see Table 5).

Table 4: How Copyleaks and Undetectable AI correctly detected all the essay sets in both English language categories of the two modules.

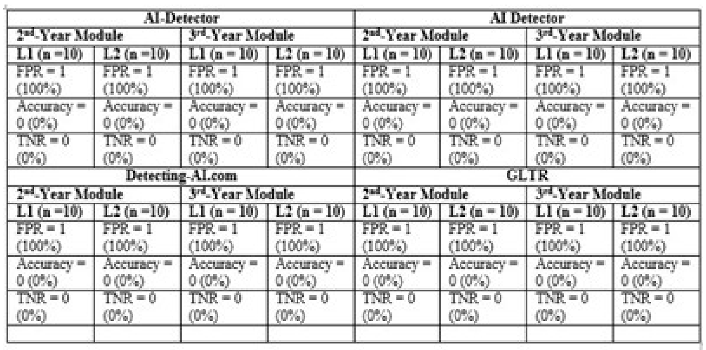


Copyleaks				Undetectable AI			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 n = 10)
FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0%	FPR = (0%)
Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)
TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)
AI Content Detector				Hive Moderation			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 n = 10)
FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0 (0%)	FPR = 0.3 (30%)	FPR = 0.1 (10%)	FPR = 0 (0%)	FPR = 0.1 (10%)	FPR = 0 (0%)
Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 1 (100%)	Accuracy = 0.7 (70%)	Accuracy = 0.9 (90%)	Accuracy = 1 (100%)	Accuracy = 0.9 (90%)	Accuracy = 1 (100%)
TNR = 1 (100%)	TNR = 1 (100%)	TNR = 1 (100%)	TNR = 0.7 (70%)	TNR = 0.9 (90%)	TNR = 1 (100%)	TNR = 0.9 (90%)	TNR = 1 (100%)

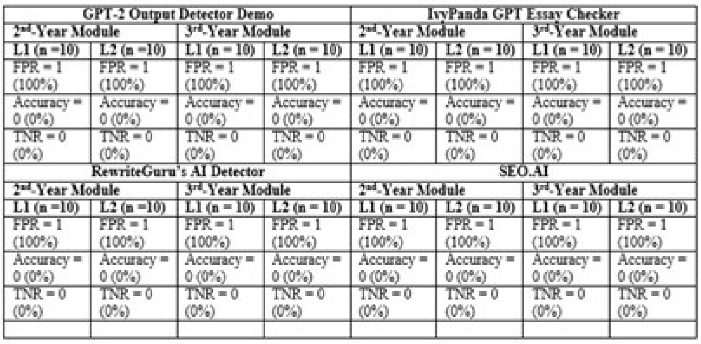
The three measures: the FPR (false positive rate), accuracy, and the TNR (true negative rate) were manually calculated based on the scan scores of the said AI detectors. The FPR was calculated using the formula, FPR = incorrectly detected AI essays/all human-written essays, or FP/FP + TP, where

FP and TP stand for false positives and true positives, respectively. This is related to each essay set (see Table 3). In the same breadth, accuracy was calculated by utilising the formula, accuracy = correctly detected essays/all essays, or TP + TN/TP + TN + FP + FN. In this case, TN and FN stand for true negatives and false negatives. For its part, the TNR was calculated through this formula: TNR = correctly detected human-written essays/all human-written essays, or TN/TN + FP (see Table 3). For example, Table 4 depicts the FPR, the accuracy, and the TNR of each of the L1 and L2 essay sets of both the 2nd-year module and the 3rd-year module for Writer and ZeroGPT. On one hand, for the 2nd-year module's L1 and L2, Writer had the following sets of scores for each of these two English language categories: FRP = 0.2, Accuracy = 0.8, and TNR = 0.8; and FRP = 0.3, Accuracy = 0.7, and TNR = 0.7. Its 3rd-year module's L1 and L2 scores for these three measures were as follows: FRP = 0.4, Accuracy = 0.6, and TNR = 0.6; and FRP = 0.5, Accuracy = 0.5, and TNR = 0.5.

