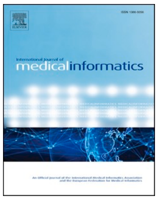
International Journal of Medical Informatics 190 (2024) 105562

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: A comparative study

Robert Olszewski [a,b], Klaudia Watros [a], Małgorzata Mańczak [a], Jakub Owoc [a], Krzysztof Jeziorski [a,c], Jakub Brzeziński [a,*]

[a] Gerontology, Public Health and Education Department, National Institute of Geriatrics, Rheumatology and Rehabilitation, Warsaw, Poland
[b] Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences
[c] Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland

ARTICLE INFO

Keywords:
Chatbots
Readability
Cardiovascular health
Oncology
Psoriasis

ABSTRACT

Background: Chatbots using the Large Language Model (LLM) generate human responses to questions from all categories. Due to staff shortages in healthcare systems, patients waiting for an appointment increasingly use chatbots to get information about their condition. Given the number of chatbots currently available, assessing the responses they generate is essential.

Methods: Five chatbots with free access were selected (Gemini, Microsoft Copilot, PiAI, ChatGPT, ChatSpot) and blinded using letters (A, B, C, D, E). Each chatbot was asked questions about cardiology, oncology, and psoriasis. Responses were compared to guidelines from the European Society of Cardiology, American Academy of Dermatology and American Society of Clinical Oncology. All answers were assessed using readability scales (Flesch Reading Scale, Gunning Fog Scale Level, Flesch-Kincaid Grade Level and Dale-Chall Score). Using a 3-point Likert scale, two independent medical professionals assessed the compliance of the responses with the guidelines.

Results: A total of 45 questions were asked of all chatbots. Chatbot C gave the shortest answers, 7.0 (6.0 – 8.0), and Chatbot A the longest 17.5 (13.0 – 24.5). The Flesch Reading Ease Scale ranged from 16.3 (12.2 – 21.9) (Chatbot D) to 39.8 (29.0 – 50.4) (Chatbot A). Flesch-Kincaid Grade Level ranged from 12.5 (10.6 – 14.6) (Chatbot A) to 15.9 (15.1 – 17.1) (Chatbot D). Gunning Fog Scale Level ranged from 15.77 (Chatbot A) to 19.73 (Chatbot D). Dale-Chall Score ranged from 10.3 (9.3 – 11.3) (Chatbot A) to 11.9 (11.5 – 12.4) (Chatbot D).

Conclusion: This study indicates that chatbots vary in length, quality, and readability. They answer each question in their own way, based on the data they have pulled from the web. Reliability of the responses generated by chatbots is high. This suggests that people who want information from a chatbot need to be careful and verify the answers they receive, particularly when they ask about medical and health aspects.

1. Introduction

Chatbots are computer programs that, through the use of machine learning algorithms and natural language processing, can comprehend and respond to human language, both spoken and written, in a human-like manner [1]. Using large language models (LLMs), they have the ability to simulate conversational reasoning, generating responses that are contextually correct, but do not fully understand the context as humans do [2].

Using the available capabilities and technologies, more and more companies are creating their own chatbots that can communicate on a variety of issues [3]. A significant breakthrough came in 2022 with the introduction of ChatGPT (short for Chat Generative Pretrained Transformer), a tool that was made available under certain restrictions. This caused a huge stir and interest in chatbots and further work on artificial intelligence (AI) [4]. The newly released ChatGPT tool offered extensive natural language processing (NLP) capabilities, including text analysis, machine translation and answering questions in a way that simulates

* Corresponding author at: Spartanska 1, 02-637, Warsaw, Poland, Gerontology, Public Health and Education Department, National Institute of Geriatrics, Rheumatology and Rehabilitation, Poland.

E-mail addresses: robert.olszewski@spartanska.pl (R. Olszewski), klaudia.watros@spartanska.pl (K. Watros), m.manczak@op.pl (M. Mańczak), kowoc@wp.pl (J. Owoc), krzysztof.jeziorski@spartanska.pl (K. Jeziorski), jakub.brzezinski@spartanska.pl (J. Brzeziński).

https://doi.org/10.1016/j.ijmedinf.2024.105562
Received 21 May 2024; Received in revised form 11 July 2024; Accepted 18 July 2024
Available online 19 July 2024

human conversation.

The emergence of ChatGPT has prompted the creation of more chatbots using LLMs. Many companies have developed their own AI chatbots to answer questions, analyze texts and perform translations [5].

ChatGPT, short for "Chat Generative Pre-trained Transformer," is a chatbot launched by OpenAI in November 2022. It is built on top of the OpenAI GPT-3.5 family of large language models (LLMs). Designed to generate human-like text in response to user queries and prompts, ChatGPT can be used for a wide range of applications, including dialog systems, language translation, and content generation [6]. Microsoft Copilot, originally known as Bing Chat, is another chatbot based on the same technology. Both models utilize the evolving GPT model developed by OpenAI, but they differ in functionality. Copilot has access to GPT-4, which enables better language understanding, reasoning, and other advanced capabilities. Unlike ChatGPT, Copilot can also search the web for information and update its knowledge base [7]. Google Gemini is an advanced AI model introduced by Google. It was developed through collaborative efforts across various Google teams, including Google Research. The primary goal was to create a multimodal model capable of understanding and combining different types of information, such as text, code, audio, images, and videos. Gemini, formerly known as Bard, serves as a writing, planning, and learning assistant. Users can engage in chat with Google AI to enhance creativity, productivity, and problem-solving [8]. These models represent significant advancements in artificial intelligence, allowing them to generate content and answer questions in a manner similar to human communication.

AI has brought revolutionary changes to many areas of science, including healthcare [9]. In clinical medicine, artificial intelligence systems are primarily used to analyze the genome of the human body [10], as well as diagnose, support to treatment diseases and predict clinical outcomes based on patient data [11,12,13]. The huge impact that artificial intelligence has had on medicine has led to the increasing introduction of AI technologies into other medical disciplines and the introduction of more novel solutions [14,15].

Continuing advances in AI are driving further research into chatbots, and this trend is set to continue to make chatbots as advanced as possible [16]. The shortage of medical staff means that more and more people are using chatbots to get advice about their medical situation. Chatbots can also help to meet the growing demand for services. However, it is important to bear in mind that with the increasing number of interactive chatbots being generated, it is difficult to control the quality of responses and ensure they are in line with current medical knowledge [17].

