Original Paper

AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study

Daniel Stephan[1*], Dr med; Annika Bertsch[1*]; Matthias Burwinkel[1], Dr med dent; Shankeeth Vinayahalingam[2], MD; Bilal Al-Nawas[1], Dr med dent, Prof Dr Med; Peer W Kämmerer[1*], MA, MSc, Dr med dent, Prof Dr Med; Daniel GE Thiem[1*], MHBA, Dr med dent, Dr med, PD

[1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Mainz, Germany

[2]Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, Nijmegen, Netherlands

[*]these authors contributed equally

Corresponding Author:
Daniel Stephan, Dr med
Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery
University Medical Centre of the Johannes Gutenberg-University Mainz
Augustusplatz 2
Mainz, 55131
Germany
Phone: 49 6131177038
Email: stephand@uni-mainz.de

Abstract

Background: Structured and standardized documentation is critical for accurately recording diagnostic findings, treatment plans, and patient progress in health care. Manual documentation can be labor-intensive and error-prone, especially under time constraints, prompting interest in the potential of artificial intelligence (AI) to automate and optimize these processes, particularly in medical documentation.

Objective: This study aimed to assess the effectiveness of ChatGPT (OpenAI) in generating radiology reports from dental panoramic radiographs, comparing the performance of AI-generated reports with those manually created by dental students.

Methods: A total of 100 dental students were tasked with analyzing panoramic radiographs and generating radiology reports manually or assisted by ChatGPT using a standardized prompt derived from a diagnostic checklist.

Results: Reports generated by ChatGPT showed a high degree of textual similarity to reference reports; however, they often lacked critical diagnostic information typically included in reports authored by students. Despite this, the AI-generated reports were consistent in being error-free and matched the readability of student-generated reports.

Conclusions: The findings from this study suggest that ChatGPT has considerable potential for generating radiology reports, although it currently faces challenges in accuracy and reliability. This underscores the need for further refinement in the AI's prompt design and the development of robust validation mechanisms to enhance its use in clinical settings.

(J Med Internet Res 2024;26:e60684) doi: 10.2196/60684

KEYWORDS

artificial intelligence; ChatGPT; radiology report; dental radiology; dental orthopantomogram; panoramic radiograph; dental; radiology; chatbot; medical documentation; medical application; imaging; disease detection; clinical decision support; natural language processing; medical licensing; dentistry; patient care

Introduction

Structured and standardized documentation plays a crucial role in health care by ensuring accurate recording and communication of diagnostic findings, treatment plans, and patient progress, thereby supporting high-quality patient care [1]. However, manual documentation is often time-consuming, error-prone, and can impede clinical workflow efficiency, especially in fast-paced medical settings. With the emergence of artificial intelligence (AI), there is a growing interest in

XSL•FO
RenderX

implementing AI technology to optimize health care workflows and improve documentation practices.

AI has proven useful in various medical applications, from diagnosing diseases to drug development [2]. In radiology, AI algorithms analyze medical images to assist in early disease detection, improve radiologists' performance, and provide clinical decision support [3,4]. Moreover, AI-driven solutions have the potential to automate repetitive tasks and reduce the workload of health care professionals [5,6].

First introduced by OpenAI in 2018, (GPT—a specific large language model developed by OpenAI) has continuously evolved and trained on extensive text data [7]. ChatGPT (implementation of GPT), an advanced large language model (a class of AI models), represents a significant advancement in natural language processing and has demonstrated remarkable capabilities in understanding and generating human-like text using deep learning techniques, like neuronal networks. GPT 3.5 showed a human-level performance across various medical exams and passed the United States Medical Licensing Exam (60.2%), Med-MCQA (57.5%), and PubMedQA (78.2%) [8-10]. With its proficiency in language generation, ChatGPT is capable of medical writing [11] and, therefore, has been increasingly integrated into medical education [12] and clinical practice, allowing it to automate the writing of examination findings, doctor's letters, or radiology reports [13].

Dental radiology, integral to dentistry, relies on the correct interpretation of x-ray images, including panoramic radiographs (OPG), to diagnose and plan numerous oral conditions or pathologies. To maintain the standard of patient care, it is, therefore, crucial to ensure high-quality training in radiology tasks during dental studies. Traditionally, radiology education involves manual interpretation of x-ray images and writing detailed medical findings reports based on visual inspection and clinical knowledge. However, the emergence of AI technologies has increased interest in alternative methods for radiology education and diagnostic reporting, including maxillofacial radiology [14,15].

The capability of AI in diagnosing medical images, including x-ray images, is well-established [3,16]. Moreover, studies have demonstrated that ChatGPT can generate clinic letters and operative notes with high correctness and readability [17,18]. Additionally, another study has proven its efficacy beyond text generation in simplifying existing radiology reports and improving patient understanding [19]. Furthermore, recent research reveals AI's capability to outperform dental students in diagnostic accuracy regarding endodontic assessments [20], highlighting its potential as a reference tool to enhance students' understanding and diagnostic skills. However, this raises concerns about the potential for overreliance on AI, considering reports about ChatGPT generating fake findings for imaginable diseases [21], which may affect the development of critical analytical and decision-making abilities. Thus, it is essential to integrate AI with human expertise and clinical judgment in dental education. ChatGPT shows promising potential in improving doctor-patient communication by simplifying complex medical information and transforming complex medical terminology into easily understandable language for patients with varying levels of health expertise [22]. While earlier versions of ChatGPT powered by GPT-3.5 generated patient-facing information lacking accuracy and important information, GPT-4 has shown improvements in appropriateness and accuracy and, despite occasional omissions, ultimately produced patient information applicable for gaining informed consent for procedures in nuclear medicine [23].

Nevertheless, the generation of radiology reports based on diagnostic findings by health care professionals remains a subject of investigation. Therefore, this study evaluated the efficacy of incorporating AI language models, specifically ChatGPT, into generating radiology reports. Dental students analyzed OPGs and provided diagnoses through checkbox lists together with written reports. A comparative analysis between radiology reports manually written by dental students and reports generated by the AI based on those prefilled checkbox lists was conducted. This study primarily investigated the readability of both report types with the null hypothesis stating no differences in readability between the 2 sets of reports. Secondary outcomes, including text accuracy and language quality, were evaluated to identify potential areas for improvement in AI-driven radiology reporting.

Methods

Overview

This study sought to investigate the efficacy of incorporating AI language models, specifically ChatGPT, in generating radiology reports from prefilled checkbox lists after analyzing OPGs. Dental students were assigned to diagnose 2 different x-ray images, providing a written radiology report for 1 and a checkbox list of diagnoses for the other. The AI then generated reports based on the diagnoses provided within the checkbox lists. Subsequently, both texts were analyzed comparatively to primarily evaluate readability, with a secondary evaluation of text quality, accuracy, similarities, and disparities between student-written and AI-written reports.

Ethical Considerations

The study adhered strictly to ethical standards and institutional guidelines, obtaining informed consent from all participants beforehand. Participants were clearly informed of the purpose of the research, the voluntary nature of participation, and their ability to withdraw at any time without any repercussions. Additionally, no compensation was provided, as the tasks were integral to students' academic training. Confidentiality and data privacy were stringently maintained throughout the research process to uphold the participant's well-being and privacy. An ethics approval was not required as the generation of radiology reports is a standard component of dental education in Germany and this type of research does not involve intervention or data collection beyond routine educational activities. The tasks performed by the students were part of their regular academic curriculum and no additional tasks outside the students' regular academic curriculum were introduced. Moreover, the analysis of data was conducted anonymously, ensuring privacy and confidentiality and preventing participating in this study resulted in either advantages or disadvantages for the students. As no

XSL•FO
RenderX

images or materials involving identifiable features were included in this study, no additional consent forms were required.

Study Setting

The study took place in the radiology section of the Department of Oral and Maxillofacial Surgery at the University Medical Centre Mainz, Germany. Certified medical monitors were provided, and all participants were supervised throughout the session without access to additional information or external help.

Participants

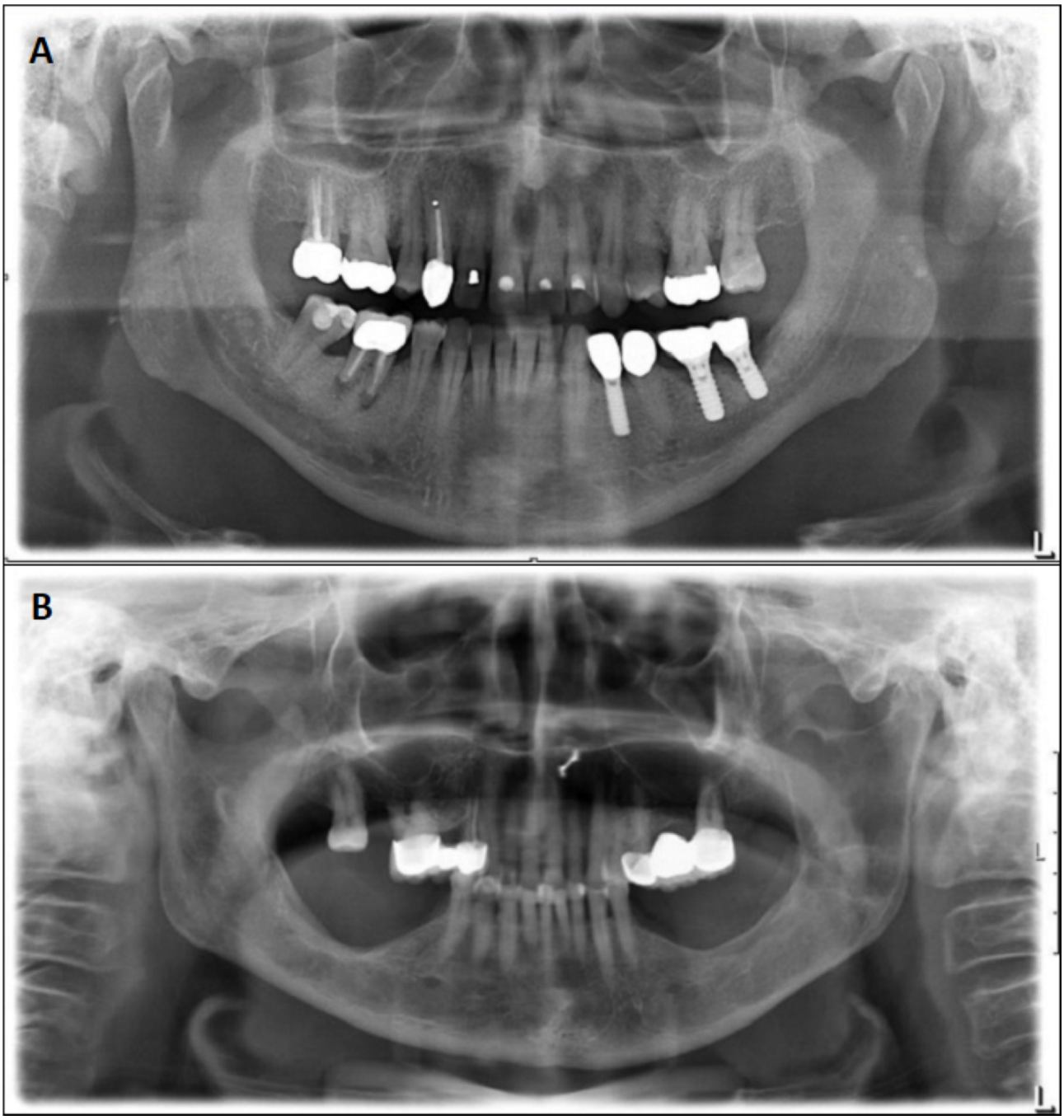
In Germany, dental education is structured into 10 semesters. The first 5 semesters focus on foundational knowledge, while the following 5 semesters (clinical semesters 1-5, corresponding to overall semesters 6-10) emphasize clinical skills. The first lesson in dental radiology was introduced in the first clinical semester and, therefore, preclinical students were excluded from this study. A total of 100 dental students from all 5 clinical semesters participated in the study with the following distribution across semesters—semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20. This equal representation across different stages of dental education highlights the progressive development in radiology report writing.

Experimental Design

Students were randomly assigned to 1 of 2 groups and presented with an unknown OPG (Figures 1A and 1B). Group A was instructed to analyze the x-ray image (Figure 1A) and compose a radiology report within 30 minutes without any external assistance. Group B received a second OPG (Figure 1B) and was tasked with completing a checkbox list (Figure 2) detailing their observations within a 10-minute time frame. These time limits were specifically chosen to investigate the potential time-saving benefits of using a structured checkbox method followed by AI-generated reporting. It was observed that all students used the entire allotted time for their respective tasks, neither exceeding the time limit nor completing early. The study, therefore, focused on the completion of the tasks within the predefined limits without measuring the exact duration for each task. Upon completion, each group was required to complete the alternate assignment with the opposite x-ray image. To minimize biases, the experimental design ensured that 1 group completed the checkbox for the same x-ray for which the other group composed the report, and vice versa. This approach reduced any influence of specific characteristics of the x-ray images (eg, the complexity of findings or difficulty of interpretation).

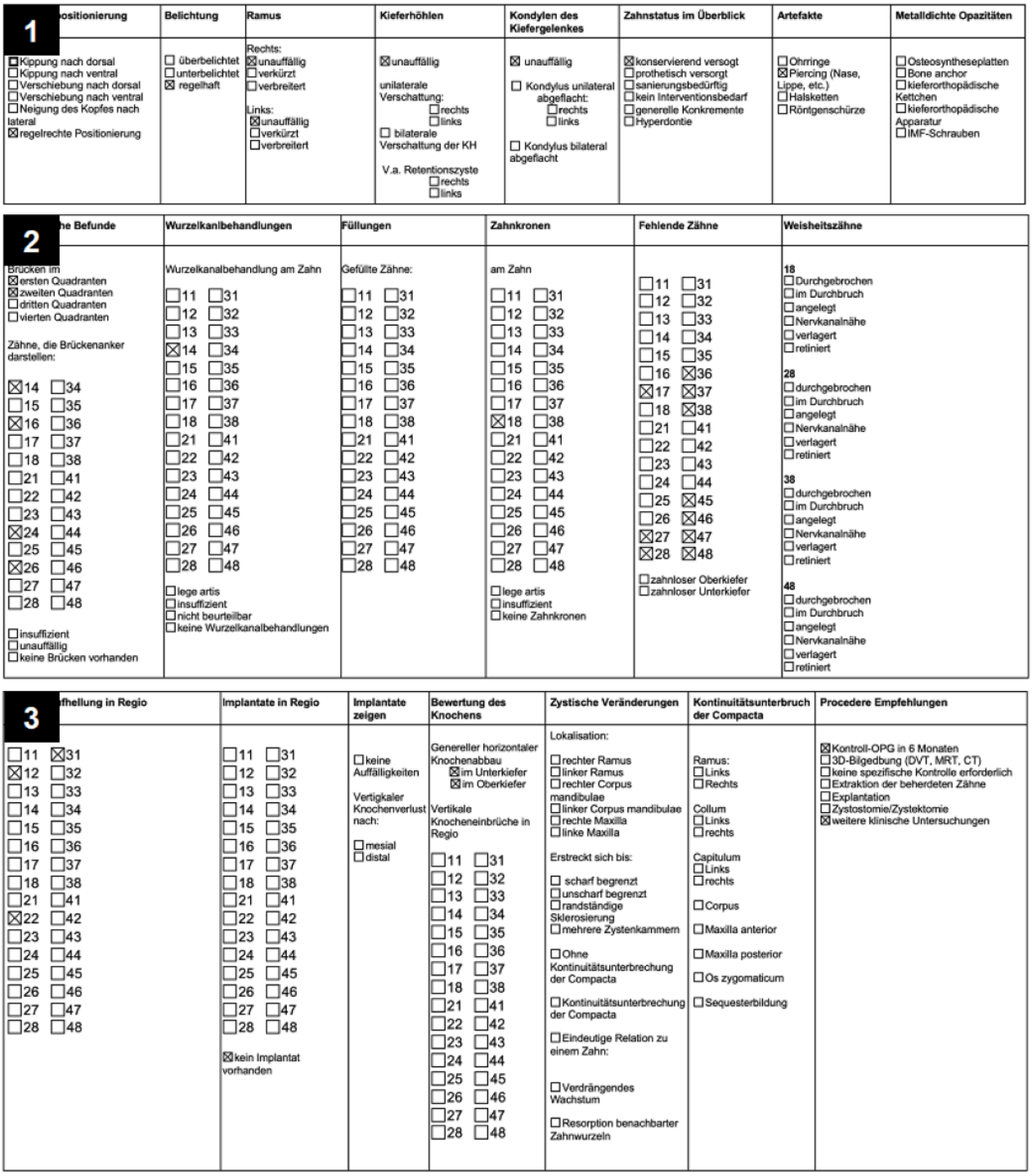
XSL•FO
RenderX

Figure 1. (A,B) Two randomly chosen panoramic radiographs featuring various pathologies to be diagnosed by dental students. Both x-ray images represented the basis of a student-written and an AI-generated radiology report. AI: artificial intelligence.



XSL•FO
RenderX

Figure 2. An example of a completed checkbox list containing 3 distinct spreadsheets (1, 2, and 3) used to generate radiology reports with ChatGPT.



1

positionierung	Belichtung	Ramus	Kieferhöhlen	Kondylen des Kiefergelenkes	Zahnstatus im Überblick	Artefakte	Metalldichte Opazitäten
☐Kippung nach dorsal ☐Kippung nach ventral ☐Verschiebung nach dorsal ☐Verschiebung nach ventral ☐Neigung des Kopfes nach lateral ☒regelrechte Positionierung	☐ überbelichtet ☐unterbelichtet ☒ regelhaft	Rechts: ☒unauffällig ☐verkürzt ☐verbreitert Links: ☒unauffällig ☐verkürzt ☐verbreitert	☒unauffällig unilaterale Verschattung: ☐rechts ☐links ☐ bilaterale Verschattung der KH V.a. Retentionszyste ☐rechts ☐links	☒ unauffällig ☐ Kondylus unilateral abgeflacht: ☐rechts ☐links ☐ Kondylus bilateral abgeflacht	☒konservierend versogt ☐prothetisch versorgt ☐sanierungsbedürftig ☐kein Interventionsbedarf ☐generelle Konkremente ☐Hyperdontie	☐Ohrringe ☒Piercing (Nase, Lippe, etc.) ☐Halsketten ☐Röntgenschürze	☐Osteosyntheseplatten ☐Bone anchor ☐kieferorthopädische Kettchen ☐kieferorthopädische Apparatur ☐IMF-Schrauben

2

he Befunde	Wurzelkanalbehandlungen	Füllungen	Zahnkronen	Fehlende Zähne	Weisheitszähne
Brücken im ☒ersten Quadranten ☒zweiten Quadranten ☐dritten Quadranten ☐vierten Quadranten Zähne, die Brückenanker darstellen: ☒14 ☐34 ☐15 ☐35 ☒16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☒24 ☐44 ☐25 ☐45 ☒26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐insuffizient ☐unauffällig ☐keine Brücken vorhanden	Wurzelkanalbehandlung am Zahn ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☒14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐lege artis ☐insuffizient ☐nicht beurteilbar ☐keine Wurzelkanalbehandlungen	Gefüllte Zähne: ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	am Zahn ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☒18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☐lege artis ☐insuffizient ☐keine Zahnkronen	☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☒36 ☒17 ☒37 ☐18 ☒38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☒45 ☐26 ☒46 ☒27 ☒47 ☒28 ☒48 ☐zahnloser Oberkiefer ☐zahnloser Unterkiefer	18 ☐Durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 28 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 38 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert 48 ☐durchgebrochen ☐im Durchbruch ☐angelegt ☐Nervkanalnähe ☐verlagert ☐retiniert

3

fhellung in Regio	Implantate in Regio	Implantate zeigen	Bewertung des Knochens	Zystische Veränderungen	Kontinuitätsunterbruch der Compacta	Procedere Empfehlungen
☐11 ☒31 ☒12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☒22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48 ☒kein Implantat vorhanden	☐keine Auffälligkeiten Vertigkaler Knochenverlust nach: ☐mesial ☐distal	Genereller horizontaler Knochenabbau ☒im Unterkiefer ☒im Oberkiefer Vertikale Knocheneinbrüche in Regio ☐11 ☐31 ☐12 ☐32 ☐13 ☐33 ☐14 ☐34 ☐15 ☐35 ☐16 ☐36 ☐17 ☐37 ☐18 ☐38 ☐21 ☐41 ☐22 ☐42 ☐23 ☐43 ☐24 ☐44 ☐25 ☐45 ☐26 ☐46 ☐27 ☐47 ☐28 ☐48	Lokalisation: ☐rechter Ramus ☐linker Ramus ☐rechter Corpus mandibulae ☐linker Corpus mandibulae ☐rechte Maxilla ☐linke Maxilla Erstreckt sich bis: ☐ scharf begrenzt ☐unscharf begrenzt ☐randständige Sklerosierung ☐mehrere Zystenkammern ☐Ohne Kontinuitätsunterbrechung der Compacta ☐Kontinuitätsunterbrechung der Compacta ☐Eindeutige Relation zu einem Zahn: ☐Verdrängendes Wachstum ☐Resorption benachbarter Zahnwurzeln	Ramus: ☐Links ☐Rechts Collum ☐Links ☐rechts Capitulum ☐Links ☐rechts ☐Corpus ☐Maxilla anterior ☐Maxilla posterior ☐Os zygomaticum ☐Sequesterbildung	☒Kontroll-OPG in 6 Monaten ☐3D-Bilgedbung (DVT, MRT, CT) ☐keine spezifische Kontrolle erforderlich ☐Extraktion der beherdeten Zähne ☐Explantation ☐Zystostomie/Zystektomie ☒weitere klinische Untersuchungen

Data Transformation and AI Text Generation

Upon completion, the checkbox lists filled out by participants were carefully transcribed into an Excel (Microsoft) data sheet comprising distinct spreadsheets for each category to organize the data. Subsequently, Chat GPT 4.0, an advanced AI language model, was harnessed to generate radiology reports based on the checkbox lists. Each spreadsheet within the Excel file was sequentially analyzed, and the information marked with an "X" in the "checkbox" column was incorporated into the generated reports. The following specific prompt was used to guide the AI in formulating structured x-ray reports, ensuring consistency and completeness:

> *Formulate a structured X-ray report in the sense of an X-ray report of an OPG based on the following checkbox list of the entire Excel table, and do not omit any columns. Please mention only those statements for which a box is marked with an X in the X-ray report. The statements not marked with an X should not be included in the report. The figures given should be interpreted in the sense of an odontogram. Analyze each spreadsheet in the Excel file in the order*

XSL•FO
RenderX

given. The column with the markings (X) is marked with the term "checkbox." The report should be written in continuous text from the perspective of the treating dentist. Formulate a continuous text without subheadings.

The model settings included a temperature of 0.7 (controlling the randomness of responses), a maximum token limit of 1500 (restricting the length of the response), a frequency penalty of 0.0 (preventing repetitive word usage), and a presence penalty of 0.6 (promoting the inclusion of new topics). Those settings were shown to generate the highest output quality with the temperature setting of 0.7 being particularly important. Although lower settings are suggested to be advantageous for more deterministic tasks, preliminary tests revealed them to produce repetitive and difficult-to-read reports lacking naturalness and effectiveness in communication. In contrast, the chosen setting balanced creativity and coherence resulting in improved readability. Each report was generated using the ChatGPT web interface in a new session from September 5 to October 12, 2023, ensuring consistency and comparability across all outputs. To minimize biases associated with varying performance due to server load, which tends to be higher on weekends with higher traffic, the tasks were randomly distributed across different weekdays. This approach aimed to ensure a consistent and balanced evaluation of ChatGPT's capabilities by reducing potential variability in output quality. The checkbox lists were directly uploaded without additional preprocessing.

Readability Indices

The readability and complexity of both student-written and AI-generated texts were assessed using the Flesch reading ease (FRE) [24] score and the Lesbarhetsindex (LIX) readability index. "Readability" refers to how easily written material can be understood, determined by the complexity of the vocabulary, sentence, and word lengths used [23]. Although prior knowledge or motivation of the reader is not considered in readability formulas, especially in health care, a higher readability is associated with improved comprehension and participation of the patient.

The FRE score evaluates text readability based on its linguistic characteristics. In particular, the average sentence length (ASL) and the average number of syllables per word (ASW) are considered for the calculation using the following formula (adapted to the German language [25]):

$$\text{Flesch reading ease} = 180 - \text{ASL} - 58.5 \times \text{ASW}$$

The FRE score typically ranges between 0 and 100, with higher scores indicating greater readability and lower scores suggesting increased complexity. Due to its high reproducibility [25], validation for various text types, and correlation with other readability formulas, the FRE score is an established metric in the analysis of medical texts [26-29].

LIX index considers the ASL and the prevalence of long words with more than 6 letters to assess text readability by the following calculation:

$$\text{LIX} = \frac{\text{Total number ofwords}}{\text{Total number of sentences}} + \frac{(\text{Number of long words} \times 100)}{\text{Total number of words}}$$

A higher LIX score indicates greater complexity, whereas a lower score suggests easier comprehension. LIX has been validated as a reliable measure of readability across multiple languages, including Swedish, Danish, English, French, German, Finnish, Italian, Spanish, and Portuguese [30,31].

To assess readability, the FRE and the LIX scores were calculated for both the student-written and AI-generated reports. Differences in readability were analyzed by comparing FRE and LIX scores of AI-generated reports with student-written reports. Additionally, this analysis was conducted collectively for all texts, as well as individually for each academic semester, to evaluate the influence of the educational level on text comprehensibility in comparison to automated text generation.

Text Similarity (Bidirectional Encoder Representations from Transformers Score)

The accuracy of AI-generated texts was evaluated by comparing the number of findings diagnosed by students to those mentioned in the final AI-generated reports. Additionally, reference texts were manually created by a senior physician with extensive clinical experience in dental radiology, for each checkbox list to assess the quality of AI-generated texts. A comprehensive template was developed and carefully reviewed by all authors, serving as a standardized framework for report creation. Each reference text was individually crafted by transferring the findings from the corresponding checkbox list into the template. This standardized approach was consistently applied to each report, ensuring uniformity in content and structure while minimizing discrepancies that could bias the Bidirectional Encoder Representations from Transformers (BERT) score. These reference texts were then compared to the AI-generated text using the BERT score, a widely recognized metric for evaluating text similarity. Based on the BERT model [32], which generates high-dimensional vector representations (embeddings), capturing the BERT score measures the similarity between corresponding tokens in both texts. The BERT score includes 3 primary components—precision (P), recall (R), and F_1-score. Precision measures the proportion of words in the AI-generated text that contribute accurately to the overall meaning as compared to the reference text. Essentially, it assesses the quality of the AI's output in terms of the relevance and accuracy of the information presented. Recall evaluates the extent to which the AI-generated text covers all the relevant information contained in the reference text, highlighting how well the AI captures necessary details without omitting critical information. Finally, the F_1-score provides a harmonic mean of precision and recall, offering a single score that balances both the completeness and accuracy of the AI-generated text. The aggregated similarity scores, normalized to a range between 0 and 1, indicate overall text similarity. A higher BERT score indicates textual similarity, reflecting a higher quality of AI-generated texts. Multiple studies have already demonstrated BERT's capability of accurately predicting readability levels for various texts [33].

XSL•FO
RenderX

Text Accuracy

The accuracy of AI-generated radiology reports was assessed by comparing the number of included findings with the number of findings contained in the referring checkbox list, both overall and for each of the 3 individual spreadsheets separately.

Text Analysis

Descriptive text analysis was conducted by measurement of word count, sentence length, syllable count, diphthong count, and character count to compare AI-generated with student-written radiology reports. ASL and long word proportion (defined as words with more than 6 characters) were further assessed. Language quality across all texts was quantified by evaluating the error count including spelling, grammar, and punctuation together with the calculation of the error ratio (number of errors divided by words multiplied by 100). These metrics were analyzed collectively for all students and semesters as 1 group.

Statistical Analysis

The software packages used for statistical analysis were GraphPad Prism 9.0 (Graphpad Software, LLC), G*Power 3.1 (Heinrich-Heine-University Düsseldorf), Excel 16.76, and SPSS Statistics (version 29; IBM Corp). To assess the potential difference in readability between AI-generated reports and those written by students, an a priori power analysis was performed. This analysis was based on previously observed significant differences in the FRE scores, which showed a lower average for ChatGPT responses (mean 34.9, SD 11.2) compared to medical information on Google webpages (mean 46.5, SD 14.3), accounting for a difference of 11.6 [34]. Additionally, similar results with the LIX score have demonstrated a difference of 10 between human-written and ChatGPT-written scientific introductions [35]. To achieve a power of 80% and maintain a significance level of 5%, a minimum of 25 samples per group (study arm) is required. Significance was set at $P<.05$. All data are presented as mean (SD). Differences between student-written and AI-generated texts were analyzed using a 2-tailed student *t* test. A subsequent post hoc power analysis was conducted for each test to verify the power achieved by the *t* test.

Results

Overview

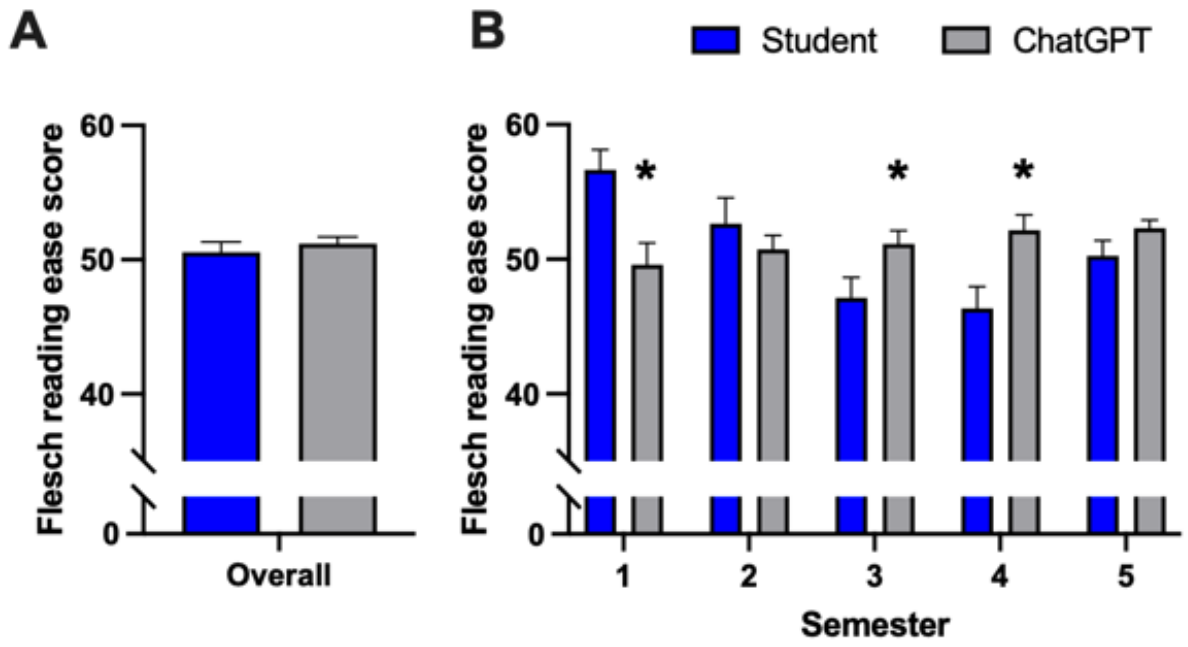
Text quality, readability, and comprehensibility of student-written and AI-generated radiology reports were compared by analysis of various language parameters. Throughout the study, students consistently used the preset time to its full extent, dedicating 30 minutes for completing the written report and 10 minutes for the competition of the checkbox list. While AI-generated radiology reports demonstrated a remarkable similarity to reference texts with no difference in readability, a significant information deficiency was observed.

AI-Generated and Student-Written Texts Possessed Identical Readability

The FRE score (Figure 3A,B) revealed no difference in readability between AI-generated and student-written texts (mean 50.55, SD 7.80 vs mean 51.19, SD 5.02) considering all reports together as demonstrated in Figure 3A ($P=.49$; $t_{198}=0.6898$). Upon examination of each semester individually, the Flesch index exhibited significant variability, with AI-generated texts demonstrating lower readability compared to texts written by the first clinical semester (mean 56.65, SD 6.70 vs mean 49.6, SD 7.17; $P=.002$; $t_{38}=3.213$) and higher readability compared to the third (mean 47.14, SD 6.97 vs mean 51.14, SD 4.55; $P=.03$; $t_{40}=2.203$) and fourth (mean 46.35, SD 7.29 vs mean 52.15, SD 5.08; $P=.006$; $t_{38}=2.918$) clinical semesters. No difference was found for second ($P=.39$; $t_{36}=0.8647$) and fifth ($P=.12$; $t_{38}=1.577$) clinical semester.

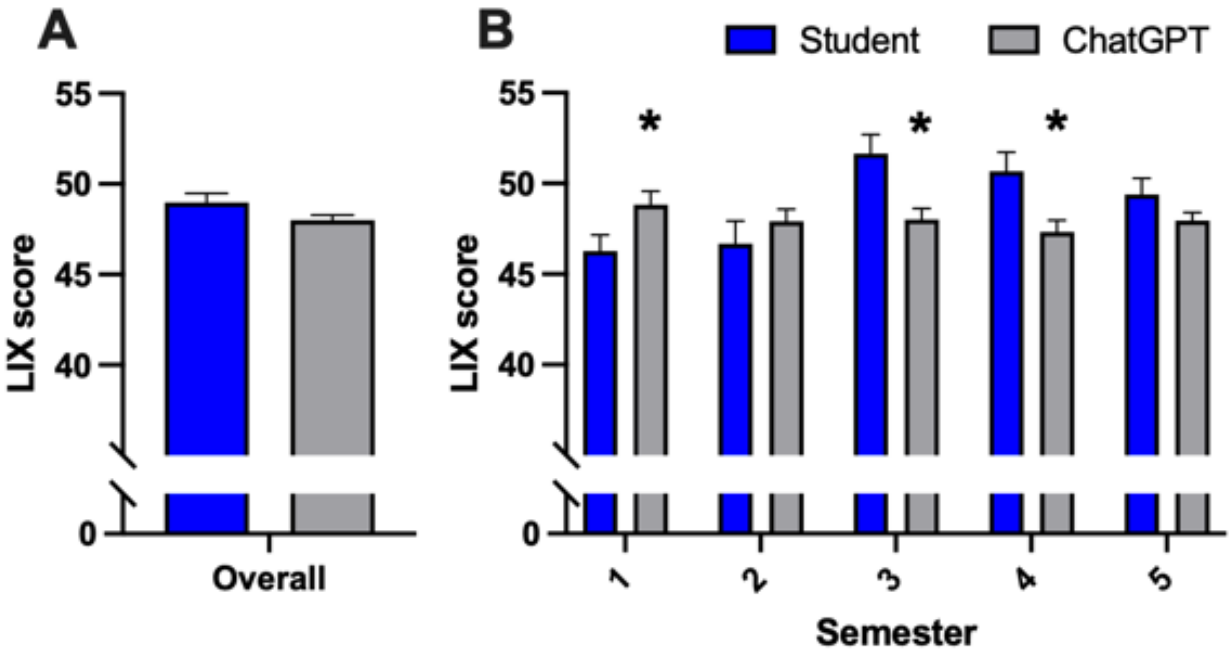
XSL•FO
RenderX

Figure 3. Metric evaluation of readability of AI-generated reports compared to student-written radiology reports (A) overall and (B) individually for each semester assessed with the Flesch readability ease score. Data represent mean (SD). Sample size: (A) n=100; (B) semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20; *P<.05; and versus students. AI: artificial intelligence.



As presented in Figure 4A, no overall difference between both groups regarding readability was found (mean 48.98, SD 5.0 vs mean 48.0, SD 2.85) as assessed with the LIX index (P=.09; t_{198}=1.699). In contrast to the FRE score, the LIX readability index exhibited opposing trends across semesters (Figure 4B), with significant differences observed in semesters 1 (mean 46.27, SD 4.0 vs mean 48.81, SD 3.44; P=.04; t_{38}=2.157); 3 (mean 51.64, SD 4.89 vs mean 48.01, SD 2.84; P=.005; t_{40}=2.944); and 4 (mean 50.67, SD 4.68 vs mean 47.32, SD 2.90; P=.098; t_{38}=2.719). No difference was observed in the second (P=.39; t_{36}=0.877) and fifth (P=.15; t_{38}=1.464) clinical semester.

Figure 4. Metric evaluation of readability of AI-generated reports compared to student-written radiology reports (A) overall and (B) individually for each semester assessed with LIX index. Data represent mean (SD). (A) Sample size: n=100; (B) semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20; *P<.05; and versus students. AI: artificial intelligence; LIX: Lesbarhetsindex.

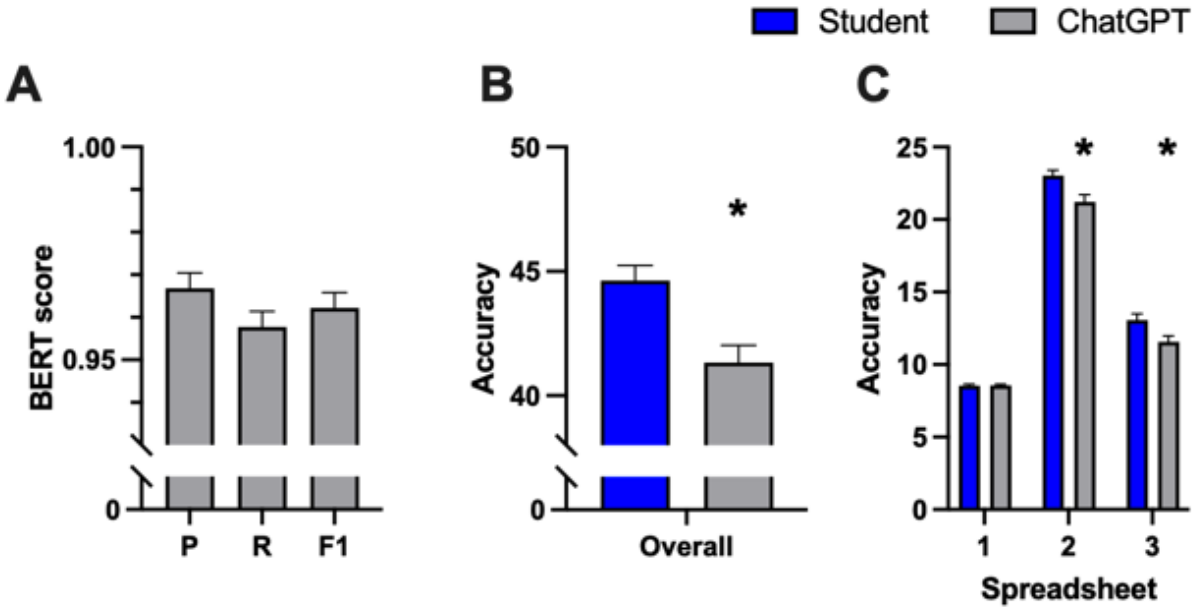


XSL•FO
RenderX

AI-Generated Reports Show Great Similarity to Reference Texts but Lack Information

As illustrated in Figure 5A, the great similarity is indicated by a high BERT score, with precision (P)=mean 0.967, SD 0.036, recall (R)=mean 0.958, SD 0.037, and F_1=mean 0.962, SD 0.036. The analysis further revealed a notable deficiency in relevant information within AI-generated texts. A significant difference was evident between the findings diagnosed by students and those mentioned in the AI-generated reports (Figure 5B), with students identifying a mean of 44.6 (SD 6.0) findings, whereas the AI reported a mean of 41.3 (SD 7.0) findings in total (P=.04; t_{198}=3.586). Specifically, as shown in Figure 5C, while no difference was observed in the first spreadsheet (mean 8.53, SD 1.06 vs mean 8.55, SD 1.02; P=.89; t_{198}=0.1361), the AI included significantly fewer findings from the second (mean 23.03, SD 3.67 vs mean 21.22, SD 4.92; P=.003; t_{198}=2.951) and third (mean 13.07, SD 4.40 vs mean 11.56, SD 4.23; P=.014; t_{198}=2.476) spreadsheets.

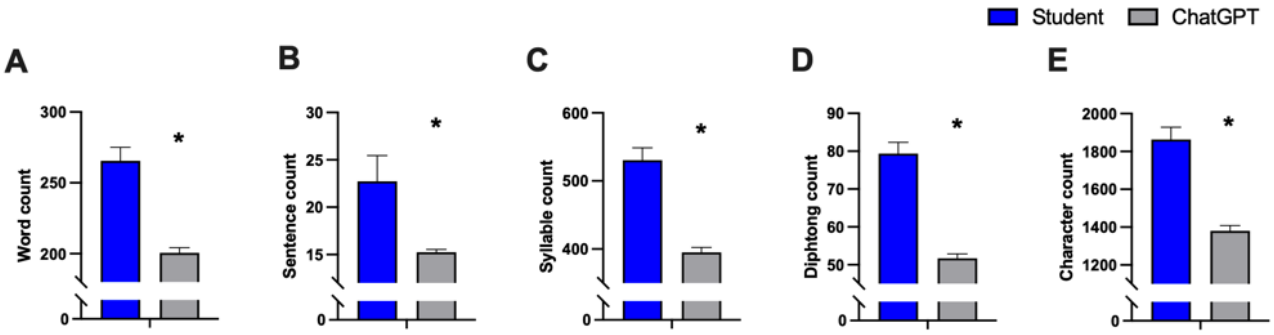
Figure 5. Evaluation of similarity compared to reference texts using the (A) BERT score with precision (P), recall (R), and F1 score (F1) representing the harmonic mean of precision and recall. The accuracy of AI-generated radiology reports was further assessed as (B) overall accuracy including the whole checkbox list and (C) individually for each spreadsheet of the checkbox list. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence; BERT: Bidirectional Encoder Representations from Transformers.



AI-Generated Significantly Shorter and Error-Free Radiology Reports

AI-generated radiology reports exhibited a significant 24% reduction in word count (mean 265.6, SD 95.4 vs mean 200.6, SD 37.3 words; P<.01; t_{198}=6.347) and sentence count (P=.007; t_{198}=2.726) accompanied by significant reductions in syllables (P<.01; t_{198}=6.823), diphthongs (P<.01; t_{198}=8.643), and characters (P<.01; t_{198}=6.841) compared to student-written texts as presented in Figure 6.

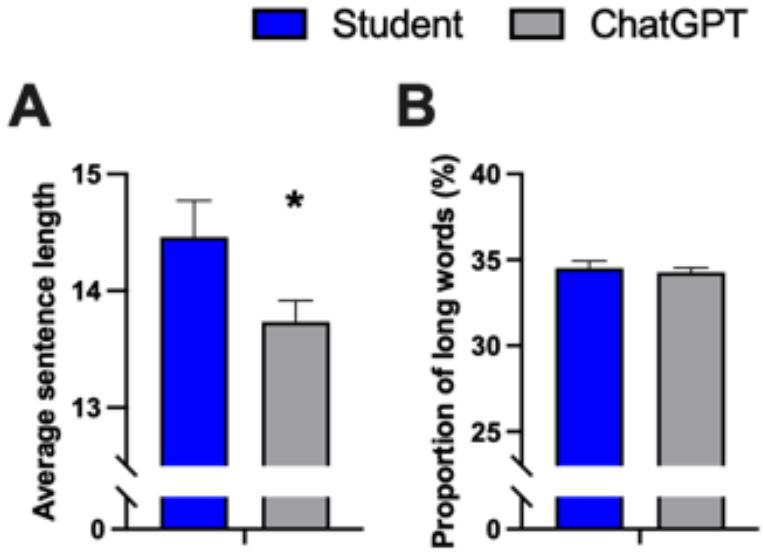
Figure 6. Analysis of (A) word count, (B) sentence count, (C) syllable count, (D) diphthong count, and (E) character count of AI-generated radiology reports compared to student-written reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.



Whereas radiology reports generated by AI showed a significant reduction in ASL compared to student-written reports (A: mean 14.5, SD 3.1 vs mean 13.7, SD 1.8 words; P=.046; t_{198}=2.007), no difference was observed regarding the proportion of long words (B: P=.61; t_{198}=0.509) and this is presented in Figure 7.

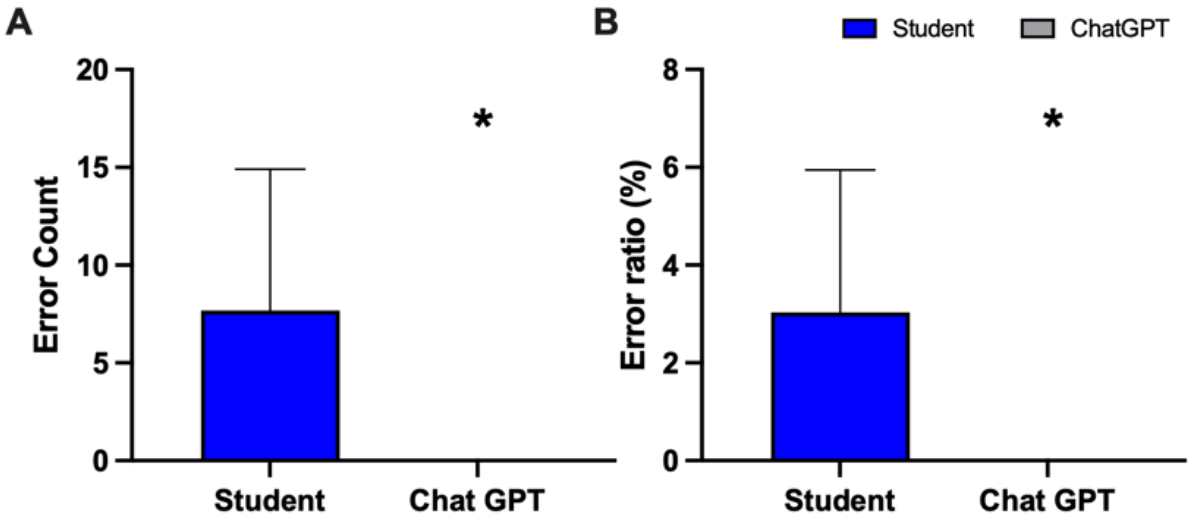
XSL•FO
RenderX

Figure 7. Analysis of sentence length and long word proportion (more than 6 characters) of AI-generated radiology reports compared to student-written reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.



Contrary to student-written reports, AI-generated texts showed a complete absence of orthographic, grammatical, and punctuation errors as presented in Figures 8A and 8B (A: mean 7.7, SD 7.2 vs mean 0; P<.01; t_{198}=10.59; B: mean 2.9, SD 2.9 vs mean 0; P<.01; t_{198}=10.41).

Figure 8. Analysis of (A) error count including grammar, spelling, and punctuation and (B) error ratio by calculating errors divided by words multiplied by 100 of student-written radiology reports and AI-generated radiology reports. Data represent mean (SD). Sample size: n=100; *P<.05 versus students. AI: artificial intelligence.

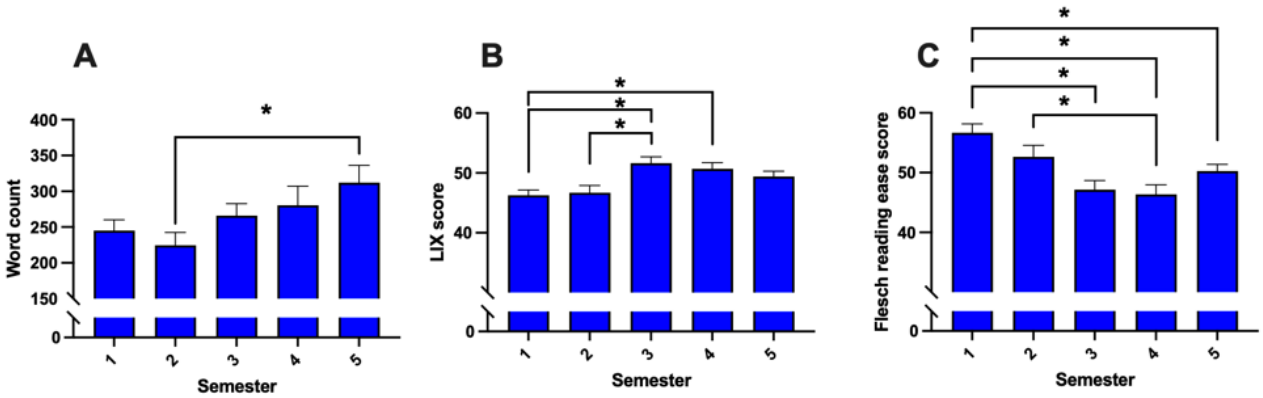


Student-Written Reports Showed Significant Differences in Length and Readability Across Semesters

Student-written reports showed a significant difference in word count across different semesters (P=.04; t_{99}=1.610), with a noticeable trend toward the use of more words in higher semesters (Figure 9A; semester 2 vs semester 5: mean 224, SD 79 vs mean 312, SD 107; P=.03; t_{95}=2.97). Additionally, significant differences in readability were observed across semesters, as presented in Figures 9B and 9C (LIX score: P=.06; t_{99}=2.315; FRE score: P<.01; t_{99}=2.762). Students from semesters 1 and 2 produced simpler and easier-to-understand reports, whereas those from higher semesters tended to write more complex and, hence, more difficult-to-read reports.

XSL•FO
RenderX

Figure 9. (A) Analysis of word count, (B) metric evaluation of readability using LIX score, and (C) Flesch reading ease score of student-written radiology reports. Data represent mean (SD). Sample size: n=100; semester 1: n=20, semester 2: n=19, semester 3: n=21, semester 4: n=20, and semester 5: n=20. *P<.05. LIX: Lesbarhetsindex.



Discussion

Principal Findings

Integrating AI into clinical workflows and medical education has attracted significant interest due to its potential to enhance efficiency. Our study, therefore, aimed to investigate the effectiveness of AI-generated radiology reports by comparing them to student-written reports. In summary, no difference regarding readability was found between the AI-generated and student-written radiology reports. Whereas AI-generated reports showed an overall high textual similarity to reference texts, they simultaneously lacked substantial diagnostic information. Noteworthy, the language quality was significantly improved compared to student-written reports, with AI-generated texts being completely error-free. The results revealed high potential in enhancing medical writing with AI while still being limited by the reliability of the transferred information [36].

AI is revolutionizing medicine across diagnosis, treatment, and administrative tasks [5]. AI algorithms analyze medical images for early disease detection, provide clinical decision support, and enable personalized treatment plans [3]. In drug development, AI accelerates processes by predicting drug interactions and screening compounds [2]. Whereas the capability of AI to diagnose x-ray images has already been proven [4], the process of diagnosing was excluded in this study to focus on generating radiology reports based on information collected by students.

Indicated by an overall high BERT score, our findings prove that AI-generated radiology reports exhibit a great level of similarity to reference texts. Integration of ChatGPT into the medical documentation process, therefore, results in high text quality as represented by high precision, recall, and F_1-score, which further highlights the overall robustness of AI in replicating the content of reference reports only based on checkbox information. Concomitantly, after the preparation of preliminaries regarding the prompt and template design, automating the process of report writing with ChatGPT results in significant time savings and, hence, highlights the potential to streamline the workflow for health care professionals. Although this study did not aim to specifically quantify time efficiency, AI-supported report generation was noticeably quicker due to the shorter time cap (10 minutes for checkbox list completion vs 30 minutes for report writing). Whereas all students were observed to use the entire allotted time, likely focusing on thoroughness and ensuring report completeness, rather than being constrained by the time limit, future research could incorporate exact time measurements and posttask surveys to gather participant feedback on time allocation to provide additional insights regarding the adequacy of the time frames. The successful use of ChatGPT in composing medical notes related to patient transfers, operative procedures, and surgical assistance [5,18,37] underscores its role in enhancing productivity within medical environments, thereby highlighting the transformative impact of AI-driven technologies in health care.

Prior to the study, the prompt design was refined extensively, with multiple versions tested to optimize the AI's output. Ultimately, only the most effective prompt was selected to continue generating reports, ensuring the highest possible accuracy in AI-generated text. However, despite their overall similarity, AI-generated reports demonstrated a significant deficiency in relevant information, indicating a crucial impact of the prompt provided to ChatGPT in determining the accuracy of the results. This discrepancy was particularly evident in identifying and incorporating findings regarding specific teeth, with AI-generated reports containing significantly fewer findings compared to the number of diagnoses documented with the checkbox lists (eg, AI-generated reports did not mention the presence of a cyst, the status of dental restorations or precise prescription of bone loss). Interestingly, while no difference was observed in the findings reported from the first spreadsheet, a significant disparity emerged in subsequent spreadsheets. The potential limitation in the AI's ability to comprehensively interpret complex odontogram data leads to inconsistencies in the inclusion of relevant findings. These findings underscore the error-prone interplay between prompt precision, image complexity, and AI performance in radiology reporting.

In contrast to missing information, another known challenge in the use of ChatGPT is its tendency to generate plausible-sounding but incorrect or fabricated information, commonly referred to as "hallucinations" [38]. However, this

XSL•FO
RenderX

study showed no indication that ChatGPT included invented findings not present in the original checkbox list, as evidenced by the high BERT score. The prompt design strictly instructed the AI to use only information from the checkbox list, thereby minimizing the risk of hallucinations. Our observations confirmed that the model adhered to these guidelines. Although the prompt instructed ChatGPT to interpret the numbers as odontogram information to identify each tooth, we encountered challenges in consistently incorporating and accurately understanding the provided data. Precision in prompt formulation emerges as a critical factor influencing the accuracy and completeness of AI-generated reports [39]. The formulation of prompts significantly influences the outcomes generated by AI systems, with precise prompts being necessary to provide clear instructions and context for the AI model, guiding it in producing relevant and accurate responses. The specificity and clarity of the prompt directly impact the quality and relevance of the AI-generated output [40]. The design of effective prompts, therefore, remains a crucial part of future research.

Regarding radiology reports, a prompt that precisely outlines the required structure, format, and content of the report will likely result in more coherent outputs. Moreover, the prompt helps the AI model understand the task and focus on relevant information. By providing detailed guidelines and constraints, the prompt narrows the scope of the AI's search and directs it toward generating responses that align with the desired objectives. Additionally, prompts can incorporate domain-specific terminology and concepts to ensure that the AI model produces contextually appropriate and clinically relevant outputs. Nonetheless, a potential bias of machine learning systems must be considered due to their susceptibility to being influenced by the training data, thereby generating biased or misleading outputs. Well-designed prompts can help mitigate these issues by guiding the AI model toward more objective and accurate responses [41]. Moreover, the challenges associated with interpreting complex diagnostic data like orthopantomograms emphasize the need for continuous refinement and optimization of AI algorithms to ensure reliable performance in a clinical setting. Addressing these challenges will require a collaborative effort between clinicians, AI developers, and educators. Enhancing prompt precision through detailed guidelines and standardized protocols can improve AI performance and reduce information deficiencies in generated reports. Notably, to realize the full potential of AI in health care, the risk of disseminating misinformation must be mitigated. The rapid spread of false or misleading content, commonly called infodemic, highlights the importance of implementing validation mechanisms to ensure the reliability of AI-generated content [42,43].

In the context of this study, the AI was not supposed to formulate diagnoses independently but rather to generate radiology reports based explicitly on the findings and diagnoses provided by the students. This approach evaluated the AI's ability to effectively translate diagnostic information into coherent and comprehensive reports, reflecting real-world clinical scenarios of radiologists interpreting images and automatically converting their findings into written reports. Overall, this evaluation of AI-generated reports underscores the reliability and consistency of AI in producing error-free content compared to student-written texts. Remarkably, AI-generated reports exhibited a considerable reduction in word count, sentence count, and various linguistic features, including syllables and diphthongs. This reduction in length was accompanied by a notable absence of orthographic, grammatical, and punctuation errors, highlighting the accuracy and precision of AI-generated text. Moreover, no discernible difference between AI-generated and student-written radiology reports was observed regarding their readability. Both sets of reports demonstrated similar readability levels as indicated by the FRE score and LIX index, with both being established and validated as reliable measures for assessing text difficulty, including medical texts [27,29,30]. However, examination of individual semesters revealed significantly lower readability for AI-generated reports than student-written ones in the first clinical semester, but higher readability compared to the third and fourth clinical semesters. A possible explanation could be the use of more advanced and specialized terminology by the AI compared to students in the first semester, resulting in lower readability scores. Reports from students in the first semester may adhere to a simpler structure, reducing difficulty and increasing readability and comprehension. In contrast, reports from the third and fourth semesters exhibit more complexity in structure and terminology to present diagnostic information due to extensive expertise and, therefore, impairing readability. These findings are supported by the significant differences in word count and readability across all semesters upon individual examination. Students in lower semesters tend to use fewer words and write reports with higher readability, whereas students from higher semesters tend to write longer, more complex, and therefore, more difficult-to-read reports. As students progress through their education, their increased clinical experience and familiarity with radiological terminology likely enhance the quality of their reports. This development is reflected by the incorporation of advanced terminology and structure, indicating a clear learning curve. The variability in skill development across semesters significantly impacts the comparison between student-written and AI-generated reports, potentially affecting the comparison in favor of later semesters. This disparity underscores the importance of considering skill levels when evaluating AI performance since differences in student proficiency could lead to variability in report quality, affecting readability and accuracy metrics. Consequently, the perceived quality of AI-generated reports may vary depending on the student cohort they are compared with, highlighting the necessity of accounting for student skill differences in the study design and analysis. However, on the other hand, the variability in student skills across semesters positively reflects the diverse real-world conditions in clinical practice, where practitioners exhibit a range of expertise. This diversity in the study cohort allows the AI-generated reports to be tested against various levels of proficiency, demonstrating the AI's potential to support users with different levels. Early-stage dental students could benefit from a structured and consistent framework provided by AI, enhancing their learning and understanding. Advanced students and experienced clinicians could use AI to reduce repetitive tasks and ensure accuracy in documentation, allowing more focus on diagnostic decision-making. AI assistance in

XSL•FO
RenderX

diagnostics has been further shown to improve performance, though radiologists often underweight AI predictions [44]. Overall, AI tools can support a wide range of users by adapting to their specific needs and improving educational and clinical outcomes. This aligns with study results proving GPT-4 to enhance productivity and quality in various tasks beyond medical use, benefiting consultants and customer support agents across all skill levels [45,46].

Nevertheless, the differentiated use of reports must be considered due to the diverse communication needs within health care settings. On the one hand, health care professionals require detailed reports for accurate clinical decision-making and effective interprofessional communication. On the other hand, patients benefit from simpler, more understandable reports to understand their medical conditions and actively engage in treatment discussions. AI has been further shown to efficiently simplify medical data for better patient understanding [22]. Hence, its implementation offers the possibility to fulfill both the detailed requirements of health care professionals and the simplified needs of patients simultaneously in response to 2 different prompts. Consequently, automated AI solutions could facilitate effective communication among health care providers and increase patient empowerment and participation beyond time-saving and more efficient documentation in health care.

Conclusions

In conclusion, AI's potential to enhance medical writing efficiency is highlighted, yet remaining challenges in ensuring reliability and comprehensiveness must be faced. The precision of prompts significantly impacts AI's accuracy, particularly in interpreting complex diagnostic data. Future research should focus on refining AI algorithms and prompt design to optimize medical reporting. Overall, integrating AI-driven solutions into routine clinical workflows offers a practical tool for enhancing productivity.

Acknowledgments

This study was conducted as part of ASB's PhD thesis.

Conflicts of Interest

None declared.

References

1. Ebbers T, Kool RB, Smeele LE, Dirven R, den Besten CA, Karssemakers LHE, et al. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. J Med Syst. 2022;46(7):46. [FREE Full text] [doi: 10.1007/s10916-022-01837-9] [Medline: 35618978]
2. Vemula D, Jayasurya P, Sushmitha V, Kumar YN, Bhandari V. CADD, AI and ML in drug discovery: a comprehensive review. Eur J Pharm Sci. 2023;181:106324. [FREE Full text] [doi: 10.1016/j.ejps.2022.106324] [Medline: 36347444]
3. Weisberg EM, Fishman EK. The future of radiology and radiologists: AI is pivotal but not the only change afoot. J Med Imaging Radiat Sci. 2024;55(4):101377. [doi: 10.1016/j.jmir.2024.02.002] [Medline: 38403516]
4. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study. Korean J Radiol. 2022;23(5):505-516. [FREE Full text] [doi: 10.3348/kjr.2021.0476] [Medline: 35434976]
5. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33. [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
6. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res. 2023;25:e47479. [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
7. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Machines. 2020;30(4):681-694. [doi: 10.1007/s11023-020-09548-1]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med. 2023;99(2):192-197. [doi: 10.1097/acm.0000000000005549]
10. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y). 2024;5(3):100943. [FREE Full text] [doi: 10.1016/j.patter.2024.100943] [Medline: 38487804]
11. 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. [FREE Full text] [doi: 10.1371/journal.pone.0297701] [Medline: 38354135]
12. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The use of ChatGPT for education modules on integrated pharmacotherapy of infectious disease: educators' perspectives. JMIR Med Educ. 2024;10:e47339. [FREE Full text] [doi: 10.2196/47339] [Medline: 38214967]

XSL•FO
RenderX

13. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
14. Flory MN, Napel S, Tsai EB. Artificial intelligence in radiology: opportunities and challenges. Semin Ultrasound CT MR. 2024;45(1):152-160.
15. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus. 2023;15(7):e42133. [FREE Full text] [doi: 10.7759/cureus.42133] [Medline: 37476297]
16. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. 2022;32(11):7998-8007. [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
17. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023;5(4):e179-e181. [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
18. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng. 2023;51(11):2353-2355. [doi: 10.1007/s10439-023-03263-5] [Medline: 37266720]
19. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817-2825. [FREE Full text] [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
20. Qutieshat A, Al Rusheidi A, Al Ghammari S, Alarabi A, Salem A, Zelihic M. Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence. Diagnosis (Berl). 2024;11(3):259-265. [FREE Full text] [doi: 10.1515/dx-2024-0034] [Medline: 38696271]
21. Yokokawa D, Yanagita Y, Li Y, Yamashita S, Shikino K, Noda K, et al. For any disease a human can imagine, ChatGPT can generate a fake report. Diagnosis (Berl). 2024;11(3):329-332. [doi: 10.1515/dx-2024-0007] [Medline: 38386808]
22. Fink MA. [Large language models such as ChatGPT and GPT-4 for patient-centered care in radiology]. Radiologie (Heidelb). 2023;63(9):665-671. [doi: 10.1007/s00117-023-01187-8] [Medline: 37615692]
23. Currie G, Robbie S, Tually P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. J Nucl Med Technol. 2023;51(4):307-313. [doi: 10.2967/jnmt.123.266151] [Medline: 37699647]
24. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
25. Amstad T. Wie verständlich sind unsere Zeitungen? German. Studenten-Schreib-Service, Zu¨rich; 1978.
26. Gajjar AA, Kumar RP, Paliwoda ED, Kuo CC, Adida S, Legarreta AD, et al. Usefulness and accuracy of artificial intelligence chatbot responses to patient questions for neurosurgical procedures. Neurosurgery. 2024;95(1):171-178. [doi: 10.1227/neu.0000000000002856] [Medline: 38353558]
27. Gajjar AA, Patel S, Patel SV, Goyal A, Sioutas GS, Gamel KL, et al. Readability of cerebrovascular diseases online educational material from major cerebrovascular organizations. J Neurointerv Surg. 2024. [doi: 10.1136/jnis-2023-021205] [Medline: 38395602]
28. Irwin SC, Lennon DT, Stanley CP, Sheridan GA, Walsh JC. Ankle conFUSION: the quality and readability of information on the internet relating to ankle arthrodesis. Surgeon. 2021;19(6):e507-e511. [doi: 10.1016/j.surge.2020.12.001] [Medline: 33451875]
29. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Educ Behav. 2006;33(3):352-573. [doi: 10.1177/1090198105277329] [Medline: 16699125]
30. Skrzypczak T, Mamak M. Assessing the readability of online health information for colonoscopy - analysis of articles in 22 european languages. J Cancer Educ. 2023;38(6):1865-1870. [FREE Full text] [doi: 10.1007/s13187-023-02344-2] [Medline: 37493981]
31. Anderson J. LIX and RIX: variations on a little-known readability index. J Reading. 1983;26(6):490-496.
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv.0.
33. Deutsch T, Jasbi M, Shieber S. Linguistic features for readability assessment. arXiv preprint arXiv.00377. [doi: 10.18653/v1/2020.bea-1.1]
34. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV information on Google versus AI (ChatGPT). Otolaryngol Head Neck Surg. 2024;170(6):1504-1511. [doi: 10.1002/ohn.506] [Medline: 37622581]
35. Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. Cureus. 2023;15(11):e49019. [FREE Full text] [doi: 10.7759/cureus.49019] [Medline: 38111405]
36. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. J Med Internet Res. 2024;26:e55037. [FREE Full text] [doi: 10.2196/55037] [Medline: 38648098]
37. Atkinson CJ, Seth I, Xie Y, Ross RJ, Hunter-Smith DJ, Rozen WM, et al. Artificial intelligence language model performance for rapid intraoperative queries in plastic surgery: ChatGPT and the deep inferior epigastric perforator flap. J Clin Med. 2024;13(3):900. [FREE Full text] [doi: 10.3390/jcm13030900] [Medline: 38337594]

XSL•FO
RenderX

38. Huang L. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv preprint arXiv.05232.
39. Nazary F, Deldjoo Y, Di Noia T. Harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT. ChatGPT-HealthPrompt. 2024. URL: https://link.springer.com/chapter/10.1007/978-3-031-50396-2_22 [accessed 2024-01-21]
40. White J, Hays S, Fu Q, Smith JS, Schmidt DC. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. arXiv preprint arXiv.07839. [doi: 10.1007/978-3-031-55642-5_4]
41. Hu X, Tian Y, Nagato K, Nakao M, Liu A. Opportunities and challenges of ChatGPT for design knowledge management. Procedia CIRP. 2023;119:21-28. [doi: 10.1016/j.procir.2023.05.001]
42. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120. [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
43. Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, et al. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. J Med Internet Res. 2023;25:e49771. [FREE Full text] [doi: 10.2196/49771] [Medline: 38096014]
44. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. NBER. 2023. URL: https://www.nber.org/papers/w31422 [accessed 2024-06-17]
45. Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN J. 2023:58. [doi: 10.2139/ssrn.4573321]
46. Brynjolfsson E, Li D, Raymond LR. Generative AI at work. arXiv:2304.11771. 2023. [doi: 10.3386/w31161]

Abbreviations

AI: artificial intelligence
ASL: average sentence length
ASW: average number of syllables per word
BERT: Bidirectional Encoder Representations from Transformers
FRE: Flesch reading ease
OPG: panoramic radiograph

Edited by Q Jin; submitted 18.05.24; peer-reviewed by A Qutieshat, R Tabari Khomeiran, S Gorthy, R Janssen; comments to author 29.06.24; revised version received 19.07.24; accepted 03.08.24; published 23.12.24

Please cite as:
Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, Thiem DGE
AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study
J Med Internet Res 2024;26:e60684
URL: https://www.jmir.org/2024/1/e60684
doi: 10.2196/60684
PMID:

©Daniel Stephan, Annika Bertsch, Matthias Burwinkel, Shankeeth Vinayahalingam, Bilal Al-Nawas, Peer W Kämmerer, Daniel GE Thiem. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX