scientific reports

Check for updates

OPEN

Identification of dental related ChatGPT generated abstracts by senior and young academicians versus artificial intelligence detectors and a similarity detector

Matheel AL-Rawas[1,2], Omar Abdul Jabbar Abdul Qader[3], Nurul Hanim Othman[1,2], Noor Huda Ismail[1,2], Rosnani Mamat[2,4], Mohamad Syahrizal Halim[2,4], Johari Yap Abdullah[5,6]✉ & Tahir Yusuf Noorani[2,4,7]✉

Several researchers have investigated the consequences of using ChatGPT in the education industry. Their findings raised doubts regarding the probable effects that ChatGPT may have on the academia. As such, the present study aimed to assess the ability of three methods, namely: (1) academicians (senior and young), (2) three AI detectors (GPT-2 output detector, Writefull GPT detector, and GPTZero) and (3) one plagiarism detector, to differentiate between human- and ChatGPT-written abstracts. A total of 160 abstracts were assessed by those three methods. Two senior and two young academicians used a newly developed rubric to assess the type and quality of 80 human-written and 80 ChatGPT-written abstracts. The results were statistically analysed using crosstabulation and chi-square analysis. Bivariate correlation and accuracy of the methods were assessed. The findings demonstrated that all the three methods made a different variety of incorrect assumptions. The level of the academician experience may play a role in the detection ability with senior academician 1 demonstrating superior accuracy. GPTZero AI and similarity detectors were very good at accurately identifying the abstracts origin. In terms of abstract type, every variable positively correlated, except in the case of similarity detectors ($p < 0.05$). Human-AI collaborations may significantly benefit the identification of the abstract origins.

Keywords ChatGPT, AI-based language model, Generative pre-trained transformer, AI-written content, Educators, Scientific abstracts

Abbreviations

AI	Artificial intelligence
HI	Human intelligence
GPT	Generative pre-trained transformer

The advent of artificial intelligence (AI)-based applications significantly affects the individuals as well as a plethora of organisations and societies. By combining linguistic and computer science models, AI aims to build computer models that can do tasks that would, otherwise, require human intelligence (HI)[1]. This includes

[1]Prosthodontic Unit, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kota Bharu, Kelantan, Malaysia. [2]Hospital Pakar Universiti Sains Malaysia, Kubang Kerian, Kota Bharu, Kelantan, Malaysia. [3]College of Dentistry, Al Mashreq University, Airport Street, Baghdad, Iraq. [4]Conservative Dentistry Unit, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kota Bharu, Kelantan, Malaysia. [5]Craniofacial Imaging Laboratory, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian Kota Bharu, Kelantan, Malaysia. [6]Dental Research Unit, Center for Transdisciplinary Research (CFTR), Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu, India. [7]Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu, India. ✉email: johariyap@usm.my; dentaltahir@yahoo.com

learning, adapting, rationalising, understanding, and grasping abstract concepts, as well as being responsive to complex human traits, such as attentiveness, emotion, and innovation[2]. As ChatGPT has gained immense popularity worldwide over the past year, it has led to widespread discussions of its implications. As ChatGPT is an AI-based language model, it has undergone extensive training on numerous text-based datasets from multiple languages. OpenAI, the developers of ChatGPT, describe it as a chatbot that uses the Generative Pre-trained Transformer (GPT) architecture to generate responses to user given text-based inputs. According to Brown et al.[3], the GPT architecture analyses natural language using a neural network and generates replies based on the input text's context. As such, when given an input by a user, it can generate text-based responses that closely resemble that of a trained human[4].

The advent of ChatGPT was met with mixed reactions from the scientific and academic communities, and further prompted the long-standing debate on the potential advantages and disadvantages of adopting cutting-edge AI-based technologies[5–7]. ChatGPT is excellent for various conversational and written tasks as it increases the speed and quality of the produced work[8]. However, many users have raised concerns over the possibility of bias due to the sets of data that were used to train ChatGPT. This is because, if bias exists, it may hinder its performance and provide erroneous answers that appear to be scientifically accurate[8]. The challenge of distinguishing between human- and AI-written content has also sparked several concerns in professional and education-related communities and renewed discussion on the importance of content written using HI[9–11]. Therefore, the controversy surrounding ChatGPT was inevitable. Nevertheless, the possibility that it could produce factually erroneous content, as well as the ethical aspect of using and abusing AI-based technologies to produce content warrant careful consideration, especially since the produced content could cause misinformation in healthcare practices and academic publications[12–14].

Furthermore, the effects of ChatGPT extend beyond the realm of academic educational-related activities. Previous scholarly publication has named ChatGPT in its publication as a "contributing" author[15]. However, many experts believe that AI does not satisfy all the requirements for authorship.

HI has several advantages over AI, namely, biological evolution, flexibility, creativity, emotional intelligence, and the capability to comprehend abstract ideas[2]. However, it could prove beneficial to combine HI and AI, solely if the latter's output can be guaranteed to be both accurate and dependable[7]. To date, no study has examined the effects of an academician's level of experience on their ability to identify the origin of scientific abstracts, nor has any study compared this aspect with the use of AI detection tools.

When using ChatGPT, or any AI model, for academic purposes, it is crucial to be precautious, particularly with regards to the ethical and societal consequences. This is because the internal operations of AI models lack transparency. Therefore, it is essential to recognize that AI models are opaque AI tools that yield outputs in response to user-inputted queries. As such, the accuracy of the outputs cannot be guaranteed[7].

The consideration of factual inaccuracies, ethical concerns, and the possibility of misuse, particularly the spread of false information, are crucial in both healthcare practice and academic writing. These risks can be mitigated by ensuring awareness of these possibilities and using appropriate tools to distinguish between human- and AI-written manuscripts.

Evaluating the detection capabilities of academicians against AI tools can determine if human expertise has a distinct advantage or if AI excels in recognizing its own outputs. Human judgment depends on experience, intuition, and contextual understanding, while AI detection tools utilize statistical patterns and probabilities. Comprehending their strengths and weaknesses can enhance detection strategies. While previous studies have explored human ability to differentiate AI-generated and authentic abstracts, few have examined the influence of experience level on content detection accuracy. Investigating whether senior academicians, with their extensive experience, outperform junior counterparts, or if both groups struggle equally, offers valuable insights into how experience level interacts with AI-generated text comprehension. Thus, the present study aimed to examine the ability and accuracy of four blinded human academicians of different experience levels to differentiate between and evaluate the quality of human- and AI-written content in conjunction with three AI output detectors and a plagiarism detector.

Methodology

A cross-sectional study was conducted at a school of dental sciences with the involvement of four (two senior and two young) academicians having completed higher education in dental specialities, three AI detectors, and a plagiarism detector. The ideal sample size of abstracts was determined to be 122. This was determined using G*power 3.1.9.6 at a 0.05 significance level (α), 0.3 effect size, and 0.8 power. A total of 160 abstracts were chosen to ensure that each of the four academicians received an equal number of abstracts to review ($n = 40$). Ethical approval was obtained from the Human Research Ethics Committee of Universiti Sains Malaysia (reference number USM/JEPeM/KK/24010127). Full informed consent was also acquired from every participant.

A total of 80 human-written titles and abstracts were gathered via random and systematic sampling of original research articles that had been published in the first five months of 2023 in eight high-impact dental-related journals, namely, the International Journal of Oral Science, the Journal of Dental Research, the Journal of Clinical Periodontology, Oral Oncology, Dental Materials, the International Endodontic Journal, Clinical Oral Implants Research, and Journal of Dentistry. The inclusion criteria were abstracts from dental-related journals, abstracts from original research papers, were written in English, and of structured or non-structured formats, while the exclusion criteria were abstracts that contained figures, were present in a case report, case series, or any type of review.

A total of 80 AI-written titles and abstracts were generated by using ChatGPT 3.5[16], to rewrite the titles and abstracts of the aforementioned 80 human-written abstracts. The query that was fed into ChatGPT 3.5 read, "*Kindly compose a research abstract for the study* [title of the human-written abstract] *in accordance with the format followed by the* [journal name] *found at* [link]."[17]. As such, the final number of abstracts was 160.

It is noteworthy that, as ChatGPT is an AI-based language model, it cannot browse the Internet on its own at the current version. Furthermore, its training data is only up to April 2023. Apart from that, it is also sensitive to the previous queries in a session. Therefore, a new session was created before entering each new query.

An online Research Randomizer[18] was used to generate unique identifiers for each abstract and ensure blinding. Any identifying information, such as journal names, titles, and other identifying information, was also removed. All the abstracts were securely stored to ensure that they could not be accessed or tampered with.

Four academicians, two young and two seniors, were selected via random sampling. For the young academicians, the inclusion criteria were ≤2 years of dental-related academic experience, an acceptable publication history and peer review experience, while the exclusion criteria were non-dental-related academic experience. For the senior academicians, the inclusion criteria were ≥10 years of dental-related academic experience, an acceptable publication history and peer review experience, while the exclusion criteria were non-dental-related academic experience. A total of four academicians was considered sufficient if compared to the previous studies[17,19].

Each academician was given an equal number of abstracts, 20 human-written and 20 AI-written, and asked to read the abstracts once only and then to determine and classify if an abstract was human- or AI-written based on the guidelines given to them. They were expected to submit their findings within seven to 10 days of receiving the abstracts. The academicians were told to give an abstract a score of 1 if they believed that it was human-written and a score of 2 if they believed that it was AI-written. They were then required to determine the quality of each abstract using a new, specially-developed rubric that scored its language complexity, cohesion, creativity, contextual understanding, grammatical accuracy, and domain-specific knowledge on a range of 1 (Lowest) to 3 (Highest) (Table 1). The total score of the six criteria was then calculated and divided by 18 to yield four different scores, with a score of 1 representing excellent quality, 0.9 to 0.7 representing good quality, 0.6 to 0.4 representing average quality, and 0.1 to 0.3 representing poor quality.

Meanwhile, Turnitin®'s Plagiarism Detector (2023)[20] was used to discover the similarity indices of the 80 AI-written abstracts. It has a scoring range of 0 to 100%, with the latter score representing that the entire abstract had been plagiarised. Three AI output detectors, namely the GPT-2 output detector (2023)[21], the Writefull GPT detector (2023)[22], and GPTZero (2023)[23], were also used to examine the 80 AI-written abstracts. More specifically, the three AI output detectors were used to ascertain the likelihood that a piece of content had been written using GPT-3, -4, or ChatGPT. All three detectors have a scoring range of 0 to 100%, with the latter score representing a higher probability. The three AI output detectors employ machine learning but vary in methodology. The GPT-2 Output Detector utilizes a fine-tuned RoBERTa model to categorize text according to statistical patterns. Writefull GPT Detector presumably employs deep learning or logistic regression to distinguish between AI-generated and human-authored content. GPTZero integrates machine learning with statistical heuristics, evaluating perplexity (word unpredictability) and burstiness (sentence variation) to identify AI-generated text. Each method employs distinct computational strategies to improve detection accuracy[21–23]. In this study, the three AI detector tools were considered acceptable when compared to another study[24].

