Original Paper

Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study

Yu Gu, PhD; Xinyi Wang, BA

School of Medical Technology, Capital Medical University, Beijing, China

Corresponding Author:
Yu Gu, PhD
School of Medical Technology
Capital Medical University
No 10 Xi Toutiao Road
Beijing, 100069
China
Phone: 86 81476234
Email: bitguyu@126.com

Abstract

Background: Artificial intelligence (AI) health care chatbots are gaining widespread adoption worldwide. It is imperative to understand the service quality of AI health care chatbots. However, there is limited guidance on how to comprehensively evaluate their service quality.

Objective: This study aimed to develop an index system based on the SERVQUAL framework for evaluating the service quality of AI health care chatbots.

Methods: An initial indicator pool was compiled through a comprehensive literature review and consultations with 4 experts. These indicators were mapped and categorized into 5 domains adapted from the SERVQUAL framework. The experts were recruited from hospital, university, and health commission settings by purposive sampling. The service quality index system was identified using a 2-round Delphi process, which included a virtual meeting between the 2 rounds. In the third round, indicator weights within each quality domain and subdomain were determined using the analytic hierarchy process.

Results: There were 26 indicators identified in the literature, based on which the 2-round Delphi process was conducted. A total of 20 experts were invited. The response rates in both rounds of Delphi and the analytic hierarchy process were 100%, and the authoritative coefficients were both >0.7. The final service quality index system for AI health care chatbots comprises 5 primary indicators and 17 secondary indicators. There were 3 (18%) indicators on assurance, 4 (24%) on reliability, 3 (18%) on human-likeness, 4 (24%) on tangibility, and 3 (18%) on responsiveness. The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167).

Conclusions: This study pioneers the development of a service quality index system for AI health care chatbots adapted from the SERVQUAL framework. The results provide a validated tool for evaluating the performance of chatbots and offer valuable insights for health service managers and developers to enhance AI-driven medical consultation services.

(J Med Internet Res 2026;28:e83051) doi: 10.2196/83051

KEYWORDS

artificial intelligence health care chatbot; artificial intelligence; AI; service quality; Delphi method; analytic hierarchy process; index system development; SERVQUAL

Introduction

Worldwide, artificial intelligence (AI) chatbots have been introduced into health care settings in recent years, where they are used by individuals as AI physicians for online medical consultations. A key innovation of AI health care chatbots lies in their ability to generate humanlike, natural language responses to diverse health-related queries anytime and anywhere, significantly improving access to medical guidance for broader populations [1]. Unlike earlier rule-based chatbots that relied on scripted replies, AI chatbots leverage advanced technologies, such as large language models (LLMs), to deliver personalized and context-aware interactions [2]. Moreover, the consultation service is often provided free of charge. AI health care chatbots show promise in delivering reliable medical advice without direct involvement from human physicians, offering a scalable solution to persistent challenges within the global health system, such as limited resources, uneven distribution, high costs, and growing demand [3]. Therefore, AI health care chatbots are playing an increasingly important role in modern health care systems [4].

AI health care chatbots represent not only a new type of service provider but also an innovative medical service model [5]. As an emerging field, chatbots have attracted growing attention from both practitioners and researchers. Despite its potential benefits, concerns remain regarding its service quality [6,7]. Efforts have been made to develop quality indicators for AI health care chatbots [8-16]. Some studies have evaluated response quality within specific disease contexts, such as labor epidurals, cardiovascular health, oncology, psoriasis, chronic hepatitis, and cancer [8-11]. Others have focused on assessing information quality [12,13] or have compared the performance of AI health care chatbots with that of human physicians [14-16]. However, existing studies primarily focus on narrow aspects of quality. Furthermore, the most commonly applied metrics—response accuracy, completeness, and consistency in closed-ended clinical questions—are predominantly defined from the health care providers' perspective rather than that of users. Therefore, a comprehensive and user-centered index system for evaluating service quality of AI health care chatbots remains underdeveloped.

Among existing service quality frameworks, SERVQUAL, developed by Parasuraman et al [17], is one of the most widely recognized frameworks for evaluating medical service quality worldwide. This framework includes 5 dimensions—tangibility, reliability, responsiveness, assurance, and empathy—and is specifically designed to assess users' expectations and perceptions of service quality [18]. Applying this classical framework enables a more comprehensive and theoretically grounded evaluation of service quality of AI health care chatbots, bridging classical service quality theory with emerging AI-driven health care contexts.

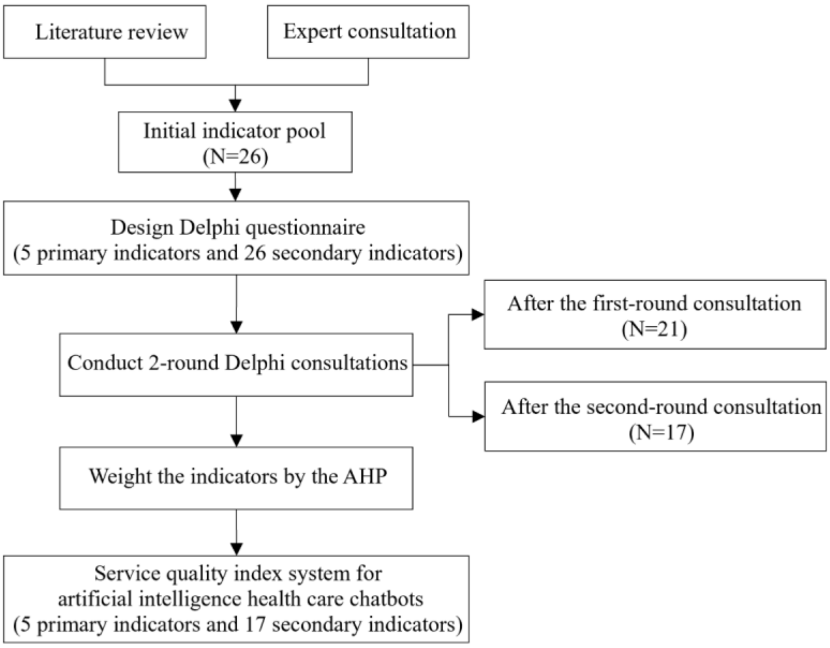
The aim of this study was to identify critical indicators that reflect the service quality of AI health care chatbots and to develop a scientifically feasible index system for its evaluation. The findings were expected to contribute to better identification of shortcomings, promote continuous quality improvement, enhance user experience, and offer new insights into the systemic evaluation of service quality of AI health care chatbots.