Consequently, the objective of this investigation was to scrutinize the reactions of five freely accessible and public chatbots in response to inquiries pertaining to the recommendations of international societies on the subjects of cardiology, oncology, and psoriasis. Cardiology and oncology, recognized as predominant epidemiological challenges, have been selected as the focal points for inquiry. The escalating prevalence of these two medical conditions could potentially lead to an increased reliance on AI as a navigational tool by patients. This trend underscores the potential of AI to serve as an informative guide in the healthcare landscape, particularly in the context of these high-incidence diseases [18,19]. The choice of questions regarding psoriasis is due to the significant health problem of psoriasis worldwide. In 2014, the World Health Organization recognized psoriasis as a serious global disease that significantly reduces the quality of life of patients and contributes to their stigmatization. Also this disease poses a challenge to healthcare systems [20,21]. This analysis was conducted with the intent of evaluating the accuracy, relevance, readability and reliability of the information provided by these artificial intelligence systems in the context of these specific medical disciplines.

Developments in artificial intelligence have brought technology into healthcare, enabling virtual conversations with chatbots. These chatbots, using extensive online knowledge, can answer a variety of questions [22]. Several preliminary studies have been published that analyze the quality of responses and readability from chatbots and compare responses between different chatbots. One of them the study by Suarez et al. assessed whether ChatGPT-4 could provide accurate and reliable answers to general dentists in oral surgery. They also evaluated its potential as an assistant for dentists [23]. The publication by Deiana et al. analyzed the responses of ChatGPT-3.5 (free) and ChatGPT-4 (paid) regarding vaccination [24]. Birkun et al. conducted a study using the Bing chatbot, developed by Microsoft, in three different countries with varying income levels. They posed the question, 'Heart attack, what to do?' and verified the answers using the International First Aid, Resuscitation, and Education Guidelines 2020. In addition, the researchers assessed the readability of the answers using the Flesch-Kincaid grade level. The study also assessed the readability of Bing responses to a question about myocardial infarction. The study was conducted in three different countries, in English. The chatbot referenced heart attack rescue guidelines; however, the quality of the responses was poor. The chatbot added various issues to the answers that were not included in the guidelines [4].

In an analysis by Cheong et al. who conducted a direct comparative evaluation of patient education materials on obstructive sleep apnoea generated by two artificial intelligence chatbots, ChatGPT and Google Bard (now Gemini). According to the response analysis, all responses generated by ChatGPT had better scores than those generated by Google Bard. The average Flesch-Kincaid score for ChatGPT was 9.0, while Google Bard's was 5.9, suggesting that Google Bard's responses are easier to read and understand than ChatGPT's [25].

The study by Koo et al. analyzed the readability of information about overactive bladder (OAB) on the Internet. The SMOG test, Dale-Chall readability formula, was used for the analysis. The authors analyzed 57 websites. Seven of them (12 %) the test was able to be read and understood by a normal adult at the 8th-grade readability level. This proves that the vast majority of online information on OAB treatment is beyond the reading ability of most adults [26].

In a Polish study conducted by Suwała and colleagues, ChatGPT answered questions from a specialist internal medicine exam. The average performance of chatbots ranged from 47.5 % to 53.44 % (with a median of 49.37 %), while the average performance of human doctors ranged from 65.21 % to 71.95 % (with a median of 69.92 %). The study revealed that despite having access to various information sources, the chatbot performed worse than humans in answering questions related to internal medicine [27].

Multiple-choice questions were obtained from the official National Board of Medical Examiners (NBME) website. These questions were sourced from NBME subject examinations in various medical fields, including medicine, pediatrics, obstetrics and gynecology, clinical neurology, ambulatory care, family medicine, psychiatry, and surgery. Each language model (LLM), namely GPT-4, GPT-3.5, Claude, and Bard, provided responses to these questions. The accuracy of each chatbot's response was compared to the answers provided by the NBME. A total of 163 questions were answered by each LLM. The results were as follows: GPT-4 scored 163/163 (100 %), GPT-3.5 scored 134/163 (82.2 %), Bard scored 123/163 (75.5 %), and Claude scored 138/163 (84.7 %) [28].

Our study revealed that the scoring of chatbot responses, as evaluated using a Likert scale in published articles, consistently indicated high levels of satisfaction. Additionally, the responses were qualitatively assessed, and readability and quality were equally considered across several publications. Notably, the vast majority of responses were either complete or partially complete, suggesting a high level of substantive quality [29,30,31].

From all available publications, it appears that chatbots can be an alternative source of information. Until there are guidelines on which information sources to use to provide answers, chatbots will vary in their level of readability, chatbots will add unnecessary content to answers or will not know how to answer the questions asked. However, chatbots already provide quality answers that are highly rated by medical professionals, which is an important benchmark for further work on chatbots in medicine. It should, however, be borne in mind that these answers provide only a general view of the given problem. This study provides valuable insights into the performance and reliability of chatbot responses in a medical context. Further research is warranted to optimize their utility in healthcare settings. The survey was aimed at general users sending questions to chatbots. Its purpose was to test whether the answers of chatbots differ even when they have access to the same data.

2. Materials and methods

2.1. Chatbots

The 43 chatbots were selected based on various articles found online [32,33,34,35]. Thirty eight chatbots were excluded due to user fees, time constraints for queries, and inability or difficulty to log in (e.g. need to create an account, double security for registration, required phone number). These are factors that may affect the availability and usability of chatbots for patients. Fig. 1. shows the chatbot selection scheme. After analyses and exclusions, 5 chatbots remained (Microsoft Copilot, Gemini, ChatGPT, PiAI, ChatSpot).At the start of the study, Microsoft Bing and Google Bard were renamed to Microsoft Copilot and Google Gemini, respectively. The companies changed their names when improvements were added. Once the chatbots were selected, they were randomly blinded using letters (A, B, C, D, E) to ensure objectivity of the researchers when assessing the readability and quality of the responses. Table 2. described each of them.