The four academicians were blinded to the plagiarism and AI detectors' results. This was to ensure that the study was conducted in a rigorous and unbiased manner, which increases the validity and reliability of the final results. The statistical analysis was performed using IBM SPSS software version 27.0. Crosstabulation was carried out to get the outcome values for all variables. Each academician's abstract type and quality assessment results were cross tabulated to determine each academician's results, and if a correlation existed between the variables. Non-parametric chi-squared analyses were then conducted to identify differences between all the academician's abstract type and quality assessment results. The bivariate correlation by Spearman of all the

Criteria	Definition	Scoring		
		1	2	3 (Highest)
Language complexity	This criterion refers to the level of technical language or jargon used in the abstract. A more technical or specialized vocabulary indicates greater complexity.	The language is simple and lacks technical terms.	The language is somewhat technical and includes some jargon.	The language is highly technical and includes many jargon terms.
Cohesion	This criterion refers to the logical connections between ideas in the abstract. A cohesive abstract presents idea in a clear, logical sequence that flows smoothly and is easy to follow.	The abstract lacks logical connections between ideas.	The abstract has some logical connections between ideas but may be disjointed.	The abstract has strong logical connections and flows smoothly.
Creativity	This criterion refers to the originality of the phrasing and content in the abstract. A more creative abstract presents ideas in a unique or innovative way.	The abstract is formulaic and lacks originality.	The abstract includes some original phrasing and content.	The abstract exhibits strong creativity and originality.
Contextual understanding	This criterion refers to the depth of understanding the abstract demonstrates about the research context and significance. A more contextually aware abstract shows a deeper understanding of the research background, purpose, and significance.	The abstract lacks depth and understanding of the research context and significance.	The abstract demonstrates some understanding of the research context and significance.	The abstract demonstrates a deep understanding of the research context and significance.
Grammatical accuracy	This criterion refers to the correctness of the grammar and syntax used in the abstract. An abstract with excellent grammatical accuracy is free of	The abstract has many grammatical errors.	The abstract has some grammatical errors.	The abstract has excellent grammatical accuracy.
	errors and follows the rules of grammar and syntax.			
Domain-specific knowledge	This criterion refers to the use of specialized terminology and knowledge specific to the research field in the abstract. An abstract with a strong understanding of domain-specific knowledge demonstrates a deep understanding of the relevant terminology and concepts.	The abstract lacks domain-specific knowledge.	The abstract shows some understanding of domain-specific knowledge.	The abstract demonstrates a strong understanding of domain-specific knowledge.

Table 1. New and specially developed rubric to help score the abstracts' type and quality.

variables was also examined to determine the direction and strength of the inter-variable correlations. A score of 0.05 was set as the point of statistical significance. Percentages of sensitivity, specificity and accuracy along with their 95% confidence intervals of all assessment methods were calculated using MedCalc Software Ltd. evaluation calculator (version 2023)[25].

Results

Figures 1, 2, 3 and 4; Tables 2 and 3 present the results of the present study. As seen in Fig. 1, every academician wrongly assumed certain number of abstracts. The highest number of wrong assumptions was 18 abstracts, while the lowest was three. No association and no significance were found in the outcomes of senior academician 1 and young academician 1 ($p > 0.05$), while the contrary was found in the outcomes of senior academician 2 and young academician 2 ($p < 0.05$).

The new rubric that was developed to assess the quality of the abstracts was effective as all the academicians rated the human-written abstracts as having either excellent or average quality, and AI-written abstracts as having either good or poor quality. The quality assessment results of all four academicians were significant and correlated with their abstract type results (Fig. 2). The results of the chi-squared analysis revealed significant differences between the abstract type and quality assessment results that each academicians reported ($p < 0.001$) (Table 2).

The cross-tabulation results of the GPT-2 output detector's findings indicated an insignificant correlation with its abstract type findings ($p = 0.063$). This could be attributed to the high number of low detection scores that it gave to 53 of the 80 AI-written abstracts (Table 3). Meanwhile, significant correlations were observed between the abstract type findings of the Writefull GPT detector and GPTZero ($p < 0.001$), which were moderately positive (phi = 0.342) and highly positive (phi = 0.951), respectively. The GPT-2 output detector and the Writefull GPT detector made a higher number of wrong assumptions as compared to GPTZero. Only GPTZero classified most of the abstracts correctly (Table 3). Meanwhile, the Turnitin® Plagiarism Detector gave all the human-written abstracts a score of 100%, while its scores varied for the AI-written abstracts. The correlation between the abstract type findings was strong and significant (Table 3).

Figure 3 shows the correlations between all study variables and between the variables and abstract type. The correlations between the abstract type findings of the GPT-2 output detector and the Writefull GPT detector were very weak to moderately positive. Meanwhile, that of the GPTZero detector and the Turnitin® Plagiarism Detector were very strong, however, the directions of their correlations differed.

The accuracy of senior academician 2 was very high (92.50%) compared to other academicians in detecting the abstract types, while for the abstract quality, one senior and one young academician achieved a high accuracy (87.5% and 82.5% respectively) in assessing the abstract quality. For the AI detectors, GPTZero scored 92.60% accuracy compared to other AI detectors (Fig. 4).

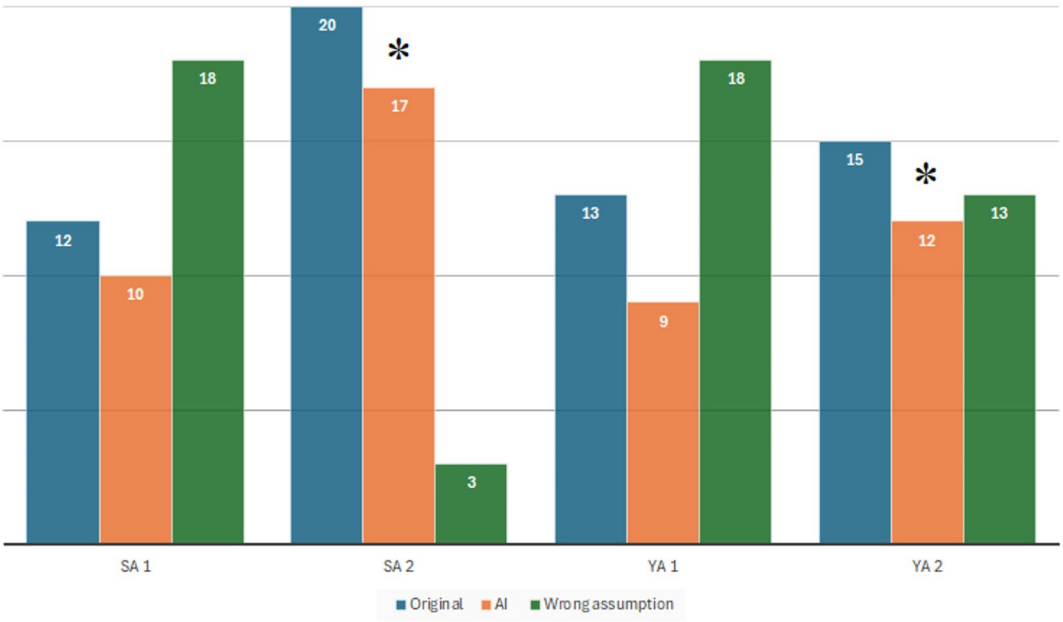

Fig. 1. The outcome of the abstract type by the four academicians [Senior Academician (SA), Young Academician (YA)]. *Statistically significant ($P < 0.05$).

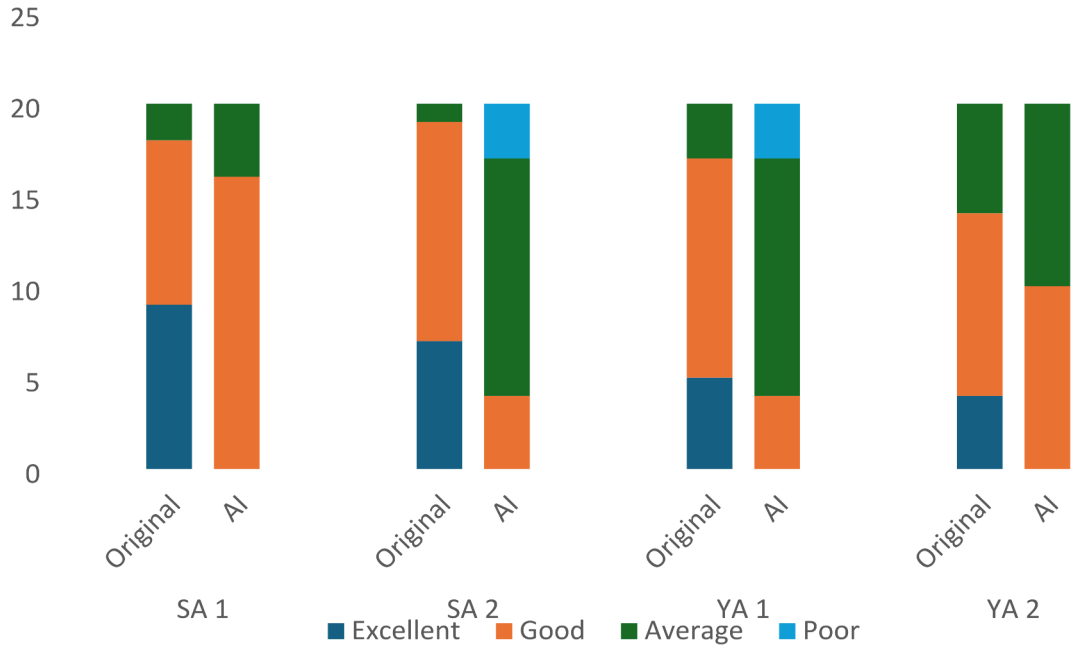

Fig. 2. The abstract quality assessment results by the four academicians using the specially designed rubric [Senior Academician (SA), Young Academician (YA)]. *Statistically significant ($P < 0.05$). All data was statistically significant.

