CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/161240/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Hassoulas, Athanasios, Powell, Ned, Roberts, Lindsay , Umla-Runge, Katja , Gray, Laurence and Coffey, Marcus 2023. Investigating marker accuracy in differentiating between university scripts written by students and those produced using ChatGPT. Journal of Applied Learning & Teaching 6 (2) 10.37074/jalt.2023.6.2.13 file

Publishers page: https://doi.org/10.37074/jalt.2023.6.2.13

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Vol.6 No.2 (2023)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

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

Investigating marker accuracy in differentiating between university scripts written by students and those produced using ChatGPT

Athanasios Hassoulas[A] A Cardiff University, Cardiff, United Kingdom

Ned Powell[B] B Cardiff University, Cardiff, United Kingdom

Lindsay Roberts[C] C Cardiff University, Cardiff, United Kingdom

Katja Umla-Runge[D] D Cardiff University, Cardiff, United Kingdom

Laurence Gray[E] E Cardiff University, Cardiff, United Kingdom

Marcus J Coffey[F] F Professor, Cardiff University, Cardiff, United Kingdom

Keywords

Assessment;
ChatGPT;
digital education;
generative AI;
higher education;
medical education.

Correspondence

HassoulasA2@cardiff.ac.uk [A]

Article Info

Received 16 July 2023
Received in revised form 21 July 2023
Accepted 22 July 2023
Available online 25 July 2023

DOI: https://doi.org/10.37074/jalt.2023.6.2.13

Abstract

The introduction of OpenAI's ChatGPT has widely been considered a turning point for assessment in higher education. Whilst we find ourselves on the precipice of a profoundly disruptive technology, generative artificial intelligence (AI) is here to stay. At present, institutions around the world are considering how best to respond to such new and emerging tools, ranging from outright bans to re-evaluating assessment strategies. In evaluating the extent of the problem that these tools pose to the marking of assessments, a study was designed to investigate marker accuracy in differentiating between scripts prepared by students and those produced using generative AI. A survey containing undergraduate reflective writing scripts and postgraduate extended essays was administered to markers at a medical school in Wales, UK. The markers were asked to assess the scripts on writing style and content, and to indicate whether they believed the scripts to have been produced by students or ChatGPT. Of the 34 markers recruited, only 23% and 19% were able to correctly identify the ChatGPT undergraduate and postgraduate scripts, respectively. A significant effect of suspected script authorship was found for script content, $X^2(4, n=34) = 10.41$, $p<0.05$, suggesting that written content holds clues as to how markers assign authorship. We recommend consideration be given to how generative AI can be responsibly integrated into assessment strategies and expanding our definition of what constitutes academic misconduct in light of this new technology.

Introduction

The use of technology in enhancing coursework submissions is by no means a new trend. From Microsoft Word's spell check and autocorrect to the more recent use of products such as Grammarly, the use of such tools has significantly improved our ability to produce well-structured written documents with the aid of inbuilt spelling and grammar assistants (Behrens et al., 2019). Artificial intelligence (AI) technology dates back decades to platforms such as ELIZA, which utilised early language models to engage in conversation, but more sophisticated generative AI technology is now capable of producing written scripts that pose a problem for higher education assessments (Rudolph et al., 2023a).

The introduction of OpenAI's ChatGPT (Generative Pre-Trained Transformer), in particular, has been viewed as a watershed moment in higher education due to the ability of the tool, through the large language model (LLM) it employs, to learn rapidly and develop sophisticated responses to a range of instructions. Objectively, ChatGPT is, therefore, the first such LLM that has captivated a global mainstream audience (Hosseini et al., 2023). The various applications of this technology for educators, researchers and students have been demonstrated impressively through a published journal article written by the chatbot on what its existence means for higher education (Bishop, 2023).

The consensus within global higher education is that the technology is here to stay and will have profound consequences for assessment strategies across all programmes of study. Immediate discussions and challenges will pertain to updating our definition, or perhaps redefining, terms such as plagiarism and academic integrity in light of this revolutionary technology (e.g., Debby et al., 2023). The advantages that this new technology also presents, however, cannot be ignored. Not only do disruptive tools such as ChatGPT provide an ideal opportunity to modernise certain outdated assessment practices, but they may, when used appropriately, significantly enhance students' learning experiences and productivity (Fauzi et al., 2023). Indeed, the technology may revolutionise the manner in which students learn and work academically.

Conversely, in the context of academic integrity, others assert that this new technology may not be as disruptive as is currently anticipated (Cotton et al., 2023), and some have suggested that this potential issue could be addressed by replacing some assessments with formats that require evidence of reflective practice by students. However, even without further evolution, it appears likely that even the current commonly available generative AI tools may be capable of deceiving coursework markers reviewing reflective student scripts as well as essay-type assessments.

To our knowledge, there has been no published study to date comparing marker accuracy in differentiating between human-written coursework submissions and AI chatbot-generated scripts in both essay-type scripts and reflective writing tasks. On this basis, we designed a study that included original student submissions and scripts generated by ChatGPT-3.5 and then investigated the performance of experienced coursework markers in terms of how they graded the assessments, as well as determining whether they could accurately differentiate between the student submissions and ChatGPT scripts.

Materials and methods

Participants

A total of 34 experienced academic and clinical academic coursework markers from a medical school in Wales were recruited to participate in this study. Participants were presented with undergraduate reflective writing submissions and postgraduate extended essays. Participants had the option to review just the reflective pieces, just the essays, or both, and were asked to review the submission formats they routinely marked. 23 participants marked the undergraduate reflective writing submissions, and 22 participants marked the postgraduate essay scripts (11 participants marked both undergraduate and postgraduate scripts). Participant confidentiality and response anonymity were assured. Consent was provided by all participants, as well as by the students whose scripts were anonymously included as examples of undergraduate reflective writing and postgraduate essays, with all identifiable information removed before being included in the survey.

Materials and procedure

The survey was designed in and disseminated using the digital survey platform Online Surveys (formerly Bristol Online Surveys). Three undergraduate reflective writing scripts were presented, along with three postgraduate extended essay scripts. Each undergraduate reflec-tive writing script was approximately 1,500 words in length, whilst the postgraduate extended essays were each approximately 3,000 words in length. Of the three reflective writing scripts, two were student submissions, and one was generated using ChatGPT-3.5. Equally, two of the postgraduate essays were student submissions, whilst one essay was generated using ChatGPT-3.5. For the undergraduate reflective writing task, the wording of the instructions provided to students was identical to the ChatGPT prompt, but the latter included additional information on a specific clinic experience to base the ChatGPT-generated reflective script on (since the undergraduate students based their reflections on actual patients that they encountered whilst on clinical placement). For the postgraduate extended essay, the wording of the ChatGPT prompt was identical to the instructions that students assigned this specific essay topic received.

Consent was captured before participants were permitted to proceed to the next part of the survey, where they considered the various scripts provided. After participants read each script, they were asked three questions. Initially, they were asked to grade each script on the basis of writing style as well as in terms of content. Four grading options were provided: Excellent, Good, Adequate, and Poor. Participants were then asked whether they suspected the script was written by a student, generated using ChatGPT-3.5, or whether they didn't know either way. An open-ended, free text item was

also included asking participants to provide a brief rationale as to why they may have felt the script was authored by a student or generated using ChatGPT-3.5. A debrief was provided at the end of the survey. Ethical approval was sought at provided by the School of Medicine Research Ethics Committee (SMREC 23/38).

Data analysis

As quantitative and qualitative data were collated using the online survey, a mixed methods cross-sectional study design was deemed appropriate. A Chi-square test was run in IBM SPSS (version 27), given the non-parametric, categorical nature of the quantitative data collated. The open-ended qualitative data collated from the free text items were analysed using content analysis; all written responses provided by participants were carefully reviewed. Content analysis has been identified as being well-suited to research in qualitative healthcare education (e.g., Downe-Wamboldt, 1992; Hassoulas et al., 2023).

Results

Analysis of quantitative data

Participants rated each script on the basis of writing style and script content on a four-point scale from Excellent to Poor. They were also asked to identify the author of each script as either human, a chatbot or to declare whether they were uncertain as to script authorship. Overall, for the undergraduate reflective writing scripts, 50% of participants correctly identified the two student submissions, whilst only 23% correctly identified the ChatGPT script. In addition, 59% of participants incorrectly attributed authorship of the student submissions to ChatGPT. This suggests that a larger proportion of markers attributed authorship of the student scripts to the generative AI tool. This further highlights the difficulty that even experienced markers may experience in differentiating between scripts that are authored by students and those prepared using such generative AI tools (see Table 1).

Table 1: Undergraduate marker responses in differentiating student reflective submissions from ChatGPT-3.5 scripts.

Actual author	Marker assessment of likely author		
	Student	ChatGPT	Uncertain
Student	50	22	28
ChatGPT	59	23	18

A similar picture emerged for the postgraduate extended essay scripts, with 50% of markers correctly identifying the two student submissions once again but with only 19% correctly identifying the ChatGPT script. The degree of uncertainty in identifying authorship, however, was higher for the postgraduate markers than those who marked the undergraduate scripts. Specifically, 37% of participants who considered the postgraduate scripts were uncertain as to whether the ChatGPT script was written by a human or by the chatbot, as compared to just 18% of participants who considered the undergraduate scripts being uncertain as to the authorship of the ChatGPT script (see Table 2).

Table 2: Postgraduate marker responses in differentiating student extended essay submissions from ChatGPT-3.5 scripts.

Actual author	Marker assessment of likely author		
	Student	ChatGPT	Uncertain
Student	50	31	19
ChatGPT	44	19	37

Categorical data collated on participants' assessment of script writing style and content were analysed using chi-square tests. There was a significant effect of author identification for content specifically, $X^2(4, n=34) = 10.41$, $p<0.05$, but not for writing style ($p>0.05$). Interesting-ly, participants graded the undergraduate reflective writing submissions slightly lower on content than they did the ChatGPT script, whilst postgraduate extended essay student content was marked higher in comparison to the content generated by ChatGPT.

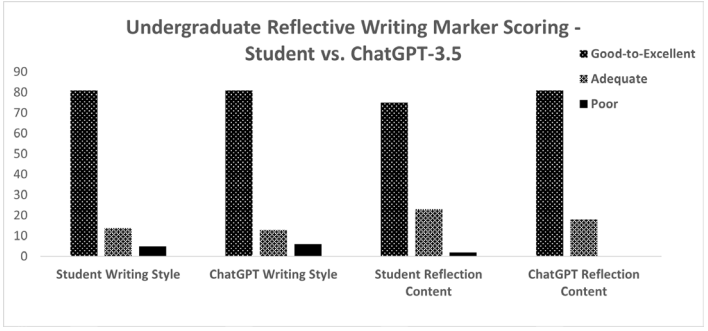

Figure 1. Marker assessment of undergraduate student submissions and the ChatGPT scripts.

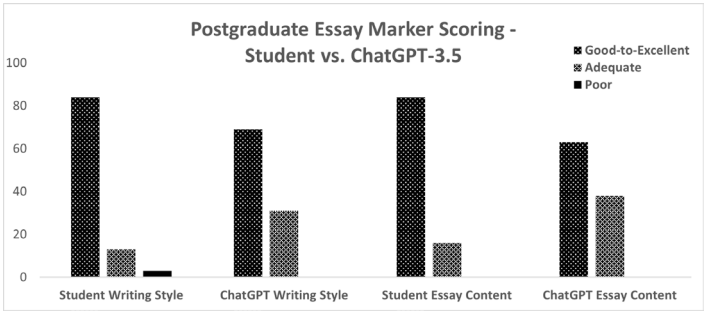

Figure 2. Marker assessment of postgraduate student submissions and the ChatGPT script.

These findings suggest that whilst writing style was statistically indistinguishable between human scripts and ChatGPT texts, script content does appear to hold certain clues as to how generative AI performed on this specific domain and whether experienced markers are able to identify clues to authorship in coursework content (see Figures 1 and 2).

Content analysis of qualitative data

Free-text responses by participants to the open-ended items included in the survey were considered in relation to the re-occurrence of key terms. As such, content analysis was performed to gain granular insight into what markers

identified as key features that influenced their responses. Four key themes emerged, with the use of language (including grammar, phrases and expressions, and syntax) accounting for more than half of all free-text responses (see Table 3).

Table 3. Content analysis frequency table of key themes identified by markers for undergraduate reflective writing scripts prepared by students.

	Use of Language	Personal & reflective	Structure and writing style	Referencing
Total Frequency (%)	*56*	*24*	*15*	*5*
Student Identified	*29*	*12*	*7*	*2.5*
ChatGPT Identified	*17*	*7*	*2.5*	*0*
Authorship Uncertain	*10*	*5*	*5*	*2.5*

In relation to the ChatGPT-3.5 constructed script, markers once again identified the use of language as a key factor influencing their decision as to whether the script was written by a human or the chatbot. The same proportion of markers who alluded to the use of language in their responses had identified the author of this script as being human, too, suggesting an inability to accurately and confidently distinguish between student-specific language and proficiency versus the language being produced by the chatbot in response to the instructions provided. Furthermore, the ChatGPT-3.5 script, in particular, revealed that fewer markers suspected the use of language within the script to be suggestive of generative AI. A larger proportion of markers, however, emphasised that they found it difficult to identify the author as being human or a chatbot based solely on inconsistencies detected in the use of language (see Table 4).

Table 4. Content analysis frequency table of key themes identified by markers for undergraduate reflective writing scripts prepared by ChatGPT-3.5.

	Use of Language	Personal vs. formulaic	Structure and writing style	Referencing
Total Frequency (%)	46	32	18	4
Student Identified	28	9	18	0
ChatGPT Identified	4	14	0	0
Authorship Uncertain	14	9	0	4

Structure and writing style were a theme identified by markers in the context of the ChatGPT-generated script as well, although all those who made reference to structure and style of writing incorrectly identified the author as being a student. This suggests that the structure of the script and style of writing deceived markers regarding the identity of the author, with a large degree of certainty, as being human. As such, generative AI may be beneficial to students as a tool to improve the structure and style of academic writing.

Markers once again identified personalised writing as a key theme. However, more reference was also made to the reflection appearing more "formulaic" in the ChatGPT script. The largest proportion of markers referring to clues identified in terms of the personal and reflective nature of the writing considered the script to have been chatbot-generated. This suggests that in the context of reflective writing, generative AI tools are yet to master the ability to deceive markers specifically in relation to the depth of reflective practice demonstrated.

Inconsistency with regard to referencing, and sources cited for which no actual reference could be located, were identified as a key theme influencing markers' suspicions as to the authorship of the respective script. For the reflective writing scripts written by students, none of the markers who alluded to referencing identified the author of the scripts as being the chat-bot, whilst for the ChatGPT-generated script, markers responded with greater uncertainty regarding authorship but did not confidently identify the script as being authored by a student. This suggests that currently, citations and referencing may hold clues as to the authorship of scripts.

Regarding the postgraduate extended essays, markers once again identified the use of lan-guage as being a key factor in considering authorship, particularly in the context of the ChatGPT-generated script (see Table 4) as opposed to the student scripts (see Table 3), where the use of language was the second most common theme. Whilst almost half (47%) of post-graduate markers referred to the use of language in the context of the ChatGPT-generated script, no marker suspected the language used as being suggestive of generative AI use, with 33% suspecting the author of being a student whilst 14% reported that they were uncertain as to the authorship of the script. This is in contrast to the student-written scripts, where those who made reference to the use of language mostly identified the scripts as being written by students, with a lower degree of uncertainty regarding authorship overall.

The structure and layout of the extended postgraduate essays were identified as the most frequent theme referred to by markers when considering the student-written script, with the majority also correctly identifying the authors of the scripts as being human. This particular theme was only the third most frequently referred to by the same group of markers in considering the ChatGPT-generated script, with no markers, however, correctly identifying the author of that particular script as being the chatbot. Whilst themes such as the use of language as well as structure and layout were commonly referred to by markers, inaccuracy in differentiating between human and ChatGPT scripts remained an issue.

Knowledge and appraisal of the literature was a key theme identified in postgraduate markers. However, as with the structure and layout theme, inaccuracy in differentiating between student scripts and AI-generated scripts was problematic on this basis too. It was in relation to citations and referencing once again (as with the undergraduate reflective writing scripts) where differentiating between student-written and ChatGPT-generated scripts did appear to yield more promising results in accurately identifying authorship. Whilst only the fourth most frequently considered theme, no markers who alluded to referencing in relation to the ChatGPT script suspected student involvement. Equally, for the student-written essays, the majority who alluded to

referencing suspected that the scripts had been written by students (see Table 5).

Table 5. Content analysis frequency table of key themes identified by markers for postgraduate extended essay scripts prepared by students.

	Structure and Layout	Use of Language	Knowledge & Appraisal	Referencing	Construction and style
Total Frequency (%)	30	19	19	19	14
Student Identified	19	11	11	11	11
ChatGPT Identified	5.5	3	8	3	3
Authorship Uncertain	5.5	5	0	5	0

An additional theme identified by markers in the context of the student-written scripts was that of construction and style, with the majority of markers considering this particular theme correctly identifying the author of the scripts as being human (see Table 3). The same cohort of markers did not refer to construction and style, however, in the context of the ChatGPT-generated script (see Table 6).

Table 6. Content analysis frequency table of key themes identified by markers for postgraduate extended essay scripts prepared by ChatGPT-3.5.

	Use of Language	Knowledge & Appraisal	Structure and Layout	Referencing
Total Frequency (%)	47	27	13	13
Student Identified	33	20	13	0
ChatGPT Identified	0	0	0	6.5
Authorship Uncertain	14	7	0	6.5

Discussion and conclusion

Our results suggest that experienced markers are currently unable to consistently differentiate between student-written scripts and text generated by natural language processing tools, such as ChatGPT. This appears to be the case for both undergraduate reflective writing tasks as well as postgraduate extended essays that form respective key components in undergraduate and postgraduate medical curricula. Whilst a significant effect of content on suspected authorship of the scripts was revealed, further analysis of the free-text qualitative data collated revealed that marker uncertainty, and even inaccuracy, in terms of which script was AI generated highlights the key difficulty that universities will face.

Whilst the application of this technology appears to be incredibly far-reaching, even in the medical sphere, from optimising clinical decision making (Liu et al., 2023) to scientific writing (Salvagno et al., 2023) as well as healthcare education and training in general (Hosseini et al., 2023), there is currently no study to our knowledge investigating human marker accuracy in differentiating between student-written scripts and generative AI produced text. Tools such as DetectGPT claim to detect the use of generative AI (on the basis of five open-source LLMs) with a 95% accuracy (Mitchell et al., 2023). These, however, remain under development and review and, as such, provide little current technological support for markers of modular coursework submissions. In an academic world rife with appeals, it is unlikely that less than 100% accuracy will be acceptable to universities, but given the stochastic nature of LLMs, this is likely to remain unachievable.

An assessment of ChatGPT's ability to accurately generate responses to complex medical que-ries has been reported by Johnson et al. (2023). There have been limitations reported, though, in regard to the robustness and reliability of using such tools in their present form in a clinical setting. Once again, whilst the outputs produced by ChatGPT may seem impressive, it is important to keep in mind that the tool currently makes use of a sophisticated model in responding to instructions and learning from its own prior responses. It is, therefore, important to healthcare professionals, students, and patients alike to continue to consult reliable sources in confirming information generated by such LLM tools. Such tools may be beguiling but also carry risks in terms of questionable source data and the perpetuation of dominant stereotypes (Bender et al., 2021). Even so, it appears likely that such tools may, over time, enhance the way in which we work, study and share information but should not be seen as accurate or reliable substitutes for human appraisal and reasoning influenced by evidence-based practice. Our findings confirm that, despite markers suspecting the use of tools such as ChatGPT at times, their suspicions were not proven to be valid on most occasions. The exception appears to be in relation to some aspects of content creation and particularly in terms of referencing, where markers were most accurate at differentiating between student-written and chatbot-generated scripts on this basis. Subsequent versions of ChatGPT as well as other LLMs such as Google's Bard, which will serve as the powerful search engine's direct interface, will undoubtedly aim to address key flaws identified in earlier versions of open-source generative AI tools (Rudolph et al., 2023b).

Given the limited accuracy demonstrated by experienced markers in differentiating between student-written scripts and those prepared by LLM tools such as ChatGPT, it would appear to be imperative that higher education assessment strategies be reconsidered to adapt to the increasing presence of such tools (Ifelebuegu, 2023). As we embark on an era where generative AI will be interwoven with, and embedded into, the student learning experience and possibly teaching provision, it is crucial that faculty work with students as partners in negotiating the responsible use of such new innovations. Knee-jerk reactions, such as the outright banning of ChatGPT that we have seen by some universities, will achieve little and will likely prove unrealistic, given the reach and implications of this technology. Furthermore, students will likely be engaging with these new technologies in the workplace. Our duty as educators has always been to ensure that students are equipped with the necessary skills to join the workforce. This now extends to the responsible use of new and emerging generative AI technologies.

Establishing trust between students and faculty, and re-evaluating what constitutes academic misconduct in light of the revolutionary shift in information creation and

dissemination, should form the cornerstone of any initial response to this technology (Mills et al., 2023). Providing clear guidance to students as to what constitutes academic misconduct in relation to the misuse of generative AI is key. Such guidance will need to align with teaching information literacy, incorporating generative AI and the appropriate use thereof. Specifically, students should demonstrate an awareness of how such LLMs generate outputs, what the advantages are of using such platforms, as well as what the limitations are of this technology (Rasul et al., 2023). Support on how to critically appraise responses generated by generative AI platforms forms a crucial part of such training, ensuring the responsible integration of these technologies in our ever-expanding toolbox of resources at our students' disposal. As such, students should be encouraged to embrace new and emerging technologies but receive the necessary training on how to appropriately apply outputs of prompts to their scholarly practices without demonstrating an overreliance on this single source of information or passing responses off as their own.

Proactive management of expectations (both student and staff) is recommended. As opposed to such generative AI tools being simply viewed as a threat, it would be preferable to instead consider how such tools can be embraced appropriately. Where transgressions of professional boundaries do occur, however, academic misconduct procedures should be updated to reflect what is considered appropriate and what is an inappropriate use of this emerging technology (Mohammadkarimi, 2023). It is no easier to ban the use of generative AI at this stage than it would have been to stop the internet from going mainstream three decades ago. Negotiating our relationships with these new tools and how they can enhance various aspects of our lives is key, without abuse of this new technology limiting our own personal and professional development.

References

Behrens, S. J., Gomez, A., & Krimgold, J. (2019). Spelling counts: The educational gatekeeping role of grading rubrics and spell check programs. *Research and Teaching in Developmental Education*, 7-20.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency (FAccT '21),* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

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

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 60,* 1-12. doi: 10.1080/14703297.2023.2190148

Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International, 13*(3), 313-321. doi: 10.1080/07399339209516006

Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A. & Hatta, H. (2023). Analysing the role of ChatGPT in improving student productivity in higher education. *Journal on Education 5*(4), 14886-14891. doi: 10.31004/joe.v5i4.2563

Hassoulas, A., de Almeida, A., West, H., Abdelrazek, M., & Coffey, M. J. (2023). Developing a personalised, evidence-based and inclusive learning (PEBIL) model of blended learning: A cross-sectional survey. *Education and Information Technologies*,1-18. doi: 10.1007/s10639-023-11770-0

Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., MacDon-ald, N., Holmes, K. L., & Kho, A. (2023). An exploratory survey about using ChatGPT in edu-cation, healthcare, and research. *medRxiv*: 2023.03.31.23287979. https://doi.org/10.1101/2023.03.31.23287979

Ifelebuegu, A. (2023). Rethinking online assessment strategies: Authenticity versus AI chatbot intervention. *Journal of Applied Learning and Teaching, 6*(2). Advance online publication. https://journals.sfu.ca/jalt/index.php/jalt/article/view/875

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., Chastain, C., … Wheless, L. (2023). Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the ChatGPT model. *Research Square,* rs.3.rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1

Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., McCoy, A. B., Sittig, D. F., & Wright, A. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association: JAMIA, 30*(7), 1237–1245. https://doi.org/10.1093/jamia/ocad072

Mills, A., Bali, M., & Eaton, L. (2023). How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process. *Journal of Applied Learning and Teaching, 6*(1), 16-30. https://journals.sfu.ca/jalt/index.php/jalt/article/view/843

Mohammadkarimi, E. (2023). Teachers' reflections on academic dishonesty in EFL students' writing in the era of artificial intelligence. *Journal of Applied Learning and Teaching, 6*(2). Advance online publication. https://doi.org/10.37074/jalt.2023.6.2.10

OpenAI. (2023). *ChatGPT.* https://chat.openai.com/chat

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

Rudolph, J., Tan., S., & Tan, S. (2023a). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1),

342-363. doi: 10.37074/jalt.2023.6.1.9

Rudolph, J., Tan, S., & Tan, S. (2023b). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://doi.org/10.37074/jalt.2023.6.1.23

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care, 27*(1), doi: 10.1186/s13054-023-04380-2

Copyright: © 2023. Athanasios Hassoulas, Ned Powell, Lindsay Roberts, Katja Umla-Runge, Laurence Gray and Marcus J Coffey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.