Artificial intelligence in dental radiology: improving the efficiency of reporting with ChatGPT – a comparative study

Daniel Stephan, Annika Sophie Bertsch, Matthias Burwinkel, Shankeeth Vinayahalingam, Bilal Al-Nawas, Peer Wolfgang Kämmerer, Daniel Gerald Eberhard Thiem

Submitted to: Journal of Medical Internet Research
on: May 18, 2024

Table of Contents

Original Manuscript..... 5

Artificial intelligence in dental radiology: improving the efficiency of reporting with ChatGPT – a comparative study

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

[1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery University Medical Centre of the Johannes Gutenberg-University Mainz Mainz DE
[2]Department of Oral and Maxillofacial Surgery Radboud University Medical Center Nijmegen NL
[*]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
DE

Abstract

Background: Structured and standardized documentation is critical for accurately recording diagnostic findings, treatment plans, and patient progress in healthcare. 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 in generating radiology reports from dental panoramic radiographs (OPG), comparing the performance of AI-generated reports with those manually created by dental students.

Methods: One hundred dental students were tasked with analyzing OPGs 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.
Clinical relevance: This underscores the need for further refinement in the AI's prompt design and the development of robust validation mechanisms to enhance its utility in clinical settings.

(JMIR Preprints 18/05/2024:60684)
DOI: https://doi.org/10.2196/preprints.60684

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Artificial intelligence in dental radiology: improving the efficiency of reporting with ChatGPT – a comparative study

Stephan D.[1#], Bertsch A.[1#], Burwinkel M.[1], Vinayahalingam S.[2], Al-Nawas B.[1], Kämmerer P.W.[1*], Thiem D.G.E.[1*]

[#]Equal contribution as first authors

* Equal contribution as last authors

Corresponding author:

Dr. med. Daniel Stephan, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Augustusplatz 2, 55131 Mainz, Germany. Phone: +49-6131-17-3080, Fax: +49-6131-17-8468, Email: daniel.stephan@unimedizin-mainz.de

Authors:

Annika Bertsch, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Dr. med. dent. Matthias Burwinkel, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Shankeeth Vinayahalingam, MD [2]Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, Nijmegen, Philips van Leydenlaan 25, 6525 EX Nijmegen, The Netherlands.

Prof. Dr. med. Dr. med. dent. Bilal Al-Nawas, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Prof. Dr. med. Dr. med. dent. Peer W. Kämmerer, MA, M.Sc. [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Priv.-Doz. Dr. med. Dr. med. dent. Daniel G.E. Thiem, MHBA [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany. Phone: +49-6131-17-3080, Fax: +49-6131-17-8468, Email: daniel.thiem@unimedizin-mainz.de

ABSTRACT

Objectives: Structured and standardized documentation is critical for accurately recording diagnostic findings, treatment plans, and patient progress in healthcare. 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. This study aimed to assess the effectiveness of ChatGPT in generating radiology reports from dental panoramic radiographs (OPG), comparing the performance of AI-generated reports with those manually created by dental students.

Materials and Methods: One hundred dental students were tasked with analyzing OPGs 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.

Conclusion: The findings from this study suggest that ChatGPT has considerable potential for generating radiology reports, although it currently faces challenges in accuracy and reliability.

Clinical relevance: This underscores the need for further refinement in the AI's prompt design and the development of robust validation mechanisms to enhance its utility in clinical settings.

Key words: artificial intelligence, ChatGPT, radiology report, dental radiology, dental orthopantomogram, panoramic radiograph

Introduction

Structured and standardized documentation plays a crucial role in healthcare 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 growing interest in implementing AI technology to optimize healthcare 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 healthcare professionals [5, 6].

First introduced by OpenAI (San Francisco, California) in 2018, generative pre-training transformer (GPT – a specific LLM model developed by OpenAI) has been continuously evolved and trained on extensive text data [7]. ChatGPT (implementation of GPT), an advanced large language model (LLM – a class of AI models), represents a significant advancement in natural language processing (NLP) 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 (USMLE) (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, 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 healthcare 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 panoramic radiographs 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 pre-filled 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 two sets of reports. Secondary outcomes, including text accuracy and language quality, were evaluated to identify potential areas for improvement in AI-driven radiology reporting.

Material and Methods

The present study sought to investigate the efficacy of incorporating AI language models, specifically ChatGPT, in generating radiology reports from pre-filled checkbox lists after analyzing panoramic radiographs. Dental students were assigned to diagnose two different X-ray images, providing a written radiology report for one 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 secondary evaluation of text quality, accuracy, similarities, and disparities between student-written and AI-written reports.

Ethical consideration

The study adhered strictly to ethical guidelines, obtaining informed consent from all participants beforehand. 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. The tasks performed by the students were part of their regular academic curriculum and no additional duties were imposed. Moreover, the analysis of data was conducted anonymously, ensuring that participating in this study resulted in neither advantages nor disadvantages for the students.

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 ten semesters. The first five semesters focus on foundational knowledge, while the following five semesters (clinical semesters 1-5, corresponding to overall semesters 6-10) emphasize clinical skills. The first lesson in dental radiology is introduced in the first clinical semester and therefore preclinical students were excluded from this study. One hundred dental students from all five 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 one of two groups and presented with an unknown

panoramic radiograph (figure 1A and B). 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 panoramic radiography (figure 1B) and was tasked with completing a checkbox list 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 utilized 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 one 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 (e.g. complexity of findings or difficulty of interpretation).

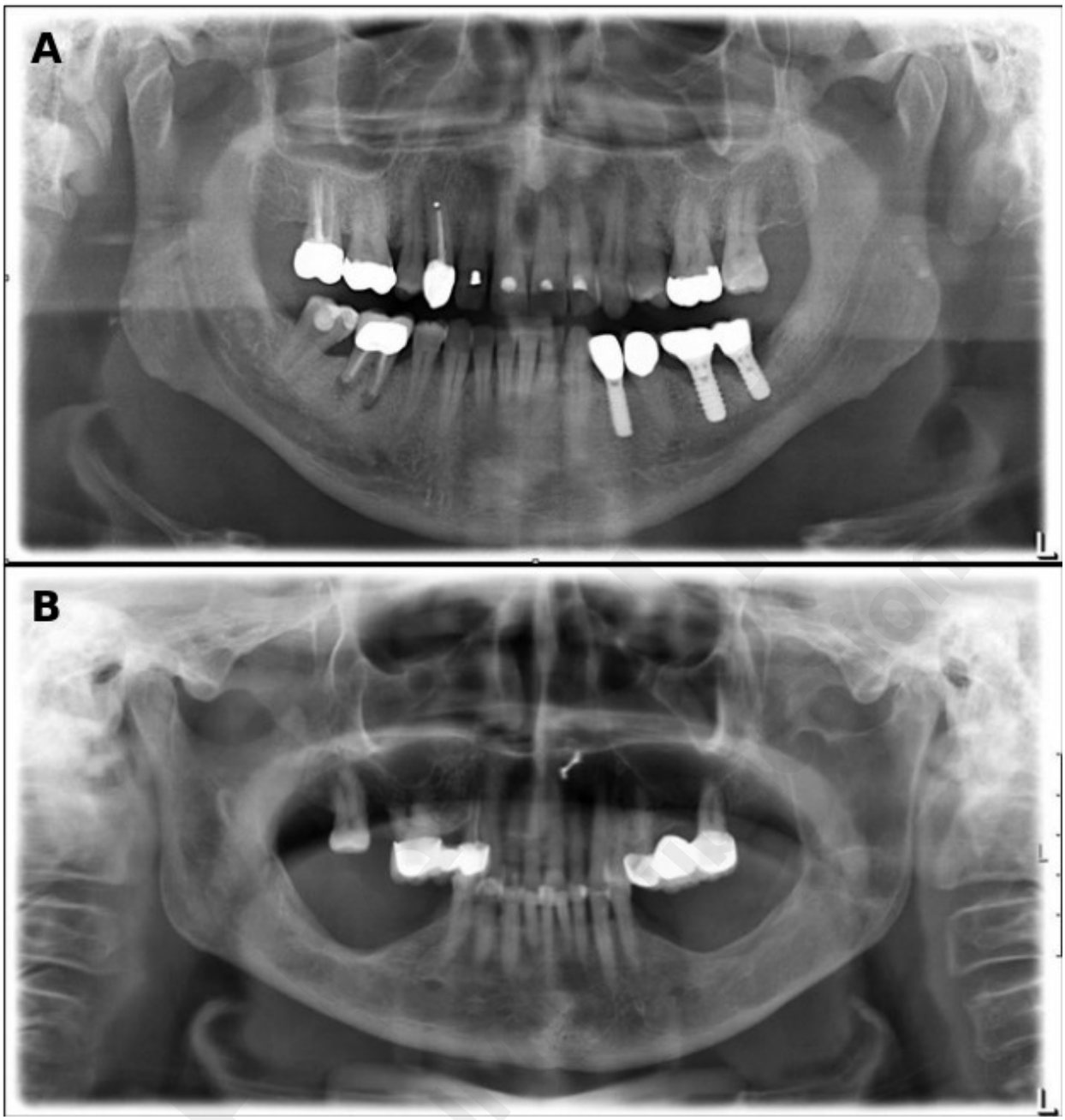

Figure 1 (A and B): Two randomly chosen panoramic radiographs featuring various pathologies to be diagnosed by dental students. Both x-ray images represented the basis of a student-written and an AI-generated radiology report.

Data transformation and AI text generation

Upon completion, the checkbox lists filled out by participants were carefully transcribed into an Excel datasheet comprising distinct spreadsheets for each category to organize the data. Subsequently, Chat GPT 4.0 (OpenAI, San Franciso, CA, USA), 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 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 test 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 5th until October 12th, 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.

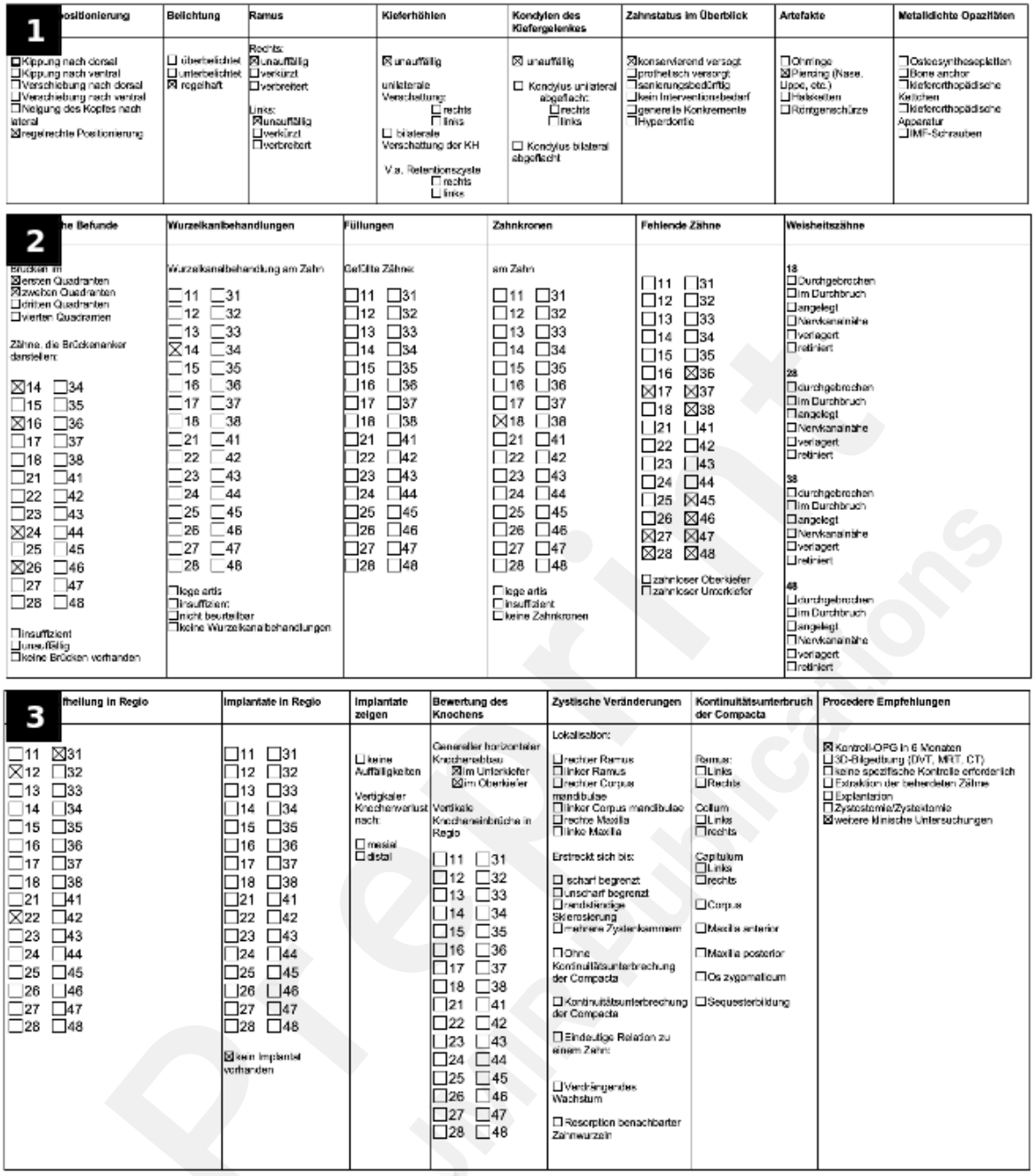

Figure 2: An example of a completed checkbox list containing three distinct spreadsheets (1, 2, and 3) used to generate radiology reports with ChatGPT.

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 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]):

$$Flesch\,Reading\,Ease = 180 - ASL - 58{,}5 * ASW$$

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: This index considers the average sentence length and the prevalence of long words with more than six letters to assess text readability by the following calculation:

$$LIX = \frac{Total\,number\,of words}{Total\,number\,of\,sentences} + \frac{(Number\,of\,long\,words * 100)}{Total\,number\,of\,words}$$

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 (BERT 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 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 (Bidirectional Encoder Representations from Transformers) 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 three primary components: precision (P), recall (R), and

F1 score (F1). 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 F1 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 higher quality of AI-generated texts. Multiple studies have already demonstrated BERT’s capability of accurately predicting readability levels for various texts [33].

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 three 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. Average sentence length and long word proportion (defined as words with more than six 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 / words * 100). These metrics were analyzed collectively for all students and semesters as one group.

Statistical Analysis

The software packages used for statistical analysis were GraphPad Prism 9.0 (GRAPHPAD SOFTWARE, LLC, Boston, USA), G*Power 3.1 (Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany), Excel 16.76 (Microsoft Corporation, Redmond, USA) and SPSS Statistics 29 (IBM Deutschland GmbH, Böblingen, Germany). 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 (34.9 ± 11.2) compared to medical information on Google webpages (46.5 ± 14.3), accounting to 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 $p < 0.05$. All data are presented as mean +/- standard deviation (SD).

Differences between student-written and AI-generated texts were analyzed using a 2-tailed Student’s t-test. A subsequent post hoc power analysis was conducted for each test to verify the power achieved by the t-test.

Results

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 utilized 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.

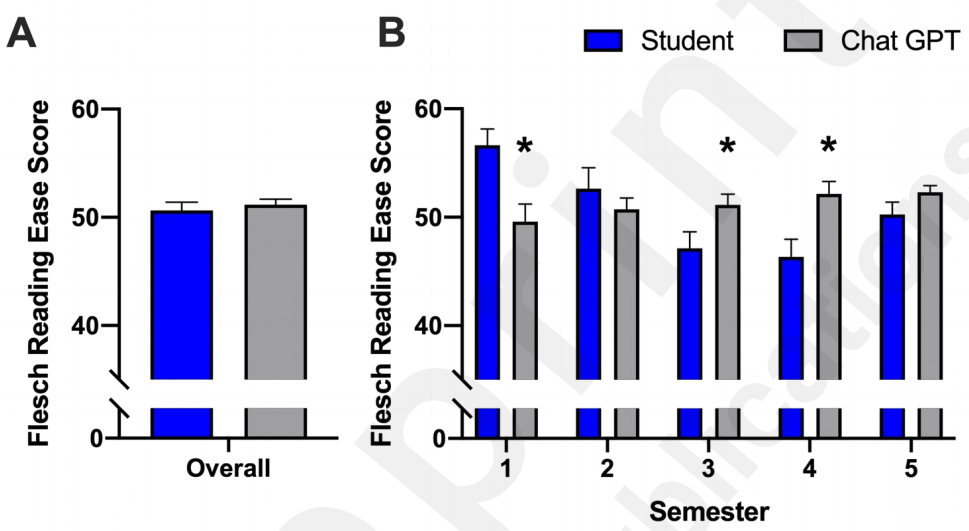

*Figure 3: Metric evaluation of readability of AI-generated reports compared to student-written radiology reports overall (A) and individually (B) for each semester assessed with the Flesch Readability Ease Score. Data represent mean +/- standard deviation. Sample size: n = 100 (A); Semester 1: n=20, Semester 2: n=19, Semester 3: n=21, Semester 4: n=20, Semester 5: n=20 (B); *p<0.05; and vs. students*

The Flesch Reading Ease Score (figure 3A and B) revealed no difference in readability between AI-generated and student-written texts (50.55 ± 7.80 vs 51.19 ± 5.02) considering all reports together as demonstrated in figure 3A (p=0.4911; t=0.6898; df=198). 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 (56.65 ± 6.70 vs 49.6 ± 7.17; p=0.002; t=3.213; df=38) and higher readability compared to the third (47.14. ± 6.97 vs 51.14 ± 4.55; p=0.033; t=2.203; df=40) and fourth (46.35. ± 7.29 vs 52.15 ± 5.08; p=0.006; t=2.918; df=38) clinical semesters. No difference was found for second (p=0.3929; t=0.8647; df=36) and fifth (p=0.123; t=1.577; df=38) clinical semester.

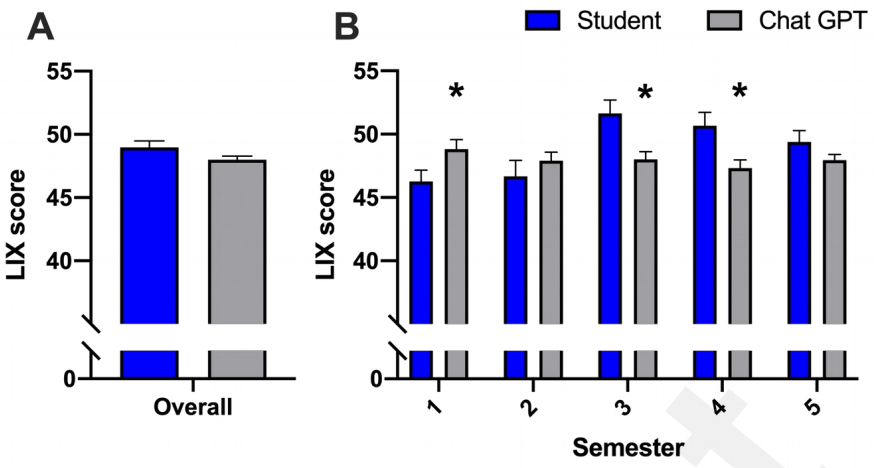

*Figure 4: Metric evaluation of readability of AI-generated reports compared to student-written radiology reports overall (A) and individually (B) for each semester assessed with LIX Index. Data represent mean +/- standard deviation. Sample size: n = 100 (A); Semester 1: n=20, Semester 2: n=19, Semester 3: n=21, Semester 4: n=20, Semester 5: n=20 (B); *p<0.05; and vs. students*

As presented in figure 4A, no overall difference between both groups regarding readability was found (48.98 ± 5.0 vs 48.0 ± 2.85) as assessed with the LIX index (p=0.091; t=1.699; df=198). In contrast to the FRE score, the LIX readability index exhibited opposing trends across semesters (figure 4B), with significant differences observed in semesters one (46.27 ± 4.0 vs 48.81 ± 3.44; p=0.037; t=2.157; df=38), three (51.64 ± 4.89 vs 48.01 ± 2.84; p=0.005; t=2.944; df=40) and four (50.67 ± 4.68 vs 47.32 ± 2.90; p=0.0098; t=2.719; df=38). No difference was observed in second (p=0.39; t=0.877; df=36) and fifth (p=0.151; t=1.464; df=38) clinical semester.

AI-generated reports show great similarity to reference texts but lack information.

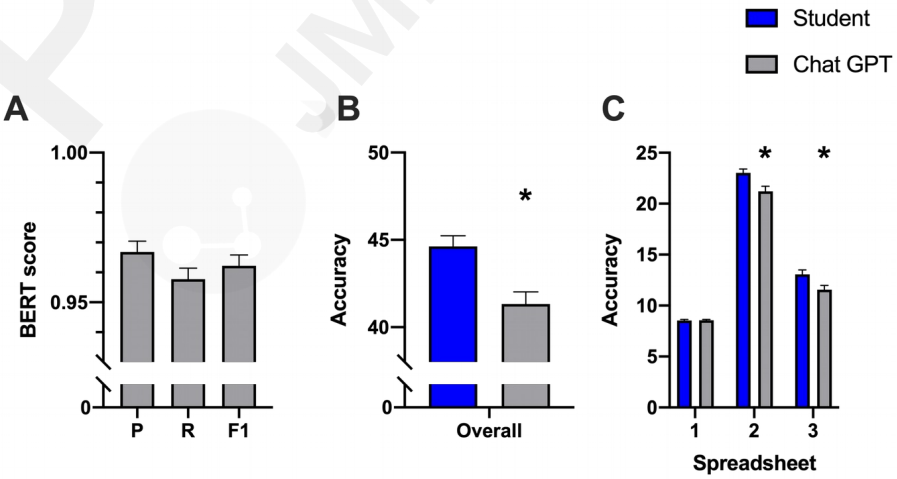

Figure 5: Evaluation of similarity compared to reference texts using the BERT score (A) with precision (P), recall (R) and F1 score (F1) representing the harmonic mean of precision and recall. The accuracy of AI-generated radiology reports was further assessed as overall accuracy including the whole checkbox list (B) and individually for each spreadsheet of the checkbox list. Data represent mean +/- standard deviation. Sample size: n = 100. *p<0.05 vs. students

As illustrated in Figure 5A, the great similarity is indicated by a high BERT Score, with precision

(P) = 0.967 ± 0.036, recall (R) = 0.958 ± 0.037, and F1 = 0.962 ± 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 44.6 ± 6.0 findings, whereas the AI reported 41.3 ± 7.0 findings in total (p=0.0004; t=3.586; df=198). Specifically, as shown in figure 5C, while no difference was observed in the first spreadsheet (8.53 ± 1.06 vs 8.55 ± 1.02; p=0.892; t=0.1361; df=198), the AI included significantly fewer findings from the second (23.03 ± 3.67 vs 21.22 ± 4.92; p=0.0035; t=2.951; df=198) and third (13.07 ± 4.40 vs 11.56 ± 4.23; p=0.0141; t=2.476; df=198) spreadsheets.

AI-generated significantly shorter and error-free radiology reports

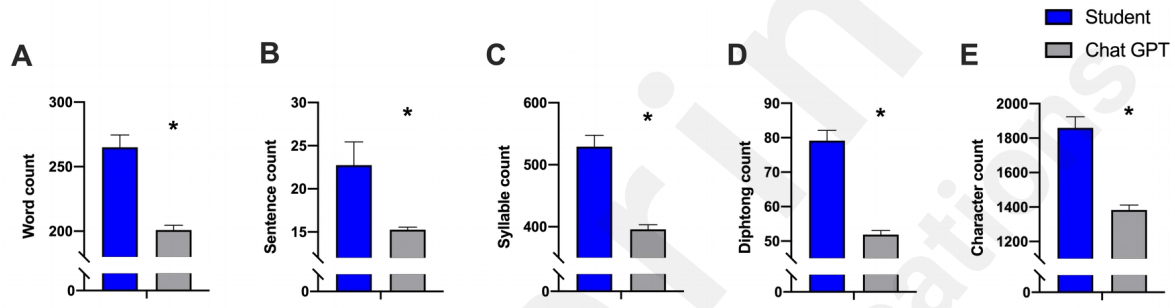

*Figure 6: Analysis of word count(A), sentence count (B), syllable count (C), diphthong count (D) and character count (E) of AI-generated radiology reports compared to student-written reports. Data represent mean +/- standard deviation. Sample size: n = 100. *p<0.05 vs. students*

AI-generated radiology reports exhibited a significant ~24% reduction in word count (265.6 ± 95.4 vs 200.6 ± 37.3 words p<0.0001; t=6.347; df=198) and sentence count (p=0.007; t=2.726; df=198) accompanied by significant reductions in syllables (p<0.0001; t=6.823; df=198), diphthongs (p<0.0001; t=8.643; df=198), and characters (p<0.0001; t=6.841; df=198) compared to student-written texts as presented in figure 6.

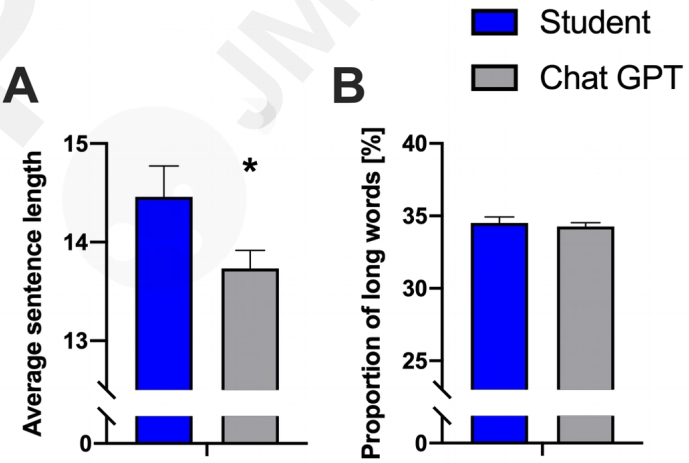

*Figure 7: Analysis of sentence length and long word proportion (more than six characters) of AI-generated radiology reports compared to student-written reports. Data represent mean +/- standard deviation. Sample size: n = 100. *p<0.05 vs. students.*

As presented in figure 7, whereas radiology reports generated by AI showed a significant reduction in average sentence length compared to student-written reports (A: 14.5 ± 3.1 vs 13.7

± 1.8 words p=0.0462; t=2.007; df=198), no difference was observed regarding the proportion of long words (B: p=0.6112; t=0.509; df=198).

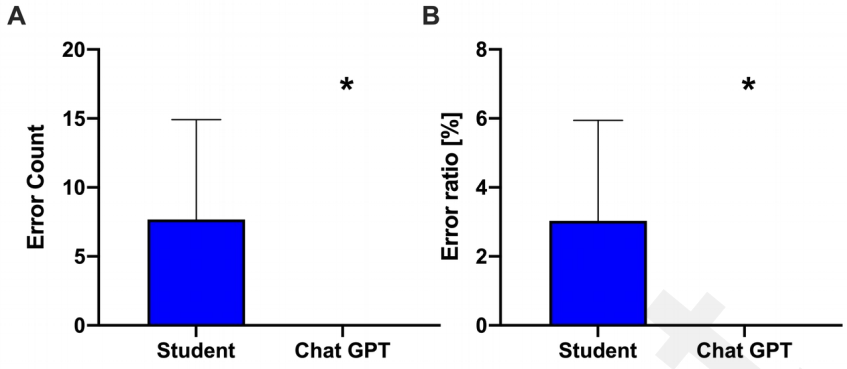

*Figure 8: Analysis of error count (A) including grammar, spelling and punctuation and error ratio (B) by calculating errors / words * 100 of student-written radiology reports and AI-generated radiology reports. Data represent mean +/- standard deviation. Sample size: n = 100. *p<0.05 vs. students.*

Contrary to student-written reports, AI-generated texts showed a complete absence of orthographic, grammatical, and punctuation errors as presented in figure 8A and B (A: 7.7 ± 7.2 vs 0; p<0.0001; t=10.59; df=198; B: 2.9 ± 2.9 vs 0; p<0.0001; t=10.41; df=198)

Student-written reports showed significant differences in length and readability across semesters

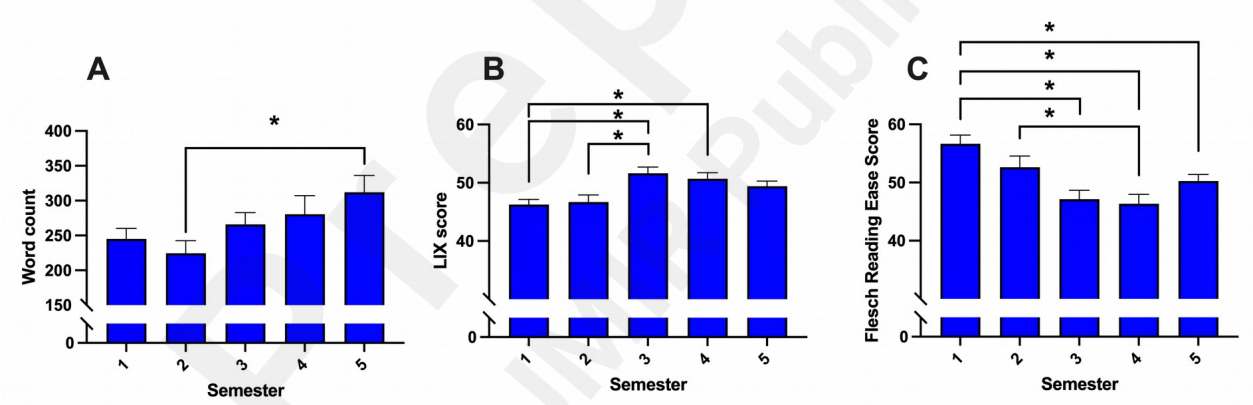

*Figure 9: Analysis of word count (A) and metric evaluation of readability using LIX score (B) and Flesch Reading Ease Score of student-written radiology reports. Data represent mean +/- standard deviation. Sample size: n = 100; Semester 1: n=20, Semester 2: n=19, Semester 3: n=21, Semester 4: n=20, Semester 5: n=20. *p<0.05.*

Student-written reports showed a significant difference in word count across different semesters (p=0.0414; t=1.610; df=99), with a noticeable trend towards the use of more words in higher semesters (Figure 9A; semester 2 vs. semester 5: 224 ± 79 vs 312 ± 107; p=0.0306, df=95; t=2,97). Additionally, significant differences in readability were observed across semesters, as presented in Figures 9B and 9C (LIX score: p=0.0006; t=2.315; df=99; Flesch Reading Ease Score: p; <0.0001 t=2.762; df=99). Students from semesters one and two produced simpler and easier-to-understand reports, whereas those from higher semesters tended to write more complex and hence more difficult-to-read reports.

Discussion

Integrating artificial intelligence (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 the 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].

Artificial intelligence 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 the present 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 F1 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 healthcare 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 (ten minutes for checkbox list completion vs. 30 minutes for report writing). Whereas all students were observed to utilize 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 post-task 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 healthcare.

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 (e.g. 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 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 towards 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 towards 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 healthcare, 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 reliability of AI-generated content [42, 43].

In the context of the presented 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 Flesch Reading Ease 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 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 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 enhances 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 healthcare settings. On the one hand, healthcare professionals require detailed reports for accurate clinical decision-making and effective inter-professional 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 fulfil both the detailed requirements of healthcare professionals and the simplified needs of patients simultaneously in response to two different prompts. Consequently, automated AI solutions could facilitate effective communication among healthcare providers and increase patient empowerment and participation beyond time-saving and more efficient documentation in health care.

Conclusion

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.

Conflicts of Interest

The authors declare no conflict of interest.

Annotation:

This study was conducted as part of Annika Bertsch’s doctoral thesis.

References

1. Ebbers, T., et al., *The Impact of Structured and Standardized Documentation on Documentation Quality; a Multicenter, Retrospective Study.* J Med Syst, 2022. **46**(7): p. 46.
2. Vemula, D., et al., *CADD, AI and ML in drug discovery: A comprehensive review.* Eur J Pharm Sci, 2023. **181**: p. 106324.
3. Weisberg, E.M. and E.K. Fishman, *The future of radiology and radiologists: AI is pivotal but not the only change afoot.* J Med Imaging Radiat Sci, 2024.
4. Lee, J.H., 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): p. 505-516.
5. Cascella, M., et al., *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* J Med Syst, 2023. **47**(1): p. 33.
6. Walker, H.L., et al., *Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument.* J Med Internet Res, 2023. **25**: p. e47479.
7. Floridi, L. and M. Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences.* Minds and Machines, 2020. **30**(4): p. 681-694.
8. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.* PLOS Digit Health, 2023. **2**(2): p. e0000198.
9. Yaneva, V., et al., *Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment.* Acad Med, 2024. **99**(2): p. 192-197.
10. Liévin, V., C.E. Hother, and O. Winther, *Can large language models reason about medical questions?* arXiv preprint arXiv:2207.08143, 2022.
11. Hwang, T., 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): p. e0297701.
12. Al-Worafi, Y.M., et al., *The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives.* JMIR Med Educ, 2024. **10**: p. e47339.
13. Dave, T., S.A. Athaluri, and S. Singh, *ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations.* Front Artif Intell, 2023. **6**: p. 1169595.
14. Flory, M.N., S. Napel, and E.B. Tsai, *AI in Radiology: Opportunities and Challenges.* Semin Ultrasound CT MR, 2024.
15. Mago, J. and M. Sharma, *The Potential Usefulness of ChatGPT in Oral and Maxillofacial Radiology.* Cureus, 2023. **15**(7): p. e42133.
16. Kelly, B.S., et al., *Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE).* Eur Radiol, 2022. **32**(11): p. 7998-8007.
17. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters.* Lancet Digit Health, 2023. **5**(4): p. e179-e181.
18. Waisberg, E., et al., *GPT-4 and Ophthalmology Operative Notes.* Ann Biomed Eng, 2023. **51**(11): p. 2353-2355.
19. Jeblick, K., et al., *ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports.* Eur Radiol, 2023.
20. Qutieshat, A., et al., *Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence.* Diagnosis, 2024.

21. Yokokawa, D., et al., *For any disease a human can imagine, ChatGPT can generate a fake report*. Diagnosis, 2024.
22. Fink, M.A., *[Large language models such as ChatGPT and GPT-4 for patient-centered care in radiology]*. Radiologie (Heidelb), 2023. **63**(9): p. 665-671.
23. Currie, G., S. Robbie, and P. Tually, *ChatGPT and Patient Information in Nuclear Medicine: GPT-3.5 Versus GPT-4*. J Nucl Med Technol, 2023. **51**(4): p. 307-313.
24. Flesch, R., *A new readability yardstick*. Journal of Applied Psychology, 1948. **32**(3): p. 221-233.
25. Amstad, T., *Wie verständlich sind unsere Zeitungen?* 1978: Studenten-Schreib-Service.
26. Gajjar, A.A., et al., *Usefulness and Accuracy of Artificial Intelligence Chatbot Responses to Patient Questions for Neurosurgical Procedures*. Neurosurgery, 2024.
27. Gajjar, A.A., et al., *Readability of cerebrovascular diseases online educational material from major cerebrovascular organizations*. J Neurointerv Surg, 2024.
28. Irwin, S.C., et al., *Ankle conFUSION: The quality and readability of information on the internet relating to ankle arthrodesis*. Surgeon, 2021. **19**(6): p. e507-e511.
29. Friedman, D.B. and L. Hoffman-Goetz, *A systematic review of readability and comprehension instruments used for print and web-based cancer information*. Health Educ Behav, 2006. **33**(3): p. 352-73.
30. Skrzypczak, T. and M. Mamak, *Assessing the Readability of Online Health Information for Colonoscopy - Analysis of Articles in 22 European Languages*. J Cancer Educ, 2023. **38**(6): p. 1865-1870.
31. Anderson, J., *Lix and Rix: Variations on a Little-known Readability Index*. Journal of Reading, 1983. **26**(6): p. 490-496.
32. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018.
33. Deutsch, T., M. Jasbi, and S. Shieber, *Linguistic features for readability assessment*. arXiv preprint arXiv:2006.00377, 2020.
34. Bellinger, J.R., et al., *BPPV Information on Google Versus AI (ChatGPT)*. Otolaryngol Head Neck Surg, 2023.
35. Sikander, B., et al., *ChatGPT-4 and Human Researchers Are Equal in Writing Scientific Introduction Sections: A Blinded, Randomized, Non-inferiority Controlled Study*. Cureus, 2023. **15**(11): p. e49019.
36. Pham, C., et al., *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**: p. e55037.
37. Atkinson, C.J., 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).
38. Huang, L., et al., *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. arXiv preprint arXiv:2311.05232, 2023.
39. Nazary, F., Y. Deldjoo, and T. Di Noia. *ChatGPT-HealthPrompt. Harnessing the Power of XAI in Prompt-Based Healthcare Decision Support using ChatGPT*. 2024. Cham: Springer Nature Switzerland.
40. White, J., et al., *Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design*. arXiv preprint arXiv:2303.07839, 2023.
41. Hu, X., et al., *Opportunities and challenges of ChatGPT for design knowledge management*. arXiv preprint arXiv:2304.02796, 2023.

42. De Angelis, L., et al., *ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health*. Front Public Health, 2023. **11**: p. 1166120.
43. Wang, G., 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**: p. e49771.
44. Agarwal, N., et al., *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*
NBER Working Paper Series. Vol. w31422. 2023, Cambridge, Mass.: National Bureau of Economic Research.
45. Dell'Acqua, F., et al., *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013; The Wharton School Research Paper, 2023: p. 58.
46. Brynjolfsson, E., D. Li, and L.R. Raymond, *Generative AI at work*. 2023, National Bureau of Economic Research.