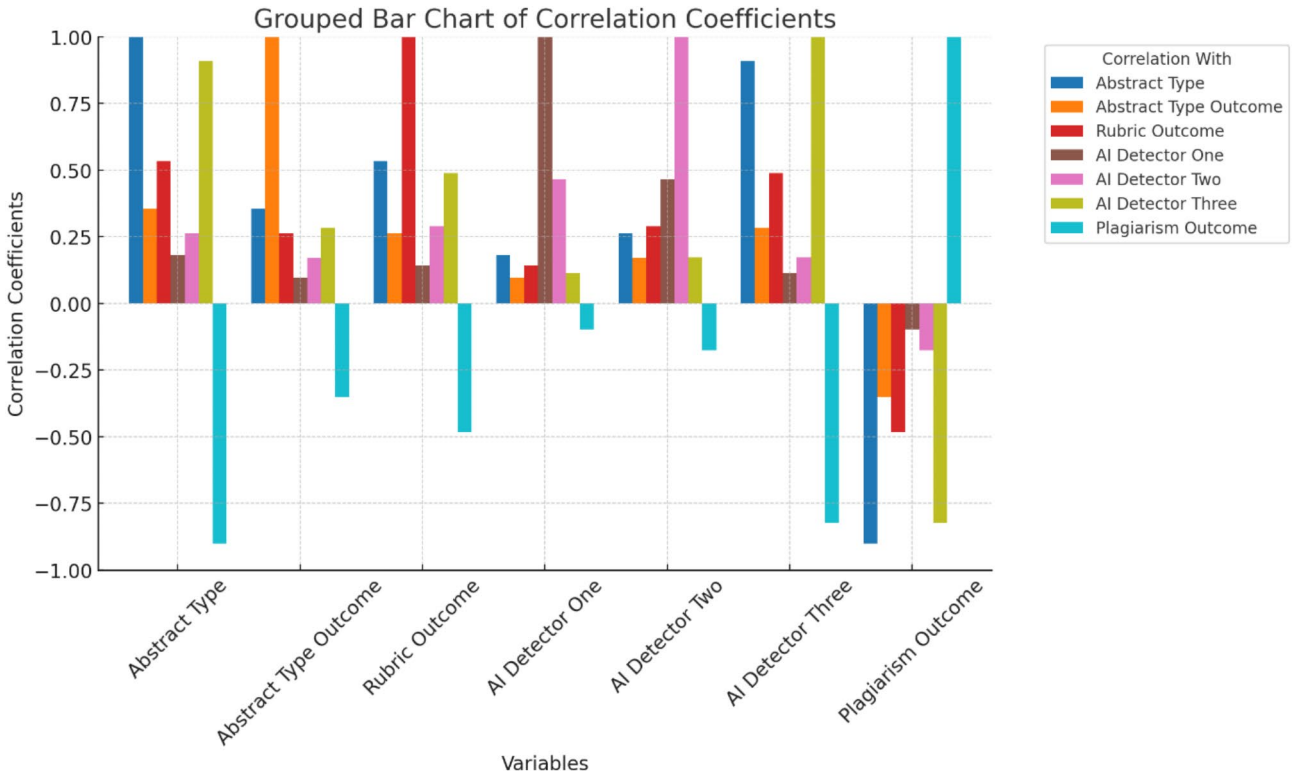

Fig. 3. The Correlation assessment among different variables. * Correlation is significant at the 0.05 level. All data was statistically significant except for the correlation between GPT-2 output detector outcomes and abstract type and rubric quality outcomes and between GPT-2 output detector outcomes and GPTzero and Turnitin similarity detector outcomes.

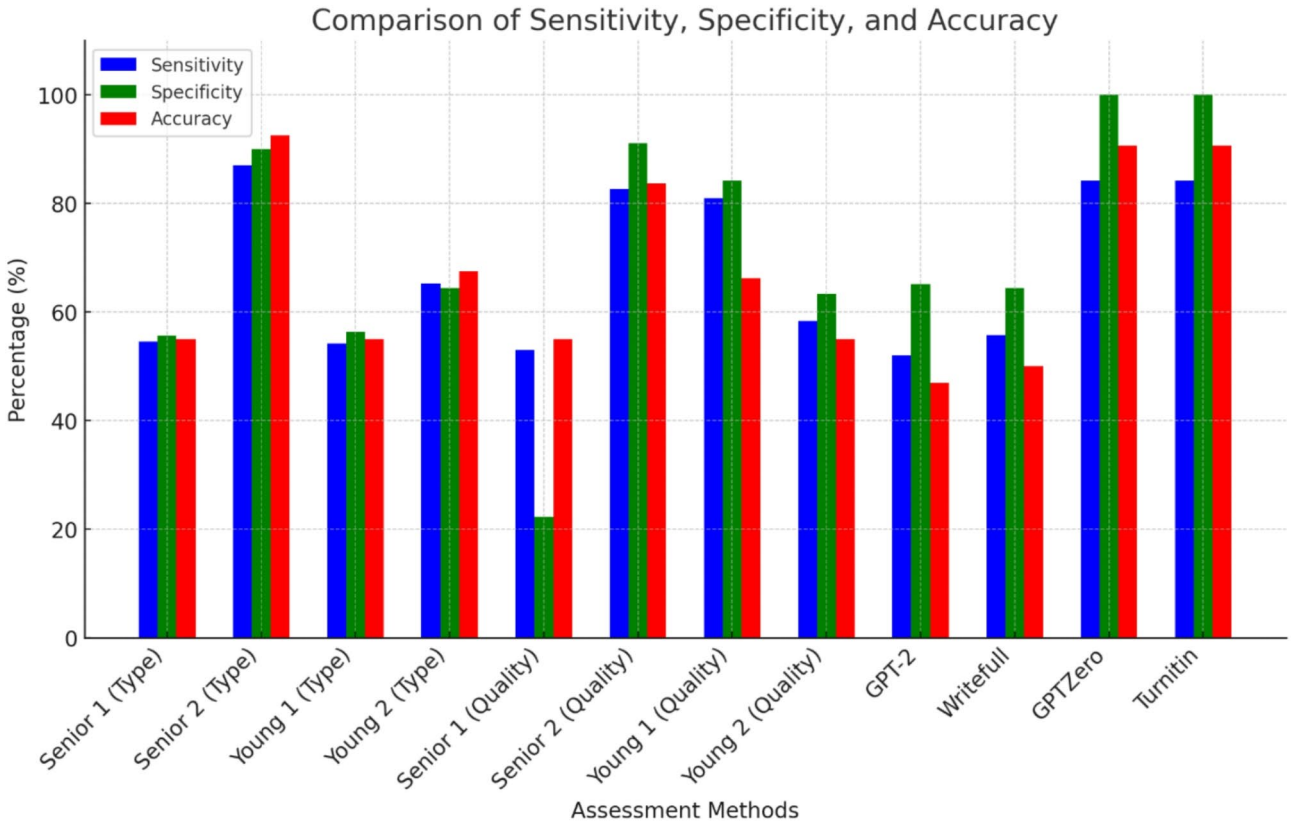

Fig. 4. Accuracy of each method of assessment of the abstract types.

Discussion

OpenAI's ChatGPT received a lot of attention when it was released in November 2022 for its ability to use AI to produce authentic and detailed text-based responses to human requests[16]. This was accomplished by first using humans to train the AI model using a set of different dialogues that they had created. It was then optimised using reinforcement learning algorithms, which incorporate human feedback[26].

ChatGPT has, undoubtedly, affected academic research and writing. At present, it is common and acceptable to use AI tools, such as Grammarly[*27] and Quillbot[*28], in rephrasing original contents and to improve academic writing. However, ChatGPT will transform the method of retrieving data for academic research and writing[9]. Nevertheless, a majority of the scholarly discussion has focused on the effects that ChatGPT will have on the education industry as it is, evidently, a significant concern[7].

The advent of ChatGPT also proves that mankind is not ready to achieve this important milestone in the field of AI. For one, it already necessitates changes to publication and research policies[7]. As such, the present study used various methods to examine AI-written abstracts of dental-related studies to assess the ability to identify AI-written content.

Gao et al.. and Bouschery et al.[17,29]. believed that ChatGPT is an effective tool for composing abstracts and, even, introduction sections. However, most of the time, the content that ChatGPT generates has been poorly rephrased, so much so that its true meaning is lost or misrepresented. A significant degree of plagiarism has also been observed. In the present study, ChatGPT was able to produce some AI-written abstracts that were so well written that all the academicians misclassified them as human-written abstracts. Although an academician's years of experience may not be a significant factor, one of the senior academicians managed to correctly identify 37 of the 40 human- and AI-written abstracts (Fig. 1). The differences between the findings of each academician were all significant (Table 2). This may be attributed to differences in the scores that the academicians gave each abstract and the number of wrong assumptions.

Senior academicians frequently outperform their junior peers in identifying AI-generated content owing to their substantial experience and refined critical analysis abilities. Their extensive engagement with academic writing allows them to identify subtle inconsistencies and unnatural patterns characteristic of AI writing[30]. Moreover, their significant scientific expertise enables them to discern factual inaccuracies that less experienced scholars may neglect. A study investigating the identification of AI-generated content in higher education assessments revealed that academic staff, particularly those with greater experience, excelled at identifying AI-generated submissions, especially when supported by AI detection tools[30].

In another study, researchers assessed the ability of novice (average age of 25) and experienced teachers (average age of 42) to identify AI-generated content in student submissions. The results indicated that both novice and experienced teachers faced challenges in distinguishing between AI-generated and human-written

Chi-square non-parametric test to test the differences			
Senior academician 1 and 2			
Variable	df	Pearson Chi square*	P-value^
Abstract type outcome	1	18.906	< 0.001
Quality outcome	3	70.010	< 0.001
Young academician 1 and 2			
Variable	df	Pearson Chi square	P-value
Abstract type outcome	1	23.256	< 0.001
Quality outcome	3	53.906	< 0.001
Senior academician 1 and young academician 1			
Variable	df	Pearson Chi square	P-value
Abstract type outcome	1	21.025	< 0.001
Quality outcome	3	74.010	< 0.001
Senior academician 1 and young academician 2			
Variable	df	Pearson Chi square	P-value
Abstract type outcome	1	18.906	< 0.001
Quality outcome	2	20.713	< 0.001
Senior academician 2 and young academician 1			
Variable	df	Pearson Chi square	P-value
Abstract type outcome	1	23.256	< 0.001
Quality outcome	3	40.625	< 0.001
Senior academician 2 and young academician 2			
Variable	df	Pearson Chi square	P-value
Abstract type outcome	1	21.025	< 0.001
Quality outcome	3	54.073	< 0.001

Table 2. Non-parametric Chi-square analysis of the differences between the academicians and their outcomes. *Cells (less than 20%) have expected count less than 5. ^Significance level is at 0.05.

Variable	GPT-2 output detector (*n*)						
Abstract type	**Low fake**	**Moderate fake**	**High fake**	**Very high fake**	**Pearson Chi-square ***	**P-value^**	**Phi value**
Original abstract	66	7	2	5	7.281	0.063	0.213
AI abstract	53	12	9	6			
Variable	Writefull GPT detector (*n*)						
Abstract type	Entirely human	Mostly human made	Partly by AI	Entirely by AI	Pearson Chi square	P-value	Phi value
Original abstract	62	2	5	11	18.705	< 0.001	0.342
AI abstract	38	13	14	15			
Variable	GPTZero detector (*n*)						
Abstract type	Low fake	Moderate fake	High fake	Very high fake	Pearson Chi-square	P-value^	Phi value
Original abstract	80	0	0	0	144.762	< 0.001	0.951
AI abstract	4	11	13	53			
Variable	Similarity outcome (*n*)						
Abstract type	Low similarity	Moderate similarity	High similarity	Very high similarity	Pearson Chi square	P-value	Phi value
Original abstract	0	0	0	80	144.762	< 0.001	0.951
AI abstract	23	42	11	4			

Table 3. The assessment of the abstract by four different methods other than academicians. *Cells (less than 20%) have expected count less than 5. ^Significance level is at 0.05.

texts[31]. However, experienced teachers demonstrated a marginally higher accuracy in their judgments for AI and student written content compared to novice teacher, suggesting that while experience offers some advantage, the sophistication of AI-generated texts makes detection difficult across all experience levels[31].

Two rationales for academician's incapacity of identifying AI-generated content: AI models succeed at mimicking human writing styles, generating content that is grammatically accurate and contextually cohesive, perhaps deceiving readers into believing it is authored by an actual human being. Secondly, academician's assessments may be erroneous because of their dependence on faulty heuristics for identifying AI-generated language[32]. Research indicates that humans can accurately recognize AI-generated text around 53% of the time[32].

The academician's reported that the rubric that was developed to help them assess the quality of the abstracts performed well. As seen in Fig. 2; Table 2, all the academicians rated most of the human-written abstracts to be of excellent to good quality, with very few deemed to be of average quality, and none to be of poor quality. However, they rated most of the AI-written abstracts to be of average quality, with very few deemed to be of good quality, and six to be of poor quality. Nevertheless, the differences in the findings that each academician reported using the rubric were significant. Besides, no comparison with previous studies cannot be drawn as no similar study has been carried out before.

Humans, generally, struggle to differentiate between human- and AI-written content, while tools, such as bot detection AIs, exhibit superior discriminatory performance. Unlike human-written content, AI-written content often lacks specificity and creativity. It also tends to over generalise specific scenarios, while the style of writing can be characterised as predominantly containing anticipated words. Artificial intelligence (AI)-based tools, like GPTZero, can quite successfully identify AI-written content[7]. The present study, similarly, found that GPTZero is very effective at differentiating between human- and AI-written content. Of the 80 AI-written abstracts examined, the GPT-2 output and Writefull GPT detectors, respectively, erroneously gave 53 and 38 abstracts low probability scores (Table 3). As seen in Fig. 3, there was a significant correlation between the abstract type findings of GPTZero and the Writefull GPT detector ($p < 0.05$) than that of the GPT-2 output detector. Therefore, GPTZero is a very useful and reliable tool that academicians can use to identify the source of the text or abstracts. It has also been proven to accurately identify 80% of AI-written content[24].

AI detectors, like GPT-2 output and Writefull GPT, encounter considerable limitations, including constraints of training data, rising sophistication of AI, and inherent biases. Their accuracy relies on diverse training datasets; however, insufficient exposure to varied writing styles may result in misclassification[33]. As AI-generated content advances, numerous detectors find it increasingly challenging to remain functional and accurate. Moreover, biases in training data may lead to the inappropriate identification of specific writing styles, thereby raising concerns regarding reliability. These challenges underscore the necessity for ongoing enhancements and human supervision, rather than complete dependence on AI detection tools[33].

The study found that the accuracy of GPTZero was notably high at 92.60% (Fig. 4), supporting the findings of Habibzadeh[24]. If GPTZero exhibits high accuracy, universities might adopt it as a standard instrument for identifying AI-generated submissions, thereby strengthening academic integrity policies. However, institutions must avoid over-reliance on this tool and consider human verification alongside AI detection tools.

Only Senior Academician 2 and the Turnitin similarity detector attained comparable accuracy values to GPTZero (Fig. 4). Nevertheless, the researchers were unable to reach a definitive conclusion regarding the accuracy of the senior academicians due to the restricted number of academicians involved.

Even though the use of ChatGPT presents particular difficulties in the education industry, the use of chatbots and other AI technologies are becoming more common and popular. As such, researchers, management, and educators must adapt to the continually evolving digital landscape. Some methods are, thankfully, currently in development to that end. The present study used Turnitin®'s plagiarism detector[20] to determine the similarity indices of 80 human- and 80 AI-written abstracts. It gave all the human-written abstracts a similarity index of 100% (Table 3). As seen in Fig. 3, the Turnitin® similarity indices significantly correlated with the abstract type findings ($p < 0.05$), with a strong negative direction indicating that, as most of the similarity indices were low, most of the abstracts were AI-written. Gao et al.[17] also reported similar findings.

Bouschery et al.[29] highlighted the role that ChatGPT played in generating abstracts with minimal involvement from authors. It is worrisome that some academicians use AI tools, such as ChatGPT, to generate content and do not disclose it to journal editors, publishers, and conference organisers. Some scholarly articles also list ChatGPT as a "contributing" author[15]. Although the validity of these studies is not being called into question, listing ChatGPT as a "contributing" author raises the issue of bias, lack of transparency, privacy, copyright, and misuse. It also calls into question the credibility of academic writing and research, which affects every field of study[7].

The evolution of publishing policies will lead to the development of newer versions of ChatGPT. However, the process of publishing scholarly articles will, most likely, continue to primarily rely on humans rather than machines to meticulously review scholarly articles prior to publication. As such, publishers, editors, and conference committees must ensure that reviewers are appropriately trained and provided with tools that can effectively address the unethical use of technology, which could compromise the quality of academic writing and the credibility of entire industries[7].

For the clinical implications, addressing factual inaccuracies, ethical dilemmas, and the potential for misuse, especially the dissemination of misinformation, are essential in healthcare and dental related practices and academic writing. These risks can be alleviated by fostering awareness of these possibilities and employing suitable tools to differentiate between human- and AI-generated manuscripts[7].

The limited number of academicians used to review the abstracts is a limitation of the present study. A newer version of ChatGPT was released after the study was completed and new features were emerged and will emerge. Another limitation is that Turnitin® had not released their AI writing detection tool at the time of writing. Therefore, it would be worthwhile to overcome those limitation for a better assessment in future studies.

Conclusion

Tools, such as ChatGPT and other related technologies, will continue to significantly affect the education industry into the foreseeable future. Therefore, research efforts must be ongoing to mitigate the risks associated with adopting the use of such tools and ensure that they are used ethically to reap their societal benefits. Methods must be developed to precisely identify AI-generated materials. While GPTzero and Turnitin similarity detector excel in accurately classifying abstract categories, both young and senior academicians must be trained on identifying AI-generated materials and use detection methods without bias. After all, combining HI and AI will significantly benefit society if the latter's outputs can be guaranteed to be both accurate and dependable.

Researchers, educators and policymakers must adapt to the continually evolving digital technologies in the research and education industry as a whole, and dentistry in particular by improve hybrid detection approaches, encourage ethical AI-assisted writing practices and promote AI regulations that balance innovation with academic integrity.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 11 December 2024; Accepted: 20 March 2025
Published online: 02 April 2025

References

1. Sarker, I. H. AI-Based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Comput. Sci.* **3**, 158–177. https://doi.org/10.1007/s42979-022-01043-x (2022).
2. Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C. & Eikelboom, A. R. Human- versus artificial intelligence. *Front. Artif. Intell.* **4**, 622364. https://doi.org/10.3389/frai.2021.622364 (2021).
3. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165 (2020).
4. OpenAI, C. G. P. T. Optimizing Language Models for Dialogue https://openai.com/blog/chatgpt/. Accessed 23 February 2023 (2023).
5. Howard, J. Artificial intelligence: implications for the future of work. *Am. J. Ind. Med.* **6**, 917–926. https://doi.org/10.1002/ajim.23037 (2019).
6. Tai, M. C. The impact of artificial intelligence on human society and bioethics. *Tzu Chi Med. J.* **32**, 339–343. https://doi.org/10.4103/tcmj.tcmj_71_20 (2020).
7. Dwivedi, Y. K. et al. So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manage.* **71**, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642 (2023).
8. Deng, J. & Lin, Y. The benefits and challenges of ChatGPT: an overview. *FCIS* **2**, 81–83. https://doi.org/10.54097/fcis.v2i2.4465 (2023).
9. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **613**, 423. https://doi.org/10.1038/d41586-023-00056-7 (2023).
10. Stokel-Walker, C. AI bot ChatGPT writes smart essays - should professors worry? *Nature* https://doi.org/10.1038/d41586-022-04397-7 (2022).
11. Stokel-Walker, C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature* **613**, 620–621. https://doi.org/10.1038/d41586-023-00107-z (2023).
12. Stokel-Walker, C. & Van Noorden, R. What ChatGPT and generative AI mean for science. *Nature* **614**, 214–216. https://doi.org/10.1038/d41586-023-00340-6 (2023).
13. Chatterjee, J. & Dethlefs, N. This new conversational AI model can be your friend, philosopher, and guide and even your worst enemy. *Patterns (N Y)*. **4**, 100676. https://doi.org/10.1016/j.patter.2022.100676 (2023).
14. Sallam, M. et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: A descriptive study at the outset of a paradigm shift in online search for information. *Cureus* **15**, e35029. https://doi.org/10.7759/cureus.35029 (2023).
15. van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. ChatGPT: five priorities for research. *Nature* **614**, 224–226. https://doi.org/10.1038/d41586-023-00288-7 (2023).
16. https://chat.openai.com. Accessed 15 Jan 2023.
17. Gao, C. A. et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digit. Med.* **6**, 75. https://doi.org/10.1038/s41746-023-00819-6 (2023).
18. https://www.randomizer.org/#randomize. Accessed 19 Jan 2023.
19. Haq, Z. U., Naeem, H., Naeem, A., Iqbal, F. & Zaeem, D. Comparing human and artificial intelligence in writing for health journals: An exploratory study. *MedRxiv* **02** (2023).
20. https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead. Accessed 05 Apr 2023.
21. https://openai-openai-detector.hf.space. Accessed 03 Feb 2023.
22. https://gptzero.me. Accessed 05 Febr 2023.
23. https://x.writefull.com/gpt-detector. Accessed 08 Feb 2023.
24. Habibzadeh, F. GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study. *J. Korean Med. Sci.* **38**, e319. https://doi.org/10.3346/jkms.2023.38.e319 (2023).
25. https://www.medcalc.org/calc/diagnostic_test.php. Accessed 20 December 2023.
26. Walsh, T. & Bard Bing and Baidu: how big tech's AI race will transform search – and all of computing. In *The Conversation*. https://theconversation.com/bard-bing-andbaidu-how-big-techs-ai-race-will-transform-search-and-all-of-computing-199501
27. https://www.grammarly.com. Accessed 13 June (2023).
28. https://www.quillbot.com. Accessed 13 June (2023).
29. Bouschery, S. G., Blazevic, V. & Piller, F. T. Augmenting human innovation teams with artificial intelligence: exploring transformer-based Language models. *J. Prod. Innov. Manage.* **40**, 139–153. https://doi.org/10.1111/jpim.12656 (2023).
30. Jakesch, M., Hancock, J. T. & Naaman, M. Human heuristics for AI-generated Language are flawed. *Proc. Natl. Acad. Sci. U S A.* **120**, e2208839120. https://doi.org/10.1073/pnas.2208839120 (2023).
31. Perkins, M., Roe, J., Postma, D., McGaughran, J. & Hickerson, H. Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. *J. Acad. Ethics.* **22**, 89–113. https://doi.org/10.1007/s10805-023-09492-6 (2024).
32. Fleckenstein, J. et al. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers Education: Artif. Intell. Jun.* **1**, 6:100209 (2024).
33. https://www.longshot.ai/blog/ai-detectors-accuracy Accessed 17 February 2025.

Acknowledgements

The authors would like to thank all participants for their contributions to this investigation and would like to thank Assoc. Prof. Ts Dr. Wan Muhamad Amir W Ahmad for his help during statistical consultation.

Author contributions

Matheel AL-Rawas, Tahir Yusuf Noorani: Investigation, Data analysis, Conceptualization, Validation, Visualization, methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration. Omar AJ Abdul Qader, Johari Yap Abdullah: Analysis, Validation, Visualization. Nurul Hanim Othman, Noor Huda Ismail: Supervision, Writing – review & editing , Writing – review & editing. Rosnani Mamat, Mohamad Syahrizal Halim: Validation, Writing – original draft, Writing – review & editing.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Ethical approval was obtained from the Human Research Ethics Committee of Universiti Sains Malaysia (reference number USM/JEPeM/KK/24010127). Full informed consent was also acquired from every participant. The study was carried out in accordance with the ethical standards and the rights and confidentiality of all participants was protected. Additionally, the study complied to all the ethical guidelines set forth by the declaration of Helsinki.

Additional information

Correspondence and requests for materials should be addressed to J.Y.A. or T.Y.N.