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*(2), 25–43.
https://doi.org/10.56297/BUKA4060/XHLD5365

ISSUES

WHO WROTE THIS ESSAY? DETECTING AI-GENERATED WRITING in SECOND LANGUAGE EDUCATION in HIGHER EDUCATION

Katarzyna Alexander[1a], Christine Savvidou[1b], Chris Alexander[2c]

[1] Department of Languages and Literature, University of Nicosia, [2] Technology Enhanced Learning Centre, University of Nicosia

Keywords: Generative AI, higher education, academic writing, academic integrity, ESL, assessment

https://doi.org/10.56297/BUKA4060/XHLD5365

Teaching English with Technology

Vol. 23, Issue 2, 2023

Recent developments in AI technologies and the increasing accessibility of AI tools, such as ChatGPT, have raised concerns about academic integrity in higher education. Thus, this research aims to shed light on the challenges faced by English as a Second Language (ESL) lecturers in identifying AI-generated texts, and highlighting the skills and resources needed to enhance their detection capabilities. In this qualitative study, data were collected from six ESL lecturers working in a higher educational institution in Cyprus. Participants evaluated four academic essays at C1 level to determine which, if any, of the sample had been fully or partially AI-generated, and these results were then compared to results from four AI detectors. Findings reveal that participants tended to exploit a deficit model of assessment that focuses on error as an indicator of L2 writing output, with high levels of technical and grammatical accuracy and sophisticated language use as indicators of AI-generated text. Moreover, findings suggest that limited awareness of the characteristics and metrics used by ChatGPT, as well as lack of attention to the veracity of facts and references generated by ChatGPT, were features of participants' evaluations. This study identifies the growing challenges encountered by ESL lecturers and underlines the need for digital literacy training, targeted professional development, the use of advanced detection tools and a review of assessment policies and practices in relation to AI. Additionally, this study highlights the importance of reviewing and reinforcing institutional policies and practices that safeguard academic integrity and ensure quality higher education.

1. Introduction

This article considers the use of the generative Artificial Intelligence (AI) language model, ChatGPT and its emerging impact on higher education (HE); specifically, it explores the implications for English as a second language educators' current professional knowledge and skills in the assessment of academic writing; it reflects on the threats that such AI language models might pose to academic integrity and considers the implications for teacher education, going forward. To this end, the study is guided by the following research questions: (1) How effective are current plagiarism detectors in providing a reliable measure of detection for identifying originality in academic writing and distinguishing human-generated text from AI-generated text? (2) What criteria are employed by ESL lecturers to assess academic writing; (3) How effective are these criteria in distinguishing human-generated text from

a alexander.k@unic.ac.cy

b savvidou.c@unic.ac.cy

c alexander.c@unic.ac.cy

AI-generated text? Based on findings from the study, suggestions are made for improving practice, policies and training opportunities that respond to the opportunities and challenges presented by this new technology. Thus, the present study aims to contribute to the emerging literature on the impact of generative AI tools in HE. As such, the mass accessibility and potential of generative language models, such as ChatGPT, highlight the need to evaluate these technologies and re-evaluate existing academic policies and practices which currently govern their use.

2. Background

Since the launch of OpenAI's ChatGPT on November 30, 2022, there has been unprecedented interest in the potentiality of large language models (LLMs) as 'transformative digital technologies' for reshaping teaching and learning in all spheres of education, in general (Kasneci et al., 2023), and ESL, in particular (Perkins, 2023). The training of ChatGPT through the use of Supervised Learning and Reinforcement Learning from Human Feedback (RLHF), means that by comparing existing data based on the most frequent and relevant responses, ChatGPT is able to respond to prompts in different languages and adapt to instructed genres, styles and tone; moreover, with over 100 million active users (Milmo, 2023), the capabilities of ChatGPT continue to develop exponentially the more it is used.

Nevertheless, interest in ChatGPT is tempered with concern. Aside from existential questions that this technology may displace humans in many sectors of the labour market (Zarifhonarvar, 2023) and change the nature of human creativity, communication and critical thinking (Haleem et al., 2022; Kasneci et al., 2023), there are also fears that such technology can enable criminal and illegal activity, threaten data privacy and endanger cybersecurity (Dash & Sharma, 2023). Moreover, from a pedagogical perspective, the creation of false data (hallucinating), the misinterpretation of information, bias and plagiarism (Borji, 2022; Fostikov, 2023; Qadir, 2022) also pose significant threats to academic integrity. However, these and other weaknesses will likely be mitigated as the technology continues to develop (Borji, 2022).

3. Literature

Given both the novelty and exponential growth of generative AI technology, the research field is still in its infancy as it attempts to keep abreast of ever-changing advancements in this technology. As such, many studies have not been subjected to lengthy peer review processes associated with academic publishing. Moreover, many of the empirical studies reported below are small-scale with findings that cannot be generalized outside the place and time in which they were conducted since the research continues to be outpaced by rapid technological developments. Thus, it might be surmised that as generative AI continues to learn and develop, the research field does not, as yet, possess sufficient longevity or stability to draw definitive insights.