Methods

Study Design

This study used a mixed methods approach, combining qualitative insights from expert opinions with quantitative metrics to develop and quantify a service quality index system for AI health care chatbots. The literature review and expert consultation were applied to construct an initial indicator pool. The 2-round Delphi consultation was then conducted to refine and establish the final index system. Subsequently, the analytic hierarchy process (AHP) was applied to determine the weight of each indicator. The process of index system development and weight determination is shown in Figure 1.

Figure 1. The research process. AHP: analytic hierarchy process.


Literature review
Expert consultation
Initial indicator pool (N=26)
Design Delphi questionnaire (5 primary indicators and 26 secondary indicators)
Conduct 2-round Delphi consultations
After the first-round consultation (N=21)
After the second-round consultation (N=17)
Weight the indicators by the AHP
Service quality index system for artificial intelligence health care chatbots (5 primary indicators and 17 secondary indicators)

Initial Indicator Pool

The initial indicator pool was compiled based on existing literature and expert opinions. A comprehensive systematic literature search was conducted in 4 databases: PubMed, Web of Science, China National Knowledge Infrastructure, and Wanfang Data. The search strategy incorporated the following key terms: ["chatbot*" OR "chat-bot*" OR "conversational agent*" OR "conversational bot*" OR "conversational system*" OR "dialogue system*" OR ChatGPT] AND ["medic*" OR "health*" OR "disease*" OR "patient*"] AND ["quality indicator*" OR "quality evaluat*" OR "quality assess*" OR "quality measure*"]. Boolean operators (AND, OR, and NOT) were used to combine or refine search terms (Multimedia

Appendix 1). The exclusion criteria were as follows: (1) studies not focused on the evaluation of AI health care chatbots; (2) commentaries, protocols, letters, editorials, and conference abstracts; and (3) studies not published in English or Chinese.

In the second phase, this study incorporated insights from 4 interdisciplinary experts specializing in intelligent health care and medical service management. These experts conducted in-depth discussions regarding the relevance, suitability, and validity of the preliminary indicators. None of them participated in the subsequent Delphi consultation rounds. Both prior literature and expert opinions emphasized that the SERVQUAL framework provides a user-centered foundation well suited for the evaluation of AI health care chatbots and highlighted anthropomorphism as a distinctive feature influencing perceived service quality in AI-driven interactions. Specifically, users desire AI health care chatbots to exhibit kindness through humanlike attributes, such as a name, image, and voice. Beyond a friendly appearance, users expect these systems to demonstrate social intelligence, including the ability to detect user emotions and respond with genuine concern, which goes beyond simple empathy. Furthermore, users expect personalized responses tailored to individual factors. Therefore, the *humanlike* dimension was introduced as an innovative replacement for the traditional "empathy" construct to better capture the emotional and interactive capabilities unique to AI health care chatbots. Accordingly, this study established 5 first-level indicators adapted from the SERVQUAL structure: tangibility, responsiveness, assurance, human-likeness, and reliability, and 26 second-level indicators were included in the initial pool.

Expert Selection

The initial expert recruitment was conducted through recommendations from our collaborators in the field of intelligent health care. They were from renowned universities, tertiary hospitals, and provincial health sectors. To broaden the reach and ensure a diverse range of perspectives, the experts initially nominated by the researchers were then asked to suggest other qualified individuals who could contribute valuable insights to the study. The inclusion criteria for experts were as follows: (1) familiarity with research areas, such as intelligent health care, medical service management, health information management, and other related fields; (2) more than 5 years of professional experience in a relevant field; and (3) willingness to actively participate in the study and provide timely responses across multiple rounds of Delphi consultation. Finally, a total of 20 experts were recruited through purposive sampling.

Delphi Process

The Delphi method is a structured communication technique designed to systematically collect expert opinions and achieve consensus [19]. It has been widely applied and validated as a robust research methodology in health care contexts [20]. This study conducted a 2-round Delphi consultation to screen, refine, and finalize the indicators.

The Delphi consultation questionnaire consists of 2 main sections: an informed consent form and the main survey. The informed consent form outlined the study's background, objectives, methodology, privacy protection measures, and contact information. The main survey collected information from five areas: (1) experts' basic information, including age, education, and years of work experience; (2) the core consultation content, in which experts scored the importance and feasibility of each indicator using a 10-point scale (1=lowest and 10=highest); (3) the familiarity scale, rated by the expert themselves using a 5-point Likert scale (1=very unfamiliar and 5=very familiar); (4) the basis of expert judgment, evaluating the impact of theoretical analysis, practical experience, literature knowledge, and instinct on scoring (rated as high, medium, or low); and (5) blank fields, allowing experts to propose additions, deletions, or modifications to the indicators. All experts completed the informed consent process, and strict confidentiality was maintained throughout the entire process.

The Delphi process was conducted between February 2025 and June 2025. In the first round, Delphi questionnaires in Microsoft Word format were distributed to 20 experts, with a 2-week response period. Experts were asked to rate both the primary and secondary indicators and to provide comments. On the basis of the results and comments from the first round, the questionnaire was revised and redistributed to the same 20 experts for the second round. The second round followed the same rating procedure as the first and achieved consensus among the experts.

Indicator Selection

To screen the indicators, this study used 3 important statistics: the mean importance score, the full-mark rate (proportion of experts assigning the highest score), and the coefficient of variation. The inclusion criteria were as follows: (1) a mean of importance score ≥7.0, (2) a full-mark rate >20%, and (3) a coefficient of variation <0.25 [21-23]. Any indicator failing to meet all 3 criteria was subject to deletion or revision based on panel discussion and qualitative feedback.

AHP Procedure

Following the 2-round Delphi consultation, the final set of indicators was confirmed. The same panel of experts was then invited to participate in a pairwise comparison process to determine indicator weights. For each pair of indicators within the same hierarchical level, judgment matrices were constructed using a 1 to 9 ordinal scale to assess their relative importance [24]. The weight of each indicator was subsequently calculated using the percentage weighting method based on the pairwise comparison matrices, with higher weight values indicating greater perceived importance.

Data Analysis

Statistical analysis was conducted using SPSS software (version 25.0; IBM Corp). The authority coefficient (Cr) represents the authority level of experts. Cr was the arithmetic mean of the experts' judgment coefficient (Ca) and the experts' familiarity coefficient (Cs) [25]. A Cr value ≥0.7 was considered acceptable [26]. The Ca value was derived from experts' self-assessment of their own judgment criteria, as detailed in Table 1. The Cs value ranges from 1.0 (very familiar) to 0.2 (unfamiliar). The coordination of expert opinions was tested using the Kendall coefficient of concordance (Kendall *W*), with a significance level of α=.05. YAAHP software (version 11.2; MetaDecision) was used to calculate the indicator weights and assess the consistency ratio. When the consistency ratio value was <0.10, it was considered acceptable, indicating sufficient consistency in expert judgments [27].

Table 1. The judgment basis and degree of influence.

Judgment basis	Degree of impact on experts' judgment		
	High	Medium	Low
Theoretical analysis	0.3	0.2	0.1
Practical experience	0.5	0.4	0.3
Reference literature	0.1	0.1	0.1
Expert intuition	0.1	0.1	0.1

Ethical Considerations

The study protocol was approved by the Ethics Committee of the Capital Medical University, Beijing, China (2025SY-071). Participants were informed of the study's purpose and procedure. Online informed consent was obtained from each participant. All research data were stored on a password-encrypted computer, and only the researchers had access to the data. No compensation was provided to participants.

Results

Characteristics of Experts

A total of 20 experts completed the 2-round Delphi consultation and AHP evaluation. The panel consisted of 12 (60%) male and 8 (40%) female experts, ranging in age from 31 to 60 years. Among the experts, 17 (85%) held a master's degree or higher. All experts possessed associate senior professional titles or higher. The panel included 10 (50%) experts from quality control departments of hospitals, 7 (35%) from universities, and 3 (15%) from national or regional health commissions. The detailed characteristics of these experts are summarized in Table 2.

Table 2. The characteristics of the Delphi consultation experts (N=20).

Characteristics	Experts, n (%)
Sex	
Male	12 (60)
Female	8 (40)
Age (years)	
31-40	6 (30)
41-50	11 (55)
51-60	3 (15)
Education	
Doctoral degree	10 (60)
Master’s degree	7 (35)
Bachelor’s degree	3 (15)
Professional title	
Senior	11 (55)
Associate senior	9 (45)
Seniority (years)	
6-10	4 (20)
11-20	9 (45)
21-30	5 (25)
>30	2 (10)
Affiliation	
Hospitals	10 (50)
Universities	7 (35)
Health commission	3 (15)
Field of expertise	
Intelligent health care	8 (40)
Medical service management	7 (35)
Health information management	5 (25)

Authority Coefficient and Degree of Coordination

The Cr values for the first and second rounds of the Delphi consultation were 0.894 (Ca=0.931; Cs=0.847) and 0.919 (Ca=0.917; Cs=0.921), respectively (Table 3). Both values exceed the accepted threshold of 0.7, indicating a high level of expert credibility and reinforcing the reliability of the consultation results.

The Kendall *W* coefficients for the 2 consultation rounds are shown in Table 3. After the second round, the coordination coefficients of the indicator increased from 0.263 to 0.339, and all associated *P* values were <.001, indicating that the experts’ opinions converged and that the degree of consensus among experts was acceptable.

Table 3. Expert authority coefficients and the degree of coordination of expert opinions.

Round	Ca[a]	Cs[b]	Cr[c]	Kendall *W*	Chi-square (*df*)	*P* value
Round 1	0.931	0.847	0.894	0.263	71.1 (25)	<.001
Round 2	0.917	0.921	0.919	0.339	123.2 (20)	<.001

[a] Cs: familiarity coefficient.

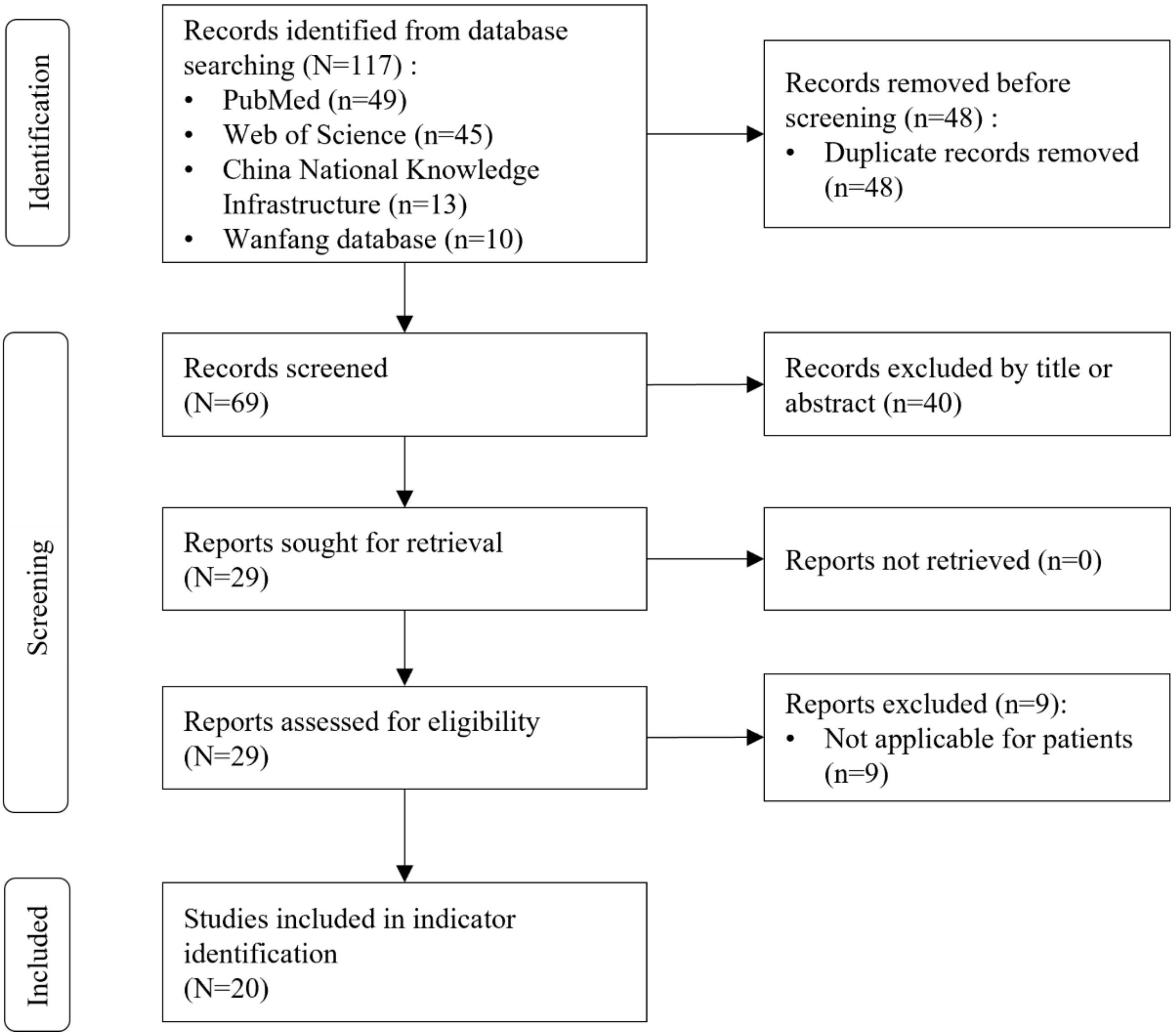
[b] Ca: judgment coefficient.

[c] Cr: authority coefficient.

Review for Initial Indicator Pool

The database search and hand searches identified 117 articles, from which 48 (41.0%) duplicates were removed. After screening the titles and abstracts, 29 (24.8%) full-text records were reviewed, of which 20 (17.1%) were included in the review [28-47]. The study selection process is illustrated in Figure 2.

Figure 2. Flowchart of the included studies.


Identification
Screening
Included
Records identified from database searching (N=117) :
• PubMed (n=49)
• Web of Science (n=45)
• China National Knowledge Infrastructure (n=13)
• Wanfang database (n=10)
Records removed before screening (n=48) :
• Duplicate records removed (n=48)
Records screened (N=69)
Records excluded by title or abstract (n=40)
Reports sought for retrieval (N=29)
Reports not retrieved (n=0)
Reports assessed for eligibility (N=29)
Reports excluded (n=9):
• Not applicable for patients (n=9)
Studies included in indicator identification (N=20)

In the 20 eligible studies, 9 (45%) focused exclusively on AI chatbots designed specifically for health care service [28-36]. To develop a comprehensive initial indicator pool, we also included 11 additional studies concerning the quality of general AI chatbots capable of delivering health care–related consultations [37-47]. Of these, 6 (30%) studies aimed to develop instruments for measuring the overall service quality of AI chatbots [28,29,34,37,38,40], with 2 (10%) grounded in the SERVQUAL model [34,37]. Another 7 (35%) studies developed instruments targeting specific dimensions of quality [30-33,35,36,42]. Additionally, 7 (35%) studies treated quality as a key determinant of acceptance or satisfaction with AI chatbots and provided detailed measurement items for AI chatbot quality [39,41,43-47]. Following a systematic sorting process and discussions with 4 domain experts, we synthesized these findings into an initial pool of 26 second-level indicators for assessing the quality of AI health care chatbots, structured according to the 5 dimensions of the SERVQUAL framework.

Indicator Selection

According to the indicator selection criteria and qualitative feedback from the experts, from a total of 26 indicators, 5 (19%) secondary indicators were removed and 21 (81%) were retained in the first round. Between round 1 and round 2, a virtual Delphi meeting was held to discuss indicators for which experts provided revision suggestions and to identify novel indicators based on perceived gaps in current indicators. Experts who had completed round 1 attended the meeting. During this panel meeting, 8 (31%) secondary indicators were merged into 4 (15%) indicators. Following the discussion, 20 experts rated the new indicators. Finally, the 2-round Delphi process reached a finalized evaluation framework comprising 5 (19%) primary indicators and 17 (65%) secondary indicators, each clearly defined in Table 4.

Table 4. The service quality index system for artificial intelligence (AI) health care chatbots.

Indicator	Definition	Weights
Assurance	The ability of AI health care chatbots to provide pertinent responses	0.239
Understandable answer	The AI health care chatbots provide an answer with a logistical structure and the right amount of information to the user's query	0.082
Accurate understanding	The AI health care chatbots understand the exact meaning of the content sent by the user in text and voice	0.079
Targeted question	The AI health care chatbots ask follow-up questions with context awareness based on the user's query	0.078
Reliability	The ability of AI health care chatbots to inspire trust and confidence	0.237
Trustworthy advice	The AI health care chatbots give medical advice, such as diagnosis, medication, and examination, and detailed explanations, which are consistent across multiple inquiries	0.071
Useful service	The AI health care chatbots are useful in addressing users' uncertainties about their health concerns	0.065
Specific risk warning	The AI health care chatbots clearly indicate the limitations of its provided answers	0.050
Protected privacy	The AI health care chatbots protect the user's privacy	0.051
Human-likeness	The social cue, personality, and empathy of AI health care chatbots	0.187
Personalized response	The AI health care chatbots tailor their answers according to the user's age, sex, and medical history	0.064
Emotional attention	The AI health care chatbots detect user emotion and makes the user feel concerned	0.063
Kind characteristic	The AI health care chatbots have a kind name, image, and voice	0.060
Tangibility	The hardware or software manifestations of AI health care chatbots	0.170
Accurate recognition	The AI health care chatbots recognize and transfer the speech of the users into accurate text	0.046
Compatible operation	It is convenient for the user to obtain the service of AI health care chatbots on a mobile app, WeChat mini program, or website	0.043
Friendly layout	The layout of AI health care chatbots is clear and easy to operate	0.041
Stable service	The AI health care chatbots provide the same smooth service in any situation	0.040
Responsiveness	The response ability of AI health care chatbots	0.167
Anytime response	The AI health care chatbots are available 24 hours a day for 365 days	0.043
Prompt response	The AI health care chatbots always give timely feedback when it is needed	0.042
Coherent response	The AI health care chatbots can communicate with the user seamlessly by maintaining records within the personal account	0.042

Indicator Weights

On the basis of the AHP and percentage weighting method, the weights for all indicators were calculated (Table 4). The primary indicators, ranked from highest to lowest weight, were assurance (0.239), reliability (0.237), human-likeness (0.187), tangibility (0.170), and responsiveness (0.167). Assurance received the highest weight. For the secondary indicators, weights ranged from 0.040 to 0.082. "Understandable answer" had the highest secondary weight (0.082), followed by "Accurate understanding" (0.079), "Targeted question" (0.078), and "Trustworthy advice" (0.071).

Discussion

Principal Findings

The development of AI health care chatbots is on the rise, and their adoption is becoming increasingly vital in modern health care. Providing AI health care chatbots with high service quality is critical to facilitating their broader diffusion and addressing contemporary health care challenges. Although previous studies have attempted to evaluate the quality of AI chatbots in responding to queries related to specific diseases, a comprehensive and user-centered index system for evaluating the service quality of AI health care chatbots has remained lacking. To our knowledge, this is the first study to develop a comprehensive service quality index system for AI health care chatbots from a patient perspective, using SERVQUAL as the theoretical framework. Through a 2-round Delphi process, a finalized set of 5 primary indicators and 17 secondary indicators was derived, specifically designed to capture both the technical functionality and interactive experience unique to AI health care chatbots. Subsequently, the indicator weights were obtained using the AHP.

Among the 5 primary indicators, assurance was identified as the most important dimension, which refers to the ability of AI health care chatbots to provide pertinent responses. Unlike consultations with a clinician, AI health care chatbots lack the capacity to perform physical examinations to support diagnosis. Therefore, it is essential for AI health care chatbots to deliver goal-oriented and unambiguous conversations, accurately understand user queries, ask follow-up questions with contextual awareness, and provide understandable answers [39,48]. Previous studies [8-10] have taken understandability, often reflected through situation-appropriate response length and information quantity, as a sole metric to measure the quality of AI health care chatbots. Consistent with this emphasis, the secondary indicator "Understandable answer" received the highest weight among secondary indicators. Although aspects such as completeness and consistency have been associated with answer readability in previous studies [11,48], they were not included in this framework. This omission stems from their highly specialized and profession-centric evaluation criteria, which may not align with patient-centered usability expectations.

The reliability dimension ranked second, following assurance. Current AI technologies remain fallible and necessitate oversight by health professionals to ensure the applicability and safety advice of AI health care chatbots [7]. For users, it is critical that AI health care chatbots provide not only trustworthy medical advice regarding diagnosis, medication, and examinations but also clear explanations that enhance transparency and facilitate informed decision-making, thereby fostering trust and promoting sustained engagement [4,28]. Both users and clinicians have underscored the importance of clearly indicating the limitations of AI-generated medical advice [3-5]. In contrast to earlier studies [9,49], this study deliberately excluded diagnostic accuracy, a frequently used metric, from this index system, as patients generally lack the specialized medical knowledge required to evaluate this aspect.

