Article

Application of Conversational AI Models in Decision Making for Clinical Periodontology: Analysis and Predictive Modeling

Albert Camlet [1], Aida Kusiak [1] and Dariusz Świetlik [2,*]

[1] Department of Periodontology and Oral Mucosa Diseases, Medical University of Gdansk, Orzeszkowej 18 St., 80-208 Gdansk, Poland; acamlet@gumed.edu.pl (A.C.); akusiak@gumed.edu.pl (A.K.)

[2] Division of Biostatistics and Neural Networks, Medical University of Gdansk, Debinki 1 St., 80-211 Gdansk, Poland

* Correspondence: dariusz.swietlik@gumed.edu.pl

Abstract: (1) Background: Language represents a crucial ability of humans, enabling communication and collaboration. ChatGPT is an AI chatbot utilizing the GPT (Generative Pretrained Transformer) language model architecture, enabling the generation of human-like text. The aim of the research was to assess the effectiveness of ChatGPT-3.5 and the latest version, ChatGPT-4, in responding to questions posed within the scope of a periodontology specialization exam. (2) Methods: Two certification examinations in periodontology, available in both English and Polish, comprising 120 multiple-choice questions, each in a single-best-answer format. The questions were additionally assigned to five types in accordance with the subject covered. These exams were utilized to evaluate the performance of ChatGPT-3.5 and ChatGPT-4. Logistic regression models were used to estimate the chances of correct answers regarding the type of question, exam session, AI model, and difficulty index. (3) Results: The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the Spring 2023 session in Polish and English were 40.3% vs. 55.5% and 45.4% vs. 68.9%, respectively. The periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 for both sessions ($p < 0.05$). For the ChatGPT-4 spring session, it was significantly more effective in the English language ($p = 0.0325$) due to the lack of statistically significant differences for ChatGPT-3.5. In the case of ChatGPT-3.5 and ChatGPT-4, incorrect responses showed notably lower difficulty index values during the Spring 2023 session in English and Polish ($p < 0.05$). (4) Conclusions: ChatGPT-4 exceeded the 60% threshold and passed the examination in the Spring 2023 session in the English version. In general, ChatGPT-4 performed better than ChatGPT-3.5, achieving significantly better results in the Spring 2023 test in the Polish and English versions.

Keywords: artificial intelligence; periodontology; dental education; ChatGPT; deep learning-based language model (LLM)

check for updates

Academic Editors: Demos T. Tsahalis and Ioannis Kakkos

Received: 11 November 2024
Revised: 10 December 2024
Accepted: 17 December 2024
Published: 2 January 2025

Citation: Camlet, A.; Kusiak, A.; Świetlik, D. Application of Conversational AI Models in Decision Making for Clinical Periodontology: Analysis and Predictive Modeling. *AI* **2025**, *6*, 3. https://doi.org/10.3390/ai6010003

1. Introduction

Language has had a profound impact on hominid evolution, brain development, and cultural expansion. Spoken language outperforms physical gestures and enables humans to constitute a shared perception of reality [1]. Language is believed to be one of the most vital abilities in human beings. It enables individuals to communicate and cooperate and has an immense potential to evolve over a lifespan [2]. Natural language processing is a scientific discipline focused on enabling computers to comprehend human language effectively. NLP is applicable in machine translation, summarization, text classification, and named entity recognition [3]. Natural language processing can be divided into two major groups: natural language understanding and natural language generation [4]. Language modeling (LM) is an essential aspect of natural language processing (NLP), providing a framework to understand the contextual and statistical patterns of language. Researchers have developed language modeling as a powerful utility to describe the probability distribution of natural language [5]. The most recent development stage of LM is large language models (LLMs). LLMs are large-sized pre-trained language models (PLMs), which are more effective in solving complex tasks compared to the original PLMs [6]. The GPT (Generative Pretrained Transformer) model was the first LLM released in 2018 [7]. Its capabilities were generating human-like text and performing tasks, such as translation and summarization [8]. Concurrently, Google Research introduced BERT. One year later, the company released RoBERTa in cooperation with OpenAI and Facebook AI [9,10]. From 2018 until the time of writing this publication, there have been four versions of GPT. The GPT-1 model was trained on 117 million parameters and its learning target was unsupervised learning [11]. For the next generations of GPT, the numbers of model parameters were 1.5 billion and 175 billion, respectively, for GPT-2 and GPT-3 [12]. The number of model parameters for GPT-4 is unpublished. The learning target for each version of GPT is different, including multi-task learning for GPT-2, in-context learning for GPT-3, and multimodal learning for GPT-4 [12,13].

ChatGPT is an AI chatbot that uses the architecture of the GPT language model. Interaction with ChatGPT is carried out through prompts, which are text inputs. The principal feature of ChatGPT is that inputs and outputs reflect natural language, enabling users to conduct a dialog in a conversational way. ChatGPT is a powerful tool and its utility is being constantly discovered by companies, researchers, and millions of people worldwide. The accessibility and multilingual features of ChatGPT provide unique opportunities in education, including the medical sciences [14]. ChatGPT can support automated essay analysis, language translation, personalized education, and adaptive and interactive learning [15]. ChatGPT has the potential to change the way medical education is implemented. In a recent study, ChatGPT's performance on the United States Medical Licensing Examination (USMLE) was tested and the achieved accuracy was 64.4% on the NBME-Free-Step 1 dataset. The result could be assessed as commensurate with the performance of a third-year medical student [16]. German researchers evaluated if ChatGPT can pass state medical examinations (M1 and M2—first and second state examinations) in the native language. The percentage of correct answers for M1 was 60.01% and for M2 it was 66.7%. Based on these results, ChatGPT was given a passing grade [17]. ChatGPT can also bring educational benefits to fields of medicine that are incomprehensible for patients. For example, ChatGPT was used to simplify radiology reports. The authors found that the simplified reports were generally correct and did not present harmful content for patients [18].

AI is increasingly vital in medicine and dentistry today, offering versatile solutions across various scenarios [19–21]. Its application spans multiple medical disciplines, such as radiology, pathomorphology, oncology, cardiology, psychiatry, nuclear medicine, and more [22–24]. Exploring the intricacies of the nervous system, which are inaccessible through conventional research methods, can be achieved through computer models of neural networks [24–26]. These in silico methods have gained extensive use, particularly in cancer, autoimmune diseases, and neurodegenerative conditions, aiding in the discovery of potential groundbreaking pharmaceutical treatments [27,28]. In dentistry, ChatGPT could potentially be a source of knowledge for patients and professionals. Alsayed et al. reviewed the quality of ChatGPT responses in the fields of oral surgery, oral pathology, and preventive dentistry. While ChatGPT provided reliable information in preventive dentistry, the outcome in the remaining topics was recognized to be less accurate. Researchers suggested that medical advice retrieved from ChatGPT needs to be monitored by medical authorities [29]. After the ChatGPT release in 2022, the number of AI-assisted dental research papers significantly increased. ChatGPT acts effectively as an assistant in text processing, data summarization, and searching scientific information, but it lacks critical analysis [30]. ChatGPT can enhance scientific performance but can also lead to bias consolidation [31].

Regarding the elevated level of interest in assessing ChatGPT's capabilities and limitations, there are still branches of medicine that are insufficiently studied. One of them is periodontology. Periodontology is a branch of dentistry focused on periodontal disease (periodontitis), which affects approximately 3.5 billion people worldwide [32]. The common aspect of the disease makes it a vital social issue. ChatGPT is a natural candidate to be one of the pillars of raising awareness of oral health and oral therapy. ChatGPT is applicable in answering frequently asked questions related to periodontal disease. Alan et al. rated that ChatGPT responses about periodontitis were "good" (according to the DISCERN instrument); however, the performance in "treatment options" was significantly lower. ChatGPT might be a substantial part of patient education; nevertheless, it cannot manage in more sophisticated cases [33]. ChatGPT may be a useful option for clinicians to accelerate the classification of periodontitis in patients. According to Eroglu et al., ChatGPT correctly determined the parameters of periodontitis (stage, grade, and extent) in 59.5%, 50.5%, and 84.0% of cases, respectively [34].

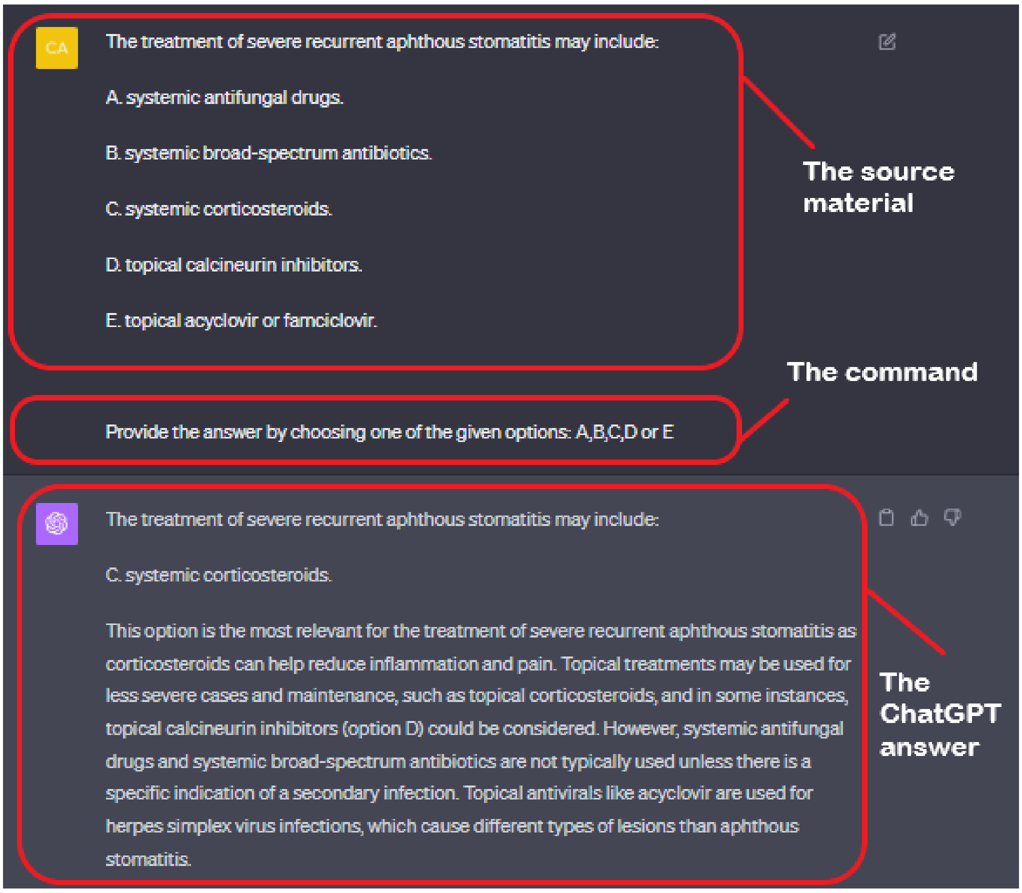
The Specialty Certificate Examination in Periodontology combines topics from various medical fields. The examination covers scientific, epidemiological, as well as strictly clinical issues. The aim of this paper is to evaluate the effectiveness of ChatGPT-3.5 and the latest version, ChatGPT-4, in answering questions derived from the Specialty Certificate Examination in Periodontology. The results will help to identify the usefulness of ChatGPT in supporting postgraduate education at the specialist level as well as verify the ability to process complex multiple-choice questions.

2. Materials and Methods

2.1. Research Strategy and Questions

Polish dentists aspiring to obtain a specialist degree in periodontology are obliged to pass the Specialty Certificate Examination (SCE). The educational institution in charge of conducting medical exams is the Center for Medical Examinations in Poland (CEM). Since the spring of 2023, CEM has started publishing the questions and answers 7 days after administering the test. We selected the two latest examinations performed in the spring and autumn of 2023. Each test consisted of 120 questions, among which there was one question involving graphic content. The research was developed through collaborative dialog with ChatGPT-3.5 Dec 15 Version (OpenAI Inc., San Francisco, CA, USA) and AI ChatGPT-4 March 14 version (OpenAI Inc., San Francisco, CA, USA). We excluded the graphic-related question from each test due to ChatGPT-3.5 limitations. Questions were single-best-answer, multiple-choice type and provided with five possible answers. Both tests were manually translated into English by a medical professional with no automated solutions. The translation was peer-reviewed by a specialist in periodontology. The medical terminology used in the translation was checked for consistency with the available literature. The questions were divided into five types according to the topics covered: periodontitis (type 1), regenerative and resective surgery (type 2), muco-gingival surgery (type 3), clinical pharmacology (type 4), and oral mucosa disorders (type 5). ChatGPT-3.5 and ChatGPT-4 were asked both Polish and English versions of the tests. We performed the study from 19 October 2023 to 7 November 2023. Considering the timeframe of the research, the knowledge cutoffs for ChatGPT-3.5 and ChatGPT-4 were September 2021 and April 2023, respectively.

The general research strategy was to ask ChatGPT questions directly exported from the public CEM database in their original form. Every question was entered into a separate dialog box, the response was acquired, documented (including the source question), and the conversation was deleted. Implementation of the above procedure reduced the risk of internal, unrecognizable data interaction. In the case of a correct or false answer (valid answer), the outcome was saved in the dedicated file. However, when equivocal answers (invalid; more than one option or no option selected) were produced, we conducted a second attempt involving the original query modification. The special standardized command was added beneath the source text: 'Provide the answer by choosing one of the given options, A, B, C, D, or E' to enforce the valid response. An example of second attempt input and output is shown in Figure 1.



The treatment of severe recurrent aphthous stomatitis may include:

A. systemic antifungal drugs.

B. systemic broad-spectrum antibiotics.

C. systemic corticosteroids.

D. topical calcineurin inhibitors.

E. topical acyclovir or famciclovir.

Provide the answer by choosing one of the given options: A,B,C,D or E

The treatment of severe recurrent aphthous stomatitis may include:

C. systemic corticosteroids.

This option is the most relevant for the treatment of severe recurrent aphthous stomatitis as corticosteroids can help reduce inflammation and pain. Topical treatments may be used for less severe cases and maintenance, such as topical corticosteroids, and in some instances, topical calcineurin inhibitors (option D) could be considered. However, systemic antifungal drugs and systemic broad-spectrum antibiotics are not typically used unless there is a specific indication of a secondary infection. Topical antivirals like acyclovir are used for herpes simplex virus infections, which cause different types of lesions than aphthous stomatitis.

Figure 1. Example of second attempt input and output.

The second attempt outputs were obtained, documented, and recognized as final answers (correct or false), as shown in Figure 2. After completing the process of obtaining questions and answers, the collected material was adequately assigned to 4 groups: Spring_ChatGPT_3.5, Spring_ChatGPT_4.0, Autumn_ChatGPT_3.5, and Autumn_ChatGPT_4.0. Each group included data (version Polish and English) in the form of screenshots arranged in a specific order: the original question (on a white background, horizontal page layout) → first attempt (on a black background, vertical page layout) → second attempt (under certain circumstances; on a black background, vertical page layout), as shown in the Supplementary Materials.

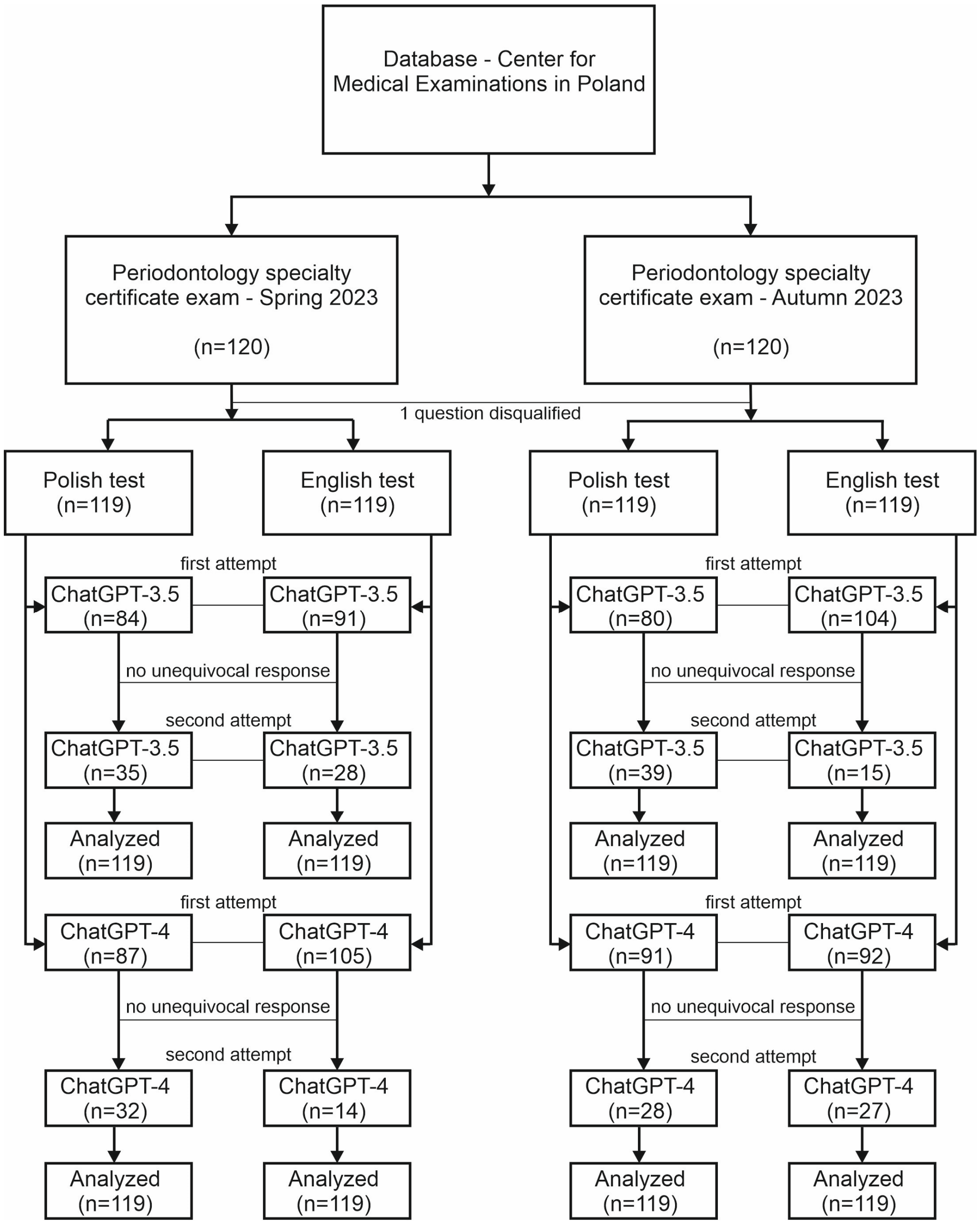


Figure 2. Flow diagram of question selection and exclusion—the overview of the research approach.

2.2. Statistical Analysis

All statistical calculations were performed using the statistical package Statistica (data analysis software system) version 13 by TIBCO Software Inc.; Palo Alto, CA, USA (2017) (http://statistica.com), accessed on 27 November 2023. Chi-square tests were applied to examine relationships between qualitative variables, while the Mann–Whitney test was employed to compare distributions of quantitative variables. Cohen's Kappa coefficient was used to assess the agreement of incorrect responses obtained in the Specialty Certificate Examination in Periodontology between humans and ChatGPT. The coefficient ranges from −1 to 1, with 1 indicating perfect agreement, 0 indicating agreement equivalent to chance, and negative values suggesting agreement worse than chance. In the statistical analysis, the item difficulty index (DI) was utilized, as follows:

$$\mathrm{DI} = (\mathrm{A} + \mathrm{B})/2\mathrm{n}, \quad (1)$$

where n represents the number of examinees in each of the extreme groups (extreme groups comprise the top 27% of performers and the bottom 27% of performers in the entire test), A is the number of correct responses to the analyzed task in the top-performing group, and B is the number of correct responses to the analyzed task in the bottom-performing group. In all calculations, the level of significance was set at $\alpha = 0.05$.

3. Results

3.1. ChatGPT Prompts—The First and Second Attempts

The percentages of answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 Polish session on the first attempt were 70.6% vs. 73.1%. There were no statistically significant differences between ChatGPT versions on the first attempt (p = 0.6655). Similar results were obtained for the autumn 2023 session, respectively, 67.2% vs. 76.5%. There were no statistically significant differences between ChatGPT versions (p = 0.1129). Whereas the percentages of answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 English session were 76.5% vs. 88.2%. ChatGPT-4.0 answered significantly more questions compared to ChatGPT-3.5 (p = 0.0173). It was the opposite in the autumn session, where the percentages of answers obtained by ChatGPT-3.5 and ChatGPT-4 were 87.4% vs. 77.3%. ChatGPT-3.5 answered significantly more questions compared to ChatGPT-4 (p = 0.0413).

3.2. Comparison of ChatGPT-3.5 and ChatGPT-4

The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 Polish session were 40.3% vs. 55.5%. The periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 (p = 0.0195). Whereas the percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn 2023 Polish session were 37.0% vs. 46.2%. No statistically significant differences were found (p = 0.1480), see Table 1. The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 English session were 45.4% vs. 68.9%. The periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 (p = 0.0002). Whereas the percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn 2023 English session were 44.5% vs. 57.1%. No statistically significant differences were found (p = 0.0518), with the results yielding borderline statistical significance in Table 1.

Table 1. Comparison of ChatGPT-3.5 and ChatGPT-4 performance on the periodontology specialty examination test in English and Polish.

Specialty Certificate Exam	Session	ChatGPT-3.5 (n = 119)	ChatGPT-4 (n = 119)	*p*-Value *
Polish	Spring 2023	48 (40.3%)	66 (55.5%)	0.0195
	Autumn 2023	44 (37.0%)	55 (46.2%)	0.1480
English	Spring 2023	54 (45.4%)	82 (68.9%)	0.0002
	Autumn 2023	53 (44.5%)	68 (57.1%)	0.0518

* Chi-square.

The percentages of correct answers on the first and second attempts obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 Polish session were 44.1% vs. 69.0% and 31.4% vs. 18.8%. The periodontology specialty examination test accuracy of ChatGPT-4 on the first attempt was significantly better than that of ChatGPT-3.5 (p = 0.0010). It was the opposite on the second attempt, with ChatGPT-3.5 performing better than ChatGPT-4, but it was not statistically significant (p = 0.2336). The percentages of correct answers on the first and second attempts obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn

2023 Polish session were 37.5% vs. 57.1% and 35.9% vs. 10.7%. The periodontology specialty examination test accuracy of ChatGPT-4 on the first attempt was significantly better than that of ChatGPT-3.5 (p = 0.0103). It was the opposite on the second attempt, with ChatGPT-3.5 performing better than ChatGPT-4, and it was statistically significant (p = 0.0195). See Tables 2 and 3.

Table 2. Comparison of ChatGPT-3.5 and ChatGPT-4 performance on the periodontology specialty examination test in English and Polish on the first attempt.

Specialty Certificate Exam	Session	ChatGPT-3.5	ChatGPT-4	*p*-Value *
Polish	Spring 2023	n = 84 37 (44.1%)	n = 87 60 (69.0%)	0.0010
	Autumn 2023	n = 80 30 (37.5%)	n = 91 52 (57.1%)	0.0103
English	Spring 2023	n = 91 46 (50.6%)	n = 105 77 (73.3%)	0.0010
	Autumn 2023	n = 104 50 (48.1%)	n = 92 62 (67.4%)	0.0064

* Chi-square.

Table 3. Comparison of ChatGPT-3.5 and ChatGPT-4 performance on the periodontology specialty examination test in English and Polish on the second attempt.

Specialty Certificate Exam	Session	ChatGPT-3.5	ChatGPT-4	*p*-Value *
Polish	Spring 2023	n = 35 11 (31.4%)	n = 32 6 (18.8%)	0.2336
	Autumn 2023	n = 39 14 (35.9%)	n = 28 3 (10.7%)	0.0195
English	Spring 2023	n = 28 8 (28.6%)	n = 14 5 (35.7%)	0.6369
	Autumn 2023	n = 15 3 (20.0%)	n = 27 6 (22.2%)	0.8664

* Chi-square.

The percentages of correct answers on the first and second attempts obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 English session were 50.6% vs. 73.3% and 28.6% vs. 35.7%. The periodontology specialty examination test accuracy of ChatGPT-4 on the first attempt was significantly better than that of ChatGPT-3.5 (p = 0.0010), and there were no statistically significant differences between ChatGPT versions on the second attempt in the first and second attempts (p = 0.6369). The percentages of correct answers on the first and second attempts obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn 2023 English session were 48.1% vs. 67.4% and 20.0% vs. 22.2%. The periodontology specialty examination test accuracy of ChatGPT-4 on the first attempt was significantly better than that of ChatGPT-3.5 (p = 0.0064), and there were no statistically significant differences between ChatGPT versions on the second attempt in the first and second attempts (p = 0.8664). See Tables 2 and 3.

3.3. Comparison of Examination Languages—Polish and English

The percentages of correct answers obtained by ChatGPT-3.5 in the spring and autumn 2023 sessions in the Polish and English versions were 40.3% vs. 45.4% and 37.0% vs. 44.5%. There were no statistically significant differences between the language used ($p > 0.05$). The percentages of correct answers obtained by ChatGPT-4 in the spring and autumn 2023 sessions in the Polish and English versions were 55.5% vs. 68.9% and 46.2% vs. 57.1%.

For the ChatGPT-4 spring session, it was significantly more effective in the English language ($p = 0.0325$). There were no statistically significant differences between the language used for the autumn 2023 session ($p = 0.0918$) Table 4.

Table 4. Performance of ChatGPT-3.5 and ChatGPT-4 depending on the language used.

ChatGPT	Session	Polish (n = 119)	English (n = 119)	*p*-Value *
ChatGPT-3.5	Spring 2023	48 (40.3%)	54 (45.4%)	0.4319
	Autumn 2023	44 (37.0%)	53 (44.5%)	0.2351
ChatGPT-4	Spring 2023	66 (55.5%)	82 (68.9%)	0.0325
	Autumn 2023	55 (46.2%)	68 (57.1%)	0.0918

* Chi-square.

3.4. Performance of ChatGPT-3.5 and ChatGPT-4 Based on the Difficulty Index of Questions

The median (Q1–Q3) values of the difficulty index of questions in the spring 2023 Polish test for incorrect and correct answers obtained by ChatGPT-3.5 were 0.75 (0.75–1.00) vs. 1.00 (0.75–1.00), and for ChatGPT-4 they were 0.75 (0.50–1.00) vs. 1.00 (0.75–1.00). Incorrect answers were characterized by significantly lower values of the difficulty index for both versions of ChatGPT ($p < 0.05$). Similar results were obtained for ChatGPT-4 in the Autumn 2023 Polish test ($p < 0.05$). However, statistical significance was not observed for ChatGPT-3.5 in the Autumn 2023 Polish test ($p > 0.05$). See Table 5.

Table 5. Performance of ChatGPT-3.5 and ChatGPT-4 based on the difficulty index of questions in Polish test.

Group	Incorrect Answers	Correct Answers	*p*-Value *
ChatGPT-3.5 Spring 2023	n = 71	n = 48	
mean (SD)	0.79 (0.22)	0.86 (0.21)	
range (min–max)	0.25–1.00	0.25–1.00	
median (Q1–Q3)	0.75 (0.75–1.00)	1.00 (0.75–1.00)	0.0477
ChatGPT-4 Spring 2023	n = 53	n = 66	
mean (SD)	0.75 (0.25)	0.87 (0.17)	
range (min–max)	0.25–1.00	0.50–1.00	
median (Q1–Q3)	0.75 (0.50–1.00)	1.00 (0.75–1.00)	0.0125
ChatGPT-3.5 Autumn 2023	n = 75	n = 44	
mean (SD)	0.73 (0.27)	0.76 (0.29)	
range (min–max)	0.00–1.00	0.00–1.00	
median (Q1–Q3)	0.83 (0.50–1.00)	0.83 (0.67–1.00)	0.5308
ChatGPT-4 Autumn 2023	n = 64	n = 55	
mean (SD)	0.71 (0.27)	0.79 (0.28)	
range (min–max)	0.00–1.00	0.00–1.00	
median (Q1–Q3)	0.83 (0.50–1.00)	0.83 (0.67–1.00)	0.0414

* U Mann–Whitney, SD—Standard Deviation, Q1—Lower quartile, Q3—Upper quartile.

For sessions in English, incorrect answers were characterized by significantly lower values of the difficulty index in the Spring 2023 and Autumn 2023 sessions for both versions of ChatGPT except for ChatGPT-3.5 in the Autumn 2023 session, where statistical significance was not observed. See Table 6.

Table 6. Performance of ChatGPT-3.5 and ChatGPT-4 based on the difficulty index of questions in the English test.

Group	Incorrect Answers	Correct Answers	*p*-Value *
ChatGPT-3.5 Spring 2023	n = 65	n = 54	
mean (SD)	0.78 (0.24)	0.87 (0.17)	
range (min–max)	0.25–1.00	0.50–1.00	
median (Q1–Q3)	0.75 (0.75–1.00)	1.00 (0.75–1.00)	0.0487
ChatGPT-4 Spring 2023	n = 37	n = 82	
mean (SD)	0.72 (0.23)	0.86 (0.20)	
range (min–max)	0.25–1.00	0.25–1.00	
median (Q1–Q3)	0.75 (0.50–1.00)	1.00 (0.75–1.00)	0.0009
ChatGPT-3.5 Autumn 2023	n = 66	n = 53	
mean (SD)	0.73 (0.27)	0.76 (0.29)	
range (min–max)	0.00–1.00	0.00–1.00	
median (Q1–Q3)	0.83 (0.50–1.00)	0.83 (0.67–1.00)	0.2867
ChatGPT-4 Autumn 2023	n = 51	n = 68	
mean (SD)	0.68 (0.29)	0.80 (0.25)	
range (min–max)	0.00–1.00	0.00–1.00	
median (Q1–Q3)	0.67 (0.50–1.00)	0.83 (0.67–1.00)	0.0213

* U Mann–Whitney, SD—Standard Deviation, Q1—Lower quartile, Q3—Upper quartile.

3.5. Assessment of the Agreement of Incorrect Responses in the Periodontology Specialty Certificate Examination for Dentists and ChatGPT-3.5/ChatGPT-4

The Cohen’s Kappa coefficients were 0.38 for ChatGPT-3.5 and 0.41 for ChatGPT-4, which can be interpreted as minimal agreement in the Polish Spring 2023 session. The confidence intervals were overlapping, and the difference in effect estimates between the two Cohen’s Kappa coefficients was statistically insignificant. We obtained similar results for the remaining configurations of ChatGPT versions, exam sessions, and languages. See Table 7.

Table 7. Cohen’s Kappa coefficients for incorrect responses of dentists and ChatGPT-3.5/ChatGPT-4.

Specialty Certificate Exams	Session	ChatGPT-3.5 Cohen’s Kappa [95% CI *]	ChatGPT-4 Cohen’s Kappa [95% CI]
Polish	Spring 2023	0.38 [0.24, 0.52]	0.41 [0.22, 0.60]
	Autumn 2023	0.55 [0.41, 0.70]	0.51 [0.33, 0.69]
English	Spring 2023	0.53 [0.38, 0.68]	0.51 [0.30, 0.72]
	Autumn 2023	0.49 [0.34, 0.64]	0.38 [0.19, 0.58]

* CI—Confidence interval.

3.6. Comparison of Examination Languages, Polish and English, by Question Type

The percentages of correct answers obtained by ChatGPT-3.5 in the spring 2023 sessions in the Polish and English versions by question type 1 were 44.2% vs. 48.8%; for type 2 they were 6.7% vs. 6.7%; for type 3 they were 50.0% vs. 66.7%; for type 4 they were 38.5% vs. 61.5%; and for type 5 they were 47.6% vs. 47.6%. There were no statistically significant differences between the language used ($p > 0.05$). The percentages of correct answers obtained by ChatGPT-3.5 in the autumn 2023 sessions in the Polish and English versions by question type 1 were 37.3% vs. 39.2%; for type 2 they were 14.3% vs. 14.3%; for type 3 they were 14.3% vs. 28.6%; for type 4 they were 75.0% vs. 75.0%; and for type 5 they were 44.2% vs. 60.5%. There were no statistically significant differences between the language used ($p > 0.05$). The percentages of correct answers obtained by ChatGPT-4 in the spring 2023 sessions in the Polish and English versions by question type 1 were 53.5% vs. 69.8%; for type 2 they were 33.3% vs. 53.3%; for type 3 they were 50.0% vs. 50.0%; for type 4 they were 84.6% vs. 84.6%; and for type 5 they were 57.1% vs. 71.4%. There were no statistically significant differences between the language used ($p > 0.05$). The percentages of correct answers obtained by ChatGPT-4 in the autumn 2023 sessions in the Polish and English versions by question type 1 were 39.2% vs. 49.0%; for type 2 they were 21.4% vs. 42.9%; for type 3 they were 14.3% vs. 14.3%; for type 4 they were 100.0% vs. 75.0%; and for type 5 they were 62.8% vs. 76.7%. There were no statistically significant differences between the language used ($p > 0.05$). See Figure 3.

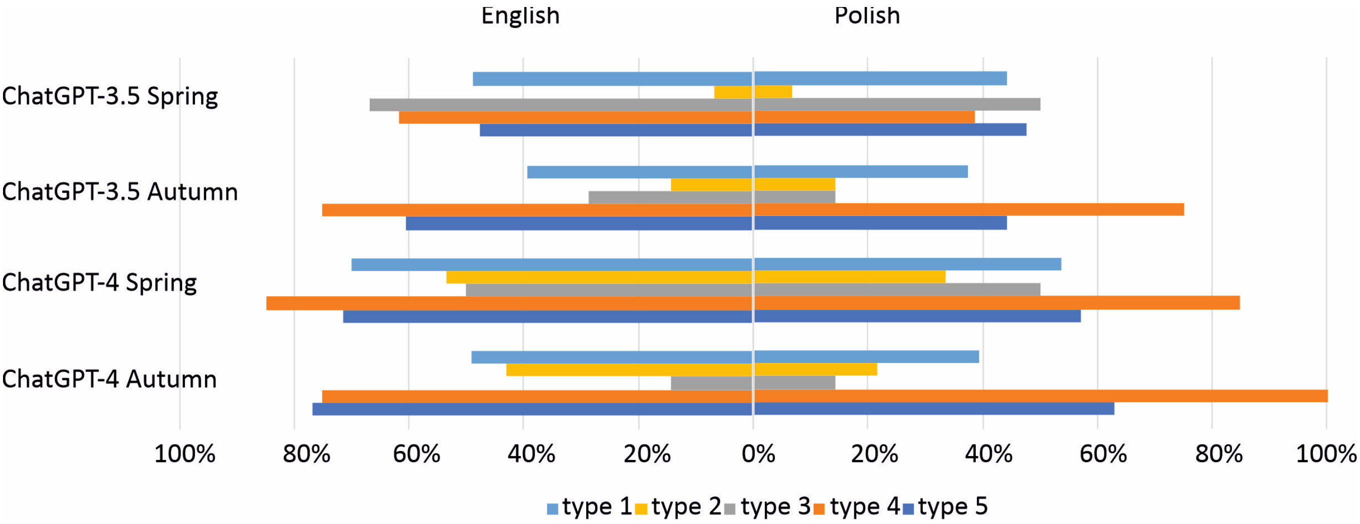


Figure 3. Performance of ChatGPT-3.5 and ChatGPT-4 depending on the language used by question type.

3.7. Comparison of ChatGPT-3.5 and ChatGPT-4 by Question Type

The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 Polish session by question type 1 were 44.2% vs. 53.5%; for type 2 they were 6.7% vs. 33.3%; for type 3 they were 50.0% vs. 50.0%; for type 4 they were 38.5% vs. 84.6%; and for type 5 they were 47.6% vs. 57.1%. No statistically significant differences were found ($p > 0.05$) except for question type 4, where the periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 ($p = 0.0156$). The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn 2023 Polish session by question type 1 were 37.3% vs. 39.2%; for type 2 they were 14.3% vs. 21.4%; for type 3 they were 14.3% vs. 14.3%; for type 4 they were 75.0% vs. 100.0%; and for type 5 they were 44.2% vs. 62.8%. There were no statistically significant differences between type of ChatGPT ($p > 0.05$). The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 English session by question type 1 were 48.8% vs. 69.8%, and the periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 ($p = 0.0482$). The percentages of correct answers by question type 2 were 6.7% vs. 53.3%, and the periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5 ($p = 0.0053$). The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the spring 2023 English session by question type 3 were 66.7% vs. 50.0%, and for type 4 they were 61.5% vs. 84.6%; no statistically significant differences were found ($p > 0.05$). The percentages of correct answers by question type 5 were 47.6% vs. 71.4%, and the periodontology specialty examination test accuracy of ChatGPT-4 was significantly better than that of ChatGPT-3.5; $p = 0.0262$). The percentages of correct answers obtained by ChatGPT-3.5 and ChatGPT-4 in the autumn 2023 English session by question type 1 were 39.2% vs. 49.0%; for type 2 they were 14.3% vs. 42.9%; for type 3 they were 28.6% vs. 14.3%; for type 4 they were 75.0%

vs. 75.0%; and for type 5 they were 60.5% vs. 76.7%. No statistically significant differences were found ($p > 0.05$). See Figure 4.

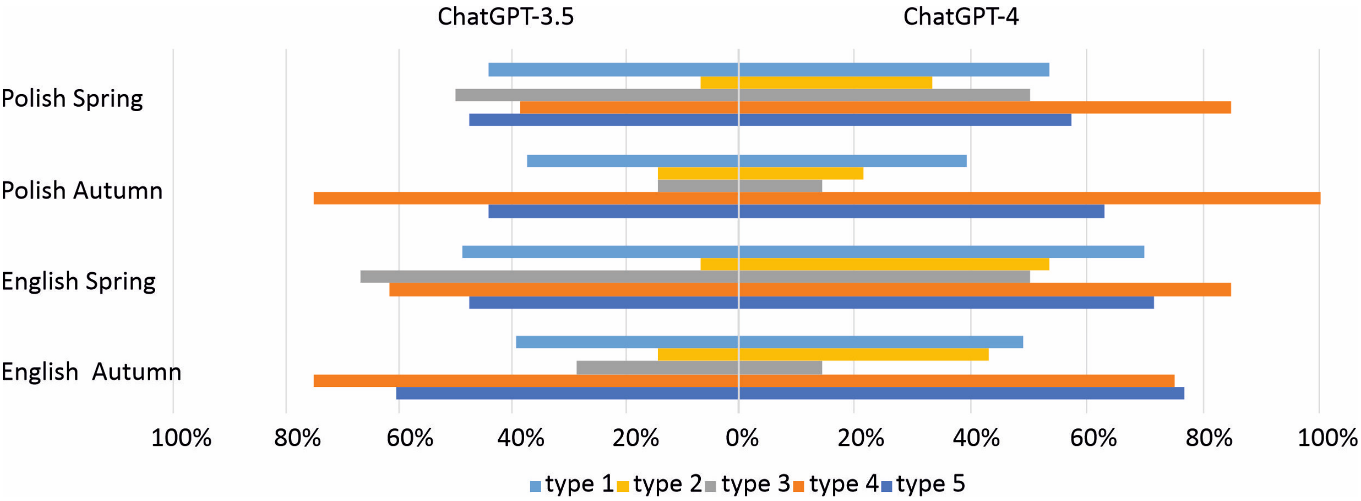


Figure 4. Performance of ChatGPT-3.5 and ChatGPT-4 depending on question type.

3.8. Application of Logistic Regression in Performance of Conversational AI Models in Decision Making for Clinical Periodontology

The logistic regression models (univariate and multivariate) derived from the results are presented in Table 8. A negative relationship existed between type 2 questions, AI Model ChatGPT-3.5, and the chances of correct answers, while a positive relationship existed between type 4 and 5 questions, AI Model ChatGPT-4, difficulty index, and the chances of correct answers for the univariate logistic regression model. Whereas in the multivariate logistic regression model, a negative relationship existed between type 2 questions, AI Model ChatGPT-3.5, and the chances of correct answers, while a positive relationship existed between type 4 and 5 questions, AI Model ChatGPT-4, difficulty index, and the chances of correct answers. Logistic regression used the following equation for the Polish version:

$$\text{correct answer} = -1.69 + (-1.24) \times \text{type 2} + 1.09 \times \text{type 4} + 0.44 \times \text{type 5} + (-0.54) \times \text{chatGPT-3.5} + 0.54 \times \text{chatGPT-4} + 1.40 \times \text{difficulty index}, \quad (2)$$

Table 8. Results of logistic regression with odds ratio of independent variable (Polish version).

Variable	Univariate		Multivariate	
	OR (95% CI)	*p*-Value	OR (95% CI)	*p*-Value
Type of question				
1	0.90 (0.62–1.30)	0.5556		
2	0.25 (0.13–0.50)	0.0001	0.29 (0.14–0.60)	0.0008
3	0.53 (0.23–1.25)	0.1461		
4	2.77 (1.32–5.83)	0.0071	2.98 (1.35–6.58)	0.0068
5	1.67 (1.15–2.44)	0.0076	1.55 (1.02–2.34)	0.0397
Session				
Autumn	0.78 (0.54–1.11)	0.1671		
Spring	1.29 (0.90–1.85)	0.1671		
AI Model				
ChatGPT-3.5	0.61 (0.42–0.88)	0.0077	0.58 (0.40–0.86)	0.0057
ChatGPT-4	1.64 (1.14–2.36)	0.0077	1.71 (1.17–2.51)	0.0057
Difficulty index	3.57 (1.66–7.69)	0.0011	4.07 (1.86–8.92)	0.0005

Based on the first regression coefficient (Equation (2)), we expected the log of the chances of correct answers to go down by 1.24 for type 2questions and to go up by 1.09 for type 4 questions. We expected the log of the chances of correct answers to go up by 0.44 for type 5 questions and to go down by 0.54 for the type of AI Model—ChatGPT-3.5. Whereas we expected the log of the chances of correct answers to go up by 0.54 for the type of AI Model—ChatGPT-4 and by 1.40 for every unit increase in difficulty index.

The logistic regression models (univariate and multivariate) derived from the results of the English version are presented in Table 9. A negative relationship existed between type 2 questions, AI Model ChatGPT-3.5, and the chances of correct answers, while a positive relationship existed between type 4 and 5 questions, AI Model ChatGPT-4, difficulty index, and the chances of correct answers for the univariate logistic regression model. Whereas in the multivariate logistic regression model, a negative relationship existed between type 2 questions, AI Model ChatGPT-3.5, and the chances of correct answers, while a positive relationship existed between type 4 and 5 questions, AI Model ChatGPT-4, difficulty index, and the chances of correct answers. Logistic regression used the following equation for the English version:

$$\text{correct answer} = -0.92 + (-1.03) \times \text{type } 2 + 1.09 \times \text{type } 4 + 0.22 \times \text{type } 5 + (-0.81) \times \text{chatGPT-3.5} + 0.81 \times \text{chatGPT-4} + 1.72 \times \text{difficulty index}, \quad (3)$$

Table 9. Results of logistic regression with odds ratio of independent variable (English version).

Variable	Univariate		Multivariate	
	OR (95% CI)	***p*-Value**	**OR (95% CI)**	***p*-Value**
Type of question				
1	0.82 (0.57–1.19)	0.3007		
2	0.31 (0.17–0.56)	0.0001	0.36 (0.19–0.68)	0.0018
3	0.51 (0.23–1.16)	0.1077		
4	2.51 (1.15–5.51)	0.0212	2.97 (1.27–6.92)	0.0118
5	1.91 (0.30–2.80)	0.0010	1.82 (1.19–2.80)	0.0063
Session				
Autumn	0.78 (0.54–1.11)	0.1680		
Spring	1.29 (0.90–1.85)	0.1680		
AI Model				
ChatGPT-3.5	0.48 (0.33–0.69)	0.0001	0.44 (0.30–0.65)	<0.0001
ChatGPT-4	2.09 (1.45–3.01)	0.0001	1.82 (1.19–2.80)	<0.0001
Difficulty index	4.59 (2.16–9.76)	0.0001	5.58 (2.53–12.27)	<0.0001

Based on the first regression coefficient (Equation (2)), expected the log of the chances of correct answers to go down by 1.03 for type 2 questions and to go up by 1.09 for type 4 questions. We expected the log of the chances of correct answers to go up by 0.22 for type 5 questions and to go down by 0.81 for the type of AI Model—chatGPT-3.5. Whereas we expected the log of the chances of correct answers to go up by 0.81 for the type of AI Model—ChatGPT-4 and by 1.72 for every unit increase in difficulty index.

4. Discussion

In our study, we found that ChatGPT-3.5 and ChatGPT-4 sometimes provided equivocal answers (invalid-inconsistent with the principle of selecting a single answer). The two-step method of obtaining answers was introduced: the first attempt and the second attempt. The second attempt was conducted with a specifically modified question and was focused on the group of invalid outputs. We evaluated the percentage of questions that were valid (the rate of questions successfully obtained in the first—and so the final—attempt). In the Polish version of the Spring 2023 and Autumn 2023 sessions, no statistical significance was observed between ChatGPT editions. In the English version, however, ChatGPT-4 had a significantly higher rate of obtained answers in the Spring 2023 session, while ChatGPT-3.5 answered significantly more queries in the Autumn 2023 session. The diverse number of received answers in the English version among ChatGPT-3.5 and ChatGPT-4 could be related to their different architectures, language proficiency, and knowledge base.

The passing threshold for the Specialty Certificate Examination in Periodontology is 60% of correct answers. ChatGPT-4 scored 68.9% of correct answers in the English version of the Spring 2023 test; thus, it passed the examination. In any other case, the prerequisite for passing was not obtained. Considering the ChatGPT edition, the percentage of correct answers obtained by ChatGPT-4 was higher than that obtained by ChatGPT-3.5 in every session—Spring 2023 and Autumn 2023, both in the Polish and English versions. However, a statistically significant difference was observed only in the Spring 2023 session, in both the Polish and English versions. Considering the type of question, ChatGPT-4 performed significantly better than ChatGPT-3.5 for question types 1, 2, and 5 (Spring 2023, English version) and question type 4 (Spring 2023, Polish version). There were no significant differences for the Autumn 2023 session, both in the English and Polish versions. Based on these results, we can conclude that ChatGPT-4 performs slightly better in periodontology exams than ChatGPT-3.5 not only in English but also in the Polish language. Similar trends, but substantially greater, have been observed by Polish scientists in nephrology specialty exams. ChatGPT-3.5 and ChatGPT-4 were asked 1560 single-answer questions. ChatGPT-4.0 reached 69.5% of correct answers, which was equivalent to passing 11 out of 13 tests. In addition, ChatGPT-4.0 has performed better than humans on average. The accuracy of ChatGPT-3.5 was significantly lower and did not exceed 45.7% of correct answers [35]. Other studies showed that ChatGPT-4 could noticeably outperform its older edition in other fields of medicine. Ali et al. demonstrated the capabilities of ChatGPT 4.0 in the American Board of Neurological Surgery Self-Assessment Examination 1 (500 questions in single-best-answer, multiple-choice format) compared to those of ChatGPT-3.5 and question bank users. The results were 83.4%, 73.4%, and 72.8%, respectively. ChatGPT-4 reported the highest rate of correct answers and its performance was significantly better than the others [36]. Lahat et al. compared the efficiency of ChatGPT-4 and ChatGPT-3.5 in responding to clinical questions in emergency medicine, internal medicine, and ethics. The AI outputs were evaluated by eight medical professionals regarding accuracy, clarity, utility, comprehensiveness, and relevance. ChatGPT-4 performed better than ChatGPT-3.5 in each category. The highest percentage of right answers was noted for ethics. ChatGPT-4 presents a promising outlook for physicians in facilitating clinical work [37].

Regarding the correct answer subdivision in the first and the second attempts, ChatGPT-4 performance in the first attempt was significantly better than ChatGPT-3.5 performance in every session, both in the Polish and English versions. Conversely, in the second attempt, the results were more heterogenous. ChatGPT-3.5 performed better than ChatGPT-4 in the Spring 2023 and Autumn 2023 sessions in the Polish version (statistical significance only in Autumn 2023 session). ChatGPT-4 had a higher rate of correct answers in the Spring 2023 and Autumn 2023 sessions in English (no statistical significance).

Considering the language criterion, the percentage of correct answers was higher for the English language in the Spring 2023 and Autumn 2023 sessions, for both ChatGPT-3.5 and ChatGPT-4 (statistical significance only for ChatGPT-4 in the Spring 2023 session). The higher rate of correct answers obtained in the English version of the tests shows that ChatGPT cooperates more effectively in the English language, but its built-in translation capabilities potentially reduce language-related obstacles. However, among the five types of questions, there were no statistically significant differences in the rates of correct answers in English and Polish (for both ChatGPT-3.5 and ChatGPT-4 in both sessions). Khorshidi et. al., in their study, assessed the ChatGPT-4 performance in the Iranian Residency Entrance Examination. ChatGPT-4 was asked questions in Persian (native language) and in English, French, and Spanish (questions were translated by ChatGPT-4). The percentage of correct answers was 161 (out of 198) for Persian and 167, 162, and 166 for English, French, and Spanish, respectively. No statistical differences between languages were observed [38]. Another study showed a different tendency. The ChatGPT-3.5 performance in the Taiwanese Pharmacist Licensing Examination was evaluated regarding the Chinese and English language. The English test scores were greater than the Chinese test scores in all subjects, with statistical significance in clinical pharmacy, dispensing pharmacy, and therapeutics [39]. According to Ando et al., ChatGPT encountered difficulties with medical questions in Japanese. Bilingual experts appraised that outputs in English presented significantly higher quality in communication and quality [40]. Liu et al. conducted a meta-analysis of 45 studies on the efficiency of ChatGPT in medical licensing exams. Scientists discovered that ChatGPT-3.5 performed better when the questions were translated to English. This correlation was not applicable to ChatGPT-4.0 [41]. Diversified results among the studies could be the effect of specific language features, distinctive terms that vary depending on the context, or structural nuances in questions. The factors that have an impact on performance in different languages could be also ChatGPT-related—more features are added by human trainers (possibly largest representation of English-speaking users) and the constantly increasing range of publicly available data (last update to ChatGPT-4 was released in April 2023, but the next update is highly predictable).

For the Polish version of the Spring 2023 session, the difficulty index (range between 0.0 and 1.0; the higher the value, the easier the question) of incorrect answers was significantly lower than that of the correct ones for ChatGPT-3.5 and ChatGPT-4. For the Polish version of the Autumn 2023 session for ChatGPT-4, similar results were obtained (no statistical significance for ChatGPT-3.5). The observable tendencies of the difficulty index of incorrect answers for the English version of the Spring 2023 and Autumn 2023 sessions corresponded to the ones for the Polish version. Hence, these results proved that ChatGPT-3.5 and ChatGPT-4 encountered difficulties in challenging questions in periodontology, regardless of the language version. In the literature, there are more examples of ChatGPT performance correlation with difficulty levels. Lewandowski et al. tested the ChatGPT knowledge level in the Specialty Certificate Examination in Dermatology. They found that the parameter of incorrect ChatGPT-4 and ChatGPT-3.5 answers was significantly correlated with a lower difficulty index [42]. In a study in which ChatGPT performed a radiology board-style examination (including 150 multiple-choice questions), the researchers found that the performance was better on clinical management questions and lower-order questions but ChatGPT encountered difficulties with higher-order thinking questions, such as application of concepts, classification, imaging findings, and calculation [43]. Li, in his letter to the editor, suggested that ChatGPT performance was insufficient in more complex clinical issues, such as adjuvant chemotherapy regimens for stage II colon cancer [44].

We assessed the agreement of incorrect responses for dentists and for ChatGPT-3.5 and ChatGPT-4. The lowest value of Cohen's Kappa coefficient was observed for ChatGPT-3.5 in the Spring 2023 session, Polish version, and for ChatGPT-4 in the Autumn 2023 session, English version (both could be interpreted as minimal agreement). The highest value was recorded for ChatGPT-3.5 in the Autumn 2023 session, Polish version (parameter could be interpreted as moderate agreement). The data showed that physicians answered incorrectly to questions other than the ones that ChatGPT answered. Periodontology specialist education is highly standardized, based on the specialty program developed by Center for Medical Examinations in Poland (CEM). Certain scientific literature is designated as the knowledge base for physicians. These data are not necessarily available for ChatGPT to access, which could be the reason for the value of agreement of incorrect answers.

ChatGPT could be considered as a statistical engine that evaluates the patterns and extrapolates the most likely conversational responses; it needs an immense amount of data, in contrast to a human mind, which works efficiently with a small information volume [45]. It undergoes a process of constant learning from human-generated data and, because of that, its outputs are more related to human perception than to entirely rational representatives and could mirror equal cognitive biases [46]. Artificial hallucinations could occur as a combination of true and fabricated data. In fields where integrity and credibility are essential (e.g., scientific papers), the deployment of ChatGPT raises concerns [47]. However, ChatGPT's performance repeatedly demonstrates resistance to unproven claims and statements. Sallam et al. conducted a descriptive study to evaluate if ChatGPT could be a reasonable source of information in terms of COVID-19 vaccine conspiracies [48]. They found that ChatGPT declined theories about COVID-19 vaccine conspiracies and provided impartial responses about compulsory vaccination [48]. Deiana et al. suggested the AI tools may be useful in healthcare fields, but with no reliable medical support, there is a notable risk of deceptive responses [49].

In the field of periodontology, there is still an insufficient number of scientific publications regarding the use of ChatGPT in medical examinations, making it impossible to compare our results with those of other authors. Nevertheless, Babayiğit et al. evaluated the usability of ChatGPT in patient information on periodontology. They requested ChatGPT to create 10 most frequently asked questions by patients on seven issues (tooth sensitivity, gingival recessions, periodontal diseases, peri-implant diseases, halitosis, dental implants, and periodontal surgery), then the questions were submitted to ChatGPT. The responses were rated in terms of accuracy (0—Likert scale) and completeness (0—scale). The authors obtained a median accuracy score of 6 for all responses and a completeness score of 2 (the mean values for accuracy and completeness were 5.50 ± 0.23 and 2.34 ± 0.24, respectively). The highest accuracy score was observed for peri-implant diseases and the lowest was observed for gingival recession. The highest completeness score was obtained for gingival recession and the lowest was obtained for dental implants [50]. The results showed that ChatGPT has the potential to familiarize patients with complex medical issues and raise awareness of origins and therapy of periodontal disease. Li conducted a study comparing the accuracy of periodontal surgery answers given by dental students and ChatGPT. The examination consisted of 25 multiple-choice questions. Students gave correct answers to an average of 21.51 questions. For ChatGPT 3.5 and ChatGPT 4.0, the results were 14 and 20 correct questions, respectively. The results indicated that ChatGPT generally performed worse; however, the difference between students and the latest version was minor [51]. As data volumes increase in the future, ChatGPT will potentially outperform human responders.

The syntactic structure of the input can be both a limitation and an opportunity for the efficient use of ChatGPT. It was observed in the scope of 45 studies that ChatGPT-3.5 was more effective in processing short-text queries than long-text queries. In addition, ChatGPT accuracy was higher for image-based multiple-choice questions than for open-ended questions [41]. An important criterion for evaluating studies involving ChatGPT is how the prompt is constructed. Prompt engineering is a new field of AI aimed at optimizing and increasing the efficiency of LLM outputs by creating specific instructions. The quality of output can be increased by following certain recommendations, including contextualizing the topics covered, setting boundary conditions, specifying certain goals, requesting specific information, and asking to play a role [52].

Despite the indisputable usefulness of ChatGPT, it is hardly possible to predict the entire impact of LLMs in data creation, medical education, and clinical practice. Various ethical concerns are raised about AI-generated data. Gao et al. proved that scientific abstracts produced by ChatGPT were in 32% of cases classified as original by human responders. In 14% of cases, the original works were recognized as AI generated. ChatGPT formed credible content that was predominantly synthetic [53]. In the future, society and public institutions may find it difficult to distinguish between human-generated and LLM-generated content. Verification of learning outcomes among medical students may be the next challenge in the AI era. Written essays may be partially and even entirely developed by LLMs and become undetected by plagiarism software. Eggmann et al. showed that dental knowledge is mainly verified by practical classes, oral examinations, and multiple-choice exams rather than essays [54]. However, our study indicates that ChatGPT can pass a multi-choice test in a highly specialized field of dentistry. Sustained supervision in medical education and testing is advised. Broad application of ChatGPT may contribute to a reduction in employment in healthcare and research institutions, especially among poorly qualified workers [55].

5. Limitations

We found significant limitations in our study. The number of questions obtained by ChatGPT-3.5 and ChatGPT-4 equaled 238 (119 for the Spring 2023 test and 119 for the Autumn 2023 test; 2 graphic questions excluded) and could be considered small. We used the editions of ChatGPT-3.5 and ChatGPT-4 that had their last database updates in September 2021 and April 2023, respectively. In terms of continuous development of ChatGPT, it is possible that on the day of this paper publication, a new version of ChatGPT will be released. Additionally, the source material translation from Polish to English could impact the research.

6. Conclusions

There is a high probability that ChatGPT will change the way medical knowledge is acquired and will make it more accessible to all people. The main findings of this paper are:

1. ChatGPT-4 can pass the Specialty Certificate Examination in Periodontology.
2. ChatGPT-4 significantly outperforms its older version in answering periodontology-related queries.
3. ChatGPT-3.5 and ChatGPT-4 operate more efficiently in English than in Polish.
4. ChatGPT sometimes provides invalid answers.
5. ChatGPT encounters difficulties with different questions than dentists.
6. The efficiency of ChatGPT is lower for more difficult queries.

The results show how ChatGPT is evolving and shed light on new opportunities in medical education. Dental students and physicians struggling with acquiring certain knowledge may benefit from individualized courses of study. Nevertheless, regarding the higher effectiveness of ChatGPT in English, non-native speakers may be more vulnerable to inaccurate outputs. The significant rate of incorrect and invalid answers suggests that ChatGPT cannot be an independent source of medical knowledge for the medical community. ChatGPT should be seen as one of the support tools in the process of acquiring knowledge from specialized literature and from academics. In clinical practice, ChatGPT should be used as a data processing tool rather than a source of knowledge. The risk of hallucinations and fabricated information is unacceptable in healthcare, so the supervision of medical professionals is mandatory. Further development in ChatGPT's implementation in the medical sciences is needed, particularly in periodontology.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai6010003/s1, ChatGPT screen shots of the periodontology specialty certificate exam.

Author Contributions: Conceptualization, A.C., A.K. and D.Ś.; methodology, A.C. and A.K.; software, D.Ś.; formal analysis, A.C., A.K. and D.Ś.; investigation, A.C.; resources, A.C.; data curation, A.C.; writing—original draft preparation, A.C., A.K. and D.Ś.; writing—review and editing, A.C., A.K. and D.Ś., visualization, D.Ś.; supervision, A.K. and D.Ś.; project administration, A.K. and D.Ś. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chiu, C. *Language and Culture*; Nanyang Technological University: Singapore, 2011; Volume 4. [CrossRef]
2. Hauser, M.D.; Chomsky, N.; Fitch, W.T. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science* **2002**, *298*, 1569–1579. [CrossRef] [PubMed]
3. Pereira, S.C.; Mendonça, A.M.; Campilho, A.; Sousa, P.; Lopes, C.T. Automated Image Label Extraction from Radiology Reports—A Review. *Artif. Intell. Med.* **2024**, *149*, 102814. [CrossRef] [PubMed]
4. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural Language Processing: State of the Art, Current Trends and Challenges. *Multim. Tools Appl.* **2022**, *82*, 3713–3744. [CrossRef] [PubMed]
5. Zhai, C. Statistical Language Models for Information Retrieval. In Proceedings of the Human Language Technology Conference of the NAACL, New York, NY, USA, 22–27 April 2007; Hanover: Worcester, MA, USA, 2007; pp. 3–4.
6. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
7. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 14 November 2023).
8. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
11. Khosla, M.; Setty, V.; Anand, A. A Comparative Study for Unsupervised Network Representation Learning. *IEEE Trans. Knowl. Data Eng.* **2020**, *33*, 1807–1818. [CrossRef]
12. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; Tang, Y. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [CrossRef]
13. Schreiner, M. *GPT-4 Architecture, Datasets, Costs and More Leaked*; THE DECODER. 2023. Available online: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ (accessed on 19 November 2023).
14. Curtis, N. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. *Pediatr. Infect. Dis. J.* **2023**, *42*, 275. [CrossRef]
15. Baidoo-Anu, D.; Ansah, L.O. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN J.* **2023**, *7*, 52–62. [CrossRef]
16. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef] [PubMed]
17. Jung, L.B.; Gudera, J.A.; Wiegand, T.L.T.; Allmendinger, S.; Dimitriadis, K.; Koerte, I.K. ChatGPT Passes German State Examination in Medicine with Picture Questions Omitted. *Dtsch. Arztebl. Int.* **2023**, *120*, 373–374. [CrossRef] [PubMed]

18. Jeblick, K.; Schachtner, B.; Dexl, J.; Mittermeier, A.; Stüber, A.T.; Topalis, J.; Weber, T.; Wesp, P.; Sabel, B.O.; Ricke, J.; et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *Eur. Radiol.* **2024**, *34*, 2817–2825. [CrossRef] [PubMed]
19. Ossowska, A.; Kusiak, A.; Świetlik, D. Artificial Intelligence in Dentistry—Narrative Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3449. [CrossRef]
20. Ossowska, A.; Kusiak, A.; Świetlik, D. Evaluation of the Progression of Periodontitis with the Use of Neural Networks. *J. Clin. Med.* **2022**, *11*, 4667. [CrossRef]
21. Ossowska, A.; Kusiak, A.; Świetlik, D. Progression of Selected Parameters of the Clinical Profile of Patients with Periodontitis Using Kohonen's Self-Organizing Maps. *J. Pers. Med.* **2023**, *13*, 346. [CrossRef]
22. Decuyper, M.; Maebe, J.; Van Holen, R.; Vandenberghe, S. Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Phys.* **2021**, *8*, 81. [CrossRef] [PubMed] [PubMed Central]
23. Papachristou, K.; Panagiotidis, E.; Makridou, A.; Kalathas, T.; Masganis, V.; Paschali, A.; Aliberti, M.; Chatzipavlidou, V. Artificial intelligence in Nuclear Medicine Physics and Imaging. *Hell. J. Nucl. Med.* **2023**, *26*, 57–65. [CrossRef] [PubMed]
24. Świetlik, D.; Białowąs, J.; Kusiak, A.; Cichońska, D. Memory and forgetting processes with the firing neuron model. *Folia Morphol.* **2018**, *77*, 221–233. [CrossRef]
25. Świetlik, D. Simulations of Learning, Memory, and Forgetting Processes with Model of CA1 Region of the Hippocampus. *Complexity* **2018**, *2018*, 1297150. [CrossRef]
26. Świetlik, D.; Białowąs, J.; Kusiak, A.; Cichońska, D. A computational simulation of long-term synaptic potentiation inducing protocol processes with model of CA3 hippocampal microcircuit. *Folia Morphol.* **2018**, *77*, 210–220. [CrossRef] [PubMed]
27. Świetlik, D.; Kusiak, A.; Ossowska, A. Computational Modeling of Therapy with the NMDA Antagonist in Neurodegenerative Disease: Information Theory in the Mechanism of Action of Memantine. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4727. [CrossRef] [PubMed]
28. Świetlik, D.; Kusiak, A.; Krasny, M.; Białowąs, J. The Computer Simulation of Therapy with the NMDA Antagonist in Excitotoxic Neurodegeneration in an Alzheimer's Disease-like Pathology. *J. Clin. Med.* **2022**, *11*, 1858. [CrossRef]
29. Alsayed, A.A.; Aldajani, M.B.; Aljohani, M.H.; Alamri, H.; Alwadi, M.A.; Alshammari, B.Z.; Alshammari, F.R. Assessing the Quality of AI Information from ChatGPT Regarding Oral Surgery, Preventive Dentistry, and Oral Cancer: An Exploration Study. *Saudi Dent. J.* **2024**, *36*, 1483–1489. [CrossRef]
30. Fatani, B. ChatGPT for Future Medical and Dental Research. *Cureus* **2023**, *15*, e37285. [CrossRef]
31. Uribe, S.E.; Maldupa, I. Estimating the Use of ChatGPT in Dental Research Publications. *J. Dent.* **2024**, *149*, 105275. [CrossRef]
32. Janakiram, C.; Dye, B.A. A Public Health Approach for Prevention of Periodontal Disease. *Periodontology 2000* **2020**, *84*, 202–214. [CrossRef]
33. Alan, R.; Alan, B.M.; Alan, R.; Alan, B.M. Utilizing ChatGPT-4 for Providing Information on Periodontal Disease to Patients: A DISCERN Quality Analysis. *Cureus* **2023**, *15*, e46213. [CrossRef]
34. Tastan Eroglu, Z.; Babayigit, O.; Ozkan Sen, D.; Ucan Yarkac, F. Performance of ChatGPT in Classifying Periodontitis According to the 2018 Classification of Periodontal Diseases. *Clin. Oral Investig.* **2024**, *28*, 407. [CrossRef]
35. Nicikowski, J.; Szczepański, M.; Miedziaszczyk, M.; Kudliński, B. The Potential of ChatGPT in Medicine: An Example Analysis of Nephrology Specialty Exams in Poland. *Clin. Kidney J.* **2024**, *17*, sfae193. [CrossRef] [PubMed]
36. Ali, R.; Tang, O.Y.; Connolly, I.D.; Zadnik Sullivan, P.L.; Shin, J.H.; Fridley, J.S.; Asaad, W.F.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* **2023**, *93*, 1353–1365. [CrossRef] [PubMed]
37. Lahat, A.; Sharif, K.; Zoabi, N.; Shneor Patt, Y.; Sharif, Y.; Fisher, L.; Shani, U.; Arow, M.; Levin, R.; Klang, E. Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4. *J. Med. Internet Res.* **2024**, *26*, e54571. [CrossRef] [PubMed]
38. Khorshidi, H.; Mohammadi, A.; Yousem, D.M.; Abolghasemi, J.; Ansari, G.; Mirza-Aghazadeh-Attari, M.; Acharya, U.R.; Ardakani, A.A. Application of ChatGPT in Multilingual Medical Education: How Does ChatGPT Fare in 2023's Iranian Residency Entrance Examination. *Inform. Med. Unlocked* **2023**, *41*, 101314. [CrossRef]
39. Wang, Y.-M.; Shen, H.-W.; Chen, T.-J. Performance of ChatGPT on the Pharmacist Licensing Examination in Taiwan. *J. Chin. Med. Assoc.* **2023**, *86*, 653–658. [CrossRef]
40. Ando, K.; Sato, M.; Wakatsuki, S.; Nagai, R.; Chino, K.; Kai, H.; Sasaki, T.; Kato, R.; Nguyen, T.P.; Guo, N.; et al. A Comparative Study of English and Japanese ChatGPT Responses to Anaesthesia-Related Medical Questions. *BJA Open* **2024**, *10*, 100296. [CrossRef]
41. Liu, M.; Okuhara, T.; Chang, X.; Shirabe, R.; Nishiie, Y.; Okada, H.; Kiuchi, T. Performance of ChatGPT across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **2024**, *26*, e60807. [CrossRef]

42. Lewandowski, M.; Łukowicz, P.; Świetlik, D.; Barańska-Rybak, W. ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level Based on the Specialty Certificate Examination in Dermatology. *Clin. Exp. Dermatol.* **2024**, *49*, 686–691. [CrossRef]
43. Bhayana, R.; Krishna, S.; Bleakney, R.R. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology* **2023**, *307*, 230582. [CrossRef]
44. Li, S. Exploring the Clinical Capabilities and Limitations of ChatGPT: A Cautionary Tale for Medical Applications. *Int. J. Surg.* **2023**, *109*, 2865. [CrossRef]
45. Chomsky, N.; Roberts, I.; Watumull, J.N.C. The False Promise of ChatGPT. *The New York Times*. 2023, p. 8. Available online: www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (accessed on 19 November 2023).
46. Azaria, A. ChatGPT: More Human-Like Than Computer-Like, but Not Necessarily in a Good Way. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 6–8 November 2023.
47. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef] [PubMed]
48. Sallam, M.; Salim, N.A.; Al-Tammemi, A.B.; Barakat, M.; Fayyad, D.; Hallit, S.; Harapan, H.; Hallit, R.; Mahafzah, A. ChatGPT Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study at the Outset of a Paradigm Shift in Online Search for Information. *Cureus* **2023**, *15*, e35029. [CrossRef] [PubMed]
49. Deiana, G.; Dettori, M.; Arghittu, A.; Azara, A.; Gabutti, G.; Castiglia, P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines* **2023**, *11*, 1217. [CrossRef] [PubMed]
50. Babayiğit, O.; Eroglu, Z.T.; Sen, D.O.; Yarkac, F.U. Potential Use of ChatGPT for Patient Information in Periodontology: A Descriptive Pilot Study. *Cureus* **2023**, *15*, 11. [CrossRef] [PubMed]
51. Li, C.; Zhang, J.; Abdul-Masih, J.; Zhang, S.; Yang, J. Performance of ChatGPT and Dental Students on Concepts of Periodontal Surgery. *Eur. J. Dent. Educ.* **2024**. [CrossRef]
52. Meskó, B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J. Med. Internet Res.* **2023**, *25*, e50638. [CrossRef]
53. Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers. *NPJ Digit. Med.* **2023**, *6*, 75. [CrossRef]
54. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of Large Language Models Such as ChatGPT for Dental Medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [CrossRef]
55. Zaman, M. ChatGPT for Healthcare Sector: SWOT Analysis. *Int. J. Res. Ind. Eng.* **2023**, *12*, 221–233. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.