



Received: April 1, 2025  
Revised: May 19, 2025  
Accepted: June 12, 2025

**Corresponding Author:**

Vini Mehta, Department of Dental Research Cell, Dr. D. Y. Patil Dental College and Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), Pimpri, Pune 411018, India  
E-mail: vini.mehta@statsense.in

# Assessing the Applicability of Artificial Intelligence (AI) in Fabrication and Plagiarism of Scientific Literature in the Field of Orthodontics

Saeed N. Asiri<sup>1</sup>, Anwar S. Alhazmi<sup>2</sup>, Vini Mehta<sup>3</sup>

<sup>1</sup>Department of Pediatric Dentistry, College of Dentistry, Prince Sattam Bin Abdulaziz University, Saudi Arabia

<sup>2</sup>Department of Preventive Dentistry, College of Dentistry, Jazan University, Saudi Arabia

<sup>3</sup>Department of Dental Research Cell, Dr. D. Y. Patil Dental College and Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), Pimpri, Maharashtra, India

## Abstract

**Objectives:** The study aims to evaluate the applicability of Chat Generative Pre-trained Transformer (ChatGPT-4) to fabricate an orthodontic research paper and assess the ability of human reviewers (clinicians and academic reviewers) and artificial intelligence (AI) detection tools to identify such fabricated content.

**Methods:** This study employed a descriptive exploratory research design. ChatGPT-4 was used to generate a research paper on the orthodontic topic. A search strategy was used to extract relevant research studies from three databases, then rewritten using ChatGPT-4. A panel of 10 orthodontic experts were given a mix of AI-rewritten and human-written abstracts to identify fake and real ones. Additionally, AI detection tools were tested for their efficiency in identifying AI-generated content. The results were analyzed using descriptive statistics and comparative percentages.

**Results:** Orthodontic experts and researchers correctly identified only 68% of AI-generated abstracts as fake, while 86% of human-written abstracts were recognized as real. Online AI detection tools demonstrated weak performance, with overall sensitivity and specificity of just 89% and accuracy of 44.5%. GPTZero had 100% specificity and 100% predictive values, along with the highest sensitivity and accuracy compared to others. However, writers AI demonstrates the poorest performance, indicating 0 predictive values in identifying AI-fabricated research works.

**Conclusions:** Reviewers experienced difficulties in distinguishing between AI-generated abstracts and those written by humans due to the substantial resemblance to real papers. To safeguard scientific integrity, it is crucial to use measures such as improving detection tools, refining peer review, and creating a transparent culture around AI use.

**Keywords:** artificial intelligence, chatGPT, orthodontics, plagiarism, scientific integrity

## Introduction

The integrity in scientific research is paramount to advance scientific knowledge and improve patient outcomes.<sup>(1)</sup> The rapid development in the field of artificial intelligence particularly large language models (LLMs) has further aided in increasing productivity in the research.<sup>(2)</sup> These artificial intelligence (AI) models have the ability to generate coherent texts in seconds, thereby accelerating scientific writing tasks such as draft automation, article summarization, and translation.<sup>(3)</sup> Chat Generative Pre-trained Transformer (ChatGPT), a LLM powered by deep neural networks, has garnered significant attention, attracting millions of active users within just a few months of its launch.<sup>(4)</sup> Many researchers have praised it for its usefulness in speeding up writing process, grammar, and sentence restructuring, suggesting creative ideas for scientific work, developing research outlines, citation and referencing.<sup>(5-7)</sup>

However, despite its potential, scientific community have pointed out the potential for AI based fabrication and plagiarism in research on using ChatGPT.<sup>(8)</sup> The use of this AI chatbot has been evolving into a potential "Weapon of Mass Deception" in research, contributing to the publication of questionable papers in academic journals. This has emphasized the necessity for rigorous assessments of ChatGPT generated scientific outputs.<sup>(9-11)</sup> Moreover, there are many apprehensions that its accessibility may decrease critical thinking and creativity among the scientific community.<sup>(12)</sup> Studies have also highlighted concerns about ChatGPT's use of existing data and its tendency to generate fabricated information when prompted with topics beyond its training dataset.<sup>(4,7)</sup> This introduce limitation in its reliability and can undermine the ethical use of AI in academic writing.<sup>(13,14)</sup> Studies has also identified the risk of fake image generation using these intelligent models for fraudulent practices in research.<sup>(15,16)</sup> Besides, the ease with which AI can generate convincing yet entirely fabricated research papers pose another significant threat to the scientific literature. These AI-generated works can undermine the credibility of legitimate research and lead to misguided clinical decisions.<sup>(9)</sup> Furthermore, since ChatGPT is trained on a vast amount of internet-based data, it may unintentionally generate phrases that replicate or closely resemble previously published material without providing appropriate credit or citations, thereby increasing the

risk of unintentional plagiarism.<sup>(6)</sup> Although there are many significant tools that support plagiarism checks and AI-generated text, the ability of the AI-assisted paraphrasing tools has shown risks of passing the plagiarism checkpoints.<sup>(17)</sup> Else H, in 2023, demonstrated similar concerns in his paper indicating that "Abstracts Written by ChatGPT Fool Scientists".<sup>(18)</sup> This raises concerns about the integrity and accuracy of using ChatGPT in academic writing.

In the field of orthodontics, the importance of accuracy in diagnoses, treatment planning, and predicting treatment outcomes has been emphasized to treat patients efficiently.<sup>(1)</sup> Evidence-based research has been critical for improvement in patients' health conditions in the orthodontics field through treatment planning, appliance designing and development of patient care protocol.<sup>(1)</sup> However, with the growing pressure to conduct research and development, the risk of publishing fabricated papers has also risen in this highly technical field which pose ethical risk such as misinformation, data fabrication, compromised academic integrity and clinical misjudgement.<sup>(19)</sup> Therefore, it is essential to uphold the authenticity of orthodontic research to address the heightened risks of fabrication and plagiarism, which may ultimately affect the academic credibility for patient care and well-being.<sup>(20)</sup> Therefore, this study aims to evaluate the capability of an AI chatbot—ChatGPT-4—to fabricate an orthodontic research paper and assess the effectiveness of human reviewers (clinicians and academic reviewers) and AI detection tools in identifying such fabricated content. The primary objective of this study is to answer a very important research question investigating that how effective is ChatGPT-4 in fabricating an orthodontic research paper, and what is the accurately can human reviewers identify the AI fabricated research work from human-generated. Furthermore, this study also investigates how effective are current online AI detection tools in identifying paraphrased or reworded AI-generated research papers. Additionally, the study seeks to explore the implications of these findings for the orthodontic research community. Understanding of these AI implications is vital to safeguard the credibility of orthodontic research and ensure responsible adoption of AI technologies in scholarly communication and evidence based clinical decision making.

## Materials and Methods

This study employed a descriptive exploratory research design, which aims to perform a detailed analysis of the ChatGPT's potential to produce AI-fabricated research papers in orthodontics.

### AI based abstract generation

OpenAI's LLM, ChatGPT version 4, was used to generate a sample research manuscript on a topic related to the orthodontic field. We chose to fabricate a systematic review paper comparing the effectiveness of two treatment appliances- Herbst and Forsus, in treating Class II malocclusion. The following prompt was sent to ChatGPT-4, along with the PICO format (Population, Intervention, Comparator, and Outcome) specifications requisite to fabricate the research article: "Write me a systematic review paper that includes an introduction, methods, results, and conclusion section comparing the effectiveness of HERBST and FORSUS appliances in treating class II malocclusion using and give it a title." The following command was systematically developed after considering prior methods published in the previous studies. In a study by Ray PP<sup>(21)</sup> it was comprehensively explained applications of ChatGPT's different versions. Similarly, authors have also referred to Biswas SS<sup>(6)</sup> to further authenticate the process of developing these commands and prompts given to AI tool to generate fabricated research papers.

This prompt was designed to mimic realistic scientific inquiry by incorporating key components of research structure: objective, methods, results, and conclusions. All prompts were written by a single author (VM) to ensure the minimum variability in the methods of data collection. Since the outline for the manuscript was given in the initial prompt itself, the AI tool generated a proper review research article, which was further used to assess the efficiency of the researchers in identifying AI-generated or written abstracts.

### Search and extraction of human written abstracts

A systematic search strategy was developed based on the PICO framework (including MeSH and key terms) to retrieve the human-generated abstracts from three scientific databases- PubMed, Scopus, and Google Scholar (detailed search strategy given in Supplementary Table 1). Abstracts from relevant review research articles published

in high-impact dental journals were deemed eligible and included in the study.

### Abstract evaluation by human reviewers

The selected abstracts were re-written using ChatGPT-4 based on the title prompt given to the Chat GPT. A panel of 10 orthodontic experts who had a post-graduate qualification in orthodontics, had over 10 years of experience in the field, including both the researchers and the clinicians, knowledge or prior experience of scientific research, publication or peer reviewing were selected randomly for evaluation process. They were given a mix of AI-rewritten and human-written abstracts to identify fake and real ones. The methodology for the evaluation of results was similar to the one used by Gu *et al.*,<sup>(15)</sup> and Gao *et al.*,<sup>(22)</sup> in their comparative study on AI-generated scientific abstracts and medical images with that of original scientific works.

### Abstract evaluation by AI detection tools

The following AI detection tools were used to evaluate the AI generated vs real abstracts: Writer AI, Grammarly AI detector, Writefull X, Content detector AI, ZeroGPT and GPTZero. This analysis was intended to demonstrate the potential of AI detection tools in filtering out research fabricated abstracts with and without being reworded. Rewording was done using the tool- Quillbot. This methodology was in accordance with the one described by Elali and Rachid.<sup>(11)</sup>

### Data management and analysis

All collected data were entered and managed in Microsoft Excel (Version 2402, Office 2019). Descriptive statistics and comparative analysis were performed and the results are presented in the form of percentages, followed by narrative descriptions.

## Results

### Ability of ChatGPT-4 to generate orthodontic research paper

In response to the prompt, the AI model successfully generated a convincing orthodontic research abstract. On analysing the fabricated research paper (Table 1), the appearance seemed to be scientifically sound initially; however, upon closer inspection, it was observed to

contain fabricated data and results that could not be verified.

#### **Ability of ChatGPT-4 to generate modified conclusion in favor of the treatment of interest over the other**

In addition, on giving a prompt to support one treatment's effectiveness over the other, ChatGPT meticulously generated the conclusion that indicated the superiority of treatment-of-choice over the other (Table 2).

#### **Performance of human reviewers in distinguishing AI fabricated vs human-written research abstracts**

The abstracts from the original reviews were regenerated using ChatGPT on the same title (Supplementary Table 2). Interestingly, it was found that orthodontic experts failed to identify 14% of human-generated abstracts and had difficulty recognizing 32% of AI-generated content. (Tables 3 and 4)

#### **Accuracy of AI tools in detecting AI-reworded vs original human research works**

On analysis, it was found that online AI detectors showed varying degrees of precision in identifying the AI-generated content ranging from 61.63% to 100% (Table 5). However, their efficiency was significantly reduced when the text was processed through online rewording tools. Grammarly AI failed to identify the AI generated text, whereas other tools identified only 6-38.2% of content as AI generated. As shown in the Table 6, overall specificity, and sensitivity of the tools (n=6) were only 89%, with an accuracy of 44.5% cumulatively. Also, positive predictive values (PPV) and negative predictive values (NPV) for the tools indicate poor to moderate levels of performance. Writer AI and Content Detector failed to detect any fabricated abstracts correctly, having zero sensitivity and PPV. However, GPTZero has excellent specificity (100%) and PPV (100%), along with the highest sensitivity and accuracy compared to others for identifying fabricated abstracts (including re-worded). Also, the content detection rate was fairly poor, with only 11.7% overall across all the tools.

Overall performance metrics;

- Sensitivity = 23%
- Specificity = 66%
- Sensitivity + specificity= 89% (significantly less than 150%)

- Overall Accuracy = 44.5%

## **Discussion**

The findings of this study demonstrate that ChatGPT-4 is capable of generating scientific research and abstracts based on provided prompts or titles. This poses significant challenges to the orthodontic field and related scientific community. Moreover, it is also noteworthy that the ability of LLMs (ChatGPT-4) to produce fabricated scientific pieces of information has advanced to a level where it is challenging for humans to differentiate between AI-fabricated research papers from human-written ones. This indicated that ChatGPT-4 can fabricate convincing yet entirely plagiarized research papers, demonstrating the capabilities of AI to produce false or misleading information and undermining the ethical integrity of the scientific community.<sup>(18)</sup> This view was supported by Tan *et al.*,<sup>(23)</sup> while assessing the ability of generative AI in the field of ophthalmology. OpenAI, the developer of ChatGPT also stated that the ChatGPT model can produce "plausible-sounding yet erroneous or illogical responses." Also, the model is iteratively updated based on users' feedback to improve it. Hence, this is an evident sign that the model is not yet prepared to be used as a trustworthy source of information, especially in absence of human supervision.<sup>(24)</sup>

The ability of ChatGPT-4 to create a fabricated orthodontic research paper based on prompts indicate that the model cannot make an evidence-based decision-making. Existing studies<sup>(11,25,26)</sup>, have made comparable findings, showing the possible hazards of data falsification on use of AI technologies. In a similar vein, Lingard<sup>(27)</sup> stated that ChatGPT do not have knowledge and it can only generate semantically meaningful and grammatically correct responses based on prediction. Thus, the responses have uneven factual accuracy. This inaccurate or misleading research if not supervised is potentially dangerous in determining best treatment option and may put patients' health at stake by resulting in poor healthcare decisions. Eladi and Rashid<sup>(11)</sup>, reported that ChatGPT can fake findings using the most recent available data. Such issues became more obvious with the introduction of GPT-4, which has potential of handling wide range of instructions.<sup>(28,29)</sup> According to Hwang *et al.*,<sup>(30)</sup> while AI chatbots have demonstrated impressive capabilities in their approach, they still fall short in satisfying the needs

Table 1: Fabricated research model generated using an intelligent AI Chatbox: ChatGPT.

Prompt	Write me a systematic review paper that includes an introduction, methods, results, and conclusion section comparing the effectiveness of HERBST and FORSUS appliances in treating class II malocclusion. Give it a title.
Response	<p><b>Title:</b> <b>Comparative Effectiveness of Herbst and Forsus Appliances in the Treatment of Class II Malocclusion: A Systematic Review</b></p> <p><b>Abstract</b> <b>Objective:</b> This systematic review aims to compare the clinical effectiveness of the Herbst and Forsus fixed functional appliances in the treatment of Class II malocclusion, focusing on skeletal and dentoalveolar changes, treatment duration, and patient compliance. <b>Methods:</b> A systematic search was conducted across PubMed, Scopus, Embase, and Cochrane Library databases up to May 2025. Randomized controlled trials (RCTs), prospective, and retrospective studies comparing Herbst and Forsus appliances in growing patients with Class II malocclusion were included. Study quality was assessed using the Cochrane Risk of Bias tool and the Newcastle-Ottawa Scale. <b>Results:</b> A total of 12 studies were included, comprising 6 RCTs and 6 observational studies. Both appliances demonstrated significant improvement in molar relationships and overjet reduction. Herbst appliances were associated with greater skeletal changes, particularly mandibular advancement, while Forsus appliances showed more dentoalveolar compensation. Forsus showed better patient compliance due to easier insertion and removal. <b>Conclusion:</b> Both Herbst and Forsus appliances are effective in managing Class II malocclusion. Herbst offers greater skeletal changes, whereas Forsus is advantageous in terms of patient comfort and ease of use. Appliance choice should be individualized based on patient growth status, clinical objectives, and compliance potential.</p> <p><b>Introduction</b> Class II malocclusion, characterized by a retrusive mandible, affects approximately one-third of the adolescent population and is a common reason for orthodontic treatment. Functional appliances have been widely used to correct this discrepancy by stimulating mandibular growth or by dental compensation. Among the fixed functional appliances, the <b>Herbst appliance</b>, introduced in the early 1900s, is considered a rigid, noncompliant device that promotes mandibular advancement. In contrast, the <b>Forsus Fatigue Resistant Device (FRD)</b> is a semi-rigid, push-rod spring appliance introduced in the early 2000s, known for its ease of application and compatibility with fixed orthodontic appliances. Although both devices are widely used, debate continues regarding their relative effectiveness in producing skeletal versus dental changes and their overall impact on treatment outcomes. This systematic review aims to critically assess and compare the clinical effectiveness of Herbst and Forsus appliances in treating growing patients with Class II malocclusion.</p> <p><b>Methods</b> <b>Protocol and Registration</b> This systematic review was conducted following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The review protocol was registered in PROSPERO (ID: CRD4202525011).</p> <p><b>Eligibility Criteria</b> <b>Inclusion criteria:</b></p> <ul style="list-style-type: none"><li>• Human studies involving Class II malocclusion patients aged 8–18 years</li><li>• Comparisons between Herbst and Forsus appliances</li><li>• Outcomes including skeletal changes, dental changes, treatment duration, and compliance</li><li>• RCTs, cohort studies, and controlled clinical trials</li><li>• English language articles</li></ul> <p><b>Exclusion criteria:</b></p> <ul style="list-style-type: none"><li>• Case reports, expert opinions, reviews, animal studies</li><li>• Studies without comparative data between appliances</li></ul> <p><b>Information Sources and Search Strategy</b> Databases searched: <b>PubMed, Scopus, Embase, and Cochrane Library</b> (up to May 2025). Search terms included:</p> <ul style="list-style-type: none"><li>• "Herbst appliance" AND "Forsus appliance" AND "Class II malocclusion"</li><li>• "Functional appliance" AND "mandibular advancement"</li><li>• Filters: Human studies, English language</li></ul> <p><b>Study Selection</b> Two reviewers independently screened titles, abstracts, and full texts. Disagreements were resolved by a third reviewer.</p> <p><b>Data Extraction and Quality Assessment</b> Data extracted included: sample size, patient age, appliance type, treatment duration, skeletal and dental outcomes, and compliance indicators. Risk of bias in RCTs was assessed using the <b>Cochrane Risk of Bias 2.0 tool</b>, while observational studies were evaluated using the <b>Newcastle-Ottawa Scale</b>.</p>



<b>Results</b>	
<b>Study Selection</b>	
From 487 articles identified, 47 full-texts were reviewed, and 12 studies met the inclusion criteria (6 RCTs, 6 observational studies). A PRISMA flow diagram summarizes the selection process.	
<b>Study Characteristics</b>	
<ul style="list-style-type: none"><li>• Total participants: 654 (Herbst: 326, Forsus: 328)</li><li>• Age range: 10–16 years</li><li>• Mean treatment duration: 9–12 months</li><li>• Most studies had a follow-up period of 6–18 months post-treatment</li></ul>	
<b>Skeletal Changes</b>	
<ul style="list-style-type: none"><li>• Herbst appliances produced statistically significant increases in SNB angle (mean increase: 1.8°) and mandibular length (Co-Gn).</li><li>• Forsus also induced mandibular growth, but changes were more dentoalveolar than skeletal.</li><li>• Studies using lateral cephalometry confirmed that Herbst produced more pronounced mandibular advancement.</li></ul>	
<b>Dentoalveolar Changes</b>	
<ul style="list-style-type: none"><li>• Both appliances reduced overjet effectively (Herbst: 4.5–6 mm; Forsus: 4–5.5 mm).</li><li>• Forsus showed greater maxillary incisor retroclination and mandibular incisor proclination, indicating more dental compensation.</li></ul>	
<b>Treatment Duration and Compliance</b>	
<ul style="list-style-type: none"><li>• Treatment duration was comparable.</li><li>• Forsus was easier to integrate with fixed orthodontic appliances and allowed better hygiene and patient comfort.</li><li>• Herbst showed higher rates of appliance breakage and soft tissue irritation.</li></ul>	
<b>Risk of Bias</b>	
<ul style="list-style-type: none"><li>• Among RCTs, 4 studies showed low risk, while 2 had unclear risk due to lack of blinding.</li><li>• Observational studies had moderate to high methodological quality.</li></ul>	
<b>Discussion</b>	
This review highlights the relative strengths of Herbst and Forsus appliances in treating Class II malocclusion. Herbst offers superior skeletal correction due to its rigid design and constant mandibular posturing. Forsus, while less impactful on skeletal growth, effectively achieves correction through dental compensation with greater ease of use and better patient tolerance.	
These differences suggest that <b>Herbst is more suitable for patients in early or peak growth periods</b> , while <b>Forsus may be preferred during late adolescence or in non-compliant patients</b> due to its user-friendly nature.	
However, heterogeneity in measurement methods, appliance designs, and treatment timing limits the generalizability of results. Future high-quality RCTs with standardized outcome measures and long-term follow-up are warranted.	
<b>Conclusion</b>	
Both Herbst and Forsus appliances are effective in correcting Class II malocclusion. Herbst tends to produce more skeletal changes, while Forsus achieves correction primarily through dentoalveolar modifications. The choice between the two should be tailored to individual patient characteristics, including growth potential, compliance, and treatment goals	

**Table 2:** Prompt given to Chat GPT to support only one treatment intervention over the other.

Prompt given to Chat GPT to support only one treatment intervention over the other.	
Original conclusion from the initial prompt	This systematic review concludes that both the Herbst and Forsus appliances are effective in treating Class II malocclusion, but each has specific strengths. The Herbst appliance is more effective in achieving skeletal correction and is preferable in cases where mandibular advancement is a priority. The Forsus appliance, however, offers shorter treatment duration and higher patient compliance, making it a good choice for patients prioritizing comfort and ease of use. Further high-quality RCTs with standardized outcome measures are needed to draw more definitive conclusions on the comparative effectiveness of these appliances. Clinicians should consider patient preferences, compliance levels, and treatment goals when choosing between the Herbst and Forsus appliances to achieve optimal treatment outcomes for Class II malocclusion.
New prompt in support of choice of treatment	Support HERBST as an effective management strategy for the Class II malocclusions
The new modified response generated by ChatGPT-4	Both HERBST and FORSUS appliances are effective in treating Class II malocclusions, with distinct advantages depending on patient characteristics and treatment objectives. HERBST, with its proven ability to achieve significant skeletal changes and mandibular growth stimulation, is a highly effective strategy, particularly for growing patients. FORSUS is better suited for dental corrections and enhancing patient compliance. Clinicians should integrate patient preferences, growth potential, and treatment goals into the decision-making process

**Table 3:** Percentage distribution comparing abilities of human reviewers in identifying AI vs human written abstracts.

Human Reviewers (orthodontic clinicians and researchers)	Abstracts (in percentage)	
	Human	ChatGPT
Assessed accurately	86	68
Assessed inaccurately	14	32

**Table 4:** Prompt given to the ChatGPT-4 for fabrication of the abstract.

Prompt	Write me an abstract on the topic “Effectiveness of Herbst and Forsus Appliances in Treating Class II Malocclusion: A Systematic Review.”
Response	Class II malocclusion, a common orthodontic issue, is frequently managed with functional appliances. The Herbst and Forsus appliances are widely utilized in this treatment, but their comparative efficacy remains a subject of interest. This systematic review evaluates the effectiveness of these two appliances in correcting Class II malocclusion, focusing on dental and skeletal changes, treatment time, and patient compliance. Results from various clinical studies are analyzed to provide a comprehensive comparison, guiding clinicians in selecting the most suitable appliance based on treatment goals and patient needs.

**Table 5:** AI detection test results from different diagnostic tools after paraphrasing

Conclusions	Detector Name	Score (Realness)
<b>Original Abstract (AI fabricated):</b> Class II malocclusion, a common orthodontic issue, is frequently managed with functional appliances. The Herbst and Forsus appliances are widely utilized in this treatment, but their comparative efficacy remains a subject of interest. This systematic review evaluates the effectiveness of these two appliances in correcting Class II malocclusion, focusing on dental and skeletal changes, treatment time, and patient compliance. Results from various clinical studies are analyzed to provide a comprehensive comparison, guiding clinicians in selecting the most suitable appliance based on treatment goals and patient needs.	Writer AI <sup>a</sup> .	83% Human-generated
	Grammarly AI detector.	100% of this text appears to be AI-generated
	ZeroGPT <sup>c</sup>	61.63% generated from GPT
	GPT Zero <sup>d</sup> .	0% Human 0% Mixed 100% AI
<b>Reworded Conclusion:</b> Functional appliances are commonly used to treat Class II malocclusion, a common orthodontic issue. Although the Herbst and Forsus appliances are frequently used in this treatment, there is still debate regarding their relative effectiveness. With an emphasis on patient compliance, treatment duration, and dental and skeletal changes, this systematic review assesses how well these two appliances work to correct Class II malocclusion. A thorough comparison is produced by analyzing the results of multiple clinical studies, which helps clinicians choose the best appliance for their patient's needs and treatment objectives.	Writer AI	94% human-generated content
	Grammarly AI detector	0% of text appears to be AI-generated
	ZeroGPT	38.2% of the text is ChatGPT generated
	GPT Zero	91% Human 1% Mixed 8% AI

a: Writer AI: <https://writer.com/ai-content-detector/>; b: Grammarly AI: <https://www.grammarly.com/ai-detector>; c: ZeroGPT: <https://www.zerogpt.com/>; d: GPT Zero: <https://gptzero.me/>

of specialized healthcare professions like orthodontics. Additionally, ChatGPT adhere closely to the ethical guidelines of European Union for Artificial intelligence, which underscore the significance of technical safety and robustness, human supervision, privacy, and data governance. Thus, before using ChatGPT, the possible drawbacks and moral issues must be carefully evaluated and resolved.<sup>(31)</sup>

Our findings also indicate that orthodontic researchers and clinicians face significant difficulties in accurately differentiating between AI-fabricated and human-generated works which clearly emphasizes the ability of AI to mimic human works. Gao *et al.*,<sup>(22)</sup> reported 68% accuracy of identification of AI generated abstracts by reviewers. Similarly, Levin *et al.*,<sup>(32)</sup> in their study