Against this background, the following review of recent research examines an educational landscape in which generative AI tools are becoming increasingly commonplace. Specifically, this review reports on three interrelated but distinct areas of research in relation to ChatGPT: (i) academic integrity, (ii) AI versus student performance, and (iii) stakeholders' perceptions and uses in education.

3.1. A threat to academic integrity?

A central question posed by the literature is whether the use of generative AI tools constitutes a breach of academic integrity. This concern focuses on the potential for generative AI to produce false content, plagiarize existing content, to create original content without human effort or expertise, and the potential for bias or discrimination in AI-generated text (Frye, 2022). Moreover, as early evidence suggests that a large percentage of generated text goes undetected, there are concerns that ChatGPT's ability to generate realistic human text represents a very real risk to the integrity of online exams (Susnjak, 2022), student coursework (Hong, 2023) and academic integrity (Aydın & Karaaslan, 2023). For instance, in the case of online plagiarism detection software, the literature suggests that current tools are not yet up to the task of reliably identifying AI-generated text. For example, in a study examining levels of plagiarism in essays produced by ChatGPT (Khalil & Er, 2023), two plagiarism detection tools (Turnitin and IThenticate) were used to evaluate the originality of 50 essays generated by ChatGPT. Results showed the two traditional plagiarism-detection tools used indicated high levels of originality for approximately 80% of the essays. These false negative results contrasted with the use of ChatGPT itself which showed greater accuracy at detecting AI-generated text for almost all the essays (96%) (Khalil & Er, 2023). Such results reflect a dynamic situation since online plagiarism detection systems are constantly being upgraded in response to the challenges presented by generative AI tools.

The situation is more critical in the case of human detection of AI-generated text. Analysis of a large sample of text (n=780) shows that raters' ability to identify AI-generated text decreased from 57.7% with ChatGPT2 to 49.9% with ChatGPT3 (Clark et al., 2021). In other words, the ability of human raters to identify AI-generated text is no higher than if by chance.

It is argued that, even when academic policies require students to acknowledge the use of generative AI tools such as ChatGPT, there is concern that their usage may represent academic misconduct in the form of student 'cognitive offloading' (Dawson, 2020), i.e., reducing the cognitive demands of the task through the use of technology. Essentially, it is argued that this already happens with students' use of digital writing assistants, e.g., Grammarly (Perkins, 2023). Indeed, such digital writing tools are not new (see Palmquist, 2003), and spell-checkers and grammar analysis tools are regularly used by students and accepted by educators as part of the process of improving student writing.

However, it is argued that the more substantial threat to academic integrity lies in learning objectives not being met when these are outsourced to LLMs, such as ChatGPT (Perkins, 2023).

Nevertheless, it is proposed that the ability of ChatGPT to provide satisfactory answers to academic questions merely reflects the superficiality of the questions posed by educators (Frye, 2022) and that while ChatGPT may level the playing field for non-native English language speakers, who are competing against native English speakers, it can only produce text that is as thoughtful, creative and reflective as the instructions it receives (Bishop, 2023). Thus, the need to design more academic writing tasks that require greater critical thinking is proposed as one solution for meeting learning objectives and maintaining academic integrity.

However, evidence also suggests that critical thinking can already be outsourced to generative AI tools. For example, a recent study asked ChatGPT to generate examples of questions directed to undergraduate students from various disciplines and then provide answers to those questions (Susnjak, 2022). These answers were then evaluated by ChatGPT for specific features of critical thinking (i.e., relevance, clarity, accuracy, precision, depth, breadth, logic, persuasiveness, originality). Findings showed that ChatGPT was able to generate answers that displayed high levels of critical thinking in faultless English suggesting that students are now able to generate highly competent text with relatively little input.

In sum, the extant literature suggests that concerns for academic integrity highlight the need to evaluate students' use of LLMs, in addition to reviewing institutional policies relating to academic integrity (Perkins, 2023).

3.2. Can ChatGPT outperform student writing?

Alongside issues of academic integrity, another concern is whether ChatGPT can actually outperform non-native English-speaking student writing. In a study by Bašić et al. (2023), 18 postgraduate Croatian students were divided into control (n=9) and experimental groups (n=9) to write short essays (800-1000 words). The control group wrote their essays traditionally and the experimental group used the assistance of ChatGPT. Both groups had four hours to complete their essays which were then scored by two professors using the same essay rubric. Findings showed that on average the experimental group took a slightly shorter time to write their essays (172 minutes versus 179 minutes). Both control and experimental groups obtained an average grade C with the control group gaining a slightly higher average grade. In terms of identifying text authenticity, all essays were submitted to an AI text detector which indicated that the experimental group had five cases that were identified as likely generated by AI compared to two cases likely generated by AI in the control group. Thus, this study suggests that the current quality of text detection is inadequate, identifying false positive cases in the control group and false negative cases in the experimental group. In sum, the study showed

that the experimental group did not outperform the student written output in terms of time or quality. Moreover, students in the experimental group found it more difficult than they expected to integrate AI-generated text into their writing requiring them to give consistent prompts.