2.2. Questions

The questions for the chatbots were prepared on the basis of recommendations from guidelines prepared by the:

European Society of Cardiology – *2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure* [36],
European Society of Cardiology – *2023 ESC Guidelines for the management of cardiovascular disease in patients with diabetes* [37],
American Society of Clinical Oncology – *Exercise, Diet, and Weight Management During Cancer Treatment: ASCO Guideline* [38],
American Academy of Dermatology – *Joint AADeNPF Guidelines of care for the management and treatment of psoriasis with topical therapy and alternative medicine modalities for psoriasis severity measures* [39].

The total number of questions was 45. There were 27 questions on cardiovascular disease in patients with diabetes, 5 on heart failure, 11 oncology and 3 on psoriasis. The distribution of questions was designed following established guidelines. Since the survey targeted chatbot users, we formulated questions consistent with language commonly used by patients unfamiliar with specialized terminology. Each question posed to the chatbot underwent analysis based on the guidelines and the level of simplicity required In the realm of cardiology, there has been a marked increase in inquiries pertaining to cardiovascular disease as a complication of diabetes. This trend is reflective of the guidelines, which predominantly focus on the intersection of cardiac disease and diabetes. This underscores the urgent need for comprehensive research and understanding in this area. The interplay between these two conditions presents a complex clinical challenge that warrants further exploration. The aim is to enhance patient care and outcomes through evidence-based practices and interventions. The example questions posed to the chatbots are: *What is the recommended blood pressure in patients with diabetes?* Or: *Can stress reduction improve psoriasis severity?* All questions were asked to the chatbots once. To ensure consistency in the survey, they were asked by one person using a personal computer. Questions were not asked again after a period of time. Each answer was copied for readability analysis on the tool page. To avoid the influence of previous questions on the chatbots' answers, the chat was reset before each new question to avoid possible biases in the 4 chatbots (Gemini, Microsoft Copilot, ChatGPT and ChatSpot). In PiAI, there was no option to reset the chat.

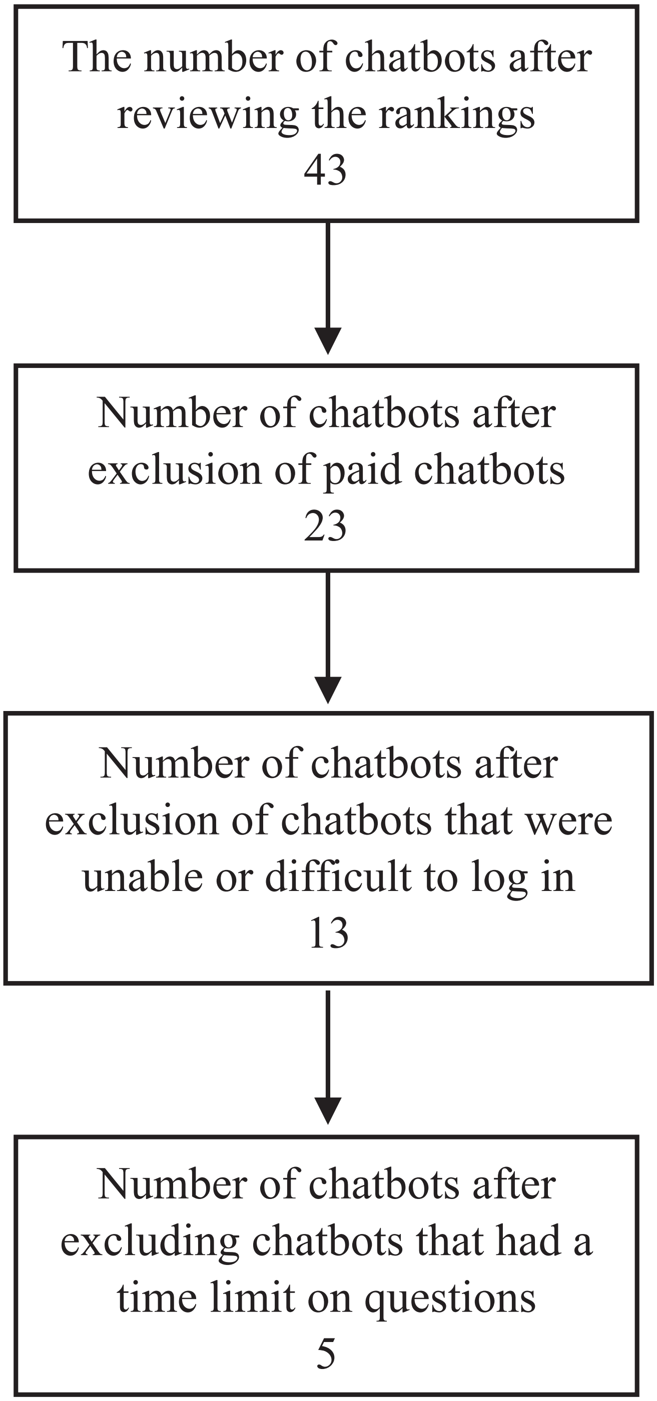

The number of chatbots after reviewing the rankings
43
Number of chatbots after exclusion of paid chatbots
23
Number of chatbots after exclusion of chatbots that were unable or difficult to log in
13
Number of chatbots after excluding chatbots that had a time limit on questions
5

Fig. 1. Diagram for choosing chatbots.

2.3. Readability analysis

Each answer was then analyzed for readability. Scales were used to assess readability: Flesch-Kincaid Grade Level, Gunning Fog Scale Level, Flesch Reading Ease Level, Dale-Chall Score. All scales were determined using the online tool Datayze. In addition, response lengths counted in sentences were also recorded [40]. These indicators are the most commonly used in assessing the readability of medical literature [41].

The Flesch-Kincaid Grade Level assesses the readability by United

States (US) grade level. The values vary from 0 to 18, where 18 represents the most difficult text. A higher score means that it is more difficult to understand the text. A score above 12 indicates that the text is written in an academic style [42].

The Gunning Fog Index is a readability metric that uses word and sentence length to determine how difficult a text is to read. Index values typically vary from 1 to 18 and more. Values correspond to the number of years of education a reader requires to understand a text. A text understood by the general public should be at a level of about 8. Texts above 17 points are intended for graduate students [43].

Flesch Reading Ease is a metric that measures the readability of a text based on sentence length determined by the average number of words in a sentence and word length determined by the average number of syllables in a word. The score ranges from 0 to 100, with higher scores indicating easier readability. A score between 70 and 80 equates to an 8th grade school level [44].

Dale-Chall score is a readability metric used to indicate how difficult a text is to read based on a predefined set of “common” words and the ratio of “difficult” words and words per sentence. A score of 4.9 or less is a score easily understood by the average 4th grade or lower student. In contrast, 9 and above is easily understood by the average college student [45].

2.4. Reliability assessment

Two independent medical specialists, (R.O. and K.J.), evaluated 217 responses from all chatbots. The experts were given questions to ask the chatbots based on the provided guidelines. They assessed the quality and accuracy of the answers according to the text. They used a three-point Likert scale from 0 to 2 in which 0 meant an incorrect answer, 1 partially correct and 2 correct [22].

2.5. Statistical analyses

Data of readability scales are presented as median (Me) and interquartile range (IQR). Comparison of the readability ratings of chatbots’ responses was made using the Friedman ANOVA test. Cohen’s kappa coefficient was calculated to assess agreement between two specialists’ assessments of the reliability of the answers provided by the chatbots. The levels of agreement for kappa were considered slight ($\kappa < 0.20$), fair ($\kappa = 0.21$ to 0.40), moderate ($\kappa = 0.41$ to 0.60), substantial ($\kappa = 0.61$ to 0.80), or almost perfect ($\kappa = 0.81$ to 1.00) [46].

3. Results

The total of 225 questions were asked to all chatbots. Chatbot A did not answer 5 questions and Chatbot E did not answered 3 questions. The term ‘No answer’ denotes instances where the chatbot responded that it lacked the necessary knowledge or resources to address a question, specifically stating that it was not a doctor. We categorized such responses as ‘No answer’ due to the absence of relevant information. Below, I have included screenshots illustrating examples of non-responses from chatbots. The total number of responses included in the analysis was 217. The values of readability indicators for all scales were significantly different among chatbots (Table 1). Chatbot A showed the most sentences per response 17.5 (13.0 – 24.5), Chatbot C the least 7.0 (6.0 – 8.0). Chatbot A had the lowest score in the Flesch-Kincaid Grade Level 12.5 (10.6 – 14.6). Chatbot D had the highest average rates 15.9 (15.1 – 17.1). Gunning Fog Scale Level was lowest in Chatbot A15.8 (13.8 – 17.7) and highest in Chatbot D 20.0 (18.5 – 21.5). Chatbot D had the lowest score 16.3 (12.2 – 21.9) in Flesch Reading Ease Level and the highest score was at Chatbot A 39.8 (29.0 – 50.4). In Dale-Chall Score, Chatbot A had the lowest score 10.3 (9.3 – 11.3) and Chatbot D had the highest score 11.9 (11.5 – 12.4). Three chatbots had an average of less than 10 sentences per answer (Chatbots B, C, E), two chatbots had more than 10 (Chatbots A, D). According to all the readability scales used, the findings demonstrate that the chatbots’ answers were at an advanced, academically based level, which may be difficult to understand for a person with less than a university degree. Table 2. shows the distribution of the results of the individual readability scales for the chatbots.

Table 1
Chatbots included in the study.

Name of chatbot	Developed by	Initial release date
Gemini	Google AI	March 21, 2023
ChatGPT	OpenAI	November 30, 2022
ChatSpot	HubSpot	March, 8, 2023
Microsoft Copilot	Microsoft	February 7, 2023
Pi AI	Inflection AI	May, 2, 2023

Figs. 2, 3, 4, 5, 6 show a comparison, including the number of sentences each chatbot assigned to a response and the values of the reading scales.

An evaluation of the reliability of the responses was conducted, involving two independent medical professionals. This process was integral to ensuring the validity and accuracy of the data collected. In the evaluation of the chatbot’s responses, the majority were deemed accurate by the researchers. Specifically, Researcher 1 classified 144 out of 225 responses as correct, constituting 64 % of the total responses. Similarly, Researcher 2 identified 141 out of 225 responses as correct, accounting for 62.6 % of the total responses. It was observed that the proportion of incorrect responses was minimal. The detailed distribution of the response ratings, as assessed by the medical professionals, is presented in Table 3. This table provides a comprehensive overview of the evaluation results.

3.1. Comparisons of responses from medical professionals

The concordance rate of the physicians’ assessments ranged from 0.73 (Chatbot C) to 0.8 (Chatbot A). The Cohen’s kappa coefficient of agreement (κ), a statistical measure of inter-rater agreement, varied from 0.40 (95 % CI: 0.09——0.71) for Chatbot D to 0.57 (95 % CI: 0.30——0.84) for Chatbot A. The slight variation in the percentage concordance among the chatbots’ responses suggests a similar response pattern across the chatbots. According to the JR Landis and GG Koch scale, the Cohen’s kappa for the chatbot responses fell within the moderate range [46]. Table 4 presents detailed results comparing responses by medical professionals.

4. Discussion

The main result of our study provides valuable insights into the performance and reliability of chatbot responses in a medical context. Artificial intelligence is increasingly present in everyday life. Chatbots provide quick answers to questions in all fields, including medicine. Due to staff shortages in the healthcare system and thus long waits for medical appointments, people needing medical advice will use chatbots more often to diagnose their symptoms [47].

The findings of the study demonstrate that the chatbots’ answers to the same questions vary in sentence length and according to different readability scales. Despite having access to the guidelines, which are publicly available online, the chatbots did not use the data contained therein. Moreover, they added information to the answers that was not included in the guidelines. This caused the lengths of the answers in the sentences to vary from chatbot to chatbot. In turn, the differences in sentence lengths account for the differences in the results of the readability scales.

We do not consider 64 % to be a good result. According to the existing literature, our findings are less favourable than other scientific publications. In a study by Neo JRE and colleagues, medical professionals rated chatbot responses as satisfactory, with ChatGPT

Table 2
Readability comparison. Me – median, IQR – interquartile range.

Variable	A		B		C		D		E		p
	n	Me(IQR)	n	Me(IQR)	n	Me(IQR)	n	Me(IQR)	n	Me(IQR)	
No. of sent.	40	17.5 (13.0 – 24.5)	45	8.0 (5.0 – 10.0)	45	7.0 (6.0 – 8.0)	45	14.0 (9.0 – 19.0)	42	7.0 (3.0 – 13.0)	<0.001
F-KGL	40	12.5 (10.6 – 14.6)	44	14.9 (12.6 – 16.7)	45	13.8 (12.5 – 15.7)	45	15.9 (15.1 – 17.1)	42	15.0 (13.6 – 15.9)	<0.001
GFSL	40	15.8 (13.8 – 17.7)	45	18.3 (16.2 – 20.2)	45	17.6 (16.3 – 19.1)	45	20.0 (18.5 – 21.5)	42	18.2 (17.1 – 20.3)	<0.001
FRES	40	39.8 (29.0 – 50.4)	45	28.6 (12.2 – 37.0)	45	31.9 (26.7 – 37.0)	45	16.3 (12.2 – 21.9)	42	23.1 (15.5 – 30.9)	<0.001
D-CS	40	10.3 (9.3 – 11.3)	45	11.6 (10.7 – 12.6)	45	10.5 (9.9 – 11.1)	45	11.9 (11.5 – 12.4)	42	11.3 (10.6 – 12.6)	<0.001

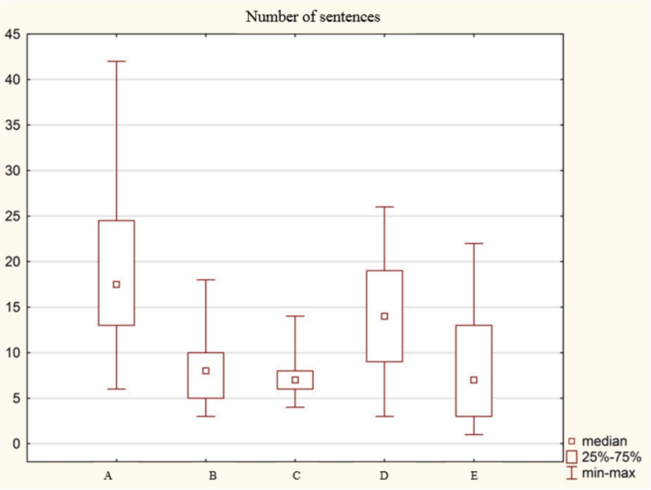


Fig. 2. Comparison of the number of sentences for all chatbots.

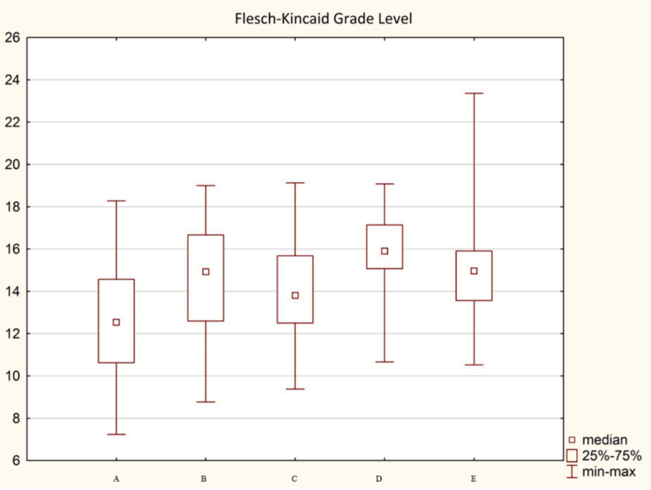


Fig. 3. Comparison of the values of readability scales: Flesch-Kincaid Grade Level.

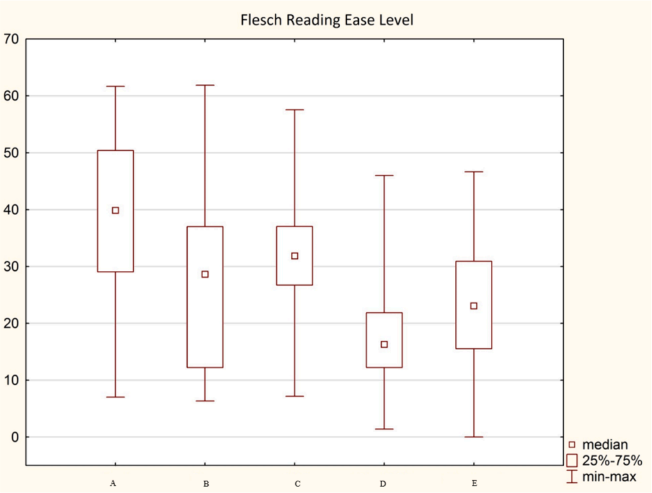


Fig. 4. Comparison of the values of readability scales: Flesch Reading Ease Level.

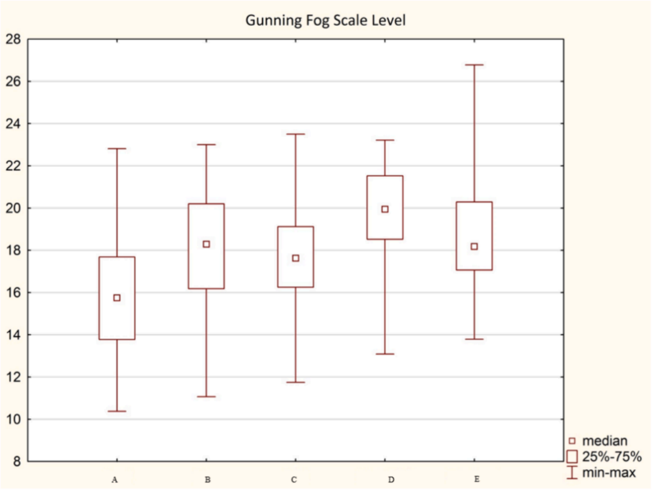


Fig. 5. Comparison of the values of readability scales: Gunning Fog Scale Level.

achieving a score of 65.8 % and Google Bard scoring 75.8 % [48]. In a separate study by Suárez A. et al., oral surgeons evaluated ChatGPT-4 responses, and the average rating for complete chatbot responses was 71.7 % [23].

The capabilities of chatbots available to users can be significantly enhanced using Retrieval-Augmented Generation (RAG). RAG is an advanced artificial intelligence technique that seamlessly combines information retrieval with text generation. By leveraging external knowledge sources, AI models can retrieve relevant information and seamlessly incorporate it into their generated responses. This approach ensures that chatbots provide more accurate and contextually relevant answers, avoiding errors or outdated data. RAG proves particularly valuable for language models that rely on general training data but require access to up-to-date and specific information [49].

5. Limitations

The study is subject to certain limitations. Firstly, not all chatbots used for the study use GPT-4. The free ChatGPT uses GPT-3.5, while the rest of the chatbots use GPT-4, which is more advanced and powerful than the basic version. Secondly, ChatSpot is primarily designed for business. This chatbot was used for the study because it was free and there was no difficulty in logging into it. Thirdly, the chatbots were

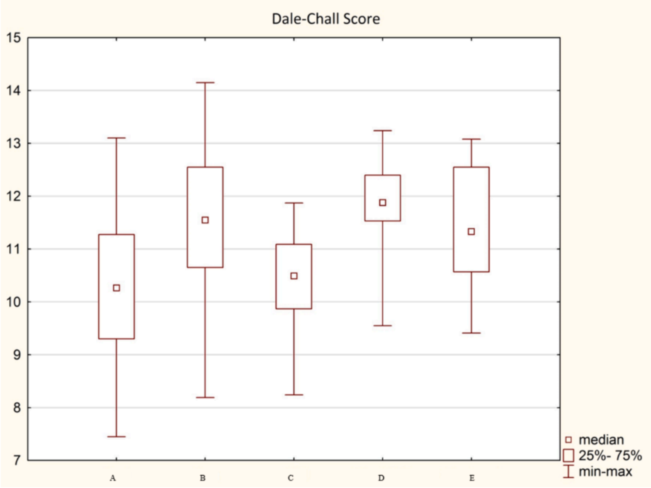


Fig. 6. Comparison of the values of readability scales: Dale-Chall Score.

Table 3
Correctness of the chatbot's answers as assessed by the Likert scale.

	Researcher 1					
	Chatbot A	Chatbot B	Chatbot C	Chatbot D	Chatbot E	%
Correct	27	34	26	34	23	64
Partially correct/ incomplete	9	7	14	7	11	21
Incorrect	4	4	5	4	8	11
No answers	5	0	0	0	3	3
	Researcher 2					
	Chatbot A	Chatbot B	Chatbot C	Chatbot D	Chatbot E	%
Correct	28	32	26	33	22	62
Partially correct/ incomplete	11	13	16	12	18	31
Incorrect	1	0	3	0	2	2
No answers	5	0	0	0	3	3

Table 4
Detailed results comparing responses by medical professionals.

	Cohen kappa	Percentage agreement
Chatbot A	0.57 (0.30 – 0.84)	80 %
Chatbot B	0.41 (0.10 – 0.72)	75 %
Chatbot C	0.51 (0.28 – 0.75)	73 %
Chatbot D	0.40 (0.09 – 0.71)	75 %
Chatbot E	0.56 (0.33 – 0.78)	74 %

quizzed on the question asked only once. This was done because for each question asked, the chatbot would answer in a different way, which could result in varying word counts, different text readability scores according to scales, and possibly varying quality according to experts.

6. Conclusions

Chatbots, underpinned by Large Language Models, respond to inquiries in the domains of cardiology, oncology, and psoriasis with a level of sophistication and academic rigor that may pose comprehension challenges for individuals lacking advanced education or familiarity with complex medical terminology. The responses generated by these chatbots exhibit variability in sentence length. Notwithstanding, the quality of the responses generated by these chatbots is high, providing a solid foundation for future enhancements aimed at optimizing the balance between comprehensibility and quality in the chatbots' responses. This endeavor is crucial in ensuring that these advanced tools are accessible and beneficial to a broad spectrum of users.

The responses provided by chatbots exhibit variability in terms of their readability, accuracy, and consistency. Chatbots often incorporate additional comments into their responses, a practice that can potentially interfere with their interpretation and lead to confusion. Even with access to a multitude of databases, including those related to medicine, chatbots infrequently make references to established guidelines. Nevertheless, the reliability of responses generated by chatbots is high.

CRediT authorship contribution statement

Robert Olszewski: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Conceptualization. **Klaudia Watros:** Writing – review & editing, Resources, Investigation. **Małgorzata Mańczak:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Jakub Owoc:** Writing – review & editing, Writing – original draft, Methodology. **Krzysztof Jeziorski:** Writing – review & editing, Supervision, Resources. **Jakub Brzeziński:** Writing – review & editing, Writing – original draft, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105562.

References

[1] A. Bohr, K. Memarzadeh, The rise of artificial intelligence in healthcare applications, Artificial Intelligence in Healthcare. (2020) 25–60, https://doi.org/10.1016/B978-0-12-818438-7.00002-2.

[2] M.D. Illescas-Manzano, N. Vicente López, N. Afonso González, C. Cristofol Rodríguez, Implementation of Chatbot in Online Commerce, and Open Innovation, J. Open Innov. Technol. Mark. Complex. 7 (2021) 125, https://doi.org/10.3390/joitmc7020125.

[3] K.I. Roumeliotis, N.D. Tselikas, ChatGPT and Open-AI Models: A Preliminary Review, Future Internet 15 (2023) 192, https://doi.org/10.3390/fi15060192.

[4] A.A. Birkun, A. Gautam, Large Language Model-based Chatbot as a Source of Advice on First Aid in Heart Attack, Curr. Probl. Cardiol. 49 (1 Pt A) (2024) 102048, https://doi.org/10.1016/j.cpcardiol.2023.102048.

[5] K.K. Nirala, N.K. Singh, V.S. Purani, A survey on providing customer and public administration based services using AI: chatbot, Multimed Tools Appl 81 (2022) 22215–22246, https://doi.org/10.1007/s11042-021-11458-y.

[6] N.A. Shayegh, D. Byer, Y. Griffiths, P.W. Coleman, L.A. Deane, J. Tonkin, Assessing artificial intelligence responses to common patient questions regarding inflatable penile prostheses using a publicly available natural language processing tool (ChatGPT), The Canadian Journal of Urology 31 (3) (2024) 11880–11885.

[7] F. Semeraro, L. Gamberini, F. Carmona, K.G. Monsieurs, Clinical questions on advanced life support answered by artificial intelligence. A comparison between ChatGPT, Google Bard and Microsoft Copilot, Resuscitation 195 (2024) 110114, https://doi.org/10.1016/j.resuscitation.2024.110114.

[8] M. Masalkhi, J. Ong, E. Waisberg, A.G. Lee, Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology, Eye (lond). 38 (8) (2024) 1412–1417, https://doi.org/10.1038/s41433-024-02958-w.

[9] S.A. Alowais, S.S. Alghamdi, N. Alsuhebany, et al., Revolutionizing healthcare: the role of artificial intelligence in clinical practice, BMC MedicineEduc 23 (2023) 689, https://doi.org/10.1186/s12909-023-04698-z.

[10] C.J. Haug, J.M. Drazen, Artificial Intelligence and Machine Learning in Clinical Medicine, 2023, N Engl. J. Med. 388 (13) (2023) 1201–1208, https://doi.org/10.1056/NEJMra2302038.

[11] K.B. Johnson, W.Q. Wei, D. Weeraratne, M.E. Frisse, K. Misulis, K. Rhee, et al., Precision Medicine, AI, and the future of Personalized Health Care, Clin Transl. Sci. 14 (1) (2021) 86–93, https://doi.org/10.1111/cts.12884.

[12] Y. Kumar, A. Koul, R. Singla, M.F. Ijaz, Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda, J. Ambient Intell Humaniz Comput. 14 (7) (2023) 8459–8486, https://doi.org/10.1007/s12652-021-03612-z.

[13] B.J. Kelly, J. Chevarria, B. O'Sullivan, G. Shorten, The potential for artificial intelligence to predict clinical outcomes in patients who have acquired acute kidney injury during the perioperative period, Perioper Med (lond). 10 (1) (2021 Dec 15) 49, https://doi.org/10.1186/s13741-021-00219-y.

[14] A.S. Ahuja, The impact of artificial intelligence in medicine on the future role of the physician, PeerJ. 4 (7) (2019 Oct) e7702.

[15] M.A. Rahman, E. Victoros, J. Ernest, R. Davis, Y. Shanjana, M.R. Islam, Impact of Artificial Intelligence (AI) Technology in Healthcare Sector: A Critical Evaluation of Both Sides of the Coin, Clinical Pathology. (2024) 17, https://doi.org/10.1177/2632010X241226887.

[16] J. Sidlauskiene, Y. Joye, V. Auruskeviciene, AI-based chatbots in conversational commerce and their effects on product and price perceptions, Electron Markets 33 (2023) 24, https://doi.org/10.1007/s12525-023-00633-8.

[17] A. Alizadehasl, A. Amin, M. Maleki, F. Noohi, A. Ghavamzadeh, M. Farrashi, Cardio-oncology discipline: focus on the necessities in developing countries, ESC Heart Fail. 7 (5) (2020 Oct) 2175–2183, https://doi.org/10.1002/ehf2.12838.

[18] M.M. Lane, E. Gamage, S. Du, D.N. Ashtree, A.J. McGuinness, S. Gauci, et al., Ultra-processed food exposure and adverse health outcomes: umbrella review of epidemiological meta-analyses, British Medical Journal (clinical Research Education) (london) 384 (2024) e077310.

[19] F.O. Nestle, D.H. Kaplan, J. Barker, Psoriasis, N Engl. J. Med. 361 (2009) 496–509, https://doi.org/10.1056/NEJMra0804595.

[20] I.M. Michalek, B. Loring, S.M. John, A systematic review of worldwide epidemiology of psoriasis, J. Eur. Acad. Dermatol. Venereol. 31 (2017) 205–212, https://doi.org/10.1111/jdv.13854.

[21] M. Paul, L. Maglaras, M.A. Ferrag, I. Almomani, Digitization of healthcare sector: A study on privacy and security concerns, ICT Express 9 (4) (2023) 571–588, https://doi.org/10.1016/j.icte.2023.02.007.

[22] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, S.S. Lee, Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science, Front. Artif Intell. 31 (6) (2023) 1237704, https://doi.org/10.3389/frai.2023.1237704.

[23] A. Suárez, J. Jiménez, M. Llorente de Pedro, C. Andreu-Vázquez, V. Díaz-Flores García, M. Gómez Sánchez, Y. Freire, Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct, Biotechnol J. 6 (24) (2023 Dec) 46–52, https://doi.org/10.1016/j.csbj.2023.11.058.

[24] G. Deiana, M. Dettori, A. Arghittu, A. Azara, G. Gabutti, P. Castiglia, Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions, Vaccines (basel). 11 (7) (2023) 1217, https://doi.org/10.3390/vaccines11071217.

[25] R.C.T. Cheong, S. Unadkat, V. Mcneillis, A. Williamson, J. Joseph, P. Randhawa, P. Andrews, V. Paleri, Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard, Eur Arch Otorhinolaryngol. 281 (2) (2024) 985–993, https://doi.org/10.1007/s00405-023-08319-9.

[26] K. Koo, K. Shee, R.L. Yap, Readability analysis of online health information about overactive bladder, Neurourol Urodyn. 36 (7) (2017) 1782–1787, https://doi.org/10.1002/nau.23176.

[27] S. Suwała, P. Szulc, A. Dudek, et al., ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn, Pol Arch Intern Med. 133 (2023) 16608, https://doi.org/10.20452/pamw.16608.

[28] A. Abbas, M.S. Rehman, S.S. Rehman, Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions, Cureus. 16 (3) (2024) e55991.

[29] B. Salam, D. Kravchenko, S. Nowak, A.M. Sprinkart, L. Weinhold, A. Odenthal, N. Mesropyan, L.M. Bischoff, U. Attenberger, D.L. Kuetting, J.A. Luetkens, A. Isaak, Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand, J Cardiovasc Magn Reson. 26 (1) (2024) 101035, https://doi.org/10.1016/j.jocmr.2024.101035.

[30] S. Garbarino, N.L. Bragazzi, Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: Comparative analysis using Google Bard and OpenAI ChatGPT-4, Journal of Sleep Research 5 (2024) e14210.

[31] R.J. Davis, O. Ayo-Ajibola, M.E. Lin, M.S. Swanson, T.N. Chambers, D.I. Kwon, N. C. Kokot, Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot, The Laryngoscope 134 (5) (2024 May) 2252–2257, https://doi.org/10.1002/lary.31191.

[32] https://www.intercom.com/learning-center/best-ai-chatbot (Access date: 08.02.2024).

[33] https://zapier.com/blog/best-ai-chatbot/ (Access date: 08.02.2024).

[34] https://www.zdnet.com/article/best-ai-chatbot/#google_vignette (Access date: 08.02.2024).

[35] https://www.forbes.com/advisor/business/software/best-chatbots/ (Access date: 08.02.2024).

[36] Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, John G F Cleland, Maria Generosa Crespo-Leiro, Dimitrios Farmakis, Martine Gilard, Stephane Heymans, Arno W Hoes, Tiny Jaarsma, Ewa A Jankowska, Mitja Lainscak, Carolyn S P Lam, Alexander R Lyon, John J V McMurray, Alexandre Mebazaa, Richard Mindham, Claudio Muneretto, Massimo Francesco Piepoli, Susanna Price, Giuseppe M C Rosano, Frank Ruschitzka, Anne Kathrine Skibelund, ESC Scientific Document Group , 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC, European Heart Journal, Volume 44, Issue 37, 1 October 2023, Pages 3627–3639, https://doi.org/10.1093/eurheartj/ehad195.

[37] Marx N, Federici M, Schütt K, Müller-Wieland D, Ajjan RA, Antunes MJ, Christodorescu RM, Crawford C, Di Angelantonio E, Eliasson B, Espinola-Klein C, Fauchier L, Halle M, Herrington WG, Kautzky-Willer A, Lambrinou E, Lesiak M, Lettino M, McGuire DK, Mullens W, Rocca B, Sattar N; ESC Scientific Document Group. 2023 ESC Guidelines for the management of cardiovascular disease in patients with diabetes. Eur Heart J. 2023 Oct 14;44(39):4043-4140. doi: 10.1093/eurheartj/ehad192. Erratum in: Eur Heart J. 2023 Dec 21;44(48):5060. doi: 10.1093/eurheartj/ehad774. Erratum in: Eur Heart J. 2024 Feb 16;45(7):518. doi: 10.1093/eurheartj/ehad857.

[38] J.A. Ligibel, K. Bohlke, A.M. May, S.K. Clinton, W. Demark-Wahnefried, S. C. Gilchrist, M.L. Irwin, M. Late, S. Mansfield, T.F. Marshall, J.A. Meyerhardt, C. A. Thomson, W.A. Wood, C.M. Alfano, Exercise, Diet, and Weight Management During Cancer Treatment: ASCO Guideline, J. Clin. Oncol. 40 (22) (2022) 2491–2507, https://doi.org/10.1200/JCO.22.00687.

[39] C.A. Elmets, N.J. Korman, E.F. Prater, E.B. Wong, R.N. Rupani, D. Kivelevitch, A. W. Armstrong, C. Connor, K.M. Cordoro, D.M.R. Davis, B.E. Elewski, J.M. Gelfand, K.B. Gordon, A.B. Gottlieb, D.H. Kaplan, A. Kavanaugh, M. Kiselica, D. Kroshinsky, M. Lebwohl, C.L. Leonardi, J. Lichten, H.W. Lim, N.N. Mehta, A.S. Paller, S. L. Parra, A.L. Pathy, M. Siegel, B. Stoff, B. Strober, J.J. Wu, V. Hariharan, A. Menter, Joint AAD-NPF Guidelines of care for the management and treatment of psoriasis with topical therapy and alternative medicine modalities for psoriasis severity measures, J. Am. Acad Dermatol. 84 (2) (2021) 432–470, https://doi.org/10.1016/j.jaad.2020.07.087.

[40] Datayze. Readability analyzer. Availavle at: https://datayze.com/readability-analyzer (Access date: 23.01.2024).

[41] A.D. Rouhi, Y.K. Ghanem, L. Yolchieva, Z. Saleh, H. Joshi, M.C. Moccia, A. Suarez-Pierre, J.J. Han, Can Artificial Intelligence Improve the Readability of Patient Education Materials on Aortic Stenosis? A Pilot Study. Cardiol Ther. 13 (1) (2024 Mar) 137–147, https://doi.org/10.1007/s40119-023-00347-0.

[42] Z.E.E. Gbedemah, M.N. Fuseini, S.K.E.J. Fordjuor, E.J. Baisie-Nkrumah, R.E. M. Beecham, K.N. Amissah-Arthur, Readability and Quality of Online Information on Sickle Cell Retinopathy for Patients, Am. J. Ophthalmol. 259 (2024) 45–52, https://doi.org/10.1016/j.ajo.2023.10.023.

[43] S. Marshall, S.J. Hanish, J. Baumann, A. Groneck, S. DeFroda, A standardised method for improving patient education material readability for orthopaedic trauma patients, Musculoskeletal Care 22 (1) (2024) e1869.

[44] Tavernier, J., Bellot, P. (2012). Flesch and Dale-Chall Readability Measures for INEX 2011 Question-Answering Track. In: Geva, S., Kamps, J., Schenkel, R. (eds) Focused Retrieval of Content and Structure. INEX 2011. Lecture Notes in Computer Science, vol 7424. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35734-3_22.

[45] O. Asupoto, S. Anwar, A.G. Wurcel, A health literacy analysis of online patient-directed educational materials about mycobacterium avium complex, J. Clin. Tuberc Other Mycobact Dis. 3 (35) (2024) 100424, https://doi.org/10.1016/j.jctube.2024.100424.

[46] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics Bulletin 33 (1) (1977) 159–174, https://doi.org/10.2307/2529310.

[47] I. Altamimi, A. Altamimi, A.S. Alhumimidi, A. Altamimi, M.H. Temsah, Artificial Intelligence (AI) Chatbots in Medicine: A Supplement, Not a Substitute, Cureus. 15 (6) (2023) e40922.

[48] J.R.E. Neo, J.S. Ser, S.S. Tay, Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers, Front Digit Health. 9 (6) (2024) 1395501, https://doi.org/10.3389/fdgth.2024.1395501.

[49] M. Alkhalaf, P. Yu, M. Yin, C. Deng, Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records, Journal of Biomedical Informatics 14 (156) (2024) 104662, https://doi.org/10.1016/j.jbi.2024.104662.