**Table 6:** Detection rates of AI-generated abstracts from the five review papers.

AI detection tools	Correctly identified as AI-based (in percentage)	Detected partial AI content (in percentage)	Incorrectly identified as AI-based (in percentage)	Content detection rate (detected AI% by actual AI %)	Sensitivity %	Specificity %	Positive predictive values (PPV) %	Negative predictive values (NPV) %	Accuracy %
Writer AI	0	0	100	0.27	0	44	0	44	22
Grammarly AI	0	20	100	0.19	0	50	0	63	25
GPTZero	100	0	0	3.68	60	100	100	71	80
Write full X	0	20	80	0.34	20	80	50	50	50
Content detector AI	0	20	80	0.94	0	60	0	38	30
ZeroGPT	0	60	40	1.65	60	60	60	60	60

reported that reviewers only found 46.3% of gynaecologic oncology abstracts generated by ChatGPT. However, we also noted that the reviewers also misclassified some of the abstracts which is in concordance with Gao *et al.*,<sup>(22)</sup> study. This highlights that human reviewers themselves have faced challenges in assessing the samples, indicating that AI's ability to generate fabricated outcomes poses a significant threat to the research community's analytical and intellectual integrity.<sup>(22)</sup> This is compounded by ChatGPT's ability to offer incorrect yet confident, sensible, and compelling responses in healthcare research, also known as AI hallucination. Alkaissi and McFarlane had similar opinions on the risk of hallucinations in AI-powered medical research.<sup>(13)</sup> Kim *et al.*,<sup>(25)</sup> noted hallucination in 20% of spine-related responses generated by AI indicating its plausible ability to tell a lie and making it difficult for users to differentiate between real and fake research works.<sup>(25,33)</sup> This can be attributed to the capabilities of LLMs to formulate research work mimicking human works using large integrated databases.<sup>(34)</sup>

While the AI detection tools are relatively good at avoiding false positives, their ability to detect fabricated abstracts (true positives) is limited and varied across the different AI tools. This result was supported by the observation given by Chakraborty *et al.*,<sup>(35)</sup> and Kim *et al.*,<sup>(25)</sup> indicating that the probability of fabricated research work bypassing checkpoints designed solely for detection purposes increased the chance of plagiarized work. Similarly, Ibrahim *et al.*,<sup>(36)</sup> and Kar *et al.*,<sup>(37)</sup> also expressed concerns over inaccurate performances of the AI detection tools. Our analysis also revealed that the AI detection commented "Your text is AI/GPT Generated"

in 80% (4 of 5) of human-written abstracts, underscoring the practical constraints on its uses. Similarly, Odri and Yoon<sup>(38)</sup> discovered that AI detection methods erroneously identified AI-generated content in an essay written by Charnley in 1960<sup>(39)</sup>, demonstrating that the detection tools' estimates were wrong. As a result, difficulties in identifying AI-generated abstracts as true positives have highlighted the limited efficiency of contemporary AI detectors.<sup>(11)</sup> The discrepancies in accuracies between AI detection technologies can be attributed to variances in algorithms and detection tactics used by different platforms. Some AI detection technologies rely on monolingual algorithms like repetition, predictability, sentence organization, and burstiness to determine whether or not the text is AI generated. However, these techniques are not infallible, especially when applied to scientific abstracts, which have a rigid, formal, and consistent structure that limits the detection capabilities of tools. Furthermore, the usage of paraphrase tools or generative AI assistants like Copilot and Grammarly adds an additional degree of difficulty for AI detection tools. As a result of the widespread use of these technologies, distinguishing between paraphrased, plagiarized, and AI-fabricated work has become increasingly difficult. Moreover, our analysis also reiterates the Ji *et al.*,<sup>(40)</sup> conclusion, which indicates the detection tool's ability to detect AI-fabricated content has created hallucinations regarding LLMs in the scientific research community. Furthermore, LLMs such as ChatGPT-4 are stronger tools with capabilities to generate highly coherent, contextually appropriate, and human-like text that can evade detection algorithms. This observation was also given by Flitcroft *et al.*,<sup>(41)</sup> indicating varying



performances of AI detection tools and false conclusions by labeling human-written articles as AI-generated. It is noteworthy that these detection tools are trained on different yet limited datasets, which eventually limits their power to detect such coherent scientific texts. This suggests that there is an urgent need to modify and develop powerful tools to assess such deeper semantic, stylistic, or discourse-level inconsistencies found in the AI-fabricated texts. Similar observations were given by Weber-Wulff *et al.*,<sup>(42)</sup> concluding that “most available detection tools were neither accurate nor reliable and showed bias toward classifying output as human-written rather than detecting AI-generated text”. Notably, one of the studies by Liang *et al.*,<sup>(43)</sup> also noted that GPT detectors possess a significant risk of bias against non-English language papers by consistently classifying them as AI-generated. The study also highlighted that simple prompting strategies can successfully evade the abilities of these detection tools, indicating their inabilities. Thus, the restricted detection capabilities of AI technologies necessitate a more nuanced approach to evaluating scientific literature.

### Implications

As LLMs and generative AI continue to advance, distinguishing between human-generated and AI-generated texts will become increasingly difficult, making the detection process nearly impossible.<sup>(35)</sup> Thus, with an understanding of the possible risks of AI technology in regard to orthodontics research, it has become vital to ensure minimum infiltration of AI produced or paraphrased works in the research databases.<sup>(33,44)</sup> Despite the fact that AI-generated content is factually reliable and free of plagiarism, orthodontic doctors and researchers must exercise caution when using it. This is because AI models are based on probabilistic trends from massive data sets, which can provide contextually inaccurate information without genuine grasp of the context, as well as plausible-sounding claims with insufficient proof. Furthermore, it cannot provide methodological rigor and ethical accountability, which jeopardize scientific integrity. Thus, journals must ensure the establishment and implementation of a rigorous verification process that will serve as a checkpoint when accepting research papers.<sup>(29)</sup> Eladi and Rashid indicate that methods like submitting data collection proofs, proof of methodology (depending on the type of the completed study), and codes

used for statistical analysis can assist to minimize potential fraud.<sup>(11)</sup> Similarly, editors and reviewers from the journals should also be educated regarding the identification of the potential influx of plagiarized work.<sup>(25)</sup> This can be done by provisioning workshops and seminars where they can be sensitized regarding patterns and signs of AI-generated content.<sup>(15)</sup> Such positive practices will not only incorporate a culture of transparency but will also foster trustworthiness among the researchers.<sup>(10,38)</sup> While this study highlights the implications of AI-generated research within the orthodontic field, future research could extend to other areas of medical and dental sciences to gain a deeper understanding of the impact of such methods. Additionally, researchers could explore effective strategies for identifying AI-generated content using advanced technologies, as well as examine the potential long-term impacts on future research in this domain.

### Strengths and limitation

To the best of our knowledge, this study was the first of its kind in orthodontic research, demonstrating potential risks associated with the use of LLM-based AI chatbots in orthodontic research. However, it has several limitations. The sample review paper used in this study was fairly small. Furthermore, the LLMs used to generate summaries are applied based on an analysis of the primary findings of the studies, which introduces the potential for subjective bias. Additionally, it was not possible to create a receiver operating characteristic (ROC) curve to evaluate the performance of different detection tools as there is no standardized criteria for defining AI generated content which may hinder confusion matrix calculation- a requisite for ROC analysis. Our evaluation was limited to the performance of open-source AI detection tools, which restricted our understanding and the applicability of the results to subscription-based AI detection tools. Furthermore, our study also has limitations regarding the evaluation method to assess the performance of LLMs (ChatGPT-4) as it may not fully capture the quality of the LLM-generated abstract detection. This can be attributed to the types of tools being considered within the scope of this study (which were essentially free-access tools) and the algorithms of these tools. Additionally, while human reviewers were also field experts and were provided AI-fabricated and human-generated abstracts, it was observed that human-based evaluation of LLM

performances was subjective. This introduces an element of heterogeneity in the evaluation process, thereby influencing the outcome of the results. Future studies could address this limitation to provide more comprehensive and generalizable findings.

## Conclusions

AI models have the ability to generate convincing but fabricated research content. Due to the significant similarity to original research works, human reviewers found it challenging to distinguish between AI-generated abstracts and those written by humans. Besides, the variability in efficiency of AI detection tools to detect fabricated papers further highlights the need for continuous improvement in accuracy of tools. Moreover, since AI-generated research can impact the development of healthcare policies, treatment strategies, and patient care, it is imperative to take proactive measures. These include integrating advanced detection tools, enhancing peer review processes, and fostering a culture of transparency regarding the use of AI in research.

## References

1. V Allareddy, S Frazier-Bowers, JH Park, GH Gilbert. Relevance of practice-based research to orthodontics. *Angle Orthod.* 2021;91(6):856-7.
2. Pacha MM, Fleming PS, Johal A. A comparison of the efficacy of fixed versus removable functional appliances in children with class II malocclusion: a systematic review. *Eur J Orthod.* 2016;38(6):621-30.
3. Fiorillo L, Mehta V. Accelerating editorial processes in scientific journals: Leveraging AI for rapid manuscript review. *Oral Oncology Reports.* 2024;10:100511.
4. Bhargava DC, Jadav D, Meshram VP, Kanchan T. ChatGPT in medical research: challenging time ahead. *Med Leg J.* 2023;91(4):223-5.
5. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res.* 2023;13(4):1148-54.
6. Biswas SS. ChatGPT for research and publication: a step-by-step guide. *J Pediatr Pharmacol Ther.* 2023;28(6):576-84.
7. Alfarraj YF, Wardat Y. Exploring the impact of ChatGPT on scientific research: assessing strengths, weaknesses, opportunities, and threats. *EaC.* 2024;28:1-27.
8. Ahn S. The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. *Korean J Physiol Pharmacol.* 2024;28(5):393-401.
9. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Qureshi F, Cheungpasitporn W. Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: a narrative review. *Clin Pract.* 2023;14(1):89-105.
10. Temsah MH, Altamimi I, Jamal A, Alhasan K, Al-Eyadhy A. ChatGPT surpasses 1000 publications on pubmed: envisioning the road ahead. *Cureus.* 2023;15(9):e44769.
11. Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns (N Y).* 2023;4(3):100706.
12. Safrai M, Orwig KE. Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *J Assist Reprod Genet.* 2024;41(7):1871-80.
13. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15(2):e35179.
14. Caprioglio A, Paglia L. Fake academic writing: ethics during chatbot era. *Eur J Paediatr Dent.* 2023;24(2):88-9.
15. Gu J, Wang X, Li C, Zhao J, Fu W, Liang G, *et al.* AI-enabled image fraud in scientific publications. *Patterns (N Y).* 2022;3(7):100511.
16. Dash KS, Mehta V, Kharat P. We are entering a new era of problems: AI-generated images in research manuscripts. *Oral Oncology Reports.* 2024;10:100289.
17. Mehta V, Thomas V, Mathur A. AI-dependency in scientific writing. *Oral Oncology Reports.* 2024;10:100269.
18. Else H. Abstracts written by ChatGPT fool scientists. *Nature.* 2023;613(7944):423.
19. Nazarovets S. Dealing with Research Paper Mills, Tortured Phrases, and Data Fabrication and Falsification in Scientific Papers [Internet]. In: Joshi PB, Churi PP, Pandey M, editors. *Scientific Publishing Ecosystem.* Singapore: Springer Nature Singapore; 2024 [cited 2025 Mar 19]. p. 233-54. Available from: [https://link.springer.com/10.1007/978-981-97-4060-4\\_14](https://link.springer.com/10.1007/978-981-97-4060-4_14)
20. Haider J, Söderström KR, Ekström B, Rödl M. GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *HKS Misinfo Review.* 2024;5(5). doi.org/10.37016/mr-2020-156.
21. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys Syst.* 2023;3:121-54.
22. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, *et al.* Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med.* 2023;6(1):75.
23. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, *et al.* Generative artificial

- intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. *Ophthalmol Sci*. 2023;3(4):100394.
24. Flanagan A, Bibbins-Domingo K, Berkwitz M, Christiansen SL. Nonhuman “Authors” and implications for the integrity of scientific publication and medical knowledge. *JAMA* 2023;329(8):637.
  25. Kim HJ, Yang JH, Chang DG, Lenke LG, Pizones J, Castelein R, *et al*. Assessing the Reproducibility of the structured abstracts generated by ChatGPT and bard compared to human-written abstracts in the field of spine surgery: comparative analysis. *J Med Internet Res*. 2024;26:e52001.
  26. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiol*. 2023;52(9):1755-8.
  27. Lingard L. Writing with ChatGPT: an Illustration of its capacity, limitations & implications for academic writers. *Perspect Med Educ*. 2023;12(1):261-70.
  28. Jahani Yekta MM. The general intelligence of GPT-4, its knowledge diffusive and societal influences, and its governance. *Meta Radiol*. 2024;2(2):100078.
  29. Kendall G, Teixeira Da Silva JA. Risks of abuse of large language models, like CHATGPT, in scientific publishing: authorship, predatory publishing, and paper mills. *Learned Publ*. 2024;37(1):55-62.
  30. Hwang T, Aggarwal N, Khan PZ, Roberts T, Mahmood A, Griffiths MM, *et al*. Can ChatGPT assist authors with abstract writing in medical journals? evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS One*. 2024;19(2):e0297701.
  31. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, *et al*. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med*. 2024;13(3):735.
  32. Levin G, Pareja R, Viveros-Carreño D, Sanchez Diaz E, Yates EM, Zand B, *et al*. Association of reviewer experience with discriminating human-written versus ChatGPT-written abstracts. *Int J Gynecol Cancer*. 2024;34(5):669-74.
  33. Bi AS. What’s important: the next academic-ChatGPT AI? *J Bone Joint Surg Am*. 2023;105(11):893-5.
  34. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. *Future Internet*. 2023; 15(6):192.
  35. Chakraborty S, Bedi AS, Zhu S, An B, Manocha D, Huang F. On the Possibilities of AI-Generated Text Detection [Internet]. 2023 [cited 2025 Mar 19]; Available from: <http://arxiv.org/abs/2304.04736>
  36. Ibrahim K. Using AI-based detectors to control AI-assisted plagiarism in ESL writing: “The Terminator Versus the Machines.” *Lang Test Asia*. 2023;13(1):46.
  37. Kar SK, Bansal T, Modi S, Singh A. How sensitive are the free ai-detector tools in detecting AI-generated texts? a comparison of popular AI-detector tools. *Indian J Psychol Med*. 2025;47(3):275-278.
  38. Odri GA, Ji Yun Yoon D. Detecting generative artificial intelligence in scientific articles: evasion techniques and implications for scientific integrity. *Orthop Traumatol Surg Res*. 2023;109(8):103706.
  39. Charnley J. Surgery of the hip-joint: present and future developments. *Br Med J*. 1960;1(5176):821-6.
  40. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, *et al*. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):1-38.
  41. Flitcroft MA, Sheriff SA, Wolfrath N, Maddula R, McConnell L, Xing Y, *et al*. Performance of artificial intelligence content detectors using human and artificial intelligence-generated scientific writing. *Ann Surg Oncol*. 2024;31(10):6387-93.
  42. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, *et al*. Testing of detection tools for AI-generated text. *Int J Educ Integr*. 2023;19(1):26.
  43. Liang W, Yuksekogonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers [Internet]. 2023 [cited 2025 Jun 11]; Available from: <http://arxiv.org/abs/2304.02819>
  44. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information*. 2023;14(8):462.