The human-likeness dimension is considered a distinctive feature of AI health care chatbots [1]. Although it is not ranked highest among the primary indicators, its associated secondary indicators, "Personalized response" and "Emotional attention," had prominent weight values. Users often perceive AI health care chatbots as an "AI doctor" and tend to evaluate it through direct comparison with human clinicians along these dimensions [3]. Enhancing humanlike attributes in interactions of AI health care chatbots remains a critical development objective. This entails increased efforts to enable AI health care chatbots to generate context-aware and individualized replies that adapt to both the conversational flow and user preferences [29].

The dimension of tangibility refers to hardware and software manifestations of AI health care chatbots. Among its secondary indicators, "Accurate recognition" was assigned the highest weight. As an information system designed to provide health guidance [50], the accuracy of recognizing and transferring user speech into precise text is the basis for correctly interpreting user queries and facilitating subsequent consultation processes [9]. The next important indicator, "Compatible operation," reflects the accessibility of the service of AI health care chatbots across diverse digital environments. A previous study has emphasized that users valued the ability to access services of AI health care chatbots through various devices, such as smartphones, tablets, or computers, which supports broader and more equitable adoption [31].

Although responsiveness is positioned as the final dimension in the framework, it constitutes a fundamental component of the service quality of AI health care chatbots [5]. This dimension is characterized by the provision of active and uninterrupted guidance 24 hours a day, real-time responses without having to wait in line, immediate accessibility from any location without the need for travel, and seamless communication across different devices [21,28]. Given that AI health care chatbots are supported by LLMs, users perceive responsiveness as their inherent capability [1].

Overall, AI health care chatbots currently represent a viable alternative to human clinicians in initial user interactions [2]. However, its performance can vary significantly depending on the underlying LLMs, knowledge bases, and health data used [31]. While this study developed a service quality index system for AI health care chatbots based on the SERVQUAL framework, most secondary indicators in this study were newly developed to reflect the unique AI context. The proposed index system offers practical value for multiple stakeholders: it enables users to better understand and assess the strengths of AI health care chatbots; supports health service managers in systematically collecting feedback and monitoring performance; and guides developers in conducting feasibility analyses, optimizing design, and implementing postlaunch evaluation. Future research will involve applying this index system in field studies with users of AI health care chatbots to validate their utility and support its ongoing refinement.

Strengths and Limitations

The strengths of this study are as follows. First, this study developed a comprehensive evaluation index system by adapting a scientific framework that incorporates both internationally validated evidence-based indicators and unique features of AI health care chatbots. Second, the mixed methods approach, combining literature review, expert consultations, a 2-round Delphi process, and AHP, ensured a rigorous and systematic development process. Third, this study provides new insights into the systemic evaluation of service quality of AI health care chatbots.

Several limitations should also be acknowledged. First, although the number of experts consulted met methodological requirements, it remained relatively limited. The panel may be subject to selection bias due to the experts' familiarity with AI health care chatbots, which could influence the selection and weighting of indicators. Therefore, the scope of expert consultation needs to be further expanded to enhance the validity of the indicators. Second, all participating experts were based in China, which may limit the generalizability of the findings to other cultural or health system contexts. Future studies should validate the proposed evaluation index across a broader range of settings of AI health care chatbots. Third, this index system has not yet been operationalized and evaluated by the users of AI health care chatbots. Further empirical research is needed to demonstrate its practical relevance and utility and to consider incorporating patient experience into the assessment process. Furthermore, although Kendall *W* was statistically significant, its value reflects only a moderate level of consensus. This implies that the findings are robust but limited in microlevel ranking. In this study, experts from diverse professional backgrounds likely held different interpretations and assigned varying weights to the indicators, which may have led to evaluation discrepancies.

Conclusions

This study developed a comprehensive, user-centered index system for evaluating the service quality of AI health care chatbots. Through the Delphi method and the AHP, a finalized framework consisting of 5 primary dimensions and 17 secondary indicators was established. These include 4 indicators for tangibility, 3 for responsiveness, 3 for assurance, 3 for human-likeness, and 4 for reliability. This index system prioritizes user needs and experiences and can practically quantify the service quality of AI health care chatbots. The proposed index system will provide valuable support for health policymakers, service managers, and developers by enabling benchmark comparisons, facilitating quality monitoring, and guiding continuous service enhancement.

Acknowledgments

The authors attest that no generative artificial intelligence technologies were used in the generation of text, figures, or other informational content of this manuscript. The authors acknowledge all experts who participated in the Delphi process and the analytic hierarchy process.

Data Availability

The datasets used or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study was supported by the Social Science Fund of Capital Medical University, Beijing, China.

