Chatbots in healthcare: A study of readability and response accuracy in answers to questions about hypertension.

Robert Olszewski, Jakub Brzeziński, Klaudia Watros, Małgorzata Mańczak, Jakub Owoc, Krzysztof Jeziorski

Submitted to: Journal of Medical Internet Research
on: October 23, 2024

Table of Contents

Chatbots in healthcare: A study of readability and response accuracy in answers to questions about hypertension.

Robert Olszewski[1, 2]; Jakub Brzezi?ski[1]; Klaudia Watros[1]; Ma?gorzata Ma?czak[1]; Jakub Owoc[1]; Krzysztof Jeziorski[1, 3]

[1]Department of Gerontology, Public Health National Institute of Geriatrics, Rheumatology and Rehabilitation Warsaw PL
[2]Department of Ultrasound Institute of Fundamental Technological Research, Polish Academy of Sciences Warsaw PL
[3]Maria Sklodowska-Curie National Research Institute of Oncology Warsaw PL

Corresponding Author:
Jakub Brzezi?ski
Department of Gerontology, Public Health
National Institute of Geriatrics, Rheumatology and Rehabilitation
Sparta?ska 1
Warsaw
PL

Abstract

Background: AI-powered chatbots, using Large Language Models, may effectively answer questions from patients with hypertension, providing responses that are accurate, empathetic, and easy to read.

Objective: This study evaluates the performance of three such chatbots in delivering quality responses.

Methods: One hundred questions were randomly selected from the Reddit forum r/hypertension and submitted to three publicly available chatbots (ChatGPT-3.5, Microsoft Copilot, Gemini), anonymized as A, B, and C. Two independent medical professionals assessed the accuracy and empathy of their responses using Likert scales. Additionally, 300 responses were analyzed with the WebFX readability tool to measure various readability indices.

Results: In total, 300 responses were evaluated. Chatbot A generated the most extensive responses, with an average of 13 sentences per reply, while Chatbot B had the shortest replies. Chatbot C achieved the highest score on the Flesch Reading Ease Scale, indicating better readability, while Chatbot A scored the lowest. Other readability metrics, including the Flesch-Kincaid Grade Level, Gunning Fog Score, and others, also showed significant differences among the chatbots, reflecting variability in readability.

Conclusions: The study indicates that while all chatbots can produce professional responses, their readability varies significantly. These findings underscore the potential of AI chatbots in patient education. However, they also highlight the urgent need for further optimization to enhance the comprehensibility of their outputs.

(JMIR Preprints 23/10/2024:67879)
DOI: https://doi.org/10.2196/preprints.67879

Preprint Settings

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

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

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Chatbots in healthcare: A study of readability and response accuracy in answers to questions about hypertension.

Assoc. Prof. Robert Olszewski. MD, PhD, FESC[1,2], Jakub Brzeziński MA[1,*], Klaudia Watros MPh[1], Małgorzata Mańczak PhD[1], Jakub Owoc PhD[1], Assoc. Prof. Krzysztof Jeziorski MD, PhD [1,3],

1 Department of Gerontology, Public Health and Didactics, National Institute of Geriatrics, Rheumatology and Rehabilitation, Spartańska 1 Street, Warsaw, Poland (robert.olszewski@spartanska.pl, jakub.brzezinski@spartanska.pl, m.manczak@op.pl, kowoc@wp.pl, krzysztof.jeziorski@spartanska.pl)

2 Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B Street, Warsaw, Poland

3 Maria Sklodowska-Curie National Research Institute of Oncology, Roentgena 5 Street, Warsaw, Poland.

* Corresponding author: jakub.brzezinski@spartanska.pl, Spartanska 1, 02-637, Warsaw, Poland, Department of Gerontology, Public Health and Didactics, National Institute of Geriatrics, National Institute of Geriatrics, Rheumatology and Rehabilitation, 22 670 92 62

Word count: 3121

Keywords: Hypertension, artificial intelligence, readability, accuracy, empathy.

Background: AI-powered chatbots, using Large Language Models, may effectively answer questions from patients with hypertension, providing responses that are accurate, empathetic, and easy to read. This study evaluates the performance of three such chatbots in delivering quality responses.

Methods: One hundred questions were randomly selected from the Reddit forum r/hypertension and submitted to three publicly available chatbots (ChatGPT-3.5, Microsoft Copilot, Gemini), anonymized as A, B, and C. Two independent medical professionals assessed the accuracy and empathy of their responses using Likert scales. Additionally, 300 responses were analyzed with the WebFX readability tool to measure various readability indices.

Results: In total, 300 responses were evaluated. Chatbot A generated the most extensive responses, with an average of 13 sentences per reply, while Chatbot B had the shortest replies. Chatbot C achieved the highest score on the Flesch Reading Ease Scale, indicating better readability, while Chatbot A scored the lowest. Other readability metrics, including the

Flesch-Kincaid Grade Level, Gunning Fog Score, and others, also showed significant differences among the chatbots, reflecting variability in readability.
Conclusions: The study indicates that while all chatbots can produce professional responses, their readability varies significantly. These findings underscore the potential of AI chatbots in patient education. However, they also highlight the urgent need for further optimization to enhance the comprehensibility of their outputs.

1. Introduction

Chatbots powered by Large Language Models (LLMs) can answer questions across a wide range of topics in a way that closely resembles human responses [1]. These chatbots can provide answers that reflect current knowledge and trends by harnessing access to constantly updated internet sources [2].
LLMs have revolutionized the field of conversational AI, enabling chatbots to interact more naturally and effectively. These models are trained on vast and diverse datasets, allowing them to generate responses that are contextually appropriate and engaging, making conversations feel more authentic and human-like [3]. Examples of such models include OpenAI's ChatGPT and GPT-4, which are highly advanced in natural language processing (NLP) capabilities and extend their use cases far beyond traditional chatbot applications.
Due to their proficiency in understanding and responding to complex queries, LLMs have become indispensable tools in various sectors. They can assist with tutoring, content creation, and personalized learning experiences in education. In research, they aid in data analysis, literature review, and hypothesis generation. In healthcare, they support diagnostics, patient communication, and the dissemination of medical information [4]. The ability of these models to provide accurate and nuanced information in real time has broadened their impact, making them integral to modern digital communication and problem-solving tools.
Due to staffing shortages within the healthcare system, chatbots may be a viable alternative for diagnosing or educating patients about diseases, particularly chronic conditions most prevalent in the community, such as hypertension, obesity, and others [5].
Hypertension is a widespread condition affecting approximately 1.28 billion adults aged 30–79 worldwide [6]. Two-thirds of these individuals live in low- and middle-income countries. Concerningly, an estimated 46% of adults with hypertension are unaware of their condition, and less than half (42%) receive proper diagnosis and treatment [7]. Hypertension is often referred to as ‘the silent killer’ because it does not cause noticeable symptoms. However, if left untreated, it can lead to

serious health issues such as heart disease, stroke, and kidney problems [8]. As we move toward 2040, it is projected that around 61% of the adult population in the United States will have hypertension [9].

A recent study demonstrated that non-physician healthcare professionals who lacked formal medical or related training but received specialized training in hypertension management achieved outstanding patient outcomes. This finding underscores the potential of task-shifting strategies in healthcare, where non-physician personnel can effectively manage chronic conditions such as hypertension, thereby improving patient health metrics and alleviating the burden on traditional healthcare providers [10]. The initial studies have already been published, suggesting that chatbots can be a valuable tool in combating hypertension. In the study conducted by Griffin et al., the objective was to design, develop, and assess the usability of a prototype chatbot for self-management of hypertension. The ready-to-use chatbot was subsequently introduced to participants, who were individuals aged 18 and older with diagnosed hypertension and prescribed medication. Overall, the chatbot received positive evaluations in tasks related to blood pressure control and medication usage, with participants describing it as beneficial [11]. The accurate evaluation of chatbot responses related to hypertension is crucial, especially for patients seeking reliable information online. Therefore, our study aims to evaluate the readability of chatbot responses using standardized scales and to assess the quality and level of empathy in the responses as judged by medical professionals.

2. Material and methods

For our study, we collected 100 random questions from the publicly accessible Reddit forum on hypertension (r/hypertension) from January 1st to July 31st 2024. The questions were posed to chatbots between August 1st and August 5th, 2024. This Reddit forum has 15,000 members who can ask questions or share hypertension-related stories [12]. Examples of this forum's utilization in scientific publications are demonstrated in the works of Nobles et al. and Zúñiga Salazar et al. [13,14].

Every 100 questions were posed to chatbots, including Microsoft Copilot, Google Gemini, and ChatGPT-3.5, blinded by letters (A, B, and C). Access to these chatbots was free of charge. After each question, we refreshed the conversation to avoid potential biases. The list of questions can be found in the Online-Only Supplementary. Once a response was generated, we copied it and assessed its readability using the webfx.com tool [15]. This tool evaluates the text based on various readability scales, including the Flesch Kincaid Reading Ease Score (FRES) [16], Flesch Kincaid Grade Level (FKGL) [17], Gunning Fog Score (GFS) [18], Smog Index (SI) [19], Coleman Liau Index (CLI) [20], and Automated Readability Index (ARI) [21]. Additionally, we calculated the average number of sentences per response.

Two independent medical professionals (physicians) subsequently evaluated the accuracy of the responses in a blinded manner using a 5-point Likert scale (ranging from 'very poor' to 'very good') as well as the level of empathy (ranging from 'not empathetic' to 'very empathetic') [22,23].

Building on the relevant sources from the He et al. article, our research team undertook a comprehensive breakdown of the questions into distinct categories and subcategories. This meticulous process led to the identification of ten primary categories of questions. From these main categories, we further delineated subcategories to provide a more granular analysis. Additionally, we calculated the percentage distribution of topics within each subcategory, offering a detailed overview of the thematic composition of the questions. This methodological approach, characterized by its thoroughness, allowed us to systematically categorize and quantify the various topics, thereby enhancing the precision and depth of our analysis [24].

2.1. Statistical analyses

The normality of the distribution of the analyzed variables was verified using the Shapiro-Wilk test, a significant step that informed our subsequent analyses. The distributions of most of the variables deviated significantly from the normal distribution; therefore, the median and IQR were used. The

Friedman ANOVA and Wilcoxon signed-rank tests were used to compare the chatbots' responses. P-value <0.017 was considered statistically significant, according to the Bonferroni correction (0.05/3=0.017). Cohen's kappa coefficient and percentage agreement of ratings were calculated to assess the agreement between experts. The interpretation of Cohen's kappa values was taken from [24]: 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 agreement ($\kappa = 0.81$ to 1.00) [25].

3. Results

A total of 300 questions were posed to all chatbots, and we received an equal number of responses. This indicates that the chatbots had no difficulty answering questions related to hypertension. The readability indicator values for all scales significantly differed among chatbots (Table 1).

Our findings indicate that Chatbot A consistently produces the most extended responses compared to other chatbots. Utilizing the Flesch-Kincaid Reading Ease scale, it is evident that all chatbot responses are crafted with advanced language. Notably, Chatbot's A responses are the most challenging to comprehend. The Flesch-Kincaid Grade Level results further reveal that Chatbot's A responses are the most sophisticated, employing language typically associated with college-level writing.

Chatbot B and Chatbot C achieved identical scores regarding the Gunning Fog Score and SMOG Index. However, Chatbot A attained the highest scores again, underscoring its propensity to generate highly advanced responses. The Coleman-Liau Index and Automated Readability Index scores also corroborate the high reading comprehension level required for Chatbot's A responses, highlighting their complexity and advanced nature.

Tab. 1 Readability comparison.

	Chatbot A	Chatbot B	Chatbot C	P
Number of sentences	13 (9 – 19)	8 (6 – 11)	10 (7 – 14)	<0.001
Flesch Kincaid Reading Ease	34.3 (26.1 – 40.9)	43.7 (35.3 – 52.4)	48.1 (38.3 – 57.5)	<0.001
Flesch Kincaid Grade Level	13.3 (12.1 – 15.0)	11.3 (10.2 – 12.5)	11.1 (9.3 – 12.7)	<0.001
Gunning Fog Score	15.5 (14.2 – 17.5)	13.6 (12.2 – 14.9)	13.6 (11.8 – 15.0)	<0.001
SMOG Index	12.0 (10.9 – 13.5)	10.3 (9.3 – 11.4)	10.3 (8.8 – 11.4)	<0.001
Coleman Liau Index	17.0 (14.9 – 18.2)	16.0 (13.9 – 17.4)	14.6 (12.8 – 16.9)	<0.001
Automated Readability Index	14.4 (12.9 – 16.0)	12.3 (10.7 – 13.8)	11.9 (10.1 – 13.6)	<0.001

The p-value in Table 1 indicates that the chatbots differ. Specifically, at least one pair of analyzed measurements shows significant differences. We compared the chatbots in pairs to identify which pair is significantly different. Table 2 presents the differences between pairs of chatbots. The results demonstrate that the differences between Chatbot A and Chatbot B, as well as between Chatbot A and Chatbot C, are statistically significant across all evaluated scales. In contrast, the statistically significant difference between Chatbot B and Chatbot C was observed in the length of their responses, Flesch Kincaid Reading Ease and Coleman Liau Index.

Tab. 2 A comparison of the differences between pairs of chatbots.

	P Chatbot A vs Chatbot B	P Chatbot A vs Chatbot C	P Chatbot B vs Chatbot C
Number of sentences	**<0.001**	**<0.001**	**<0.001**
Flesch Kincaid Reading Ease	**<0.001**	**<0.001**	**0.006**
Flesch Kincaid Grade Level	**<0.001**	**<0.001**	0.207
Gunning Fog Score	**<0.001**	**<0.001**	0.259
Smog Index	**<0.001**	**<0.001**	0.545
Coleman Liau Index	**<0.001**	**<0.001**	**0.001**
Automated Readability Index	**<0.001**	**<0.001**	0.271

The questions posed by Reddit users were analyzed in terms of the topics they addressed. Figure 1 presents a pie chart illustrating the distribution of question categories and subcategories. Table 3 shows the breakdown of categories and subcategories of questions asked by Reddit users. Based on a comprehensive analysis using the readability metrics, the chatbots' responses exhibited a high level of complexity, characteristic of academic discourse. These responses may present comprehension challenges for individuals who have a tertiary education. Figures 2-8 (available in supplementary) present a comparative analysis of chatbot responses, encompassing the number of sentences generated and corresponding readability scores.

Tab 3. Frequency of question categories

Main category	Subcategory	Frequency
Physical activity	Leisure activity	75%
Physical activity	Work activity	25%
Diet	Supplements intake	31%
Diet	Caffeine intake	23%
Diet	Eating meat	23%
Diet	Energy drinks	15%
Diet	Alcohol	8%
Laboratory test	Laboratory test	100%
Diagnosed with hypertension	BP values	45%
Diagnosed with hypertension	BP measure technique	22%

Diagnosed with hypertension	Causes of hypertension	22%
Diagnosed with hypertension	Diagnosis and previous activity	11%
Managing chronic hypertension	High blood pressure and medication	66%
Managing chronic hypertension	Hypertension and pregnancy	17%
Managing chronic hypertension	Hypertension and diet	17%
Isolate diastolic hypertension	Isolated diastolic hypertension	100%
Side effects	Angiotensin-converting enzyme side effects	31%
Side effects	Other	29%
Side effects	Weights and hypertension	8%
Side effects	Antibiotics and hypertension	8%
Side effects	BP medicaments and heat intolerance	8%
Side effects	Hypertension medicaments and heart health	8%
Side effects	Beta blocker and fatigue	8%
Symptoms	Chest tightness, pain	15%
Symptoms	Symptoms during activity	15%
Symptoms	Symptoms during sleep	15%
Symptoms	Other	15%
Symptoms	Black out, fainting	8%
Symptoms	Anxiety	8%
Symptoms	Symptoms and time of the day	8%
Symptoms	Hydration and high blood pressure	8%
Symptoms	Eyes symptoms	8%
Treatment	Other	23%
Treatment	Mistakes in taking medication	18%
Treatment	Angiotensin-converting enzyme (ACE) inhibitor	14%
Treatment	Calcium channel blocker	9%
Treatment	Sartans	9%
Treatment	Medication and lifestyle changes	4,50%
Treatment	Herbal medicine	4,50%
Treatment	Beta blockers	4,50%
Treatment	Diuretics	4,50%
Treatment	Discontinuation of medication	4,50%
Treatment	Supplements and treatment	4,50%
Other	Causes of heart disease	21,50%
Other	BP measurement worries	21,50%
Other	Hypertension life influence	11%
Other	Heart disease symptoms and other symptoms	11%
Other	Other	11%
Other	Vaping and heart disease	6%
Other	Heart disease symptoms and other symptoms	6%
Other	Oxygen therapy in hypertension	6%
Other	Surgery	6%

Fig. 1 Frequency of questions across 10 categories and their respective subcategories.

Additionally, our study evaluated the quality and empathy of the responses generated by the chatbots. Two independent medical professionals conducted the evaluations. Table 4 presents the results of the comparative analysis of the responses. The results demonstrate that the responses are high quality and exhibit a substantial degree of empathy. However, the inter-rater agreement is low: kappa values do not exceed 0.30 for quality and are close to 0.00 for empathy. Percentage agreement ranges from 42% to 74% for response quality and 32% to 40% for empathy. Notably, both quality and empathy received the highest ratings for responses generated by Chatbot A, while the lowest ratings were observed for responses generated by Chatbot B. Additionally, the supplementary materials of the article include tables and calculations related to the evaluations of responses made by two independent medical professionals.

Tab 4. Analyzes of comparative assessments by medical professionals

	Accuracy

	Expert 1 Me (IQR)	Expert 2 Me (IQR)	Cohen kappa	Percentage agreement
Chatbot A	5 (5 – 5)	5 (5 – 5)	0.18 (0.00 – 0.45)	74 %
Chatbot B	4 (4 – 5)	4 (4 – 5)	0.03 (0.00 – 0.19)	42 %
Chatbot C	5 (4 – 5)	4 (4 – 5)	0.30 (0.13 – 0.47)	61 %
	Empathy			
	Expert 1 Me (IQR)	Expert 2 Me (IQR)	Cohen kappa	Percentage agreement
Chatbot A	5 (5 – 5)	4 (3 – 5)	0.08 (0.00 – 0.23)	40%
Chatbot B	4 (4 – 5)	4 (3 – 5)	-0.03 (-0.16 – 0.10)	32%
Chatbot C	4.5 (4 – 5)	4 (4 – 5)	0.04 (0.00 – 0.20)	40%

4. Discussion

The study's findings indicate that chatbots' responses to identical questions exhibit varied sentence length and readability scores. Notably, the chatbots did not reference any medical guidelines or scientific publications. Additionally, they included extraneous information in their responses, leading to variations in sentence length across different chatbots. Consequently, these differences in sentence length contributed to the observed discrepancies in readability scores. According to the American Medical Association (AMA) and the National Institutes of Health (NIH), medical texts should be composed at a reading level corresponding to grades 6 to 8, as measured by the Flesch-Kincaid Grade Level (FKGL) scale. In our study, the average FKGL score ranged from 11.1 to 13.3, indicating that the text was produced at an academic level [26,27,28].

Scientific publications by other authors also indicate that the readability of chatbot responses varies and is generally produced at a higher understanding level. In the study conducted by Chen et al., the responses provided by chatbots to cancer-related inquiries were superior to those given by humans in terms of quality and empathy [29]. Conversely, the Flesch-Kincaid Grade Level scale revealed that the responses from both humans and chatbots were produced at an academic level. A study conducted by Mishra et al. compared five readability scales for 18 government documents on COVID-19. The findings indicated that the documents exceeded the recommended reading level, exhibited complex syntax, and employed technical terminology [30]. In the study conducted by Olszewski et al., five chatbots were queried about cardiovascular diseases, oncological conditions, and psoriasis. The results varied in terms of length, quality, and readability. All readability scales used in this study indicated that the chatbot responses were generated at an academic level, which may hinder comprehension for individuals with lower levels of education. However, the reliability of the responses was consistently high [31].

The study by Ahmed et al. evaluated the efficacy and readability of ChatGPT's responses to questions about pediatric ophthalmology and strabismus. The study found that nearly 80% of the responses were correct according to the assessment of two experienced pediatric ophthalmologists. The Flesch-Kincaid Grade Level averaged 14,49, indicating high text readability [32]. In the study conducted by Cao et al., the accuracy, reliability, and readability of responses to questions regarding liver cancer surveillance, diagnosis, and management were evaluated. The researchers utilized three chatbots: ChatGPT-3.5, Gemini, and Bing. The readability of the responses was assessed using the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level. Six liver transplant physicians evaluated the accuracy and reliability of the responses. Each chatbot was given 20 questions to answer. ChatGPT accurately and reliably answered only six questions, Gemini answered eight, and Bing answered three. The readability of the chatbot responses was at a high academic level [33].

Cohen et al. evaluated the responses of Google's search engine and ChatGPT regarding cataracts and surgery. They posed 20 questions to each and used five readability scales. ChatGPT's responses were significantly more prolonged and written at a higher reading level, with an average of 14.8. Google's

responses contained incorrect or inappropriate material in 27% of cases, compared to 6% of responses generated by the large language model (LLM). Reviewers preferred ChatGPT's responses in 66% of direct comparisons [34].

Our study also demonstrated that different chatbots' responses vary in length, readability scales, and assessments of quality and empathy. The readability scale indicated that chatbots generate text at a highly advanced academic level. This could pose a problem for individuals with lower educational backgrounds, for whom understanding specialized language is challenging. Regarding the quality and empathy of responses, two independent medical professionals found that the chatbot responses were accurate and, in most cases, empathetic. In the study by Kim et al., the responses of chatbots regarding brain tumours were evaluated, focusing on the empathy of their answers. Three chatbots were used in the study. The results showed significant differences in the empathy levels of the chatbot responses. ChatGPT scored the lowest with 75 out of 105 points, while Bing scored the highest with 86 out of 105 points [35]. Chen et al. evaluated the quality of responses and the level of empathy. They found the chatbot responses were of higher qualitative value (mean, 3.56 [95% CI, 3.48-3.63] vs 3.00 [95% CI, 2.91-3.09]; $P < .001$) compared to those of physicians. Additionally, chatbot responses demonstrated greater empathy (mean, 3.62 [95% CI, 3.53-3.70] vs. 2.43 [95% CI, 2.32-2.53]; $P < .001$) [29].

In a study by Ayres et al., responses by physicians and ChatGPT to 195 questions from internet users were compared. The study found that ChatGPT generated high-quality responses and highly empathetic ones, which were not significantly different from those of physicians. This highlights the potential of ChatGPT to generate patient responses that a physician could edit, instilling optimism about the capabilities of AI in healthcare [36].

Chatbots are also studied for the accuracy of their responses to questions from doctors or patients and their potential to make medical diagnoses. In a study by Yeung et al., a general-purpose chatbot (ChatGPT) was compared with a medical-specific chatbot (GatorTron). The study involved presenting clinical symptoms to the chatbots and asking for possible diagnoses. The results showed that LLM-based chatbots are unsuitable for generating medical diagnoses due to biases and a tendency to produce unreliable or unverified information [37].

Our analysis reveals differences in the accuracy and empathy ratings of chatbot responses. Chatbot C and, to a greater extent, Chatbot B exhibit lower scores in quality and empathy than those reported in other studies. While Chatbot A's quality score remains high, its empathy score is notably low. These findings suggest variability in the performance of chatbots across different dimensions of response quality and empathy. In a Neo JRE et al. study, medical professionals rated chatbot responses as satisfactory, with ChatGPT achieving a score of 65.8% and Google Bard 75.8% [38]. In a separate study by Suarez A. et al., oral surgeons evaluated the responses of ChatGPT-4, with the average rating of the chatbot's complete responses being 71.7%. These findings highlight the varying satisfaction levels among healthcare professionals with chatbot-generated responses across different medical specialties [39].

This study demonstrates the potential benefits and limitations of using chatbots powered by Large Language Models (LLMs) in providing medical information, specifically in the context of hypertension. While the chatbots showed a capacity for generating accurate and empathetic responses, the advanced readability level of their answers may limit accessibility for a broader audience, particularly patients with lower health literacy. Future research should focus on optimizing the readability of chatbot responses and ensuring their comprehensibility across different user groups. Additionally, the development of adaptive language models that tailor responses based on the user's background could significantly enhance the effectiveness of chatbots in healthcare settings. Exploring these directions will be crucial in realizing the full potential of chatbots as reliable and inclusive sources of medical information.

5. Strengths and limitation

Our study assessing chatbot responses has several strengths. The first strength is the inclusion of three of the most well-known and widely available chatbots on the market, which, due to frequent updates, have access to a larger dataset and, consequently, provide more accurate responses. The second strength of our study is using six readability scales, allowing for a more detailed evaluation of the readability of chatbot responses. Another strength is the number and origin of the questions posed to the chatbots. One hundred questions submitted by internet users representing potential patients can serve as a practical test for chatbots as possible assistance for other patients or healthcare professionals. The current study has several limitations that should be acknowledged. Firstly, the chatbots used in this study were free versions, which can be considered both an advantage and a disadvantage. They typically operate on less advanced language models than their paid counterparts. This discrepancy in language model sophistication could affect the quality and accuracy of the chatbot responses, thereby influencing the study's outcomes. Secondly, the chatbot responses were evaluated exclusively by medical professionals. While these experts possess a deep understanding of medical terminology and concepts, their interpretations may differ significantly from those of patients. Patients who do not have a medical background might need clarification on specific medical terms and explanations. This difference in comprehension could lead to a disparity in how different user groups perceive and understand the chatbot responses. To enhance the robustness of future studies, it would be beneficial to include a diverse group of evaluators, encompassing medical professionals and patients. This approach would provide a more comprehensive assessment of the chatbot's effectiveness in communicating medical information to a broader audience. Employing more advanced, paid chatbots could improve response quality, yielding more reliable and generalizable results.

6. Conclusions

In conclusion, our study has shown that chatbots have the potential to be a valuable source of information on hypertension. The three chatbots we evaluated in this study provided accurate and empathetic responses, each with a high academic level of readability. However, these variations in readability are significant, as they can impact the comprehensibility of the information for different user groups. While high readability levels are suitable for medical professionals, they can pose challenges for laypersons who may need to become more familiar with medical terminology. This discrepancy underscores the need for future research to evaluate and optimize the readability of chatbot responses. Such research is crucial to ensure that the information is accessible and understandable to a broader audience, including patients.

Moreover, future studies should explore developing and implementing adaptive language models that tailor responses based on the user's background and comprehension level. This approach could enhance the effectiveness of chatbots in disseminating medical information, making it more inclusive and user-friendly. Addressing these aspects can improve chatbots' utility as reliable health information sources for diverse populations.

Acknowledgments: Not applicable

Author contribution

RO: Conceptualization, Methodology, Investigation, Resources, Writing – Original Draft, Writing – Review & Editing. **JB:** Conceptualization, Methodology, Writing - Original Draft, Writing – Review & Editing, Supervision. **KW:** Investigation, Resources, Writing – Original Draft. **MM:** Methodology, Formal analysis, Writing – Original Draft. **JO:** Methodology, Writing – Original Draft, Writing – Review & Editing. **KJ**: Supervision, Resources, Writing – Review & Editing.

Availability of data and materials

The datasets during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare that there is no conflict of interest with any financial organization regarding the

material discussed in the manuscript.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

7. References

1. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, Wong WM, Wang X, Jin Tan MC, Chang Koh VT, Tham YC. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. 2023 Oct 10;26(11):108163. doi: 10.1016/j.isci.2023.108163.
2. Maia E, Vieira P, Praça I. Empowering Preventive Care with GECA Chatbot. *Healthcare (Basel)*. 2023 Sep 13;11(18):2532. doi: 10.3390/healthcare11182532.
3. Izadi S, Forouzanfar M. Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots. *AI*. 2024; 5(2):803-841. https://doi.org/10.3390/ai5020041.
4. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*. 2023; 15(6):192. https://doi.org/10.3390/fi15060192.
5. Alowais, S.A., Alghamdi, S.S., Alsuhebany, N. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 23, 689 (2023). https://doi.org/10.1186/s12909-023-04698-z
6. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. *Nat Rev Nephrol*. 2020 Apr;16(4):223-237. doi: 10.1038/s41581-019-0244-2.
7. Farhadi, F., Aliyari, R., Ebrahimi, H. et al. Prevalence of uncontrolled hypertension and its associated factors in 50–74 years old Iranian adults: a population-based study. *BMC Cardiovasc Disord* 23, 318 (2023). https://doi.org/10.1186/s12872-023-03357-x
8. Atibila F, Ten Hoor G, Donkoh ET, Kok G. Challenges experienced by patients with hypertension in Ghana: A qualitative inquiry. *PLoS One*. 2021 May 6;16(5):e0250355. doi: 10.1371/journal.pone.0250355.
9. Heaton J, Imburgio S, Mararenko A, Udongwo N, Alshami A, Schoenfeld M, Apolito R, Selan J, Sealove B, Almendral J. Potential United States Population Impact of the 2023 PREVENT Risk Calculator on Hypertension Management. *Hypertension*. 2024 Aug;81(8):e88-e90. doi: 10.1161/HYPERTENSIONAHA.124.23013.
10. Guo X, Ouyang N, Sun G, Zhang N, Li Z, Zhang X, Li G, Wang C, Qiao L, Zhou Y, Chen Z, Shi C, Liu S, Miao W, Geng D, Zhang P, Sun Y; CRHCP Study Group. Multifaceted Intensive Blood Pressure Control Model in Older and Younger Individuals With Hypertension: A Randomized Clinical Trial. *JAMA Cardiol*. 2024 Jun 18:e241449. doi: 10.1001/jamacardio.2024.1449.
11. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. *JAMIA Open*. 2023 Sep 8;6(3):ooad073. doi: 10.1093/jamiaopen/ooad073.
12. https://www.reddit.com/r/hypertension/new/ (Access: 29.07.2024).
13. Nobles AL, Leas EC, Dredze M, Ayers JW. Examining peer-to-peer and patient-provider

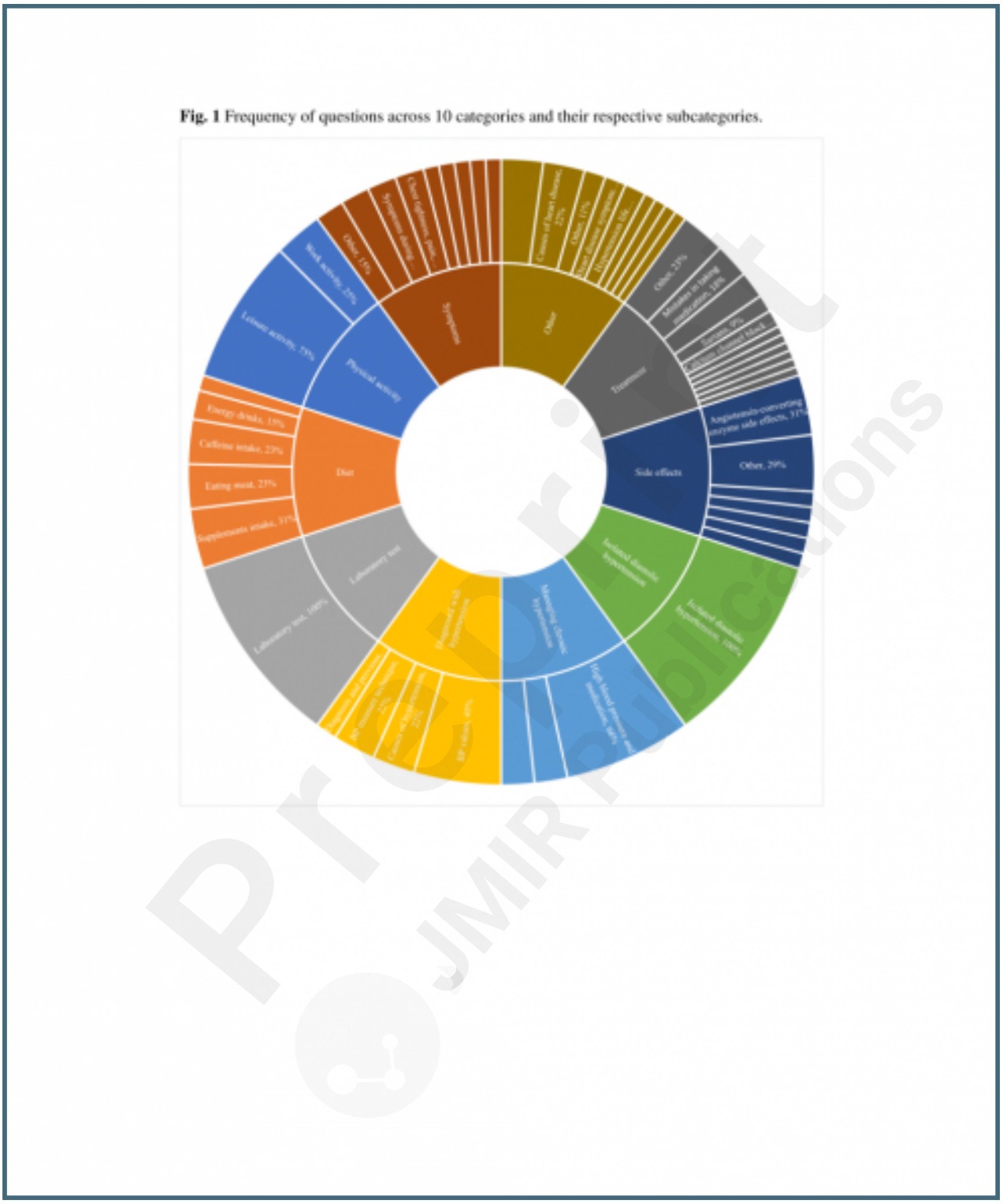
interactions on a social media community facilitating ask the doctor services. *Proc Int AAAI Conf Weblogs Soc Media*. 2020;14:464-475. doi:10. 1609/icwsm.v14i1.7315.
14. Zúñiga Salazar G, Zúñiga D, Vindel CL, Yoong AM, Hincapie S, Zúñiga AB, Zúñiga P, Salazar E, Zúñiga B. Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. *Cureus*. 2023 Sep 18;15(9):e45473. doi: 10.7759/cureus.45473.
15. https://www.webfx.com/tools/read-able/ (Access: 24.07.2024).
16. Flesch R. A new readability yardstick. *J Appl Psychol* 1948;32:221–33.
17. Kincaid P, et al. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. University of Central Florida; 1975.
18. Bogert J. In defense of the Fog Index. *Bull Assoc Bus Commun* 1985;48:9–12.
19. McLaughlin GH. SMOG grading—a new readability formula. *J Reading* 1969;12:639–46.
20. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol* 1975;60:283–4.
21. Smith EA, Senter RJ. Automated readability index. *AMRL TR* 1967:1–14.
22. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak*. 2024 Jul 29;24(1):211. doi: 10.1186/s12911-024-02619-8.
23. Small WR, Wiesenfeld B, Brandfield-Harvey B, Jonassen Z, Mandal S, Stevens ER, Major VJ, Lostraglio E, Szerencsy A, Jones S, Aphinyanaphongs Y, Johnson SB, Nov O, Mann D. Large Language Model-Based Responses to Patients' In-Basket Messages. *JAMA Netw Open*. 2024 Jul 1;7(7):e2422399. doi: 10.1001/jamanetworkopen.2024.22399.
24. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, Arguello LG, Murray P, Lu Z. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *J Med Internet Res*. 2024 Apr 17;26:e56655. doi: 10.2196/56655.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Mar;33(1):159-74. https://doi.org/10.2307/2529310
26. VanderWeele TJ, Mathur MB. SOME DESIRABLE PROPERTIES OF THE BONFERRONI CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? *Am J Epidemiol*. 2019 Mar 1;188(3):617-618. doi: 10.1093/aje/kwy250.
27. Bange M, Huh E, Novin SA, Hui FK, Yi PH. Readability of Patient Education Materials From RadiologyInfo.org: Has There Been Progress Over the Past 5 Years? *AJR Am J Roentgenol*. 2019 Oct;213(4):875-879. doi: 10.2214/AJR.18.21047.
28. Hansberry DR, John A, John E, Agarwal N, Gonzales SF, Baker SR. A critical review of the readability of online patient education resources from RadiologyInfo.Org. *AJR Am J Roentgenol*. 2014 Mar;202(3):566-75. doi: 10.2214/AJR.13.11223.
29. Chen D, Parsa R, Hope A, et al. Physician and Artificial Intelligence Chatbot Responses to Cancer Questions From Social Media. *JAMA Oncol*. 2024;10(7):956–960. doi:10.1001/jamaoncol.2024.0836.
30. Mishra V, Dexter JP. Comparison of Readability of Official Public Health Information About COVID-19 on Websites of International Agencies and the Governments of 15 Countries. *JAMA Netw Open*. 2020 Aug 3;3(8):e2018033. doi: 10.1001/jamanetworkopen.2020.18033.
31. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: A comparative study. *Int J Med Inform*. 2024 Jul 19;190:105562. doi: 10.1016/j.ijmedinf.2024.105562.
32. Ahmed HS, Thrishulamurthy CJ. Evaluating ChatGPT's efficacy and readability to common pediatric ophthalmology and strabismus-related questions. *Eur J Ophthalmol*. 2024 Aug

7:11206721241272251. doi: 10.1177/11206721241272251.
33. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, Kamaya A, Tse JR. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: accuracy, reliability, readability. *Abdom Radiol* (NY). 2024 Aug 1. doi: 10.1007/s00261-024-04501-7.
34. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: Exploring the Use of Artificial Intelligence in Ophthalmology by Comparing the Accuracy, Safety, and Readability of Responses to Frequently Asked Patient Questions Regarding Cataracts and Cataract Surgery. *Semin Ophthalmol*. 2024 Aug;39(6):472-479. doi: 10.1080/08820538.2024.2326058.
35. Kim YI, Kim KH, Oh HJ, Seo Y, Kwon SM, Sung KS, Chong K, Lee MH. Assessing the Suitability of Artificial Intelligence-Based Chatbots as Counseling Agents for Patients with Brain Tumor: A Comprehensive Survey Analysis. *World Neurosurg*. 2024 Jul;187:e963-e981. doi: 10.1016/j.wneu.2024.05.023.
36. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023 Jun 1;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838.
37. Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, Teo JT. AI chatbots not yet ready for clinical use. *Front Digit Health*. 2023 Apr 12;5:1161098. doi: 10.3389/fdgth.2023.1161098.
38. 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.
39. A. Suarez, J. Jimenez, M. Llorente de Pedro, C. Andreu-Vazquez, V. Díaz-Flores García, M. Gomez Sanchez, 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.

Supplementary Files

Figures

Frequency of questions across 10 categories and their respective subcategories.



Multimedia Appendixes

Material showing charts not used in the manuscript and a list of all questions.
URL: http://asset.jmir.pub/assets/bf45b9fbac35ee3e1be61a3ce41546a2.doc