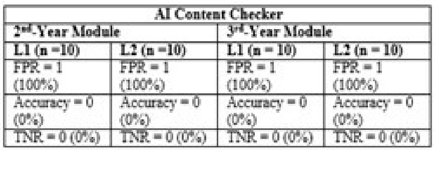
Table 5: How the nine AI incorrectly detected all the essay sets in both English language categories of the two modules.



AI-Detector				AI Detector			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n =10)	L2 (n =10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)
FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)
Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)
TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)
Detecting-AI.com				GLTR			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n =10)	L2 (n =10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)
FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)
Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)
TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)



GPT-2 Output Detector Demo				IvyPanda GPT Essay Checker			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n =10)	L2 (n =10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)
FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)
Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)
TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)
RewriteGuru's AI Detector				SEO.AI			
2nd-Year Module		3rd-Year Module		2nd-Year Module		3rd-Year Module	
L1 (n =10)	L2 (n =10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)	L1 (n = 10)	L2 (n = 10)
FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)
Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)
TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)

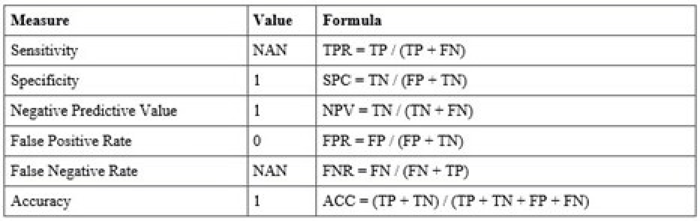


AI Content Checker			
2nd-Year Module		3rd-Year Module	
L1 (n =10)	L2 (n =10)	L1 (n = 10)	L2 (n = 10)
FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)	FPR = 1 (100%)
Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)	Accuracy = 0 (0%)
TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)	TNR = 0 (0%)

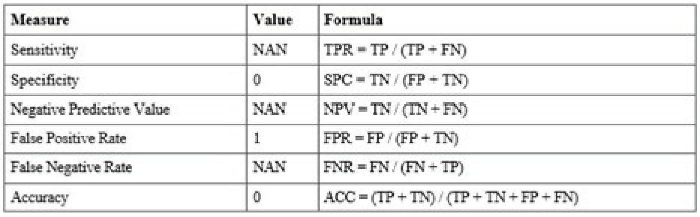
On the other hand, ZeroGPT had the following score sets for its 2nd-year module's L1 and L2: FRP = 0.1, Accuracy = 0.9, and TNR = 0.9; and FRP = 0.0, Accuracy = 1, and TNR = 1. And its score sets for the 3rd-year module's L1 and L2 were as follows: FRP = 0.6, Accuracy = 0.4 and TNR = 0.4; and FRP = 0.2, Accuracy = 0.8, and TNR = 0.8. With the exception of two essay sets (the 2nd-year module's L1 for Writer and the 3rd-year module's L2 for ZeroGPT), the two AI detectors had varying scores for these three measures in their other essay sets for these two modules. Suffice it to say that ZeroGPT correctly classified one essay set for the 2nd-year module's L2, while it incorrectly identified this module's L1 by one percentage point. Therefore, ZeroGPT performed better between the two AI detectors.

The points discussed in the preceding paragraph, lead to the calculation of the FPRs, the TNRs, the accuracy, and the specificity of the two AI detectors that correctly identified all the essay sets and of the nine AI detectors that incorrectly identified all the essay sets. Specificity is the function of TNR: it is about the proportion of correct/true negative cases correctly classified as such by an AI detector (see Elkhatat et al., 2023). In the context of the present study, this relates to the proportion of student-written essays correctly recognised by any of the 30 AI detectors out of ten student-written essays in each of the four essay sets. To calculate these four measures in the two sets of AI detectors mentioned above, an online confusion matrix calculator was used. This calculator was ideal for computing these measures. As said earlier, for Copyleaks, Undetectable AI, and the other nine AI detectors, the scores are as portrayed in Table 6.

Table 6: FPRs, TNRs, the accuracy, and the specificity of Copyleaks and Undetectable AI (top half) and of the other nine AI detectors (bottom half) for English L1 and English L2 essay sets as measured by a confusion matrix calculator.



Measure	Value	Formula
Sensitivity	NAN	TPR = TP / (TP + FN)
Specificity	1	SPC = TN / (FP + TN)
Negative Predictive Value	1	NPV = TN / (TN + FN)
False Positive Rate	0	FPR = FP / (FP + TN)
False Negative Rate	NAN	FNR = FN / (FN + TP)
Accuracy	1	ACC = (TP + TN) / (TP + TN + FP + FN)



Measure	Value	Formula
Sensitivity	NAN	TPR = TP / (TP + FN)
Specificity	0	SPC = TN / (FP + TN)
Negative Predictive Value	NAN	NPV = TN / (TN + FN)
False Positive Rate	1	FPR = FP / (FP + TN)
False Negative Rate	NAN	FNR = FN / (FN + TP)
Accuracy	0	ACC = (TP + TN) / (TP + TN + FP + FN)

As depicted in the top half of this table, the scores for the FPR, the negative predictive value (NPV) (which is also an equivalent of a true negative rate (TNR)), accuracy, and specificity for both Copyleaks and Undetectable AI were as follows: FPR = 0, NPV = 1, accuracy = 1, and specificity = 1. The acronym, NAN (not a number), or sometimes, NaN, denotes the measures whose scores could not be computed as they were not relevant for the purpose at hand. As was highlighted concerning Table 4 earlier, Copyleaks and Undetectable AI had these scores because they correctly identified all of the essay sets which consisted of the two English language categories. Inversely, as exhibited in the bottom half of Table 6, the nine AI detectors mentioned above, collectively had the score set, FPR = 1, accuracy = 0, and specificity = 0, since all of them misidentified all the essay sets of the two English language categories for both modules. Here, too, NAN signifies the measures whose scores could not be captured as they were not relevant.

All the 30 AI detectors were ranked for their accuracy in detecting if the four sets of essays (two sets of English L1 essays, n = 20; and two sets of English L2 essays, n = 20) were GenAI-generated or human-written. The accuracy and TNR scores of each AI detector were used to rank the accuracy of the 30 AI detectors (for relevant examples, see Tables 3 and 4). Based on these composite scores, many AI detectors shared joint spots when they were ranked for accuracy. For instance, two AI detectors, Copyleaks and Undetectable AI, jointly shared the first spot. They were followed by Hive Moderation and Scribbr, AI Content Detector and Plagiarism Detector, and Dupli Checker and Grammarly, which, as pairs, jointly shared the second, third, and fourth spots, respectively. ZeroGPT and Detect Bard, each notched the fifth and sixth places, while AI Checker Tool and AI Contentfy jointly occupied the seventh position. This is followed by Writer in the eighth spot and Rank Wizard AI and Sapling jointly took up the ninth position.

The spots ranging from ten to 15 were, each, occupied by different AI detectors, with GPTZero at the tenth spot and QuillBot AI Detector at the 15th place. The 16th and last spot was collectively shared by the nine AI detectors mentioned earlier.

Discussion

The results presented above is discussed in this section in response to the three research questions for this study.

The accuracy of 30 AI content detectors

As highlighted in the preceding section, of the 30 free, publicly available online AI detectors, only two of them, Copyleaks and Undetectable AI, were able to correctly identify all the essay sets of the two English language categories (English L1 and English L2) as human written. These two AI detectors also had the highest accuracy and TNR scores for all these essay sets, when their scores were manually calculated. Moreover, they did so even when their specificity and NPV was computed using a confusion matrix calculator. However, their scores in all these four measures diametrically contrasted with those of the nine AI detectors, whose scores in these measures, especially for accuracy and specificity, were zero (0%). Their FPR score of one (100%) was the polar opposite of the FPR score of zero (0%) for Copyleaks and Undetectable AI. In this sense, the nine AI detectors misidentified all four essay sets of the two English language categories. The rest of the other AI detectors had varying accuracy, FPR, and TNR scores. As such, they classified these four essay sets of English L1 and English L2 in varying degrees of accuracy, FPRs, and TNRs (see Figure 1).

In some of the previous studies conducted on the efficacy of AI detectors, Copyleaks has been the best-performing AI detector or, at least one of the best-performing AI detectors. One such study is Walters' (2023) study. This study tested the effectiveness of 16 AI detectors in identifying GenAI-generated and human-written content in three sets of first-year, undergraduate composition essays. The three sets of essays comprised 42 essays generated by ChatGPT-3.5, 42 essays created by ChatGPT-4, and 42 essays written by students. The last set was chosen from a college's English 110 (First-Year Composition) essays, which had been submitted during the 2014-2015 academic year. In this study, both Copyleaks and Turnitin had the highest accuracy rate, followed by Originality. Sapling and Content at Scale had the lowest accuracy rate among the 16 AI detectors. In the current study, Sapling and Content at Scale, occupied the 9th and 13th spots respectively.

Another study is Chaka's (2023a), which evaluated the accuracy of five AI detectors in detecting GenAI-generated content in 21 applied English language studies responses generated by three GenAI chabots: ChatGPT (n = 6), YouChat (n = 7), and Chatsonic (n = 8). The five AI detectors were GPTZero, OpenAI Text Classifier, Writer, Copyleaks, and GLTR. All the twenty-one English responses were submitted to the five AI detectors for scanning. The ChatGPT-generated responses were translated into German, French, Spanish, Southern Sotho, and isiZulu by using Google Translate. They were, then, submitted to GPTZero for scanning. The German, French and Spanish translated versions were inputted into Copyleaks for scanning. In this sense, this study utilised machine translation as an adversarial attack, which is a strategy that is related to a critical studies approach to AI as I had argued in the relevant section above. In all the different versions of the twenty-one responses, Copyleaks was the most accurate of the five AI detectors (see Chaka, 2023a). Similarly, in a literature and integrative hybrid review conducted by Chaka (2024), which reviewed 17 peer-reviewed journal articles, Copyleaks was one of the best-performing AI detectors in one of the four articles in which OpenAI Text Classifier, and Crossplag, Grammarly also topped in each of the other three articles. But, overall, in all the 17 reviewed articles, Crossplag was the best-performing AI detector, followed by Copyleaks.

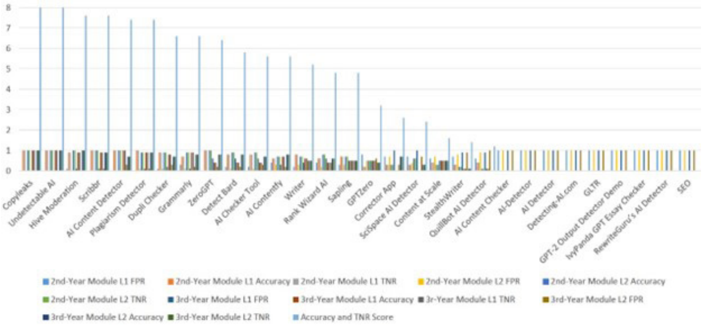


Figure 1: A graphic representation of the 30 AI detectors based on their accuracy, FPR, and TNR scores.

In Odri and Yoon's (2023) study, though, which as discussed earlier, tested 11 AI detectors and employed adversarial attacks, especially evasion techniques (e.g., improving command messages (prompts) in ChatGPT, adding minor grammatical errors, paraphrasing, and substituting Latin letters with their Cyrillic equivalents), as part of a critical studies approach to AI, Originality out-performed all the 11 AI detectors in correctly identifying the human-written-text. It, nonetheless, misclassified the final version of the AI-generated text. However, it was the AI detector that was most resistant to adversarial attacks compared to the other AI detectors (see Odri & Yoon, 2023).

Differential classification of the four sets of student essays and a language category assigned more false positives

As pointed out in the preceding section, both Copyleaks and Undetectable AI classified all the four sets of English L1 and English L2 student essays similarly and correctly by assigning the same scores for the three measures: accuracy, FPR, and TNR, to all of them. Additionally, both did so for their specificity scores for all the four essay sets. Likewise, the nine AI detectors allotted the same scores for their respective measures for the four essay sets. Even the rest of the other AI detectors, which had varying scores for these measures, did not have scores specifically skewed toward one English language category in each of the four essay sets. In fact, even in the cases where one AI detector had lower scores for essays within a given essay set of a particular language category, it had higher scores for essays within another essay set of a different language category.

In instances where a particular AI detector scored the essay sets of the one language category in a given module (e.g., the English L1 essay sets for both the 2nd-year module and the 3rd-year module) higher than the essay sets of the other language category in the same module, the differences in the scores of essay sets of these different language categories were not substantial. Or, if the scores were higher, they were not consistent for the essay sets of one language category (e.g., English L1) to the exclusion of the essay sets of the other language category (e.g., English L2) (see Table 3). So, in the present study, the AI detectors that correctly classified the student essay sets did so for both English L1 and English L2. In a similar vein, those AI detectors that misclassified the student essay sets did so for both of these English language categories. Moreover, no language category was assigned more false positives for its essay sets than those of the other language category. This means that the 30 AI detectors were not language category-biased or language category-sensitive when assigning false positives to and when classifying the essays belonging to the four essay sets. In the current study, therefore, there is no evidence suggesting that the AI detectors that were tested were consistently and invariably biased towards or against any of the student essay sets of the two English language categories.

In contrast, though, and as stated earlier, Liang et al.'s (2023) study found that the AI detectors that they evaluated tended to be biased against non-English language speakers' essays (also see Mathewson, 2023; Shane, 2023; cf. Adamson, 2023; Gillham, 2024). While this is the case, the results of the current study, nonetheless, do not nullify or invalidate those of Liang et al.'s (2023) study, as it did not use the same data sets as the ones used by that study. Instead, the present study's results are different from those of Liang et al.'s (2023) study.

Implications and recommendations

This study has implications for detecting GenAI content in student essays and for differentiating between GenAI-generated and human-written content in student essays in higher education. Firstly, detecting GenAI in student essays or distinguishing between GenAI-generated and human-written content in such essays is not simply a matter of displaying AI and human content probability scores (or percentages) and the statements accompanying them as most, if not all, AI content detectors currently tend to do. Neither is it a matter of making self-serving claims about high AI detection accuracy rates, as is the case with 28 (93%) of the 30 AI detectors tested in this study. This means that the AI detection accuracy claims made by different AI detection tools on their respective landing pages should be taken with a pinch of salt. As demonstrated in this study, such claims hardly live up to their stated expectations. Again, as shown by the results of this study, of these 28 AI detectors that did not perform as expected, nine of them completely misclassified all the human-written essays, while the remaining 19 misclassified these essays in varying degrees. Any AI content probability percentage or percentage point, however negligible it may be, that is attributed to a student essay which has no GenAI content at all, inflicts immeasurable reputational damage to that essay and to the student who produced it. This means that if this particular essay was meant for assessment purposes, then, the student concerned would be unfairly accused of a gross academic dishonesty they would not have committed. Given all of this, it is advisable for academics and for universities to which these academics belong, to exercise extreme caution when utilising any AI content detection tool for detecting GenAI content in their students' academic essays. The reason for having to be extra cautious is that most of the current AI detectors demonstrate a high degree of inaccuracy and unreliability. Importantly, it is very risky to employ one AI content detector and take its scan results as a final verdict for any given human-written text.

Secondly, the reliance of the current AI detectors on perplexity and burstiness for determining and predicting the presence or absence of GenAI content in human-written student essays results in these detectors consistently misclassifying such essays. This is one of the reasons why they keep on misclassifying student writing that has low perplexity and burstiness, such as that of non-English native speakers, as containing GenAI content portions, even when that is not the case. Repetitive word sequences and predictable lexical and syntactic parsing, as assumed by perplexity and burstiness, might work as indicators of the presence or the absence of GenAI content within the surreal world of GenAI driven by large language models. Nevertheless, in a real-world and human environment in which university students produce different forms of academic writing, informed by their diverse English language backgrounds and in response to assignment questions, perplexity and burstiness serve as weak, if not misplaced, indicators of the presence or the absence of GenAI content in student writing. The types of essays used in the current study serve as a case in point that detecting GenAI-generated content or distinguishing between it and its human-written counterpart is not merely a matter of English L1 writing versus English L2 writing. Human-produced writing cannot be reduced to robotic writing powered and aided by machine learning and GenAI large language models. Therefore, it is prudent for AI detection tools to have language training data sets that reflect the diverse, multi-dialectal, poly-racial, and pluri-ethnic speakers of a given language, in various global or geographical settings, for them to be able to capture the nuances of such a language. This is more so for a language such as English that has these types of speakers across the globe.

Conclusion

The current study had three research questions (RQs) and three corresponding objectives as stated earlier. Only two of the 30 tested, free-to-use, AI detectors, Copyleaks and Undetectable AI, did manage to correctly detect all of the student essay sets of the two English language categories (English L1 and English L2) as human-written. Nine of these 30 AI detectors (AI Content Checker, AI-Detector, AI Detector, Detecting-AI.com, GLTR, GPT-2 Output Detector Demo, IvyPanda GPT Essay Checker, RewriteGuru's AI Detector, and SEO) did the opposite: they misidentified all the essays in each of the four essay sets of the two language categories in both the 2nd-year module and the 3rd-year module. The remaining 19 AI detectors both correctly and incorrectly classified the four essay sets in varying degrees without any bias to any essay set of the two English language categories. Therefore, Copyleaks and Undetectable AI, were, jointly, the top-most accurate AI detectors that ranked first in this study, while the nine AI detectors were the most inaccurate, which collectively ranked last in the pecking order. Of the other 19 AI detectors, ten of them held joint positions, with the remaining nine notching individual accuracy slots in the ranking.

All 30 AI detectors did not assign differential classification to the four essay sets according to the English language categories to which they belonged. That is, they displayed no specific bias toward language categories in classifying or misclassifying the four essay sets. The same applies to the false positives they accorded to these essay sets. If only two AI detectors out of 30 can accurately detect all the student essay sets across the two language categories, and nine AI detectors can do the complete opposite, with the remaining AI detectors yielding variable accuracy scores for the same sets of essays in the two language categories as is the case in this study, then, university students and the universities to which they belong are in trouble concerning the presence or absence of GenAI content in student essays. Moreover, the results of the current study demonstrate that detecting GenAI-generated content or distinguishing it from its human-written counterpart is not simply a matter of perplexity and burstiness, or of English L1 writing versus English L2 writing. Human-produced writing is very complex and nuanced and thus cannot be reduced to measures of high or low perplexity and burstiness. This applies to both English L1 and English L2 writers, depending on their writing proficiency. On this basis, the present study suggests that the bulk of the currently available AI detectors are not fit for its purpose, even when the input content, such as the essays used in this study, is not manipulated through any adversarial attacks. The implication of this study, therefore, is that relying on one or even a few AI detection tools for identifying GenAI content in student essays is a risky move.

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

Adamson, D. (2023). *New research: Turnitin's AI detector shows no statistically significant bias against English language learners.* https://www.turnitin.com/blog/new-research-turnitin-s-ai-detector-shows-no-statistically-significant-bias-against-english-language-learners

AIContentfy Team. (2023). *Evaluating the effectiveness of AI detectors: Case studies and metrics.* https://aicontentfy.com/en/blog/evaluating-of-ai-detectors-case-studies-and-metrics

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

Anderson, C. (2023). *The false promise of AI writing detectors.* https://www.linkedin.com/pulse/false-promise-ai-writing-detectors-carol-anderson

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

Anthology White Paper. (2023). *AI, academic integrity, and authentic assessment: An ethical path forward for education.* https://www.anthology.com/sites/default/files/2023-09/White%20Paper-AI%20Academic%20Integrity%20and%20Authentic%20Assessment-An%20Ethical%20Path%20Forward%20for%20Education-v2_09-23_0.pdf

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

Blau, I., Goldberg, S., Friedman, A., & Eshet-Alkalai, Y. (2020). Violation of digital and analog academic integrity through the eyes of faculty members and students: Do institutional role and technology change ethical perspectives? *Journal of Computing in Higher Education, 33*(1), 157-187. https://doi.org/10.1007/s12528-020-09260-0

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

Captain Words. (2024). *Testing AI detection tools – our methodology.* https://captainwords.com/ai-detection-tools-test-methodology/

Chaka, C. (2022). Digital marginalization, data marginalization, and algorithmic exclusions: A critical southern decolonial approach to datafication, algorithms, and digital citizenship from the Souths. *Journal of e-Learning and Knowledge Society, 18*(3), 83-95. https://doi.org/10.20368/1971-8829/1135678

Chaka, C. (2023a). 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. (2023b). 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. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching, 7*(1), 1-12. https://doi.org/10.37074/jalt.2024.7.1.14

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

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1*(140216), 1-16. https:doi.org/10.1098/rsos.140216

Couldry, N., & Mejias, U. A. (2019a). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media, 20*(4), 336349. https://doi.org/10.1177/1527476418796632

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

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

Ferrara, E. (2023). *Should ChatGPT be biased? Challenges and risks of bias in large language models.* https://arxiv.org/abs/2304.03738

Gamage, K. A. A., De Silva, E. K., & Gunawardhana, N. (2020). Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences, 10*(301), 1-24. https://doi.org/10.3390/educsci10110301

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

Gillham, J. (2024). *Native English speakers?* https://originality.ai/blog/are-gpt-detectors-biased-against-non-native-english-speakers

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

Homolak, J. (2023). Exploring the adoption of ChatGPT in academic publishing: Insights and lessons for scientific writing. *Croatian Medical Journal, 64*(3), 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

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://arxiv.org/abs/2303.13408

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

Lee, D. (2023). *How hard can it be? Testing the reliability of AI detection tools.* https://www.researchgate.net/profile/Daniel-Lee-95/publication/374170650_How_hard_can_it_be_Testing_the_reliability_of_AI_detection_tools/links/6512b65237d0df2448edc358/How-hard-can-it-be-Testing-the-reliability-of-AI-detection-tools.pdf

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 1-4. https://doi.org/10.1016/j.patter.2023.100779

Lindgren, S. (2023). Introducing critical studies of artificial intelligence. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 1-19). Cheltenham: Edward Elgar Publishing.http://dx.doi.org/10.4337/9781803928562.00005

Makri, C., & Neely, A. (2021). Grounded theory: A guide for exploratory studies in management research. *International Journal of Qualitative Methods, 20,* 1-14. https://doi.org/10.1177/16094069211013654

Martin, J. L., & Wright, K. E. (2023). Bias in automatic speech recognition: The case of African American language. *Applied Linguistics, 44*(4), 613-630. https://doi.org/10.1093/applin/amac066

Mathewson, T. G. (2023). *AI detection tools falsely accuse international students of cheating*. The Markup. https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy &Technology, 33*, 659-684. https://doi.org/10.1007/s13347-020-00405-8

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

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). https://doi.org/10.53761/1.20.02.07

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22,* 89-113. https://doi.org/10.1007/s10805-023-09492-6

Ricaurte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media, 20*(4), 350-365. https://doi.org/10.1177/1527476419831640

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

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

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

Shane, J. (2023). *Don't use AI detectors for anything important*. Fortune. https://www.aiweirdness.com/dont-use-ai-detectors-for-anything-important/

Sobaih, A. E. E. (2024). Ethical concerns for using artificial intelligence chatbots in research and publication: Evidences from Saudi Arabia. *Journal of Applied Learning & Teaching, 7*(1), 1-11. http://journals.sfu.ca/jalt/index.php/jalt/index

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://journals.sfu.ca/jalt/index.php/jalt/article/view/731

Swedberg, R. (2020). Exploratory research. In C. Elman, J. Gerring, & J. Mahoney (Eds.), *The production of knowledge: Enhancing progress in social science* (pp. 17- 41). Cambridge: Cambridge University Press.

Uzun, L. (2023). ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. *Language Education & Technology, 3*(1), 45-54. https://www.researchgate.net/publication/370299956_ChatGPT_and_Academic_Integrity_Concerns_Detecting_Artificial_Intelligence_Generated_Content

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. https://doi.org/10.1515/opis-2022-0158

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

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/

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