Authors' Contributions

YG conceptualized the study, designed the consulted index system, interpreted the data, and wrote the manuscript. XW collected and analyzed the data. Both authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy.
[DOCX File , 18 KB-Multimedia Appendix 1]

References

1. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res. Jul 23, 2024;26:e56930. [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
2. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. Dec 19, 2023;6(1):236. [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
3. Iacobucci G. The AI bot will see you now: how technology is changing the doctor-patient relationship. BMJ. Mar 28, 2024;384:q711. [doi: 10.1136/bmj.q711] [Medline: 38548277]
4. Ng JY, Maduranayagam SG, Suthakar N, Li A, Lokker C, Iorio A, et al. Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: an international cross-sectional survey. Lancet Digit Health. Jan 2025;7(1):e94-102. [FREE Full text] [doi: 10.1016/S2589-7500(24)00202-4] [Medline: 39550312]
5. Mayer CJ, Mahal J, Geisel D, Geiger EJ, Staatz E, Zappel M, et al. User preferences and trust in hypothetical analog, digitalized and AI-based medical consultation scenarios: an online discrete choice survey. Comput Human Behav. Dec 2024;161:108419-108460. [FREE Full text] [doi: 10.1016/j.chb.2024.108419]
6. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" the patient's perspective: cross-sectional study. J Med Internet Res. Oct 01, 2024;26:e58831. [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
7. Moore I, Magnante C, Embry E, Mathis J, Mooney S, Haj-Hassan S, et al. Doctor AI? A pilot study examining responses of artificial intelligence to common questions asked by geriatric patients. Front Artif Intell. Jul 25, 2024;7:1438012. [FREE Full text] [doi: 10.3389/frai.2024.1438012] [Medline: 39118788]
8. Lee D, Brown M, Hammond J, Zakowski M. Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard. Int J Obstet Anesth. Feb 2025;61:104317. [doi: 10.1016/j.ijoa.2024.104317] [Medline: 39754839]

9. 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. Oct 2024;190:105562. [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105562] [Medline: 39059084]
10. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic Hepatitis B: cross-language study of English and Chinese. JMIR Med Inform. Aug 08, 2024;12:e56426. [FREE Full text] [doi: 10.2196/56426] [Medline: 39115930]
11. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. Oct 01, 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
12. Stapleton P, Santucci J, Cundy TP, Sathianathen N. Quality of information on Wilms tumor from artificial intelligence chatbots: what are your patients and their families reading? Urology. Apr 2025;198:130-134. [doi: 10.1016/j.urology.2025.01.054] [Medline: 39914668]
13. Owens OL, Leonard MS. Evaluating an AI chatbot "Prostate Cancer Info" for providing quality prostate cancer screening information: cross-sectional study. JMIR Cancer. May 21, 2025;11:e72522. [FREE Full text] [doi: 10.2196/72522] [Medline: 40397820]
14. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F, et al. The doc versus the bot: a pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. Gynecol Oncol Rep. Aug 08, 2024;55:101477. [FREE Full text] [doi: 10.1016/j.gore.2024.101477] [Medline: 39224817]
15. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res. Apr 30, 2024;26:e54706. [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
17. Parasuraman A, Zeithaml VA, Berry LL. A conceptual model of service quality and its implications for future research. J Mark. Sep 01, 1985;49(4):41-50. [FREE Full text] [doi: 10.1177/002224298504900403]
18. Karasan A, Erdogan M, Cinar M. Healthcare service quality evaluation: an integrated decision-making methodology and a case study. Socio-Econ Plan Sci. Aug 2022;82:101234. [FREE Full text] [doi: 10.1016/j.seps.2022.101234]
19. Zheng QL, Kong LN, Hu P, Liu DX. Identifying quality indicators for home care services: a modified Delphi and Analytic Hierarchy Process study. BMC Nurs. Jul 19, 2024;23(1):494. [FREE Full text] [doi: 10.1186/s12912-024-02169-4] [Medline: 39026316]
20. Lemmen C, Woopen C, Stock S. Systems medicine 2030: a Delphi study on implementation in the German healthcare system. Health Policy. Jan 2021;125(1):104-114. [FREE Full text] [doi: 10.1016/j.healthpol.2020.11.010] [Medline: 33288301]
21. Xia M, Liu Q, Ma L, Wen J, Xue Y, Hu H, et al. Developing an evaluation index system for service capability of internet hospitals in China: mixed methods study. J Med Internet Res. Jul 25, 2025;27:e72931. [FREE Full text] [doi: 10.2196/72931] [Medline: 40712159]
22. Jiang F, Liu T, Zhou H, Rakofsky JJ, Liu H, Liu Y, et al. Developing medical record-based, healthcare quality indicators for psychiatric hospitals in China: a modified Delphi-Analytic Hierarchy Process study. Int J Qual Health Care. Dec 31, 2019;31(10):733-740. [doi: 10.1093/intqhc/mzz005] [Medline: 30753601]
23. Liu M, Yu Q, Liu Y. Developing quality indicators for cancer hospitals in China: a national modified Delphi process. BMJ Open. Apr 09, 2024;14(4):e082930. [FREE Full text] [doi: 10.1136/bmjopen-2023-082930] [Medline: 38594187]
24. Saaty RW. The analytic hierarchy process—what it is and how it is used. Math Model. 1987;9(3-5):161-176. [FREE Full text] [doi: 10.1016/0270-0255(87)90473-8]
25. Geng Y, Zhao L, Wang Y, Jiang Y, Meng K, Zheng D. Competency model for dentists in China: results of a Delphi study. PLoS One. Mar 22, 2018;13(3):e0194411. [FREE Full text] [doi: 10.1371/journal.pone.0194411] [Medline: 29566048]
26. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy. Jul 23, 2024;9(1):28. [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]
27. Li Z, Guo R. Developing online medical service quality indicators in China from the perspective of online and offline integration: a modified Delphi-analytic hierarchy process study. Int J Qual Health Care. Jun 16, 2023;35(2):mzad038. [FREE Full text] [doi: 10.1093/intqhc/mzad038] [Medline: 37279543]
28. Barletta VS, Caivano D, Colizzi L, Dimauro G, Piattini M. Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. Int J Med Inform. Feb 2023;170:104951. [doi: 10.1016/j.ijmedinf.2022.104951] [Medline: 36525800]
29. Sobowale K, Humphrey DK. Evaluating the quality of psychotherapy conversational agents: framework development and cross-sectional study. JMIR Form Res. Jul 02, 2025;9:e65605. [FREE Full text] [doi: 10.2196/65605] [Medline: 40600851]
30. Yang Y, Liu S, Lei P, Huang Z, Liu L, Tan Y. Assessing usability of intelligent guidance chatbots in Chinese hospitals: cross-sectional study. Digit Health. Jun 06, 2024;10:20552076241260504. [FREE Full text] [doi: 10.1177/20552076241260504] [Medline: 38854920]

31. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan LN. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. BMC Med Inform Decis Mak. Dec 24, 2024;24(1):404. [FREE Full text] [doi: 10.1186/s12911-024-02824-5] [Medline: 39719573]
32. Motegi M, Shino M, Kuwabara M, Takahashi H, Matsuyama T, Tada H, et al. Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. Sci Rep. Jul 01, 2025;15(1):21346. [FREE Full text] [doi: 10.1038/s41598-025-06769-1] [Medline: 40596359]
33. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res. Jun 05, 2020;22(6):e18301. [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
34. Choi J, Kim JW, Lee YS, Tae JH, Choi SY, Chang IH, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep. Jan 17, 2024;14(1):1542. [FREE Full text] [doi: 10.1038/s41598-024-51531-8] [Medline: 38233511]
35. Chagas BA, Pagano AS, Prates RO, Praes EC, Ferreguetti K, Vaz H, et al. Evaluating user experience with a chatbot designed as a public health response to the COVID-19 pandemic in Brazil: mixed methods study. JMIR Hum Factors. Apr 03, 2023;10:e43135. [FREE Full text] [doi: 10.2196/43135] [Medline: 36634267]
36. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. Mhealth. Jul 2024;10:22. [FREE Full text] [doi: 10.21037/mhealth-23-55] [Medline: 39114462]
37. Noor N, Rao Hill S, Troshani I. Developing a service quality scale for artificial intelligence service agents. Eur J Mark. May 09, 2022;56(5):1301-1336. [FREE Full text] [doi: 10.1108/EJM-09-2020-0672]
38. Chen Q, Gong YM, Lu YB, Tang J. Classifying and measuring the service quality of AI chatbot in frontline service. J Bus Res. Jun 2022;145:552-568. [doi: 10.1016/j.jbusres.2022.02.088]
39. Chen S, Wang P, Wood J. Exploring the varying effects of chatbot service quality dimensions on customer intentions to switch service agents. Sci Rep. Jul 02, 2025;15(1):22559. [FREE Full text] [doi: 10.1038/s41598-025-06490-z] [Medline: 40596391]
40. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv. Preprint posted online April 15, 2017. [FREE Full text] [doi: 10.48550/arXiv.1704.04579]
41. Pattanshetti S. Extensive study: performance, metrics and usability of chatbot. Int Res J Eng Technol. Sep 2021;8(9):1527-1534. [FREE Full text]
42. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, et al. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput. Jul 21, 2021;26(1):95-119. [FREE Full text] [doi: 10.1007/S00779-021-01582-9]
43. Pelau C, Dabija DC, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav. Sep 2021;122:106855. [doi: 10.1016/j.chb.2021.106855]
44. Liu YL, Hu B, Yan W, Lin Z. Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots in mainland China and Hong Kong. Comput Human Behav. Jun 2023;143:107716. [doi: 10.1016/j.chb.2023.107716]
45. Chopra A, Ranjani KS, Narsipur S. Service quality dimensions in AI-enabled chatbots leading to customer satisfaction: a study from South Asia. IIFT Int Bus Manag Rev. 2023;1(1):26-49. [doi: 10.1177/jiift.221150355]
46. Shahzad MF, Xu S, An X, Javed I. Assessing the impact of AI-chatbot service quality on user e-brand loyalty through chatbot user trust, experience and electronic word of mouth. J Retail Consum Serv. Jul 2024;79:103867. [FREE Full text] [doi: 10.1016/j.jretconser.2024.103867]
47. Young KS, Lee SG, Hong GY. User satisfaction with the service quality of ChatGPT. Serv Bus. Sep 02, 2024;18(3-4):417-431. [FREE Full text] [doi: 10.1007/S11628-024-00566-y]
48. Fanous A, Steffner K, Daneshjou R. Patient attitudes toward the AI doctor. Nat Med. Nov 23, 2024;30(11):3057-3058. [doi: 10.1038/s41591-024-03272-4] [Medline: 39313596]
49. Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: a comparative assessment. Int J Med Inform. Sep 2025;201:105948. [doi: 10.1016/j.ijmedinf.2025.105948] [Medline: 40288015]
50. Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M. Assessing the quality and reliability of ChatGPT's responses to radiotherapy-related patient queries: comparative study with GPT-3.5 and GPT-4. JMIR Cancer. Apr 16, 2025;11:e63677. [FREE Full text] [doi: 10.2196/63677] [Medline: 40239208]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence
LLM: large language model

Edited by A Stone; submitted 28.Aug.2025; peer-reviewed by MM Al-Tehewy, K Malas; comments to author 24.Oct.2025; accepted 30.Dec.2025; published 18.Feb.2026

Please cite as:
Gu Y, Wang X
Developing a Service Quality Index System for AI Health Care Chatbots: Mixed Methods Study
J Med Internet Res 2026;28:e83051
URL: https://www.jmir.org/2026/1/e83051
doi: 10.2196/83051
PMID: