

# Effectiveness of artificial intelligence-driven chatbot responses in diabetes knowledge: a readability and reliability assessment

Muhammad Thesa Ghozali<sup>1</sup>, Woro Supadmi<sup>2</sup>, Fadya Bella Suci Maharani<sup>1</sup>, Aloina Jean Rassyifa<sup>1</sup>

<sup>1</sup>Department of Pharmaceutical Management, School of Pharmacy, Faculty of Medicine and Health Sciences,  
Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

<sup>2</sup>Faculty of Pharmacy, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

## Article Info

### Article history:

Received Aug 11, 2024

Revised Feb 28, 2025

Accepted Mar 15, 2025

### Keywords:

Artificial intelligence

Chatbot

Diabetes knowledge

Natural language processing

Readability

Reliability

## ABSTRACT

Patient education is vital in diabetes management, empowering patients with necessary knowledge and skills to manage condition effectively. However, traditional educational methods often face challenges such as limited access to healthcare professionals and variability in information quality. This study aimed to assess the reliability and readability of artificial intelligence (AI)-driven chatbot responses in disseminating diabetes knowledge. Technically, the diabetes knowledge questionnaire (DKQ-24) was administered to evaluate the effectiveness of AI-driven chatbot in disseminating diabetes-related information. Responses were evaluated for reliability and quality applying the modified DISCERN (mDISCERN) scale and global quality scale (GQS), and readability was assessed using the Flesch reading ease (FRE) score, Flesch-Kincaid grade level (FKGL), gunning fog index (GFI), Coleman-Liau index (CLI), and simple measure of gobbledygook (SMOG). The mean mDISCERN score was  $31.50 \pm 2.89$ , indicating generally reliable responses. The median GQS score was 4, reflecting the high overall quality. The readability assessment revealed a mean FRE score of 66.30, indicating that the text was fairly easy to read. FKGL mean score was  $6.54 \pm 3.19$ , suggesting the text was suitable for readers at a sixth-grade level. In conclusion, AI-driven chatbot provides reliable and high-quality information on the diabetes self-management, but it requires improvements to enhance accessibility.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Muhammad Thesa Ghozali

Department of Pharmaceutical Management, School of Pharmacy, Faculty of Medicine and Health Sciences  
Universitas Muhammadiyah Yogyakarta

Brawijaya St., Geblagan, Bantul, Special Region of Yogyakarta 55183, Indonesia

Email: ghozali@umy.ac.id

## 1. INTRODUCTION

Diabetes-described as a long-term (chronic) condition characterized by the elevated levels of blood glucose-remains one of the most common public health challenges globally. According to the International Diabetes Federation, approximately 463 million adults were living with diabetes in 2019, and this number is projected to rise to 700 million by 2045 [1]. The condition not only affects the individual's health and quality of life but also substantial economic burden on healthcare systems. Effective management and prevention of the condition are crucial to mitigating all these effects and necessitate a comprehensive understanding of the disease among patients and healthcare providers. Patient education is a cornerstone in diabetes management.

It strengthens diabetic patients with the knowledge and skills required to manage its symptoms and condition effectively, adhere to treatment regimens, and make informed lifestyle choices [2]. The education

extends beyond patients to include their caregivers and general publics, supporting an environment that supports better health outcomes. A comprehensive diabetes education covers a wide range of topics, such as pathophysiology of disease, dietary management, physical activities, medication adherence, and monitoring of blood glucose levels as well. By enhancing knowledge, the patient education facilitates early detection and intervention, which are essential for preventing complications associated with diabetes.

Despite the critical role of education in diabetes management, traditional educational methods face a number of challenges. One of many obstacles is the limited access to healthcare professionals, particularly in remote and underserved areas [3]. The gap in access means that a number of diabetic patients do not receive the necessary education and support to manage their condition effectively. Additionally, the time constraints within the clinical settings often limit the amount of information that healthcare providers can deliver during consultation. Moreover, the variability in the quality and accessibility of information across different sources lead to misinformation and confusion among patients [4]. The proliferation of information available online has also contributed to the spread of inaccurate or conflicting advice while beneficial in some respects. The patients may struggle to DISCERN reliable sources from those that provide misleading information. This issue is compounded by complexity of medical information, which can be relatively difficult for individuals without a healthcare background to fully understand. Therefore, there is a requirement for consistent, accurate, and easily accessible educational resources to support diabetes management.

Artificial intelligence (AI) technology has emerged as a transformative force in the healthcare, offering innovative solutions to enhance service delivery and patient care. In particular, AI-driven chatbots have shown significant promise in addressing some of the main challenges associated with diabetes education [5]. The chatbots utilize natural language processing and machine learning algorithms to interact with users, providing instant, personalized responses to a wide range of health-related queries. One of primary advantages of AI-driven chatbots is their ability to deliver consistent and scalable information. Unlike human healthcare providers, such chatbots can operate continuously without fatigue, offering the patients immediate access to information at any time. Such technology is particularly beneficial for individuals living in remote areas or those who have limited access to the healthcare services. Furthermore, chatbots can be programmed to deliver information that adheres to established medical guidelines, therefore reducing the risk of misinformation [6]. ChatGPT®, developed by OpenAI, is an example of an AI-driven chatbot that has showed potential in various domains, including medical information dissemination. By training on vast datasets, ChatGPT can generate human-like responses and provide valuable support in education. However, its effectiveness in specific medical contexts, such as diabetes education, remains to be thoroughly evaluated.

AI-driven chatbots have shown promise not only in diabetes education but also in managing other chronic diseases such as cardiovascular conditions, asthma, and hypertension. Studies have demonstrated that these tools can assist patients in medication adherence, symptom tracking, and lifestyle modifications [7]–[9]. For instance, AI chatbots have been used to monitor blood pressure and provide tailored dietary advice for hypertensive patients [10]. The scalability and personalization of AI make it a versatile tool for addressing the unique challenges posed by various chronic diseases, further emphasizing the need for targeted research to optimize their use in specific contexts like diabetes education. A number of studies have investigated the capabilities of AI-driven chatbots in delivering healthcare information. All these studies have yielded mixed results, with some highlighting the accuracy and utility of chatbot responses, while the others have identified limitations [11]–[13]. For instance, a study of narrative review showed an AI-driven chatbot designed to answer questions related to breast cancer [14]. This review found high user satisfaction, and many have shown efficacy in improving patient-centered communication, accessibility to cancer-related information, and access to care. Currently, chatbots are mainly limited by the needs for extensive user-testing and iterative improvement before widespread implementation. Conversely, another research has highlighted that such chatbots may sometimes generate responses that lack the necessary depth or context required for complex medical queries [15]. In the context of diabetes education, it is crucial that chatbots not only deliver accurate information but also present it in a manner easily understandable and actionable for patients. The readability of information is a key factor in ensuring that patients can effectively use the knowledge provided to manage their condition.

Despite the growing interest in AI-driven chatbot topics, there is a gap in the research—specifically assessing its effectiveness in disseminating the diabetes-related knowledge. Most existing studies focused on other medical conditions or have not comprehensively evaluated the readability and reliability of AI-driven chatbot responses in the context of diabetes education. Given the unique educational needs of patients, it is essential to investigate how well AI-driven chatbots can meet these requirements. The current literature does not provide an assessment of chatbot's performance using standardized diabetes knowledge questionnaires, (i.e., diabetes knowledge questionnaire (DKQ-24)). This gap outlines the need for research that examines both accuracy and readability of AI-generated responses to ensure that they are suitable for patient education. Therefore, the main aim of this study is to assess the reliability and readability of AI-driven chatbot's (in this case, ChatGPT®) responses in disseminating diabetes knowledge using DKQ-24 questionnaire. By assessing

its performance, this study seeks to determine its potential as one of effective educational tools for diabetes self-management.

## 2. METHOD

### 2.1. Questionnaire

This study utilized the DKQ-24, a recognized instrument for assessing diabetes knowledge, including its management, complications, and prevention strategies (Table 1). The questionnaire has been extensively validated in the previous studies, demonstrating robust reliability and validity across diverse populations [16]–[18]. The DKQ-24's internal consistency, as measured by the Cronbach's alpha method, typically exceeds 0.80, indicating high reliability. The questionnaire's validity has also been supported by its strong correlations with other established measures of the diabetes knowledge and clinical outcomes. In this study, the DKQ-24 was administered to evaluate diabetes knowledge dissemination effectiveness of ChatGPT-4 version.

Table 1. A 24-items of patient's DKQ-24 (n=24)

No	Question items (n=24)	Yes	No	Don't know
1	Eating too much sugar and other sweet foods is a cause of diabetes.			
2	The usual cause of diabetes is lack of effective insulin in the body.			
3	Diabetes is caused by failure of the kidneys to keep sugar out of the urine.			
4	Kidneys produce insulin.			
5	In untreated diabetes, the amount of sugar in the blood usually increases.			
6	If I am diabetic, my children have a higher chance of being diabetic.			
7	Diabetes can be cured.			
8	A fasting blood sugar level of 210 is too high.			
9	The best way to check my diabetes is by testing my urine.			
10	Regular exercise will increase the need for insulin or other diabetic medication.			
11	There are two main types of diabetes: type 1 (insulin-dependent) and type 2 (non-insulin-dependent).			
12	An insulin reaction is caused by too much food.			
13	Medication is more important than diet and exercise to control my diabetes.			
14	Diabetes often causes poor circulation.			
15	Cuts and abrasions on diabetics heal more slowly.			
16	Diabetics should take extra care when cutting their toenails.			
17	A person with diabetes should cleanse a cut with iodine and alcohol.			
18	The way I prepare my food is as important as the foods I eat.			
19	Diabetes can damage my kidneys.			
20	Diabetes can cause loss of feeling in my hands, fingers, and feet.			
21	Shaking and sweating are signs of high blood sugar.			
22	Frequent urination and thirst are signs of low blood sugar.			
23	Tight elastic hose or socks are not bad for diabetics.			
24	A diabetic diet consists mostly of special foods.			

### 2.2. ChatGPT-4 interaction

For this study, ChatGPT-4 was used due to its reported superior performance in generating human-like responses across various domains, including healthcare [19]. ChatGPT is a large language model trained on a vast dataset encompassing a wide range of various topics, enabling the chatbot to provide detailed and contextually relevant answers to user queries. Technically, each of the 24 questions from the DKQ-24 was posed to ChatGPT-4 on two separate occasions, with one-week interval between the testing sessions, see Figure 1. By testing each question twice, this study sought to identify variations in the responses that might impact the reliability and perceived quality of the information provided.

### 2.3. Evaluation criteria

The reproducibility of ChatGPT-4's responses was a critical focus of the study, and responses were categorized based on their comprehensiveness and accuracy. Specifically, each response was classified into one of two primary categories, as applied in the previous study [20], namely: "comprehensive and correct" or "some correct and some incorrect". The classification was intended to capture the degree to which chatbot's responses adhered to established medical guidelines and provided complete and accurate information. Any responses containing minor inaccuracies or omissions but were otherwise broadly correct were categorized as the "some correct and some incorrect".

The reliability and quality of ChatGPT-4's responses were independently evaluated by two health professionals (two general practitioners). The evaluations were performed applying the modified DISCERN (mDISCERN) scale as shown in Table 2 and the global quality scale (GQS) as shown in Table 3. The

mDISCERN scale, adapted from the original DISCERN instrument, mainly focuses on the reliability of provided healthcare information by assessing criteria such as the clarity of aims, the comprehensiveness of information, the transparency of the sources, and the balance and bias of the content [21]. Each criterion was scored on a 5-point scale, with higher scores indicating better reliability. The GQS was used to evaluate the overall quality of the responses, with a focus on their usefulness and the logical flow of information. The GQS also applies a five-point scale, with higher scores reflecting higher quality [22].

The readability of ChatGPT-4's responses was assessed using many established readability metrics, such as: Flesch reading ease (FRE) score, the Flesch-Kincaid grade level (FKGL), the gunning fog index (GFI), the Coleman-Liau index (CLI), as well as the simple measure of gobbledygook (SMOG). These tools collectively provide a comprehensive evaluation of how easily the text can be read and understood by general public (in this case, diabetic patients): i) FRE score ranges from 0 to 100, with higher scores indicating easier readability. Texts with a score of 60-70 are considered easily understandable by 8<sup>th</sup> or 9<sup>th</sup> graders [23]; ii) FKGL metric translates the FRE score into the United States (U.S.) school grade level, indicating the minimum grade level required to understand all the text. For example, a score of 8.0 means the text is understandable by an 8th grader [24]; iii) GFI index estimates the years of formal education a person needs to understand the text on the first reading. A score of 12 is equivalent to high school level [25]; iv) CLI measures the readability of the text based on the average number of letters per 100 words and the average number of sentences per 100 words. It outputs the U.S. grade level [26]; v) SMOG model estimates the years of education needed to understand a piece of writing. It is often used for texts that require careful consideration of complex topics [27].

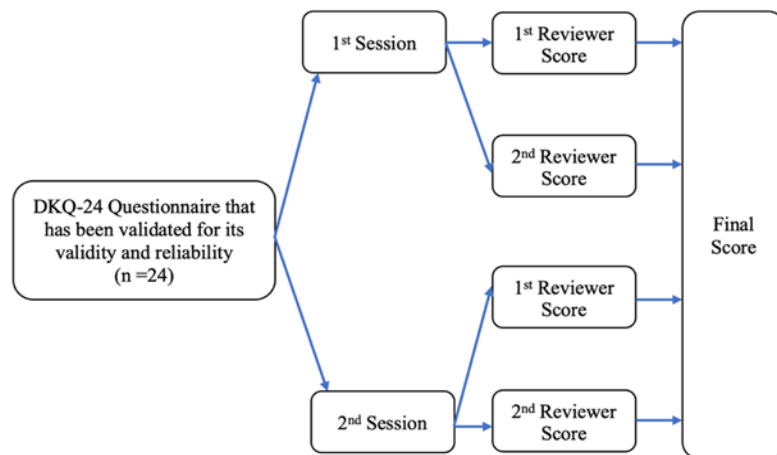


Figure 1. Flowchart of the reliability, quality, and readability assessment of this study

Table 2. Question items of the mDISCERN questionnaire (n=8)

No	mDISCERN questionnaire	Score
1	Are the aims clear?	1-5
2	Does it achieve its aims?	1-5
3	Is it relevant?	1-5
4	Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1-5
5	Is it clear when the information used or reported in the publication was produced?	1-5
6	Is it balanced and unbiased?	1-5
7	Does it provide details of additional sources of support and information?	1-5
8	Does it refer to areas of uncertainty?	1-5

Table 3. GQS utilized in the study (n=5)

No	GQS	Score
1	Poor quality, poor flow of the site, most information missing, not at all useful for patients	1
2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients	2
3	Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients	3
4	Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients	4
5	Excellent quality and excellent flow, very useful for patients	5

### 3. RESULTS AND DISCUSSION

#### 3.1. Reliability and quality

In the contexts of consensus scores, reliability and quality of ChatGPT-4's responses were evaluated using the mDISCERN scale and QQS. The mean mDISCERN score was  $31.50 \pm 2.89$ , indicating a generally good level of reliability. The mDISCERN scale, which assesses the reliability of health information, revealed that most of chatbot's responses were reliable, with a majority scoring above the threshold for fair reliability. The median QQS score was 4, with scores ranging from 1 to 5, suggesting that overall quality of responses was high, reflecting high overall quality but also highlighting room for improvement.

Regarding the response consistency, the responses provided by ChatGPT-4 were consistent across multiple testing sessions. Each response was analyzed twice on different days to evaluate reproducibility, and the results indicated a high degree of consistency. Responses included advice to consult an endocrinologist, ensuring that the users were reminded of the importance of professional guidance. Importantly, no misleading information was found in any of the responses, underscoring the reliability of ChatGPT-4 as the information source. In terms of reliability classification, 75% of responses were rated as high quality, 25% as moderate quality, and only a small fraction (8.3%) was considered poor.

The consistency of the responses was another critical aspect of this study. Each question was tested twice on two different days, and the responses demonstrated high consistency, with no significant variations. This consistency is important for ensuring that users receive dependable information irrespective of the time of interaction. All responses included advice to consult a healthcare professional, strengthening the necessity of professional medical guidance-as well as underscoring the responsible design of the AI system.

The findings of this study are consistent with previous studies indicating that the AI chatbots can be effective tools for disseminating health information. For example, a study found that the AI chatbot provided reliable and satisfactory responses to patient's questions about breast cancer [28]. Similarly, a review of the psychiatric landscape reported that AI-driven conversational agents could provide appropriate and accurate responses to inquiries on the mental health and interpersonal violence [29]. All previous studies highlight the potential of AI chatbots to serve as valuable resources for health education and information dissemination.

#### 3.2. Readability

The readability of ChatGPT-4's responses was assessed by multiple readability indexes, including FRE, FKGL, GFI, CLI, and SMOG scores as shown in Table 4. The median FRE score was 69.88, with a range of 6.39 to 97.02, indicating that all the text varied from very difficult to read to quite easy to read. On average, the readability metrics suggested that the text was somewhat challenging, with an FKGL mean of  $6.54 \pm 3.19$ , indicating that the text was suitable for its readers at a sixth-grade reading level. The GFI mean was  $11.92 \pm 1.55$ , and the CLI mean was  $12.52 \pm 3.53$ , both scores suggesting that the text required a level of education corresponding to high school or early college years. The SMOG index, which measures years of education needed to understand a text, averaged  $5.61 \pm 3.20$ , meaning that the text was moderately complex.

In terms of reading levels, 20.8% of AI-generated responses were rated as fairly easy to read, 29.1% as standard or average, and 16.6% as fairly difficult to read. Only 4.1% of the texts were classified as very difficult or extremely difficult to read. This distribution suggests that while a significant portion of responses was accessible to a broad audience, a considerable number required a higher level of reading proficiency. The reader's age analysis showed that the majority of responses were suitable for individuals with a reading level corresponding to high school and above. Specifically, 50% of the texts were appropriate for readers aged 14-17 years, and 12.5% for college-level readers aged 18-22 years.

As shown in Table 5, the evaluation of ChatGPT-4's responses according to mDISCERN scale and QQS scale revealed notable insights into the chatbot's performance in disseminating diabetes knowledge. Out of the 24 responses assessed, 2 responses (8.3%) were classified as poor, 17 responses (70.8%) as fair, and 5 responses (20.8%) as good based on DISCERN criteria. In terms of overall quality classification, 6 responses (25%) were deemed to be of moderate quality, while 18 responses (75%) were rated as high quality.

These findings suggest that while a majority of the responses provided by ChatGPT-4 were of fair or high quality, there remains a small proportion that highly requires improvement to ensure the reliability and comprehensiveness in health information dissemination. Previous studies also highlighted the importance of readability in health communication. For instance, this study emphasized that effectiveness of conversational agents in the healthcare depends significantly on the readability of the information they provide [30]. If the information is too complex, it may not be useful to the general public, particularly individuals with lower health literacy levels. Therefore, improving the readability of AI-generated health information is essential to maximize its impact.

To improve the accessibility of AI-driven chatbots, it is essential to address health literacy and cultural sensitivity. Simplifying medical terminology and tailoring messages to align with patient's cultural and linguistic contexts can significantly enhance comprehension and engagement. For example, incorporating culturally relevant dietary advice or providing responses in regional languages can make the

information more relatable and actionable. Future development of chatbots should prioritize inclusivity by incorporating diverse datasets and consulting with multidisciplinary teams to ensure that the generated content meets the needs of various populations.

Table 4. Summary of mDISCERN, GQS, and readability scores of ChatGPT responses

Reliability, Quality, and Readability	n = 24
Reliability	
mDISCERN score (mean ± SD)	31.50±2.89
Quality	
GQS score [median (min–max)]	4 (1-5)
Readability indexes	
FRE [median (min–max)]	[69.88 (6.39–97.02)]
FKGL (Mean ± SD)	(6.54±3.19)
GFI (Mean ± SD)	(15.92±11.52)
CLI (Mean ± SD)	(12.52±5.23)
SMOG (Mean ± SD)	(5.61±3.20)
Reading level	
Fairly easy to read n (%)	5 (20.8%)
Standard/average n (%)	7 (29.1%)
Fairly difficult to read n (%)	4 (16.6%)
Difficult to read n (%)	4 (16.6%)
Very difficult to read n (%)	1 (4.1%)
Extremely difficult to read (%)	1 (4.1%)
Reader's age	
8–9 years old (fourth and fifth graders) n (%)	0 (0%)
10–11 years old (fifth and sixth graders) n (%)	1 (4.1%)
11–13 years old (sixth and seventh graders) n (%)	4 (16.6%)
12–14 years old (seventh and eighth graders) n (%)	4 (16.6%)
13–15 years old (eighth and ninth graders) n (%)	3 (12.5%)
14–15 years old (ninth and tenth graders) n (%)	0 (0%)
15–17 years old (tenth and eleventh graders) n (%)	5 (20.8%)
17–18 years old (twelfth graders) n (%)	0 (0%)
18–19 years old (college level entry) n (%)	0 (0%)
21–22 years old (college level) n (%)	3 (12.5%)
College graduate n (%)	2 (8.3%)
Professional (%)	1 (4.1%)

Table 5. Score distribution of ChatGPT according to the DISCERN scale

Parameters	n=24 (%)
mDISCERN criteria	
Poor	2 (8.3)
Fair	17 (70.8)
Good	5 (20.8)
Quality classification	
Low quality	0 (0.00)
Moderate quality	6 (25.0)
High quality	18 (75.0)

### 3.3. Correlation analysis

The correlation analysis of this study examined the relationship between the reliability, quality, and readability of the ChatGPT-4's responses. The score of mDISCERN was found to be moderately positively correlated with the GQS score, suggesting that responses rated as more reliable were also perceived to be of higher quality. However, the mDISCERN scores did not show a significant correlation with the readability metrics (FRE, FKGL, GFI, CLI, SMOG). This indicates that the reliability and quality of the responses were independent of their readability, meaning that reliable and high-quality responses were not necessarily easier to read. The FRE score was significantly negatively correlated with other readability formulas such as FKGL, CLI, SMOG, and GFI. This negative correlation is expected since these formulas measure different aspects of readability: while the FRE score increases with easier text, the other indexes increase with more complex text. The FKGL, GFI, CLI, and SMOG scores were all positively correlated with each other, indicating that texts considered difficult by one measure were also likely to be rated as difficult by the others.

### 3.4. Strengths and limitations of the study

One of the main strengths of this study is its comprehensive evaluation of ChatGPT-4 using DKQ-24. This standardized questionnaire allowed for reliable comparisons across different interaction

sessions and provided a robust measure of the AI-driven chatbot's effectiveness as an educational resource. Additionally, various readability metrics implemented in this study provided a nuanced understanding of the accessibility of responses, highlighting areas for improvement in AI-driven health communication. Another strength lies in the reproducibility assessment. By testing each question twice on different days, the study confirmed the consistency of ChatGPT-4's responses, one of vital factors for the reliability of AI-driven educational tools. Furthermore, the independent evaluation by experienced endocrinologists using the mDISCERN and GQS scales ensured a comprehensive and objective assessment of the reliability and quality of the information provided. Additionally, the methodology applied in this study aligns with best practices in evaluating health information tools. The use of validated instruments like DKQ-24 and mDISCERN scale provides a reliable framework for assessing both the quality and reliability of health information. This method is consistent with previous studies that have utilized similar methodologies to evaluate such tools [16], [17].

This study, despite its strengths, also has many limitations. Firstly, it relied on a specific version of the AI-driven chatbot, ChatGPT-4, which may not fully represent the capabilities of newer versions, such as ChatGPT-4. Future studies should consider evaluating these newer versions to provide a more comprehensive understanding of advancements in the AI-driven health communication. Secondly, while the study assessed the readability of responses, it did not directly measure the comprehension and retention of the information by users. Future research should include feedbacks from its users and comprehension assessments to provide a more comprehensive evaluation of the educational impact of AI-driven chatbots.

While AI-driven chatbots offer immense potential, its implementation in healthcare raises important ethical considerations. Issues such as data privacy, security, and the potential for misinformation must be carefully addressed. Ensuring that patient data used to train AI models is anonymized and handled in compliance with regulatory standards is critical. Moreover, reliance on AI tools should not replace professional medical guidance but rather complement it. The ethical challenge of ensuring equitable access to AI tools for underserved populations also warrants attention. These considerations highlight the importance of developing AI systems that are transparent, inclusive, and aligned with established medical guidelines.

### 3.5. Implications for practice and future research

The findings of this study have significant implications for the utilization of AI-driven chatbots in healthcare education. The high reliability and quality of ChatGPT-4's responses suggest that AI tech can be a good medium for disseminating health information. However, the variability in readability suggests a need for ongoing improvements to ensure that the information is accessible to a wider audience. Future research should focus on enhancing the readability of AI-generated responses without compromising their accuracy. This could involve refining the algorithms to generate simpler language or providing additional context and explanations for more complex terms. Additionally, the studies should explore the use of AI-driven chatbots in various healthcare contexts and languages to generalize all the findings and identify the best practices for implementing these tools in diverse settings.

The potential of the AI-driven chatbots to serve as educational tools in healthcare is promising. By providing reliable and high-quality information, these tools can support patients in managing their conditions and making informed health decisions. However, ensuring that the information is easily understandable is essential for better impacts. Additionally, addressing the readability challenges identified in this study is very important for making the AI-driven tools more effective.

To maximize the utility of AI-driven chatbots in diabetes education, their integration into existing healthcare systems is essential. Chatbots can be embedded in electronic health record (EHR) systems to provide context-sensitive support during consultations or linked with telemedicine platforms to assist patients in remote areas. Collaboration between AI developers, healthcare providers, and policymakers is crucial for ensuring seamless integration and interoperability. Additionally, training healthcare professionals to use and monitor these tools will enhance their adoption and effectiveness. By embedding AI tools into routine healthcare workflows, we can bridge gaps in education and support for diabetes management.

### 3.6. Comparison with other large language models

Future studies should investigate the performance of alternative large language models, such as Llama, Claude, and Gemini, alongside ChatGPT. A comparative evaluation would provide insights into the relative strengths and limitations of these models in disseminating diabetes-related knowledge. This exploration could help identify the most effective model for addressing the educational needs of diabetic patients, particularly in terms of accuracy, reliability, readability, and contextual adaptability. Such comparative research would also reveal whether any models demonstrate unique advantages in tailoring responses to specific patient demographics or healthcare settings. This broader investigation would strengthen the understanding of AI's role in healthcare education and guide the selection of optimal tools for clinical implementation.

### 3.7. Expanding scope with additional questionnaires and datasets

While the DKQ-24 provides a validated framework for assessing diabetes knowledge, future research should incorporate additional instruments to explore new dimensions of patient education. For example, the Michigan diabetes knowledge test (MDKT) or the summary of diabetes self-care activities (SDSCA) could assess broader aspects of diabetes management, including self-care behaviors and treatment adherence. General health literacy tools, such as the test of functional health literacy in adults (TOFHLA) and the rapid estimate of adult literacy in medicine (REALM), would allow researchers to evaluate the effectiveness of chatbot responses for individuals with varying health literacy levels. Furthermore, utilizing multilingual datasets could help determine the accessibility and adaptability of AI-driven chatbots for diverse linguistic and cultural populations. These expansions would provide a more comprehensive evaluation of AI's potential in healthcare education.

### 3.8. Implications for future AI development in healthcare

To maximize the effectiveness of AI-driven chatbots in patient education, future research should focus on optimizing their readability and tailoring responses to individual user needs. Enhancing chatbot algorithms to generate simpler, more comprehensible language could improve accessibility for populations with lower health literacy. Additionally, dynamic updates to AI models through domain-specific fine-tuning may enhance the contextual relevance and accuracy of medical responses. Exploring the use of these models in real-world settings, combined with direct feedback from patients and healthcare providers, could provide practical insights into usability and engagement. Cost-effectiveness analyses of implementing such tools in large-scale health education programs would also inform healthcare systems about their potential economic benefits and scalability.

## 4. CONCLUSION

The findings of this study indicate that ChatGPT-4 provides reliable and high-quality information on diabetes management. Nevertheless, the readability of the responses varied, with some texts requiring higher levels of education to comprehend fully. The study findings underscore the potential of AI-driven chatbots as educational tools in healthcare while underlining the need for ongoing improvements to improve or enhance the accessibility of the information provided.

## ACKNOWLEDGEMENTS

The authors would like to extend their deepest gratitude to the School of Pharmacy, Faculty of Medicine and Health Sciences, Universitas Muhammadiyah Yogyakarta, for their unwavering support and resources that significantly contributed to the success of our research. Additionally, the heartfelt appreciation goes to the 1984-EL research team for their invaluable contributions, dedication, and collaborative efforts throughout the project.

## FUNDING INFORMATION

This study was supported by Universitas Muhammadiyah Yogyakarta.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Thesa Ghozali	✓	✓	✓	✓	✓	✓			✓	✓		✓		✓
Woro Supadmi		✓				✓			✓	✓	✓	✓	✓	
Fadya Bella Suci Maharani	✓		✓	✓			✓			✓	✓		✓	✓
Aloina Jean Rassyifa	✓			✓		✓	✓	✓		✓	✓		✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition



## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

## ETHICAL APPROVAL

The study protocol was reviewed and approved by the Institutional Review Board (IRB) of Faculty of Medicine and Health Sciences, Universitas Muhammadiyah Yogyakarta, Indonesia (approval no. 288/EC-KEPK FKIK UMY/XII/2023). All written informed consent was obtained from all participants prior to the administration of the questionnaire. Meanwhile, participant's anonymity and confidentiality were maintained throughout the study with data securely stored and accessible only to the research team.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




## REFERENCES

- [1] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, 2019, doi: 10.1016/j.diabres.2019.107843.
- [2] S. S. Chawla *et al.*, "Impact of health education on knowledge, attitude, practices and glycemic control in type 2 diabetes mellitus," *Journal of Family Medicine and Primary Care*, vol. 8, no. 1, 2019, doi: 10.4103/jfmpe.jfmpe\_228\_18.
- [3] S. Sharma *et al.*, "A critical review of ChatGPT as a potential substitute for diabetes educators," *Cureus*, vol. 15, no. 5, May 2023, doi: 10.7759/cureus.38380.
- [4] M. A. Boroumand, S. Sedghi, P. Adibi, S. Panahi, and A. Rahimi, "Patients' perspectives on the quality of online patient education materials," *Journal of Education and Health Promotion*, vol. 11, no. 1, 2022, doi: 10.4103/jehp.jehp\_1127\_21.
- [5] A. Nakhleh, S. Spitzer, and N. Shehadeh, "ChatGPT's response to the diabetes knowledge questionnaire: implications for diabetes education," *Diabetes Technology & Therapeutics*, vol. 25, no. 8, pp. 571–573, Aug. 2023, doi: 10.1089/dia.2023.0134.
- [6] L. Balcombe, "AI chatbots in digital mental health," *Informatics*, vol. 10, no. 4, 2023, doi: 10.3390/informatics10040082.
- [7] T. Sutikno, "The future of artificial intelligence-driven robotics: applications and implications," *IAES International Journal of Robotics and Automation (IJRA)*, vol. 13, no. 4, 2024, doi: 10.11591/ijra.v13i4.pp361-372.
- [8] N. Tangri and C. Sabanayagam, "Artificial intelligence approaches to enable early detection of CKD," *Nature Reviews Nephrology*, vol. 21, no. 3, pp. 153–154, Mar. 2025, doi: 10.1038/s41581-025-00933-6.
- [9] Amisha, P. Malik, M. Pathania, and V. K. Rathaur, "Overview of artificial intelligence in medicine," *Journal of Family Medicine and Primary Care*, vol. 8, no. 7, pp. 2328–2331, Jul. 2019, doi: 10.4103/jfmpe.jfmpe\_440\_19.
- [10] Y. Shaikh, V. Parvati, and S. R. Biradar, "Early disease prediction algorithm for hypertension-based diseases using data aware algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 2, 2022, doi: 10.11591/ijeecs.v27.i2.pp1100-1108.
- [11] E. Grassini, M. Buzzi, B. Leporini, and A. Vozna, "A systematic review of chatbots in inclusive healthcare: insights from the last 5 years," *Universal Access in the Information Society*, vol. 24, no. 1, pp. 195–203, Mar. 2025, doi: 10.1007/s10209-024-01118-x.
- [12] I. Altamimi, A. Altamimi, A. S. Alhumimidi, A. Altamimi, and M.-H. Temsah, "Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute," *Cureus*, vol. 15, no. 6, Jun. 2023, doi: 10.7759/cureus.40922.
- [13] M. Ghazali and I. W. Murni, "Knowledge and attitude among community pharmacists regarding pharmacovigilance—a cross sectional survey," *Indonesian Journal of Pharmacy*, vol. 34, no. 4, Oct. 2023, doi: 10.22146/ijp.5814.
- [14] A. Wang, Z. Qian, L. Briggs, A. P. Cole, L. O. Reis, and Q.-D. Trinh, "The use of chatbots in oncological care: a narrative review," *International Journal of General Medicine*, vol. 16, pp. 1591–1602, May 2023, doi: 10.2147/IJGM.S408208.
- [15] R. S. Goodman *et al.*, "Accuracy and reliability of chatbot responses to physician questions," *JAMA Network Open*, vol. 6, no. 10, Oct. 2023, doi: 10.1001/jamanetworkopen.2023.36483.
- [16] M.-H. Hsieh, Y.-C. Chen, C.-H. Ho, and C.-Y. Lin, "Validation of diabetes knowledge questionnaire (DKQ) in the Taiwanese population-concurrent validity with diabetes-specific quality of life questionnaire module," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 15, pp. 2391–2403, Aug. 2022, doi: 10.2147/DMSO.S369552.
- [17] J. A. Zuñiga *et al.*, "Revision and psychometric evaluation of the diabetes knowledge questionnaire for people with type 2 diabetes," *Diabetes Spectrum*, vol. 36, no. 4, pp. 345–353, Nov. 2023, doi: 10.2337/ds22-0079.
- [18] A. Zakiudin, G. Irianto, A. Badrujaludin, H. Rumahorbo, and S. Susilawati, "Validation of the diabetes knowledge questionnaire (DKQ) with an Indonesian population," *KnE Medicine*, vol. 2, no. 2, Jun. 2022, doi: 10.18502/kme.v2i2.11072.
- [19] P. F. Funk *et al.*, "ChatGPT's response consistency: a study on repeated queries of medical examination questions," *European Journal of Investigation in Health, Psychology and Education*, vol. 14, no. 3, pp. 657–668, Mar. 2024, doi: 10.3390/ejihpe14030043.
- [20] C. E. Onder, G. Koc, P. Gokbulut, I. Taskaldiran, and S. M. Kuskonmaz, "Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-023-50884-w.
- [21] D. Charnock, S. Shepperd, G. Needham, and R. Gann, "DISCERN: an instrument for judging the quality of written consumer health information on treatment choices," *Journal of Epidemiology & Community Health*, vol. 53, no. 2, pp. 105–111, 1999, doi: 10.1136/jech.53.2.105.
- [22] E. Furukawa *et al.*, "Evaluating the understandability and actionability of online educational videos on pre-dialysis chronic kidney disease," *Nephrology*, vol. 28, no. 11, pp. 620–628, Nov. 2023, doi: 10.1111/nep.14226.




- [23] D. Eleyan, A. Othman, and A. Eleyan, "Enhancing software comments readability using flesch reading ease score," *Information*, vol. 11, no. 9, Sep. 2020, doi: 10.3390/info11090430.
- [24] S. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky, "A large-scaled corpus for assessing text readability," *Behavior Research Methods*, vol. 55, no. 2, pp. 491–507, Mar. 2022, doi: 10.3758/s13428-022-01802-x.
- [25] G. Minervini, M. M. Marrapodi, and M. Cicciù, "Online bruxism-related information: can people understand what they read? a cross-sectional study," *Journal of Oral Rehabilitation*, vol. 50, no. 11, pp. 1211–1216, Nov. 2023, doi: 10.1111/joor.13519.
- [26] J. Tran and E. Tsui, "Assessment of the readability, availability, and quality of online patient education materials regarding uveitis medications," *Ocular Immunology and Inflammation*, vol. 29, no. 7–8, pp. 1507–1512, Nov. 2021, doi: 10.1080/09273948.2020.1737144.
- [27] M. Shneyderman, R. Davis, G. Snow, S. Dhar, and L. M. Akst, "Zenker's diverticulum: readability and quality of online written education materials," *Dysphagia*, vol. 37, no. 6, pp. 1461–1467, Dec. 2022, doi: 10.1007/s00455-022-10406-8.
- [28] J.-E. Bibault *et al.*, "A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial," *Journal of Medical Internet Research*, vol. 21, no. 11, Nov. 2019, doi: 10.2196/15787.
- [29] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape," *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, Jul. 2019, doi: 10.1177/0706743719828977.
- [30] C. Liu, D. Zowghi, G. Peng, and S. Kong, "Information quality of conversational agents in healthcare," *Information Development*, May 2023, doi: 10.1177/02666669231172434.

## BIOGRAPHIES OF AUTHORS






**Muhammad Thesa Ghozali**    is a dedicated researcher and lecturer at the School of Pharmacy, Faculty of Medicine and Health Sciences, Universitas Muhammadiyah Yogyakarta. With a strong foundation in pharmaceutical management, he specializes in integrating advanced information and communication technologies, including the internet of things, mobile health, artificial intelligence, and machine learning, into health promotion and patient education. His work aims to enhance the effectiveness of patient education and improve healthcare outcomes by leveraging these cutting-edge technologies. He can be contacted at email: ghozali@umy.ac.id.






**Woro Supadmi**    is a researcher and lecturer specializing in Community Pharmacy, Clinical Pharmacy, and Pharmacoeconomics. Her work focuses on evaluating drug utilization in patients, assessing patient compliance and quality of life, and conducting cost-of-illness evaluations. Her expertise in these areas contributes to improving healthcare outcomes through better understanding of medication use and associated costs. She is part of the Pharmacoeconomics research group. She can be contacted at email: wsupadmi@yahoo.com.



**Fadya Bella Suci Maharani**    holds a Bachelor of Pharmacy degree, specializing in pharmaceutical management. Her academic and professional interests are centered around patient education and health promotion, areas in which she is deeply committed to improving healthcare outcomes. With a strong foundation in pharmaceutical management, she is focused on developing and implementing strategies that enhance patient understanding of their medications and overall health. Her work aims to bridge the gap between healthcare providers and patients, ensuring that patients are well-informed and empowered to manage their health effectively. She can be contacted at email: fadya.bella.fkik20@mail.umy.ac.id.



**Aloina Jean Rassyifa**    is currently an undergraduate student pursuing a degree in pharmaceutical management, with a specialization in pharmacy informatics. Her academic focus is on the intersection of technology and pharmacy, where she aims to contribute to the optimization of healthcare processes through the effective use of information systems. As she continues her studies, she is developing a strong foundation in both pharmaceutical management and informatics, preparing her to play a vital role in the evolving field of healthcare technology. She can be contacted at email: aloina.jean.fkik21@mail.umy.ac.id.