scientific reports

Check for updates

OPEN

Human versus artificial intelligence: investigating ability of young academics from research and non-research institutions to identify ChatGPT-generated dental research abstracts

Matheel AL-Rawas[1], Omar Abdul Jabbar Abdul Qader[2], Galvin Sim Siang Lin[3], Yew Hin Beh[4], Muhammad Annurdin Sabarudin[5], Yee Ang[6], Jun Fay Low[7], Johari Yap Abdullah[8,9]✉ & Tahir Yusuf Noorani[9,10]✉

The rapid adoption of generative artificial intelligence (AI) tools such as ChatGPT in academic writing raises concerns about research integrity and authorship transparency, including in dentistry. The aim of this study was to investigate whether young dental academicians from research and non-research universities can differentiate original abstracts from ChatGPT-generated abstracts, and to compare their performances, and accuracy with three AI-output detectors, and a similarity detector. In this study, six early-career academicians (≤ 2 years of academic experience) from 6 different universities reviewed 150 dental research abstracts (75 original and 75 ChatGPT-generated) under blinded conditions and assessed abstract quality using a previously developed rubric. The same abstracts were also evaluated using the GPT-2 Output Detector, Writefull GPT Detector, GPTZero, and Turnitin similarity detection. Blinded human reviewers and most AI tools made variable wrong assumptions. Correlation analyses showed significant positive associations between abstract type and all assessment variables, while similarity detection demonstrated an inverse relationship ($p < 0.05$). Overall, young academicians, regardless of institutional category, had difficulty identifying the origin of AI-generated abstracts, whereas GPTZero showed the highest discrimination accuracy (90.0%). This indicates that early-career status and current level of training/exposure to AI-assisted writing may hold greater significance than the institutional category alone. These findings suggest that relying on human judgment alone is insufficient for identifying AI-assisted academic text and that selected detection tools may support academic integrity safeguards as AI writing technologies continue to evolve.

Keywords Dentistry, AI-generated text, Academic ethics, Early career educators, AI detection tools, Research integrity

Abbreviations

AI	Artificial intelligence
HI	Human intelligence
RUs	Research universities
GUYA	Governmental university young academicians
PUYA	Private university young academicians

[1]Prosthodontic Unit, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kota Bharu, Kelantan, Malaysia. [2]Deanship, College of Dentistry, Al Mashreq University, Airport Street, Baghdad, Iraq. [3]Department of Restorative Dentistry, Kulliyyah of Dentistry, International Islamic University Malaysia, Kuantan Campus, 25200 Kuantan, Pahang, Malaysia. [4]Department Restorative Dentistry, Faculty of Dentistry, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia. [5]Department of Periodontology

and Community Oral Health, Faculty of Dentistry, Universiti Sains Islam Malaysia, Jalan Pandan Utama, 55100 Kuala Lumpur, Malaysia. [6]Department of Restorative Dentistry, Faculty of Dentistry, MAHSA University, Bandar Saujana Putra, Jenjarom, Selangor, Malaysia. [7]Department of Restorative Dentistry, Faculty of Dentistry, Lincoln University College, 47301 Petaling Jaya, Selangor, Malaysia. [8]Craniofacial Imaging Laboratory, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kubang Kerian, 16150 Kota Bharu, Malaysia. [9]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. [10]Conservative Unit, School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kubang Kerian, Kota Bharu, Kelantan, Malaysia. ✉email: johariyap@usm.my; dentaltahir@yahoo.com

Artificial intelligence (AI) is an exceptionally debated subject nowadays, with little consensus about the differences and overlaps between human intelligence (HI) and AI[1]. ChatGPT is an AI-driven large language model (LLM) that has been trained on extensive text datasets in many languages. ChatGPT is a chatbot developed by OpenAI (OpenAI, L.L.C., San Francisco, CA, USA)[2]. It is capable of generating text responses that closely resemble those of a human when given input by using a neural network to analyze natural language[3].

ChatGPT was met with a variety of reactions in the scientific and academic communities, reflecting the long-standing debate about the potential advantages and disadvantages of cutting-edge AI-based technologies[4–6]. ChatGPT and other chatbots are increasingly used for drafting and editing academic text, including in the health sciences. Although they may improve efficiency and language quality, they raise concerns about research integrity, authorship transparency, and the spread of content that can appear convincing while containing factual errors or fabricated elements[4,7]. As a result, it's easy to see why controversy and alarm surfaced so quickly once ChatGPT became widely available. Multiple fields contributed to the widespread interest in ChatGPT[8–10].

There are difficulties associated with ChatGPT that relate to the conducting of research. These tools for AI can be employed throughout the research process to generate hypotheses, plan experiments, compose papers, and comprehend scientific findings, which could be advantageous in resource-constrained environments where financial resources and human knowledge may be restricted[11]. Users of those AI chatbots have pointed out issues such as incomplete citations or references to research studies that are non-existent, hence underscoring certain limitations of the extensive language model[12].

AI-assisted writing may evade conventional plagiarism cues because it can produce low text similarity while still obscuring the degree of machine involvement and masking problems such as incomplete or nonexistent citations[12]. Evidence from scientific publishing indicates that AI-generated scientific abstracts can be difficult to distinguish from human-authored ones, highlighting a growing need for empirical research[13–15].

A key question is whether human judgment—particularly among early-career academics—can reliably identify AI-assisted scientific writing, and how this compares with AI detection tools. Prior works suggested that humans struggled to detect AI-generated text and that AI detectors accuracies varied by tool, domain, and the model version that generated the text[13–17]. Although an academician's years of experience may not be a critical element, academic professionals, particularly those with extensive experience, demonstrated proficiency in detecting AI-generated texts compared to their junior peers[14,18]. These studies support the need for young academics' specific evaluation to inform education and academia[14,18].