These findings align with another study by Fyfe (2022), which shows that when students were asked to weave AI-generated text using ChatGPT2 into their essays, they reported numerous challenges, including the time it took to provide prompts that would elicit content that not only sounded 'natural' but was focused and accurate. Indeed, it is observed that without specific instruction, AI-generated text tends to be overly literal, have difficulty using idiomatic speech and generate over-detailed and comprehensive text. Additionally, answers were found to be overly neutral, rather than taking a more partial stance as more usual in human-generated text (Borji, 2022). Moreover, students also reported the capacity of ChatGPT to generate fabricated data, often providing quotes from non-existent sources. Incorrect, false and fabricated citations have also been reported in other studies (e.g., see Perkins, 2023).

3.3. Perceptions and uses in education

In the few months since its launch, a metanalysis of social media tweets (over 300,000) suggests that ChatGPT is overwhelmingly perceived and discussed with positive sentiment (Leiter et al., 2023). However, in the same study, an analysis of 150 academic papers in the scientific community indicated that, while ChatGPT is viewed as an opportunity for some domains, such as medicine, it is considered a threat for others, such as education (Leiter et al., 2023). Within education, there are also varying perceptions by different stakeholders.

Regarding the perceptions of academics, a small-scale qualitative study in Indonesia by Firaina and Sulisworo (2023) found that, overall, lecturers were positive about ChatGPT primarily using it to address limitations in their English proficiency, translating scientific articles, and searching for ideas to meet their specific requirements. However, limitations were also cited including the need to verify information and the need to use it selectively and critically.

For educators, Hong (2023) argues that ChatGPT provides many opportunities for language teaching and learning through its use as a personalized tool for learners to develop confidence and aptitude in language skills, its integration into classroom teaching as a way to provide authentic language examples, its ability to brainstorm and create outlines for classroom use, and as an automated grading systems that can provide useful feedback (see also Sok & Heng, 2023). For example, as an automated essay-scoring tool, ChatGPT can overcome issues of fatigue, inconsistency and unreliability associated with human scoring. One study used ChatGPT to grade 12,100 essays from a corpus of non-native written essays previously graded and stored

on a database (Mizumoto & Eguchi, 2023). Results showed high degrees of reliability with human raters (almost 90% agreement), with only a one-to-two-point difference. Moreover, it was also shown that when prompted, ChatGPT provided detailed and useful feedback for ESL educators and learners (Mizumoto & Eguchi, 2023).

For students, perceptions of ChatGPT are overwhelmingly positive with over 100 university students in Ghana citing their intentions to use it based on convenience, accuracy and improved academic performance (Bonsu & Baffour-Koduah, 2023). Significantly, students' knowledge of ChatGPT derives not from educational contexts but rather from social media and, specifically, TikTok. In an analysis of the top-100 most liked TikTok videos with the hashtag #chatgpt, videos focused on the uses of ChatGPT for essay writing, answering questions and writing code (Haensch et al., 2023). The content of these videos was overwhelmingly positive, ignoring any academic concerns such as cheating, plagiarism, inaccuracies in content and failure to meet learning objectives. It was also noted that a considerable number of these videos also alerted viewers on how to avoid detection by AI software. Haensch et al. (2023) argue that simply banning ChatGPT from educational settings will not circumvent any of these academic concerns. Rather, educational settings need to engage in greater discussion with students educating them of the responsible use, limitations, biases, inaccuracies and ethical considerations of ChatGPT.

In sum, in relation to academic integrity, emerging research suggests that a significant amount of AI-generated text remains undetected by plagiarism detection tools and the amount is even greater in the case of human raters. Additionally, the concept of students failing to achieve learning objectives by 'cognitive offloading' represents a very real threat to the foundations of HE. Moreover, in regards to whether or not AI-generated text can outperform student writing, the early literature is inconclusive with students citing reasons such as the time required to give effective prompts that meet task requirements, as well as the effort required to achieve 'human-sounding' text with reliable and verifiable content. Furthermore, scores awarded to AI-generated essays were also broadly comparable to those given to student writing. Finally, a review of this early literature shows that both academics and students view ChatGPT in a mostly positive light, recognizing its potential to support research, teaching and learning. However, while academics are generally cautious, acknowledging its limitations, students are overwhelmingly positive, valuing its convenience and ability to improve academic performance. Moreover, the literature suggests that most student knowledge comes from social media, highlighting the absence of classroom discussion about the responsible uses, risks, biases and inaccuracies related to the use of such AI tools.

4. Methodology

4.1. Research design and research questions

The present qualitative study is both exploratory and small-scale in its design. While there are limitations in terms of generalizability, it is hoped that such a study can provide early insights into this emerging field of technology and generate hypotheses that may be used to explore larger-scale projects in the future.

Taking into consideration the recent developments in LLMs and their potential influence on education, this study is guided by the following questions: (1) How effective are current plagiarism detectors in providing a reliable measure of detection for identifying originality in academic writing and distinguishing human-generated text from AI-generated text?; (2) What criteria are employed by ELS lecturers to assess academic writing?; and (3) How effective are these criteria in distinguishing human-generated text from AI-generated text?

With the view of answering the research questions, a study was devised in which a group of university ESL lecturers teaching C1 level English academic writing courses was asked to analyse four sample essays and determine whether they included any elements generated by ChatGPT. The study comprised three stages: (1) essay sample preparation, (2) analysis of sample essays with AI detectors, and (3) the analysis of texts by the ESL lecturers.

4.2. Generating essay samples

The study was based on a sample of four essays: three essays included varying amounts of AI-generated text and one essay that was fully written by a university student at a C1 level of English language proficiency during the Fall semester of 2022. All four essays covered subjects related to environmentalism and sustainability. Essays I, II and parts of III were generated using the free version of ChatGPT accessible between 17 and 18 January 2023.

Essay I, entitled 'The Impact of Deforestation on Russia's Environment: An Analysis of Current Trends and Possible Solutions', was fully generated with ChatGPT after prompting it to write an essay on deforestation in Russia (the subject was inspired by a theme of an existing essay by a university student). The prompt defined the length of the essay (1500-2000 words), the minimum number of sources to be cited and referenced and specified the general essay structure and format to be followed (e.g., introduction, thesis statement, cohesive devices, in-text citations, conclusion, APA format). The generated essay (681 words long) was used in the form in which it was generated by ChatGPT. The essay included five references.

Essay II, entitled 'Building a Sustainable Future: The Importance of Using Sustainable Materials in Construction', was generated based on an outline of an existing student essay on green buildings. ChatGPT was prompted to

Table 1. Study participants' background data

	Age	Years of teaching experience	Years of experience teaching C1	Highest qualification	Used ChatGPT
M1	35	5	5	PhD	No
M2	35	12	2	PhD	No
M3	41	5	5	MA TESOL	Yes
F1	41	20	20	PhD	No
F2	40	17	12	MA TESOL	No
F3	- [a]	20	6	Med	Yes

[a]The study participant did not include the information on her age in the background questionnaire.

generate each section separately and it was asked to focus each section on themes that were included in corresponding sections of the existing essay. The used prompts also asked for in-text citations to be included in each section. The generated sections of the essay were combined resulting in a 1151-word essay with a list of four references.

Essay III, entitled 'Sustainability in the Restaurant Industry: From Dish Preparation to the Restaurant', was a mixture of student-written and AI-generated text. More precisely, an essay written by a C1 student in an academic writing course was used as a base and some sections of the essay were replaced with ChatGPT-produced text. When prompted, ChatGPT was asked to create sections of text which corresponded to sections from the original essay. The prompts also asked for the text to include in-text citations. The final version of the essay was 1925 words long, of which 1164 (60.5%) was student work and 761 (39.5%) was ChatGPT-generated. The essay included seven references.

Essay IV, entitled 'Forest Fires: The Effects on Biodiversity', was an essay fully written by a C1 level student. The essay was 1416 words in length and included nine references.

It should also be noted that all four essays were formatted for uniformity (font and alignment, line spacing, etc.).

4.3. The participants

The data were collected from six participants, three males and three females (see Table 1). All six were educators with experience of working in higher education and teaching courses in academic writing. Their age ranged from 35 to 41 (average 38.4), years of teaching experience from five to 20 years (average 13.1 years) and years of experience teaching C1 level from two to 20 years (average 8.3 years). Three participants were PhD holders and the remaining three participants had a master's degree. Two participants declared prior experience of using ChatGPT before and four did not; however, one of the lecturers who stated that he did not use ChatGPT (M2) indicated that he was familiar with AI-generated texts. In order to preserve anonymity, all participants were given an identifier M1, F1, etc..

4.4. The task

The study participants were presented with the four sample essays and told that some of the essays were written, fully or partially, with AI. They were instructed to determine which essays, or parts of essays, were written with the use of AI. They were asked (1) to add comments using the Review tool in Word throughout each essay and (2) to leave comments at the end of each essay relating to the likelihood of the work being fully or partially student-written versus partially computer-generated giving reasons for the evaluations. The participants were not required to give a grade to any of the essays.

The participants were also provided with the requirements the students had been given before completing the essays and with the assessment criteria to be followed. The assessment criteria included the following: task achievement, coherence/cohesion, technical accuracy, research skills and critical thinking. They were instructed to spend approximately the same time that they would normally spend when grading essays at the C1 of the Common European Framework of Reference level and to evaluate the essays independently and to avoid consulting others.

Participants were asked to submit feedback within two weeks of receiving the four essays (in an electronic form). The data were collected in February 2023.

5. Findings

The following findings are organised according to the three research questions which guide this study and focus on the efficacy of several AI detectors, the criteria ESL lecturers use when assessing whether a text is created by a human or an AI and the accuracy of the criteria used by the lecturers.

5.1. Analysis of essay samples with AI detection software

In relation to the first research question (see 4.1), the essay samples were submitted for analysis to four software packages: (1) Turnitin (similarity report and AI detection tool), (2) OpenAI Detector, (3) GPTZero and (4) Crossplag (see Table 2). Each of the detectors generated a report which gives the likelihood of work being generated by AI. Table 2 presents the evaluations generated for each of the four essays by all four software packages using their terminology for identifying AI-generated texts (e.g. 'Fake', 'Real', 'highly likely to be human').

The analysis shows that all four detectors were accurate at detecting the use of AI work in Essay I (100% AI-generated). Crossplag, OpenAI Detector and Turnitin suggest between 97%, 99.3% and 100% of AI output respectively while GPTZero states 'Your text is highly likely to be written entirely by AI'.

AI detectors showed varying degrees of accuracy in the analyses of Essay II (100% AI). While OpenAI Detector stated the essay was 99.98% 'Fake' and Crossplag judged it was 100% written by an AI, Turnitin detected only 82% of AI work and GPTZero stated 'Your text includes parts written by AI'.

Table 2. Analysis with AI detection software packages

Essay	1	2	3	4
Description	100% AI	100% AI, written in separate sections according to instructions	39.5% AI mixed with 60.5% natural	100% Natural instance (student-generated)
Topic	Deforestation	Green buildings	Sustainable restaurants	Forest fires
Wordcount including references	758	1215	1203	1571
Turnitin similarity rating	5%	6%	61% (Detected only the student-written sections which are stored in the repository having been previously submitted)	2%
Turnitin AI detection	100%	82%	12%	0%
OpenAI detector Real/Fake	99.3% Fake	99.98% Fake	Gets stuck on 'predicting'	99.98% Real
GPTZero	Your text is highly likely to be written entirely by AI	Your text includes parts written by AI	Your text is most likely human-written but there are some sentences with low perplexities	Your text is most likely human-written but there are some sentences with low perplexities
Crossplag	97% Mainly written by an AI	100% Mainly written by an AI	1% Mainly written by a human	1% Mainly written by a human

In Essay III (39.5% AI), Turnitin detected only 12% of AI work, GPTZero assumed that most of the essay was 'most likely human-written' and added that 'there are some sentences with low perplexities', Crossplag noticed 1% of the AI-generated text and OpenAI Detector reported 'getting stuck' on processing the text and never provided any judgement.

Finally, Essay IV (100% natural) was assumed to be written (mainly) by humans by all detection packages, but GPTZero also pointed out that there were some sentences with low perplexities (see section 6) in the text.

These findings indicate that, while the tested AI detectors were accurate in evaluating text fully written by a human or fully produced by ChatGPT, their accuracy dropped significantly when assessing text generated fully by AI but using a human outline (e.g., Essay II) and text which was a mix of human and AI-generated text (Essay III). Overall, findings demonstrate that the efficiency of all tested packages was even lower when dealing with mixed (AI-generated and human-written) text.

5.2. Analysis by ESL lecturers

With respect to research question 2 (see 4.1), while the participants managed to identify some features of AI writing in the sample essays, overall, findings indicate that as a group, participants were less successful at distinguishing between AI and human-generated writing with accuracy rates between

Table 3. Summary of essay evaluations by participant

	Essay I: 100% AI	Essay II: 100% AI based on human-generated outline	Essay III: 39.5% AI mixed with 60.5% natural	Essay IV: 100% Natural instance (student-generated)
M1	AI	Mix	Mix	Mix (mostly student)
M2	Student[a]	Mix	Student	Mix (mostly AI)
M3	AI	AI	Student	Mostly student
F1	Student	Mix	Mix	Student
F2	Mix	Student	Mix	Student
F3	Mix	AI	Mix	Student
Accuracy	33%	33%	66%	50%

[a]Unless the AI was trained to write with an Intermediate level English

33-66%. As shown in Table 3, participants experienced more difficulties with the accurate identification of fully AI-generated essays than of the AI/human mix and a fully human-written essay. In sum, they tended to assess human writing based on technical accuracy and error and AI writing based on verbal complexity. The following discussion illustrates the ways participants determined which (sections of) texts were generated by ChatGPT in all four essay samples.

ESSAY I — AN ESSAY FULLY GENERATED BY CHATGPT

When assessing Essay I, which was fully generated by ChatGPT ('Deforestation'), out of the six study participants, two participants (M2, F1) stated that the text was fully written by a student, two assumed that it was generated by AI, and two believed that it was partially generated by AI. When explaining why they believed the essay or parts of it were written by a student, participants referred to weaknesses of the essay such as excessive repetition of vocabulary and ideas, lack of cohesion and proper transitions, and lack of a central thesis. The analyses in the essay were described by M2 as 'hasty and superficial'. The exclusive use of electronic sources was also one of the cited issues. M2 described their impression of the essay stating *'... this essay shows efficient use of very limited resources. This may indicate that the author is still in the process of mastering English as a means to write academically.'*

Participants (F2, F3) who judged the essay as a mix of human and AI-generated text believed that the AI-like features included the use of advanced vocabulary and the absence of technical errors. Among the weak elements giving the impression of AI being used were insufficient formatting and inadequate length of the text. Elaborating on the length and lack of depth being the indicators of AI work, F3 stated '*There is no discussion for each main point but rather a very direct answer to a question*.'

The features that made participants (M1, M3) believe the essay was fully written by AI were a high level of language used, the lack of typical writing mistakes made by the local students, and at the same time superficial analysis of the presented arguments, inadequate formatting and exclusive use of online

sources. M3 stated *'The combination of the complete lack of errors in grammar, spelling, syntax, the vocabulary used and the correct in-text citations, along with the very limited and generic sources used, make me believe that this paper was not written by a student. Most probably, the student provided chatgpt with separate points of the thesis statement.'*

ESSAY II – AN ESSAY FULLY GENERATED BY CHAT GPT BASED ON A HUMAN-GENERATED OUTLINE

The essay which was produced using ChatGPT based on a human-written outline ('Green Buildings') was assumed to be fully AI-generated by two participants, fully student-written by one participant and a mix of AI and human work by the remaining three.

The participant (F2) who assumed the essay was fully written by a student, stated that what convinced her was the fact that the essay was shorter than expected, did not use enough sources and lacked headings. Other features that seemed human-like were cohesive writing and in general, it displayed '*writing expected from a good student*' (F2).

Participants (M1, M2, F1) who believed the essay was partly prepared by a human cited 'human-like' characteristics, such as the presence of language errors and errors in in-text citations, lack of focus on main points, but also a clear purpose guiding each paragraph and the attempt to give each paragraph the required structure. The elements that made those respondents suspect a partial use of AI were the use of correct citations along incorrect citations and the use of complex sentences along simple (supposedly student-generated) sentences. Other specific cues were not mentioned, only general suspicions were noted such as '*I suspect the author might have used the help of AI, but the essay is not entirely composed by it*' (M2). Participants (M3, F3) who attributed the authorship solely to AI, drew attention to the fact that the task instructions given by the lecturer were not followed by the essay author. Participants also, observed the absence of language errors, a lack of consistency and accuracy relating to in-text citations, and a limited number of references. One of the participants also pointed out a similar pattern in the structure of the majority of paragraphs which started with a transitional phrase, a topic sentence and then a citation. Finally, according to F3 the text *'provides information in a general manner with no specific concentration. Writing is "robotic".'*

ESSAY III – AN ESSAY GENERATED PARTLY BY A HUMAN AND PARTLY BY CHATGPT

The paper which was only partly generated by ChatGPT ('Sustainable Restaurants') was assessed as a mix of student and AI work by 4 participants (M1, F1, F2, F3) and as a fully student-generated work by 2 (M1, M2).

Participants who judged the essay as a mix of student and AI writing cited the mix of sections written with good language and sections with language errors. The former sections were correctly identified as generated by AI and the latter

sections written by a student. M1 wrote '*There are a lot of language errors throughout the essay. But, surprisingly, some sections do not include errors, which makes me believe that the essay is partly-student and partly computer-generated*.' Some respondents correctly attributed examples of incompletely developed ideas and repetition as features of student writing in human-generated sections. The use of advanced vocabulary in the AI sections was ascribed to ChatGPT or to a student paraphrasing external sources. As an example, when discussing one of the AI-generated sections, F1 stated '*The vocabulary is more advanced here, so it could be through the ChatGPT or because the student is paraphrasing, therefore language from the original source might be used*.'

Participants who were more inclined to think that the essay was written by a student based their evaluations on some linguistic and technical weaknesses of the paper such as language mistakes, messy structure, cohesion issues, the use of citations that did not logically match the rest of the content and overgeneralisations. The authorship of this paper was also ascribed to student writing on the length and depth of the essay which was evaluated as adequate, despite the use of few sources. Commenting on Essay III, M2 wrote '*this attempt has a relatively low likelihood of having been composed by AI. I can see an attempt to deliver an idea on the topic of sustainable restaurants, but the author is all over the place. The subsections are not sufficiently cohesive and the citations does not make much sense in some places*.'

ESSAY IV – A HUMAN-WRITTEN ESSAY

Half of the participants (F1, F2, F3) classified the student-written paper ('Forest Fires') as fully written by a human. Out of the remaining three, two respondents (M1, M3) felt the essay was a mix of human and AI writing, whereas one participant (M3) believed the essay was mostly generated by AI. Those who believed the paper was fully or partially authored by a human as their cues pointed out spelling mistakes, inadequate quoting methods, repetition, problems with coherence and undeveloped vocabulary. Some mentioned the lack of a thesis statement and the use of slightly outdated sources. The way the subject was analysed was also assessed: *'Most arguments are simply stated rather than analysed or compared' (F2)*. The fact that a lecturer's instructions were followed and that the subject referred to Cyprus (data collected in Cyprus) made some attribute the authorship to a person.

The two respondents who judged the paper as mostly student-written but possibly with some use of AI pointed out language errors, issues with punctuation, repetition, and poor paraphrasing as characteristics of student writing. The features that gave them an impression of AI writing were the use of 'ibid.' in in-text citations and sudden changes in writing style.

The respondent who described the paper as mostly written by AI quoted the superficiality of arguments used in the essay. He stated '*My experience with AI-generated texts is that they are not usually wrong in terms of language use, but they meander through topics in a somewhat superficial manner*' (M2). He also

wrote that the arguments in that essay were explored artificially. '*The author meanders through loosely related topics without a concise focus, but without being too bad at it either, just like the AI texts I'm familiar with*' (M2).

5.3. Hallucinations

Findings suggest that whilst analysing the texts, the participants paid attention to features of language ranging from spelling mistakes to the ways sentences were formed and the way the subjects were investigated; however, none of the participants reported on fact-checking. It has been widely reported in the literature, that ChatGPT has the tendency to hallucinate, i.e. make up facts and references that do not exist (for example, see Borji, 2022; Fostikov, 2023; Qadir, 2023). It should be noted that participants did not report whether they fact-checked the information included in the essays and/or what they found out, if they did.

As a matter of fact, several references listed in the sample essays could not be verified (see Table 4). Namely, in Essay I, 100% of referenced sources could not be found, in Essay II, 75%, Essay III, 14% (the one that was not found is one of the two references provided by ChatGPT) and in Essay IV, 11%. These hallucinated references were neither identified nor reported in participants' evaluation of the essay samples.

Table 4. Number of existing and non-existing (i.e. 'hallucinated') sources referred to in sample essays

Essay	Number of references	Existing	Non-existing	
I	5	0 (0%)	5 (100%)	
II	4	1 (25%)	3 (75%)	
III	7	6 (86%)	1 (14%)	The one that was not found is one of the two references provided by ChatGPT
IV	9	8 (89%)	1 (11%)	

6. Discussion

The data collected facilitated investigation into (1) the effectiveness of some of the available AI detectors (Turnitin, OpenAI Detector, GPTZero and Crossplag); (2) criteria employed by ESL lecturers to assess the originality of academic writing; and (3) the effectiveness of the criteria in distinguishing between human and AI-generated texts.

The brief analysis of four AI detectors' performance based on the four essay samples indicates that they are highly accurate, although to varying degrees, in recognising fully AI-generated texts. In this small sample, no false-positive or false-negative results were detected. However, they are not as accurate in detection when working on texts that are a combination of human and AI-produced samples. While this has serious implications for the reliability of AI detectors in education, technological advances are currently being made in this specific area. Regarding the participants' analysis of sample texts, the results clearly show that participants' attempts to differentiate between AI

and human-generated texts were not as accurate as the AI detection software and that their accuracy ranged between 33% and 66%. As regards the criteria employed by participants, these mainly focused on language features ranging from spelling to sentence structure, and also on the depth of analyses included in essays.

Their feedback shows that there was also a prevalent expectation among the participants, all experienced lecturers, that the texts written by AI would be flawless while inadequacies and errors were largely attributed to human authors. Hence, on the whole, AI was associated with the use of advanced vocabulary, complex sentences, high level of language skills, lack of errors and correct citations. Conversely, student writing was associated with limited language skills, repetition in vocabulary and sentence structure, poor cohesion, inadequate essay structure and mistakes in referencing. This outlines lecturers' expectations of AI being a flawless and sophisticated writer while L2 students have limited writing skills and language resources.

These expectations are in contrast to how GPTZero, an AI detector, is reported to function. While assessing writing, GPTZero analyses two specific variables – perplexity and burstiness. Perplexity measures the predictability of strings of words in texts, whereas burstiness assesses the level of complexity of sentences. Based on these metrics, GPTZero assumes that AI writing will have a lower level of perplexity and burstiness, thus more predictable word combinations and predictable, repetitive sentence structures. This is believed to be different for human writing, which is more likely to have higher levels of perplexity and burstiness (Singh, 2023). In contrast, findings suggest that the participants in the current study seemed to have contrary expectations of AI and human performance. As an example, repetition in sentence structures and vocabulary used was identified as a human-like error. On the contrary, such repetition could be classified as showing lower levels of perplexity (repetitive vocabulary) and burstiness (repetitive sentence structure), which is considered typical of AI-generated writing. There was a tendency amongst participants to assign high levels of perplexity and burstiness in writing to the use of AI when, typically, the opposite tends to be the case.

It must be highlighted that one out of all six participants noticed a link between lower levels of perplexity and burstiness to the use of AI. F3 described writing in Essay II (100% AI-generated) as 'robotic', which might be linked to low burstiness. She was one of the two participants who had used ChatGPT before the experiment and may therefore have been more familiar with ChatGPT output. Such ability of a lecturer to recognise AI features in writing, possibly due to her previous exposure to ChatGPT and its output, indicates that it could be beneficial for other ESL lecturers to become familiarised with AI-generated texts and their features, including those referring to the levels of burstiness and perplexity. Yet, this assumption should be treated with caution

as it was arrived at based on a small sample. Whether having experience with AI indeed is a factor in making it easier for some to spot AI's work is not certain, but this is an observation that needs a follow-up with further research.

Another criterion that was raised in participants' reports was the superficiality of discussion in certain essays. F3, a participant who was familiar with ChatGPT output prior to taking part in the study, linked the superficiality of writing to AI use and she did that when discussing a fully AI-generated work. Interestingly, M2, who stated he was familiar with AI-written texts, attributed the superficiality of discussions to AI work in Essay IV which was fully written by a human. Thus, in this sample, being familiarised with some features of AI writing did not prevent a participant from mistaking student work for AI-generated text.

The results show that the study participants did not comment on possible AI hallucinations or fake sources used in the text samples. While conclusions might be drawn that the participants were not aware of AI's tendency to hallucinate, it is unclear what is the cause for fake sources not being reported. The task did not explicitly ask for comments on the accuracy of the information included in the essays; rather, it generally focused on assessing student performance and detecting the use of AI tools. It was felt that too detailed instructions might have influenced the evaluations given by the study participants. It is also possible that the participants may have noticed the inaccuracies but did not report them.

Conclusion

To sum up, the study indicated that currently there seems to be no fully reliable way of establishing whether a text was written by a human or generated by an AI. Neither humans nor AI detectors proved able to detect AI efficiently and reliably although AI detectors appeared to be more reliable than human raters. It was also established that human evaluators' expectations of AI texts differ from what in reality is generated by ChatGPT. There appears to be a prevalent expectation that an AI-generated essay will be flawless and sophisticated. Not only were participants, who were all experienced and specialized writing educators, unable to recognise all AI-written sections, but they are also likely to confuse student-written work for AI-generated text and to produce unreliable evaluations.

While the small scale nature of this exploratory study does not allow any serious generalisations about human and/or AI detectors' potential for assessing the origin of written texts, its results indicate implications regarding further research and teaching policies. Hence, it seems essential to carry out research focusing on reviewing, evaluating and revising existing assessment policies and procedures and implementing policies and procedures fit for purpose. Furthermore, there is a need for large scale studies across higher education institutions examining necessary changes to policies and practices in regard to developments in AI and perceived risks to academic integrity. Finally, it

seems essential to examine attitudes towards and use of AI by stakeholders in education – students, educators and administrators. Concerning the implications for policy and practice, the results demonstrate that to improve practice and to safeguard academic integrity, there is a need for digital training for educators with respect to the use of AI tools in the ESL classroom and the use of advanced detection tools. There is also a need for revision of assessment policies and procedures and the development of suitable rubrics and assessment criteria. Lastly, lecturers' tendency to confuse student-written essays for AI-generated text points out the need for lecturers to be trained on the features of AI-created texts and the differences between them and human-written essays.

Acknowledgement

We would like to thank the six ESL lecturers who participated in this study. We are also grateful to the anonymous reviewers for their constructive comments on our paper.

REFERENCES

Basic, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). *Better by you, better than me? ChatGPT-3 as writing assistance in students' essays*. arXiv. https://arxiv.org/abs/2302.04536

Bishop, L. (2023). A computer wrote this paper: What ChatGPT means for education, research, and writing. *SSRN*. https://doi.org/10.2139/ssrn.4338981

Bonsu, E., & Baffour-Koduah, D. (2023). From the consumers' side: Determining students' perception and intention to use ChatGPTin Ghanaian higher education. *SSRN*. https://doi.org/10.2139/ssrn.4387107

Borji, A. (2022). *A categorical archive of ChatGPT failures*. arXiv. https://arxiv.org/abs/2212.09292

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv*. https://doi.org/10.48550/arXiv.2107.00061

Dash, B., & Sharma, P. (2023). Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? *A Review. International Journal of Engineering and Applied Sciences*, *10*(1). https://www.ijeas.org/download_data/IJEAS1001001.pdf

Dawson, P. (2020). Cognitive Offloading and Assessment. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining University Assessment in a Digital World* (pp. 37–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-41956-1_4

Firaina, R., & Sulisworo, D. (2023). Exploring the usage of ChatGPT in higher education: Frequency and impact on productivity. *Buletin Edukasi Indonesia*, *2*(01), 67–74. https://doi.org/10.56741/bei.v2i01.310

Fostikov, A. (2023). First impressions on using AI powered chatbots, tools and search engines: ChatGPT, perplexity and other–possibilities and usage problems. *Humanities Commons*. https://hcommons.org/deposits/item/hc:51415

Frye, B. L. (2022). Should using an AI text generator to produce academic writing be plagiarism? *Fordham Intellectual Property, Media & Entertainment Law Journal*. https://ssrn.com/abstract=429228

Fyfe, P. (2022). How to cheat on your final paper: Assigning AI for student writing. *AI & Society*, 1–11. https://doi.org/10.1007/s00146-022-01397-z

Haensch, A.-C., Ball, S., Herklotz, M., & Kreuter, F. (2023). Seeing ChatGPT through students' eyes: An analysis of TikTok data. *arXiv*. https://doi.org/10.48550/arXiv.2303.05349

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research. *Journal of Educational Technology and Innovation*, *5*(1). https://jeti.thewsu.org/index.php/cieti/article/view/103

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*(102274), 102274. https://doi.org/10.1016/j.lindif.2023.102274

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

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). *ChatGPT: A meta-analysis after 2.5 months*. arXiv. https://arxiv.org/abs/2302.13795

Milmo, D. (2023). ChatGPT reaches 100 million users two months after launch. *The Guardian*. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *SSRN*. https://doi.org/10.2139/ssrn.4373111

Palmquist, M. (2003). A brief history of computer support for writing centers and writing-across-the-curriculum programs. *Computers and Composition*, *20*(4), 395–413. https://doi.org/10.1016/j.compcom.2003.08.013

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

Qadir, J. (2022). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv*. https://doi.org/10.36227/techrxiv.21789434.v1

Singh, A. (2023). A comparison study on AI language detector. *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 0489–0493. https://doi.org/10.1109/ccwc57344.2023.10099219

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *SSRN*. https://doi.org/10.2139/ssrn.4378735

Susnjak, T. (2022). *ChatGPT: The End of Online Exam Integrity?* arXiv. https://arxiv.org/abs/2212.09292

Zarifhonarvar, A. (2023). Economics of ChatGPT: A labor market view on the occupational impact of artificial intelligence. *SSRN*. https://doi.org/10.2139/ssrn.4350925