Early-career academics' or young academics' research competencies are shaped by research skills development and institutional support, which may influence how they evaluate scholarly writing[19,20]. Those in research-intensive governmental universities often receive better training in critical analysis, academic writing, and methodology, enhancing their ability to assess abstract authenticity. In contrast, non-research private institutions may prioritize teaching over research, limiting exposure to rigorous scholarly evaluation. Additionally, universities with strong research cultures provide greater access to mentorship, funding, and collaborative opportunities, further refining analytical skills[20]. These factors suggest that the institutional environment plays a crucial role in shaping young academics' research competencies. Therefore, the aim of this study was to investigate the ability of six blinded young academicians from three research-based governmental and three non-research-based private universities to differentiate and evaluate the quality of original abstracts versus ChatGPT generated abstracts and to investigate the influence of those universities on the young academicians' performance.

Human judgment on AI content relies on experience, intuition, and contextual comprehension, whereas AI detection tools employ statistical patterns and probabilities[5]. Understanding their strengths and weaknesses can improve detection strategies. Furthermore, limited studies have assessed the impact on the accuracy of HI versus AI on content detection. Therefore, we evaluated how HI compares with AI-output detectors and similarity detection. The main hypothesis proposes that young academics from research-based universities will outperform their counterparts from non-research-based institutions.

Methodology

This was a cross-sectional study involving six young academic reviewers and three AI detectors, and a similarity detector. The study was carried out in the School of Dental Sciences and involved only academicians with higher education in dental specialties. Utilizing the G*power version 3.1.9.6 software (Heinrich-Heine-Universität Düsseldorf), a sample size of 135 abstracts was determined based on a significance level (α) of 0.05, power of 0.8, and effect Size of 0.3. Ultimately, a sample size of 150 abstracts was selected to guarantee that all reviewers receive an equal number of abstracts. The Human Ethics and Research Committee of Universiti Sains Malaysia granted ethical approval; reference number USM/JEPeM/KK/24,010,127. Full informed written informed consent was acquired from every participant.

A total of 75 titles and original abstracts were collected using random sampling from recent issues (published in the first five months of 2023) of 8 high-impact dental journals using systematic sampling. Abstracts only were selected because they are the most visible and frequently screened component of a paper and have been used

in prior human-versus-detector comparisons[13]. The inclusion criteria were: the abstracts from dental-related journals; abstracts from original research papers; abstracts written in English only; abstracts that were structured and non-structured. The exclusion criteria were: the abstracts that contained figures or abstracts that were present in a case report, case series, or any type of review. Moreover, we created 75 abstracts by using ChatGPT (Version 3.5)[21] by using the same titles from the extracted abstracts of the original scientific dental papers. Finally, the total number of abstracts were 150. ChatGPT was given instructions to create a research abstract for the study titled "Title of the original abstract" available at [link]. ChatGPT lacks autonomous internet browsing capabilities and its training data is limited to the period until April 2023. To account for ChatGPT's sensitivity to prior requests within the same chat, we executed each prompt in a separate session[13]. These abstracts produced will henceforth be referred to as AI-abstracts.

Following the identification of the abstracts to be used in the study, we assigned each abstract, regardless of whether it is original or AI-Abstract, to six young academic assessors at random. Their task was to evaluate the abstracts and determine their originality. To guarantee blinding, all identifying information from the abstracts was deliberately eliminated. This comprised the name, title, and related details of the journal. Each abstract was assigned a distinct identifier to guarantee blinding by using an online research randomizer[22] that produced unique numbers for each abstract. Each abstract was securely stored to prevent any unauthorized access or tampering.

Six academic investigators, with less than two years of academic experience in dentistry, were recruited from 5th March 2023 till 16th March 2023, and were divided into three research based governmental university young academicians (GUYA) and three non-research based private university young academicians (PUYA). Those investigators were selected according to random sampling. The inclusion criteria of young academicians were: academicians with experience of less than 2 years in the dental academia; academicians with acceptable publication history and peer reviewing. The exclusion criteria were: Non academicians and academicians from non-dental academia. Six young reviewers were selected and considered sufficient to ensure feasibility of blinded text detection and rubric-based scoring and to be consistent with similar studies with blinded detection designs[13,14,23]. Each assessor was assigned an equal number of original and AI-generated abstracts (25 original and AI-Abstracts for each chosen academician).

The researchers assessed the originality of the abstracts and determined if they were produced by AI following specific predefined criteria. Following the assessment, each abstract was assigned a score of either 1 or 2, with 1 representing an original abstract. Consequently, the abstracts were evaluated using a previously developed rubric[14], as presented in Table 1, which considered specific criteria including language complexity, coherence, creativity, contextual understanding, grammatical accuracy, and domain-specific knowledge. They assigned a number ranging from 1 (Lowest) to 3 (Highest) to each criterion, depending on the level of perfection shown by the abstract in that particular criterion. The six criteria scores were computed and then divided by 18 to obtain either one of the four distinct ratings: 1 for excellent quality, 0.9 to 0.7 for good quality, 0.6 to 0.4 for average quality, and 0.1 to 0.3 for poor quality.

Upon completing the initial phase of data collection, we assessed the AI abstracts for similarity detection using Turnitin software (2023 version)[24]. This software provides a similarity score ranging from 0% to 100%, where 100% indicates complete detection of similarity. Turnitin software was used because it was the institutionally available similarity-checking tool during the study period.

An evaluation of the abstracts was conducted using three AI output detectors: the GPT-2 Output Detector (2023 Version 2023)[25], the Writefull GPT detector (Version 2023)[26], and GPTZero (Version 2023)[27]. The GPT-2 Output Detector is an AI tool that assigns a probability estimate to abstracts, ranging from 0 ('real') to 100% ('fake'). A higher score suggests that the abstract is more likely to have been produced by an AI tool. The Writefull GPT detector and GPTZero detector provide results expressed as a percentage indicating the likelihood that the text originates from GPT-3, GPT-4, or ChatGPT. This can facilitate the evaluation process and distinguish different levels of originality in abstracts.

Each AI detector method employs distinct computational strategies to improve detection accuracy[25–27]. In this study, the three AI detector tools were considered acceptable when compared to another study[28]. Using

Criteria	Definition	Scoring		
		1	2	3 (Highest)
Cohesion	It refers to the logical connections between ideas in the abstract.	The abstract lacks logical connections between ideas.	The abstract has some logical connections between ideas.	The abstract has strong logical connections.
Creativity	It refers to the originality of the phrasing and content in the abstract.	The abstract is formulaic and lacks originality.	The abstract includes some original phrasing and content.	The abstract exhibits strong creativity and originality.
Contextual understanding	It refers to the depth of understanding the abstract demonstrates about the research context and significance.	The abstract lacks depth and understanding of the research context.	The abstract demonstrates some understanding of the research.	The abstract demonstrates a deep understanding of the research.
Grammatical accuracy	It refers to the correctness of the grammar and syntax used in the abstract.	The abstract has many grammatical errors.	The abstract has some grammatical errors.	The abstract has excellent grammatical accuracy.
Language complexity	It refers to the level of technical language or jargon used in the abstract.	The language is simple.	The language is somewhat technical.	The language is highly technical.
Domain-specific knowledge	It refers to the use of specialized terminology and knowledge specific to the research field in the abstract.	The abstract lacks domain-specific knowledge.	The abstract shows some understanding.	The abstract demonstrates a strong understanding.

Table 1. Newly formulated rubric designed to evaluate the quality of abstracts[20].

multiple detectors is recommended because AI-detection performance is known to vary by tool, domain, and model version, and no single detector is consistently reliable across contexts[14,28].

Prior to the completion of the analysis, assessors weren't provided with any data analysis. Each reviewer was allotted a period of 7 to 10 days to submit their findings. Reviewers were given instructions not to use AI detection tools or perform internet searches for the sources of abstracts throughout the evaluation process. By following these procedures, one can carry out a study with meticulousness and impartiality, thus enhancing the credibility and dependability of the findings.

Outcome values for all variables were obtained by crosstabulation. A crosstabulation was conducted on the abstract type results and quality assessment results from each reviewer of both GUYA and PUYA to determine their results and explore any potential correlation. A non-parametric Chi-Square analysis was conducted to examine the differences among reviewers in terms of their abstract type results and quality assessment results. A bivariate correlation analysis using Spearman's correlation coefficient was conducted to evaluate the magnitude and direction of the association between all variables. The statistical significance level was established at 0.05. Percentages of sensitivity, specificity, and accuracy of all assessment methods were calculated using MedCalc Software Ltd. evaluation calculator (version 2023)[28,29].

Results

Identification of the abstract origin by the young academics

Six young academics (GUYA1–3, PUYA1–3) evaluated a total of 150 abstracts (75 original and 75 ChatGPT-generated). As shown in Fig. 1, the number of incorrectly classified abstracts varied across reviewers, ranging from 6 (lowest) to 14 (highest). When comparing the results of the reviewers, only three of them (GUYA3, PUYA1, and PUYA3) had statistically significant differences in their categorization outcomes ($P < 0.05$). The other reviewers did not have any significant differences ($P > 0.05$).

Rubric-based quality assessment

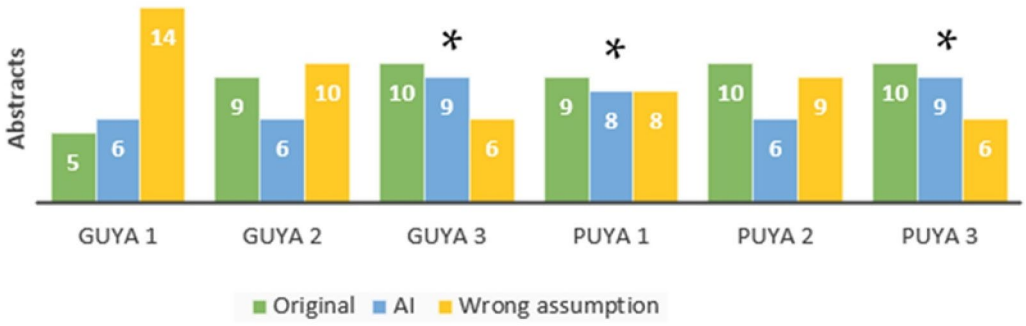
Using the previously developed rubric[14], reviewers mostly rated original abstracts as excellent to average, and none were classified as poor. AI-generated abstracts, on the other hand, were mostly rated as average, with a few scored as poor (Fig. 2). The majority of reviewers' rubric scores were not statistically significant, with the exception of GUYA3 and PUYA3 ($P < 0.05$). The non-parametric Chi-square analysis indicated statistically significant differences among reviewers regarding both abstract-type classification and quality ratings (Table 2; $P < 0.05$).

Performance of AI-output detectors and similarity detection

Cross-tabulation analysis revealed that all three AI-output detectors were significantly associated with abstract type (Table 3; $P < 0.05$), but with different strengths of association. The GPT-2 Output Detector had a weak correlation (Phi = 0.236), the Writefull GPT Detector had a moderate correlation (Phi = 0.357), while GPTZero showed a very strong correlation with abstract type (Phi = 0.923), indicating greater differentiation between original and AI-generated abstracts within this dataset. For similarity detection (Turnitin), all original abstracts showed 100% similarity (very high similarity); however, AI-generated abstracts were mostly in the low to moderate similarity category (Table 3). There was a very high correlation between the type of abstract and the similarity scores (Phi = 1.000, $P < 0.001$).

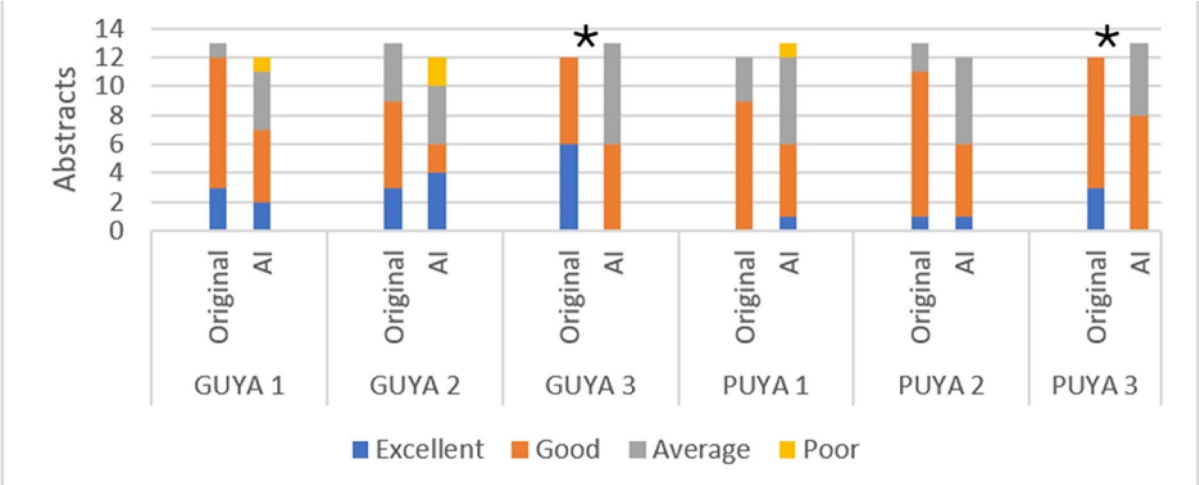
Correlation analysis

Spearman correlation analysis (Fig. 3) revealed statistically significant associations between abstract type and the study variables. The correlations were positive (very weak to weak) for the outcomes of human reviewers, and for the GPT-2 output and Writefull detectors. However, GPTZero had a very strong association with abstract type.



GUYA = governmental university young academician (research-based); PUYA = private university young academician (non-research-based). *P < 0.05 indicates a statistically significant difference in reviewer outcomes.

Fig. 1. Classification outcomes of the six young dental academicians when identifying abstract origin (original vs. ChatGPT-generated). Bars show the number of abstracts classified as original and AI-generated and the number of incorrect classifications ("wrong assumptions") for each reviewer



GUYA = governmental university young academician (research-based); PUYA = private university young academician (non-research-based). *P < 0.05 indicates a statistically significant difference in quality outcomes.

Fig. 2. Rubric-based quality ratings assigned by the six young dental academicians for original and ChatGPT-generated abstracts. Bars represent the number of abstracts rated by each reviewer and as excellent, good, average, or poor for abstract type

Similarity detection exhibited a strong association in the opposite direction, indicating that original abstracts clustered at the highest similarity score, whereas AI-generated abstracts displayed lower similarity values.

Accuracy of the study variables

The overall accuracy of human reviewers in detecting abstract origin varied from 44% to 76% (Fig. 4). The accuracy of quality rating based on rubrics was between 64% and 76%. GPTZero was the most accurate AI-output detector, with an accuracy rate of 90%. This was better than GPT-2 Output (55%) and Writefull GPT (58%). The similarity detector was 94% accurate (Fig. 4).

Governmental research-based university versus private non-research-based university

Figure 1 indicates that both GUYA (research-based) and PUYA (non-research-based) reviewers made many incorrect classifications. This pattern showed that inter-individual variability outweighed group-level separation between GUYA and PUYA. In abstract quality evaluation, Fig. 2 shows a broadly similar rating pattern in both groups. This suggests that performance differed mostly by reviewer rather than institution.

Table 2 shows substantial differences in abstract-type classification and quality results among reviewers ($P < 0.05$). The direction of these differences did not translate into research-based institutional advantage; instead, it supported heterogeneous performance within each group. Therefore, institutional setting appeared to be an inconsistent predictor of detection capability at an early career stage, while individual differences (e.g., prior exposure to research writing, reviewing experience, and familiarity with AI-assisted writing) may play a larger role.

Discussion

Since its public release in late 2022, ChatGPT has gained immense popularity worldwide, particularly throughout 2023, leading to widespread discussions about its implications. OpenAI recognizes certain constraints of the chatbot, and stakeholders are actively engaging and conducting experiments to discover further capabilities[15]. The ramifications of ChatGPT and other open AI platforms in higher education are the subject of intense discussion, with immediate concerns expressed over the impact on education due to the lack of complete transparency in the internal operational mechanisms of ChatGPT[15]. It is important to note that these platforms can be utilized by students to produce assignments and dissertations, giving rise to concerns of plagiarism and academic integrity[30].

According to recent research conducted[14,19,31], ChatGPT has the potential to be a valuable instrument for composing abstracts or even introduction sections. Through this study, ChatGPT effectively produced abstracts, leading to all reviewers rating some of the AI-generated abstracts as authentic abstracts.

Concurrent review of generative AI guidelines related to dental education underscored the importance of transparency, ethical utilization, academic integrity, and detection/verification practices, while also acknowledging the scarcity of dental-specific guidance and studies emphasizing a deficiency that predominantly impacts early-career professionals[32]. The challenges faced by young dental academicians in differentiating AI-generated abstracts from original ones[14] underscored the necessity for focused institutional initiatives that integrate AI literacy (critical evaluation of AI-assisted writing, disclosure standards, and ethical principles) with practical verification processes and mentorship-driven research skills enhancement[32]. To date, only one

Chi-Square non-parametric test for differences			
GUYA 1, 2 and 3			
	df	Pearson chi square*	p-value
Reviewer outcome	1	13.500	< 0.001
Quality outcome	3	63.344	< 0.001
PUYA 1, 2 and 3			
	df	Pearson chi square	p-value
Reviewer outcome	1	21.660	< 0.001
Quality outcome	3	92.411	< 0.001
GUYA 1 and PUYA 1			
	df	Pearson chi square	p-value
Reviewer outcome	1	6.250	0.012
Quality outcome	3	48.867	< 0.001
GUYA 1 and PUYA 2			
	df	Pearson chi square	p-value
Reviewer outcome	1	9.610	0.002
Quality outcome	3	55.217	< 0.001
GUYA 1 and PUYA 3			
	df	Pearson chi square	p-value
Reviewer outcome	1	6.250	0.012
Quality outcome	3	65.617	< 0.001
GUYA 2 and PUYA 1			
	df	Pearson chi square	p-value
Reviewer outcome	1	13.690	< 0.001
Quality outcome	3	30.917	< 0.001
GUYA 2 and PUYA 2			
	df	Pearson chi square	p-value
Reviewer outcome	1	18.490	< 0.001
Quality outcome	3	36.367	< 0.001
GUYA 2 and PUYA 3			
	df	Pearson chi square	p-value
Reviewer outcome	1	13.690	< 0.001
Quality outcome	3	43.967	< 0.001
GUYA 3 and PUYA 1			
	df	Pearson chi square	p-value
Reviewer outcome	1	11.560	< 0.001
Quality outcome	3	44.517	< 0.001
GUYA 3 and PUYA 2			
	df	Pearson chi square	p-value
Reviewer outcome	1	16.00	< 0.001
Quality outcome	2	28.844	< 0.001
GUYA 3 and PUYA 3			
	df	Pearson chi square	p-value
Reviewer outcome	1	11.560	< 0.001
Quality outcome	2	15.940	< 0.001

Table 2. Non-parametric Chi-square analysis of the differences in the abstract type outcomes and quality assessment between and among young academic reviewers. *Cells (less than 20%) have expected count less than 5. ^Significance level is at 0.05

published study has directly compared young versus senior dental academicians in identifying ChatGPT-generated dental abstracts[14]. To our knowledge, no prior dental study has focused specifically on early-career dental academicians across different institutional settings (research vs. non-research) while benchmarking their performance against multiple AI-output detectors and similarity detection.

In Malaysia, higher education universities are classified into three primary categories, with research universities (RUs) recognized as the principal category that gained focus from the Ministry, including grants, funding, and support compared to non-research universities (non-RUs)[33]. The RU's high-impact outcome is more extensive and produces better output than non-RUs. Therefore, academicians in RUs are more engaged

Variable	GPT-2 output detector (*n*)						
Abstract type	Low fake	Moderate fake	High fake	Very high fake	Pearson Chi-square *	P-value^	Phi
Original abstract	62	6	2	5	8.327	0.04	0.236
AI abstract	48	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
Original abstract	58	1	5	11	19.167	< 0.001	0.357
AI abstract	35	12	13	15			
Variable	**GPTZero detector (n)**						
Abstract type	Low fake	Moderate fake	High fake	Very high fake	Pearson Chi square	P-value	Phi
Original abstract	75	0	0	0	127.778	< 0.001	0.923
AI abstract	6	9	13	47			
Variable	**Similarity detector outcome (n)**						
Abstract type	Low similarity	Moderate similarity	High similarity	Very high similarity	Pearson Chi square	P-value	Phi
Original abstract	0	0	0	75	150.000	< 0.001	1.000
AI abstract	29	38	8	0			

Table 3. The assessment outcomes of the abstracts by AI and similarity detectors. *Cells (less than 20%) have expected count less than 5. ^Significance level is at 0.05.

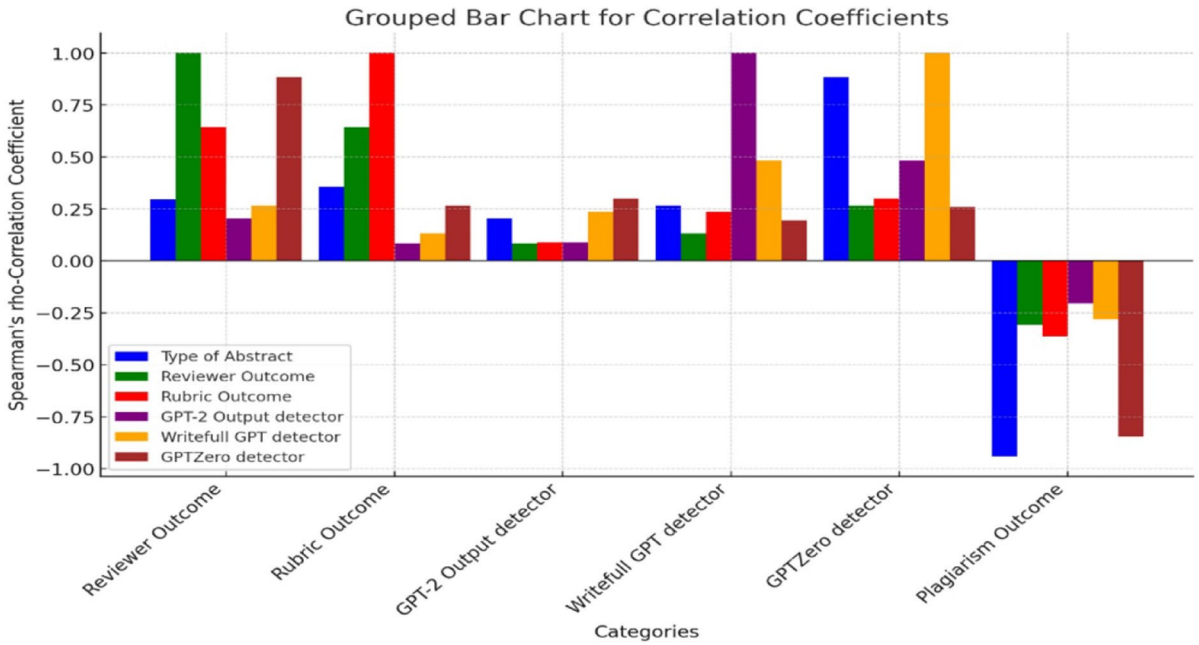


Fig. 3. Spearman correlation analysis among study variables, including abstract type, human reviewer outcome, rubric quality outcome, AI-output detector scores (GPT-2 Output, Writefull GPT, GPTZero), and similarity detection outcome. Values indicate the direction and strength of correlations between variables.

in research-related matters. Research engagement among young academicians can be categorized into several activities, including intensive publication in high impact peer-reviewed journals involving professional bodies, presenting papers in conferences, and many other responsibilities appointed by the institution[20]. Location, whether in a government (research based) or private school (non-research based), may not be a determining factor as proven in this study. However, GUYAs are more likely to be engaged in research activities because of the financial resources available for research in research-based universities. Moreover, GUYA has more expertise in writing research papers and peer reviewing for academic journals compared to PUYA. One GUYA and one PUYA in this study achieved a score of 19 out of 25 abstracts correctly identified, as indicated in Fig. 1. Table 2 demonstrates that all the differences among the young academic reviewers were statistically significant ($P < 0.05$). One possible explanation for this could be the variation in scores provided by the reviewers for each abstract and the level of incorrect assumptions. However, no comparison with other previous studies could be made as no similar study has been carried out before.

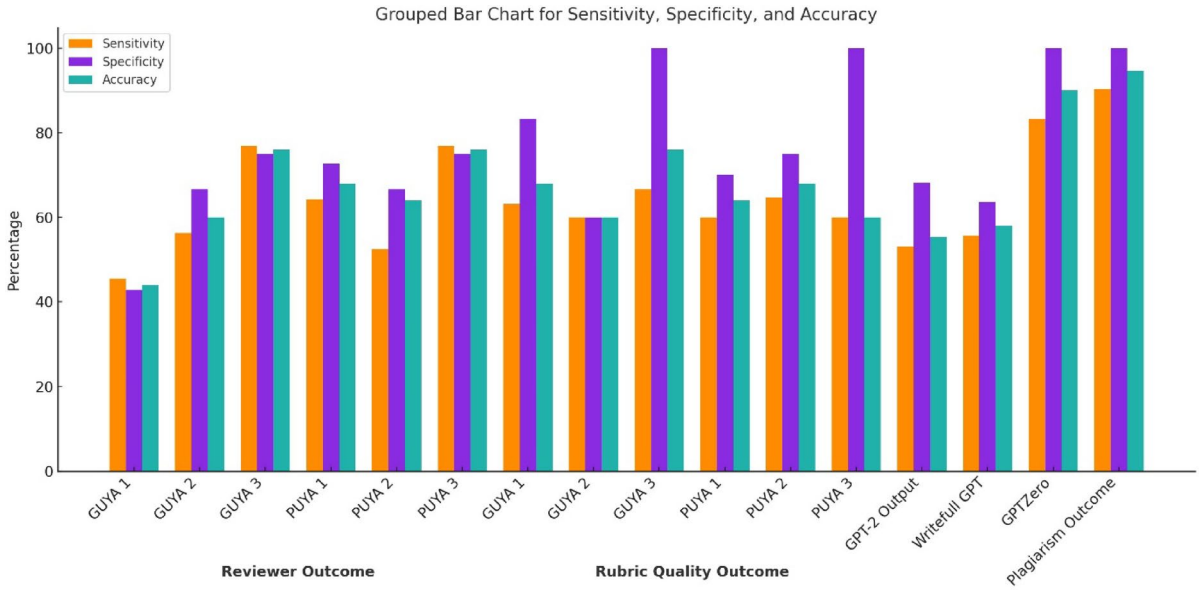


Fig. 4. Sensitivity, specificity, and accuracy of each assessment method for differentiating original versus ChatGPT-generated abstracts, including human reviewers (GUYA/PUYA), rubric-based quality assessment, AI-output detectors, and similarity detection.

Researchers conducted a study evaluating the capacity of novice teachers (average age 25) and experienced teachers (average age 42) to identify AI-generated content in student submissions. The findings revealed that both novice and experienced teachers encountered difficulties in differentiating between AI-generated and human-authored texts[16]. Another study examined the ability of young academicians (with two years or less of experience) versus experienced academicians (with ten years or more of experience) in recognizing ChatGPT-generated abstracts. The findings revealed that all participants exhibited varying degrees of erroneous assumptions during the identification process, highlighting the significant difficulties faced by academicians in distinguishing between human-generated and AI-generated abstracts[14]. Two explanations for scholars' inability to recognize AI-generated content: AI models excel at emulating human writing styles, producing content that is grammatically correct and contextually coherent, potentially misleading readers into thinking it is composed by a genuine human author. Secondly, scholars' evaluations may be flawed due to their reliance on erroneous heuristics for recognizing AI-generated language[17]. One study demonstrated humans can correctly identify AI-generated text approximately 53% of the time[17]. In this study, the abstract identification accuracy of all young academics ranged from 44% to 76%, as shown in Fig. 4, which is in line with the other study[17].

The rubric, which was specifically designed to help reviewers evaluate the quality[14], was successfully used by the reviewers, with good feedback from them. As shown in Fig. 2, all the reviewers scored the original abstracts with excellent to good quality; very few abstracts were scored with average quality, with no poor quality detected. On the other hand, they scored the AI abstracts mostly with average quality, to a lesser extent with good quality. Poor quality was the outcome of four AI abstracts out of 75. Regarding the differences between the academic reviewers utilizing the rubric, all the differences were significant ($P < 0.05$), as shown in Table 2. The accuracy of all young academics in using the rubric in identifying the abstracts' quality was in a range from 64% to 76%, as shown in Fig. 3.

Generally, research has shown that humans have difficulty in differentiating between text produced by AI and text generated by humans. Automated systems, specifically AI bot detection, excel at distinguishing content. AI-generated composition often lacks specificity, creativity, tends to overstate particular instances, and exhibits a distinct writing style characterized by the use of predominantly predicted words, in contrast to human writing. AI systems such as GPTZero have achieved notable success in probabilistically detecting AI-generated text[4,14,28]. In this study, the GPTZero detector was very effective in differentiating AI abstracts from the original abstracts compared to the other two detectors, and in our observation, there were more false outcomes by the GPT-2 Output and Writefull GPT detectors, scoring 48 and 35 AI abstracts, respectively, out of 75 as abstracts with low fake content, as shown in Table 3. In Fig. 3, all detectors showed significant correlations with the abstracts' type ($p < 0.05$) with a stronger positive strength of the GPTZero detector as compared to others. AI detectors like GPT-2 output and Writefull GPT face training data constraints and biases. Diversity in training datasets is necessary for accuracy, but insufficient exposure to different writing styles may lead to inaccuracy. Many detectors struggle to stay functional and accurate as AI-generated content advances[34].

In our opinion, GPTZero is a very useful tool for the academicians to identify the source of the text or abstracts. One study stated the GPTZero had an accuracy of 80% in detecting AI generated texts[28]. In this study, GPTZero outperformed other AI detectors (GPT-2 Output 55% and Writefull GPT 58%) with 90% accuracy. Only the similarity detector accuracy was close to GPTZero. If GPTZero demonstrates high accuracy, universities may implement it as a standard tool for detecting AI-generated submissions, thereby enhancing

academic integrity policies. Nevertheless, institutions should refrain from excessive dependence on this tool and incorporate human verification in conjunction with AI detection mechanisms[16]. AI detectors should only be used as decision-support instruments within a transparent, policy-driven human evaluation framework. This is because AI detectors might give false positives and negatives, and their accuracies change over time and in different situations as ChatGPT and other AI chatbots evolve over time[16].

The increasing prevalence and establishment of chatbots and other AI tools in higher education necessitate university educators, researchers, and management to adjust to the fast-evolving digital scene[4]. In this study, Turnitin was used, and it gave a similarity of 100% for all original abstracts, as shown in Table 3. While for the AI abstracts, it gave a score between 25% and 50% similarity rates for 38 AI abstracts out of 75. Only eight AI abstracts were scored with an above 50% similarity rate. This reflected the ability of the ChatGPT to generate abstracts that have low to moderate similarities to texts written by humans. In Fig. 3, the similarity detector by Turnitin showed a significant correlation with the abstract type ($P < 0.05$) and a negative strong power, which means that as the score of similarity is low, the abstract type is mostly AI generated.

The authors of a recently published paper in a respected journal recognized the assistance of ChatGPT in composing the abstract with minimal direct involvement from the authors[31]. There is a legitimate concern that a small number of scholars may use ChatGPT to generate material without properly acknowledging publishers, journal editors, or conference organizers. Another instance involved the publication of an academic article where ChatGPT has been included as a co-author[35]. Although publishers are currently revising their publication policies to tackle these problems, the inclusion of ChatGPT as a co-author gives rise to two profound concerns regarding the credibility of academic research and writing that affects all fields of study[4].

Due to its extensive availability, ChatGPT is generating considerable concern. Several conferences and journals have already established policies that explicitly forbid the use of ChatGPT in their research output. In the midst of publishers' efforts to revise their policies, it is the responsibility of all human authors to understand that any violation of these policies will be considered scientific misconduct, similar to plagiarism of existing research[4]. Furthermore, as the thorough assessment of manuscripts is carried out by humans, not robots, then publishers, editors, and conference committees have a duty to guarantee that reviewers receive appropriate training to assist in reducing the risk of technologies (when used unethically) that could compromise the quality of academic writing and the credibility of our fields[4,16].

In regards to our objectives, the comparison between research-based governmental universities (GUYA) and non-research private universities (PUYA) did not reveal a distinct institutional advantage for identifying AI-generated abstracts. Some reviewers had very different results, but both groups made mistakes when they tried to identify the source of the abstracts. This indicates that early-career status and current level of training/exposure to AI-assisted writing may hold greater significance than the institutional category alone. This suggests that the level of research intensity may not be directly related to the ability to recognize AI-generated scientific writing at the beginning of a career. This means that both types of universities need to offer structured training in AI literacy.

The AI detectors had different levels of accuracy and detection because they used different algorithms, training sets, and decision thresholds. Their performance is also known to be affected by the writing patterns of the domain and the model/version of AI chatbots that generated the data. In our research, GPTZero exhibited the highest level of accuracy, while GPT-2 Output and Writefull demonstrated a greater susceptibility to incorrect categorization. This aligns with findings indicating that detector performance is dependent upon the specific tool used and may deteriorate as AI generative models advance. In practice, these findings suggest that institutions should not depend on a single detector; instead, they should adopt a layered strategy: AI literacy training for young dental academicians and the combined use of detection tools with human oversight for interpretation and ethical judgment.

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. 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 these limitations for a better assessment.

In conclusion, it is evident at present that ChatGPT and its related tools and technologies will persistently impact the field of education. Emphasizing continuous research to mitigate the possible hazards will guarantee the practical application of the advantages of this technology[4,16].

Conclusion

ChatGPT has brought to light the fact that we were not well equipped to attain this AI milestone. As a result of the most recent changes, the regulations governing research and publishing will need modifications. While GPTZero and Turnitin similarity detectors exceled in accurately classifying abstract categories, young academicians from either research and non- research based governmental or private universities struggled in the identification of HI and AI-generated abstracts. This indicates that early-career status and current level of training/exposure to AI-assisted writing may hold greater significance than the institutional category alone. The collaboration between HI and AI and the combined use of detection tools with human oversight have the potential to provide optimal benefits, provided that the contents of the texts created by AI are reliable and correct and that ethical standards and guidelines are adhered to.

Data availability

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

Received: 21 November 2025; Accepted: 26 February 2026
Published online: 05 March 2026

References

1. 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).
2. OpenAI. ChatGPT: Optimizing language models for dialogue [Internet]. 2023 [cited 2023 Feb 23]. Available from: https://openai.com/blog/chatgpt/
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. 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).
5. Howard, J. Artificial intelligence: Implications for the future of work. *Am. J. Ind. Med.* **62**, 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_76_20 (2020).
7. Deng, J. & Lin, Y. The benefits and challenges of ChatGPT: An overview. *FCIS* **2**, 81–83 (2023).
8. 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).
9. 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).
10. 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).
11. Thorp, H. H. ChatGPT is fun, but not an author. *Science* https://doi.org/10.1126/science.adg7879 (2023).
12. Kendrick, C. The efficacy of ChatGPT: Is it time for librarians to go home? [Internet]. Scholarly Kitchen; 2023 [cited 2023 Jan 26]. Available from: https://scholarlykitchen.sspnet.org/2023/01/26/guest-post-the-efficacy-of-chatgpt-is-it-time-for-the-librarians-to-go-home/
13. 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).
14. Al-Rawas, M. et al. Identification of dental related ChatGPT generated abstracts by senior and young academicians versus artificial intelligence detectors and a similarity detector. *Sci. Rep.* **15** (1), 11275. https://doi.org/10.1038/s41598-025-11275-0 (2025).
15. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **613** (7944), 423. https://doi.org/10.1038/d41586-023-00001-9 (2023).
16. Perkins, M. et al. 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/s10805-023-09460-7 (2024).
17. Fleckenstein, J. et al. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Comput. Educ. Artif. Intell.* **6**, 100209 (2024).
18. 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).
19. Javier, C. To teach or not to teach? Junior academics and the teaching-research relationship. *High. Educ. Res. Dev.* **41**, 1417–1435. https://doi.org/10.1080/07294360.2021.1933395 (2022).
20. Fauzi, M. A. Research vs. non-research universities: Knowledge sharing and research engagement among academicians. *Asia Pac. Educ. Rev.* **24**, 25–39. https://doi.org/10.1007/s12564-021-09719-4 (2023).
21. OpenAI. ChatGPT [Internet]. Available from: https://chat.openai.com [cited 2023 Jan 15].
22. Randomizer.org [Internet]. Available from: https://www.randomizer.org/#randomize [cited 2023 Jan 19].
23. 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* https://doi.org/10.1101/2023.02.22.23286322 (2023).
24. Turnitin The launch of Turnitin's AI writing detector and the road ahead [Internet]. 2023 [cited 2023 Apr 5]. Available from: https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead
25. OpenAI AI Detector [Internet]. Available from: https://openai-openai-detector.hf.space [cited 2023 Feb 3].
26. Writefull, G. P. T. & Detector [Internet]. Available from: https://x.writefull.com/gpt-detector [cited 2023 Feb 5].
27. GPTZero [Internet]. Available from: https://gptzero.me [cited 2023 Feb 8].
28. Habibzadeh, F. GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study.. *J. Korean Med. Sci.* https://doi.org/10.3346/jkms.2023.38.e319 (2023).
29. MedCalc Diagnostic Test Evaluation Calculator [Internet]. Available from: https://www.medcalc.org/calc/diagnostic_test.php [cited 2023 Dec 20].
30. 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).
31. Bouschery, S. G., Blazevic, V. & Piller, F. T. Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *J. Prod. Innov. Manag.* **40**(2), 139–53. https://doi.org/10.1111/jpim.12655 (2023).
32. Uribe, S. E., Maldupa, I. & Schwendicke, F. Integrating generative AI in dental education: A scoping review of current practices and recommendations. *Eur. J. Dent. Educ.* **29**(2), 341–355. https://doi.org/10.1111/eje.13074 (2025).
33. Ministry of Higher Education Malaysia (MOHE). *Enhancing University Transformation Programme, Silver Book: Academic and Cost Efficiency* (MOHE, 2017).
34. LongShot, A. I. AI detectors accuracy [Internet]. Available from: https://www.longshot.ai/blog/ai-detectors-accuracy [cited 2025 Feb 13].
35. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. ChatGPT: Five priorities for research. *Nature* **614**, 224–6 (2023).

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, Galvin Sim Siang Lin, Johari Yap Abdullah: Analysis, Validation, Visualization, Project administration. Yew Hin Beh, Muhammad Annurdin Sabarudin: Supervision, Writing—review & editing, Writ-

ing—review & editing. Yee Ang, Jun Fay Low: 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 Ethics and Research Committee (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.