Metrics of Success: Evaluating User Satisfaction in AI Chatbots

Møller, Cecilie Grace; Ang, Ke En; Bongiovanni , Maria de Lourdes; Khalid, Md Saifuddin; Wu, Jiayan

Published in:
Proceedings of the 8th International Conference on Advances in Artificial Intelligence (ICAAI 2024)

Link to article, DOI:
10.1145/3704137.3704182

Publication date:
2025

Document Version
Peer reviewed version

Metrics of Success: Evaluating User Satisfaction in AI Chatbots

Cecilie Grace Møller
Department of Digitalization
Copenhagen Business School
Fredriksberg, Denmark
ceciliegrace@hotmail.com

Ke En Ang*
Department of Digitalization
Copenhagen Business School
Fredriksberg, Denmark
joanake@me.com

María de Lourdes Bongiovanni*
Department of Digitalization
Copenhagen Business School
Fredriksberg, Denmark
lourdes.bongio@gmail.com

Md Saifuddin Khalid
Department of Applied Mathematics
and Computer Science
Danmarks Tekniske Universitet
Lyngby, Denmark
skhalid@dtu.dk

Jiayan Wu
Department of Applied Mathematics
and Computer Science
Danmarks Tekniske Universitet
Lyngby, Denmark
jiawu@dtu.dk

Abstract

The rapid advancement of Artificial Intelligence (AI), particularly through Large Language Models (LLMs), has catalysed a technological revolution, leading to the widespread adoption of AI-driven chatbots across industries. OpenAI's customisable generative pre-trained transformer (GPT) offerings have popularised generative AI, enabling organisations of all sizes to implement chatbots for customer support. This development presents an opportunity for businesses to offer 24/7, cost-efficient customer service that can overcome the historical limitations of chatbots that lack a "human element." However, despite the proliferation of AI chatbots, there remains a crucial need to evaluate their effectiveness in meeting user needs and preferences for human-like interaction. Current service quality assessment tools, such as SERVQUAL and E-SERVQUAL, are unable to evaluate AI-specific capabilities like language intelligence and recognition. Existing research also lacks information on factors that affect user satisfaction and the continued use of AI chatbots. Based on a mixed-methods study, this paper proposes a new instrument for measuring user satisfaction with AI chatbots, specifically for customer support roles. Using the Stanford five-step Design Thinking Process, this study devised a customer support AI chatbot evaluation instrument through a literature review, Cheatstorming, and SCAMPER techniques, followed by testing in a Danish company. The research employs Prentice and Nguyen's three-stage scale development process to ensure content, reliability, and construct validity, addressing gaps in current scholarship and advancing understanding of AI chatbot user satisfaction.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Human-centered computing** → **User studies**;

ICAAI 2024, October 17–19, 2024, London, United Kingdom

ACM ISBN 979-8-4007-1801-4/24/10
https://doi.org/10.1145/3704137.3704182

Keywords

artificial intelligence, chatbots, user satisfaction, scale development, AI chatbot evaluation

ACM Reference Format:
Cecilie Grace Møller, Ke En Ang, María de Lourdes Bongiovanni, Md Saifuddin Khalid, and Jiayan Wu. 2024. Metrics of Success: Evaluating User Satisfaction in AI Chatbots. In *2024 The 8th International Conference on Advances in Artificial Intelligence (ICAAI 2024), October 17–19, 2024, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3704137.3704182

1 Introduction

Recent advancements in AI, particularly with the emerging popularity of LLM-driven solutions, have resulted in a technological "space race" to develop and adopt LLMs that power Generative AI chatbots such as ChatGPT powered by OpenAI. The availability of customisable offerings like GPTs introduced by OpenAI in November 2023, has popularized Generative AI for the masses, leading to the adoption of GPT-enabled chatbots for both internal and external use by many private corporations.

While chatbots have existed in various forms for some time, the recent advancement in AI has removed the initial limitations of chatbot effectiveness, being marked by "an absence of the human element" [17]. It is crucial to evaluate if AI chatbots implemented by organisations can serve the needs of their users, by evaluating the ability of the AI to interact and resolve issues with users' satisfaction. In addition, it is important to evaluate the degree to which AI chatbots can replicate interactions with humans, to resolve user preference for humans over chatbots.

However, existing instruments used to evaluate the quality of service, such as SERVQUAL and E-SERVQUAL, are unclear. These measurement scales fail to assess and capture novel capabilities like language intelligence and recognition, which are unique to AI chatbots [7]. In addition, current scholarship lacks sufficient insights into which factors affect user satisfaction levels and also their intent to continue using AI chatbots [8].

Ashfaq et al. [3] further claim that researchers have paid little attention to understanding the antecedents that influence user satisfaction and retention intention. Recognising the extensive potential of AI chatbot implementation and current limitations in existing

academic knowledge, the following research study is proposed to answer the question, "How can customer support teams evaluate AI chatbot user satisfaction?".

2 Literature Review

This review synthesizes existing empirical studies on design cues, adoption and user experience, and service quality measurement of AI chatbots. In addition, evaluation instruments from software engineering and product feature prioritization are briefly reviewed, which were possible directions for this study.

2.1 Dialogic Communication and Anthropomorphic Design Cues

A survey conducted via Amazon Mechanical Turk collected 958 responses from US residents about the satisfaction of their use of over 30 different chatbot services from Fortune 500 companies [14]. Participants answered 37 items that were scored on a seven-point Likert scale measuring six different variables—Responsiveness, A Conversational Tone (ACT), Satisfaction with Chatbot Services (SCS), Customers' Social Media Engagement (CSME), Price Premium (PP) and Purchase Intention (PI). The findings indicated that Dialogic Communication efforts were identified as positive predictors of customers' satisfaction, particularly Responsiveness and ACT.

An experiment involving a chatbot that simulates customer service scenarios in online banking, was conducted with 153 participants [1]. Four variables were measured—Anthropomorphic Design Cues (ADC), Foot-in-the-Door Technique (FITD), Social Presence (SP) and User Compliance (UC) through 17 items rated on a seven-point Likert scale. Findings showed that participants confronted with an anthropomorphically design chatbot or exposed to FITD are significantly more inclined to consent to the chatbot's request. Although SP is a significant mediator of ADC on UC, it is not a moderator of the effect of FITD on UC. Participants knew that they were interacting with a chatbot, but continued to apply similar social expectations as they would with humans. Therefore, companies should inform customers when they are interacting with chatbots, which may help manage expectations and enhance satisfaction. Lastly, chatbot dialogues should also be designed with as much care as user interfaces to mimic human-human communications.

2.2 Adoption and User Experience of AI Chatbots

A survey was conducted with 370 chatbot users from the United States, consisting of 33 items that measured seven variables: Information Quality (IQ), Service Quality (SQ), Perceived Enjoyment (PE), Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Need for Interaction with a Service Employee (NFI-SE), Satisfaction and continuity Intention (CI) [3]. All variables were evaluated using a seven-point Likert scale, except for satisfaction.

The variables were derived from existing models including the Expectation-Confirmation Model (ECM), Information System Success (ISS) Model, Technology Acceptance Model (TAM) and NFI-SE, was incorporated as a moderator to account for individualist differences. The findings highlighted that IQ and SQ have a positive impact on user satisfaction, which affects CI on the usage of chatbots, aligning with existing research. Furthermore, PU and PE impact both satisfaction and CI, whereas PEOU only has a significant impact on CI. The inclusion of NFI-SE validated that users with a preference for human service will be less satisfied with the chatbot as compared to users who do not, indicating that individual preferences can significantly impact Satisfaction.

A survey on adoption was conducted on 1,064 US consumers who interacted with chatbots from 30 leading US brands[8]. 28 items were scored on a seven-point Likert scale measuring eight variables—Information, Entertainment, Media Appeal, Social Presence, Perceived Privacy Risk, User Satisfaction, Customer Loyalty, and Continued Use. The first four variables originated from the Uses and Gratification (U&G) theory that differentiates user gratification into four individual aspects corresponding to the respective four variables. Findings indicated that all four gratifications are positively related to user satisfaction, with technology being the most prominent gratification form. High quality and efficiency of service provided by chatbots result in higher satisfaction, which in turn results in continued use intentions and increased brand loyalty. Thus, making Technology and Utilitarian gratifications more crucial contributors to user satisfaction. However, Perceived Privacy Risk was also an important determinant that could reduce User Satisfaction and Continued Use, if users were concerned about information misuse.

2.3 Measuring Service Quality of AI Chatbots

Chen et al. [7] developed a scale to assess AI Chatbot Service Quality (AICSQ) consisting of 95 items measuring seven constructs and 18 sub-constructs—Semantic Understanding, Close Human-AI Collaboration, Human-like, Continuous Improvement, Personalisation, Culture Adaption and Efficiency, through a three-phase process. The final refined scale and success model ran through a nomological test with responses collected from 597 participants across 14 countries, who had prior AI chatbot use experience. The findings supported that all seven dimensions of AICSQ positively influence user satisfaction, perceived value, and the intention to continue using AI chatbots, while the perceived value and satisfaction also influence continuance intention. In contrast to the ISS model, where service quality comprises information and system quality, for AI chatbot services, the system information, and service quality aspects are inseparable. This distinguishes AICSQ from ISS and establishes a need for differentiated evaluation of AI chatbot services.

A survey of 219 participants in Taiwan was conducted to assess e-service quality from customer service chatbots, scored on a five-point Likert scale measured via 47 items [13]. The study consisted of three variables—Core AI Bot Service Quality (measured using E-S-QUAL), AI Bot Recovery Quality (measured using E-RecS-QUAL), and AI Bot Conversational Quality, of which the final variable proposed comprised three factors from the Input-Process-Output (IPO) model: understanding humanness (input), perceived contingency (process) and response humanness (output). The findings showed that AI Bot Recovery Quality and AI Bot Conversational Quality significantly influence user satisfaction, with the latter having the

strongest impact, thus emphasising the importance of human-like interactions performed by AI chatbots. On the other hand, Core AI Bot Service Quality, which does not influence satisfaction, has a direct influence on loyalty. This indicates that chatbots, which fail to provide fundamental functionalities (hygiene factors that prevent dissatisfaction), will lead to user abandonment.

2.4 Motivators vs. Hygiene Factors

The concept of hygiene factors was first proposed by Herzberg [11] who classified the presence of job factors that result in satisfaction as Motivators, in contrast to job factors that prevent dissatisfaction as Hygiene factors. This was adapted to customer satisfaction research as Kano's Two-Factor Theory [5], which classifies requirements as Attractive, Must-be and One-dimensional, where Attractive elements correspond to Motivation factors which result in satisfaction, and Must-be elements correspond to Hygiene factors which prevents dissatisfaction.

2.5 Evaluating User Satisfaction

Existing literature presented two potential pathways for evaluating user satisfaction with AI chatbots—(1) evaluation for functionality prioritisation, testing and other common usability metrics [4, 5], and (2) furthering theory development on evaluation of user satisfaction post-AI chatbot implementation. Although existing research offers instruments tailored for the evaluation of user satisfaction, they are generally industry or context-specific and there is no consensus on what factors influence user satisfaction. By focusing on synthesising a generalisable instrument that can be used to evaluate user satisfaction, this study addresses an opportunity to further theory development through critical analysis of existing scales developed for AI chatbots.

3 Methodology

This research employed a mixed-method exploratory approach, valuable for clarifying uncertain phenomena and formulating or redrafting questionnaire items [16]. The Stanford five-step Design Thinking Process was applied throughout the research process [18], where a literature review on existing chatbot evaluation instruments was conducted during the Empathise and Define stages, providing defined constructs and items for scale development. During the Ideation stage, existing research was synthesized into a generic scale using Cheatstorming and SCAMPER methods [9, 10]. In the Prototype phase, the instrument was implemented on Microsoft Forms and tested, by conducting interviews with (1) two key managers of the case company, and (2) four individuals selected by convenience sampling. During scale development, Prentice and Nguyen's three-stage process was adapted for methodological presentations, with Hinkin's Scale Development Process consistently cross-referenced throughout [12, 15].

4 Developing and Testing the Instrument

The three-stage process is employed, involving a mix of qualitative and quantitative methods, to ensure the reliability, content and construct validity of the scale [12, 15]. The questionnaire received 16 employee responses, of which 11 were included in the analysis, as the remaining five respondents indicated that they had never used the AI service chatbot before.

4.1 Stage 1: Item generation

4.1.1 ***Domain specification.*** In this phase, a literature review was done to specify the domain of construct, which resulted in the selection of six key papers, with instruments measuring various dimensions of AI chatbots linked to user satisfaction, totalling in 26 dimensions and 256 questionnaire items for refinement.

4.1.2 ***Initial content validity assessment.*** The 256 items were analysed for conceptual consistency, with poorly worded and ambiguous questions removed to prevent validity issues and ensure accurate conceptual capture. The 26 dimensions were further evaluated, retaining those with the most rigorous reliability and validity results.

4.1.3 ***Item generation.*** The relevance of items and dimensions found in the literature was reviewed iteratively. Dimensions that measured the same construct, despite variations in wording, were combined. The phrasing of items was revised to eliminate ambiguity and avoid negatively worded questions [12]. This shortlisting process served as a second content validity assessment, with the research group evaluating items as 'essential', 'useful but not essential' or 'not necessary' to provide adequate coverage of the investigative questions [16]. Thus reducing to 41 items measuring eight constructs—*Humanness (or ADC)*, *Dialogic Communication*, *Information Quality*, *Perceived Privacy Risk*, *Perceived Usefulness*, *Human-AI Collaboration*, *Satisfaction*, and *Continuance Intention*. Of the eight shortlisted constructs, the first six are quality dimensions influencing *Satisfaction* and resulting in *Continuance Intention*. New items for the *Satisfaction* dimension were derived from hedonic qualities from Benyon [4] as existing items lacked comprehensiveness. Items for each construct were measured on a five-point Likert scale. As shown in Table 1, question 1 is a five-point Likert scale ranging from "Almost Never" (1) to "Almost Always" (5); question 2 is a Yes/No question; question 3 onwards are all five-point Likert scale that ranges from "Strongly Disagree" (1) to "Strongly Agree" (5). It ensured sufficient variance in responses for statistical analyses as Lissitz and Green demonstrated that reliability stabilizes at five points which makes higher scales unnecessary [12].

4.2 Stage 2: Scale refinement

4.2.1 ***Instrument Evaluation.*** In the second stage, a Think Aloud Test [2] was conducted, involving four participants who shared their thoughts while completing the questionnaire. This exercise aimed to identify any questions that may be misconstrued or ambiguous. As a result, the scale was refined into a 40-item questionnaire, as shown in Table 1, with the elimination of an item within the *Satisfaction* dimension. Additionally, nine questions in the *Humanness*, *Perceived Privacy Risk*, *Perceived Usefulness*, and *Satisfaction* dimensions were also refined.

4.2.2 ***Questionnaire Administration.*** A leading Nordic automotive case company was secured to provide validation data and real-world feedback. The company has developed its own internal AI service chatbot, which functions as a knowledge management tool, helping employees find information to address their queries.

To align the questionnaire with the company's needs, several modifications were made. Optional questions on age and gender were added, along with inquiries about the frequency of use and proficiency with two AI chatbots (Microsoft Copilot and the company's internal AI chatbot. Item 24 was revised to state "The AI chatbot is a useful tool in my work life", which adjusted the questionnaire item to focus on the chatbot's utility in the workplace. Additionally, adjustments were made to items related to human-AI collaborations, as the company's chatbot was unable to redirect users to human employees. As a result, items 29 and 31 in the questionnaire were revised to enquire about attitudes and preferences, changing the wording from "I prefer" to "If I had the opportunity" or "I would like to". Lastly, a final open-ended question was included to gather general feedback, though it was excluded from Cronbach's alpha analysis.

4.3 Stage 3: Scale validation

4.3.1 ***Data collection and survey design.*** The quantitative phase involved the validation of the questionnaire by collecting real-world data, which was distributed as an online survey via the case company's intranet. Data was collected on Microsoft Forms anonymously, with the survey being distributed to employees who have access to the company's AI chatbot. All questions in the survey were mandatory and branching was included to redirect respondents who had not used the AI chatbot before, to the end of the survey.

4.3.2 ***Testing validity and reliability.*** Two properties were assessed to evaluate the instruments' empirical measurement: reliability (to ensure consistency) and validity (to measure intended constructs) [19]. These factors are crucial in the evaluation process to ensure the quality and trustworthiness of the instrument. To assess the reliability of the scale, the standardised Cronbach's alpha (Cronbach's α) method was calculated in Stata, which assesses the internal consistency among all items in the scale.

The standardised Cronbach's alpha formula is defined as:

$$\alpha_{std} = \frac{K\bar{r}}{(1 + (K-1)\bar{r})} \tag{1}$$

where K is the number of items and $\bar{r}$ is the average of all coefficients between the items (i.e. the mean of K). The value of Cronbach's denoted as α, ranges from zero to one, with values between 0.6 and 0.8 indicating high level of reliability. Due to sample size limitations, it was not possible to assess construct validity (comprising of convergent and discriminant validity), thus it is unknown if the instrument is indeed measuring the intended constructs [20].

The standardised Cronbach's alpha was applied to the responses received. The *Humanness* subscale consisted of 5-items (α= 0.59), *Dialogic Communication* subscale consisted of 5-items (α= 0.51), *Information Quality* subscale consisted of 8-items (α= 0.85), *Perceived Privacy Risk* subscale consisted of 3-items (α= 0.92), *Perceived Usefulness* subscale consisted of 6-items (α= 0.82), *Human-AI Collaboration* subscale consisted of 5-items (α= 0.81), *Satisfaction* subscale consisted of 5-items (α= 0.83) and *Continuance Intention* consisted of 2-items (α= 0.93).

Cronbach's alpha values for the measure of *Humanness* and *Dialogic Communication* are lower than 0.70, which presents low internal consistency and requires revision. Meanwhile, *Information Quality, Perceived Privacy Risk, Perceived Usefulness, Human-AI Collaboration* and *Continuance Intention* show acceptable values and high reliability, with Cronbach's alpha values higher than 0.8.

5 Discussion and Conclusion

The proposed instrument seeks to address the current gaps by evaluating the unique capabilities of AI chatbots and providing a tool to use in a post-implementation context. This research has identified various factors affecting user satisfaction with AI chatbots as noted in academic literature and aggregates dimensions that were overlooked in some research [3] but addressed in others [13, 14]—such as perceived risk and privacy concerns. Eight constructs were synthesised in a comprehensive scale that measures user satisfaction levels with AI chatbots—comprising of refinement of 256 items found in existing literature and new items assessing hedonic qualities to measure satisfaction, which was not covered in existing research. This generic instrument is designed to evaluate user satisfaction of AI customer support comprehensively and can be applied across industries and contexts. Future research can apply our instrument within organizations to guide the identification of key factors in the development process and to evaluate user satisfaction following the implementation of AI chatbots. The generalisability and applicability of the scale may be affected by differences in customer service practices across various sectors. Distinct factors may exert a stronger influence on satisfaction in certain industries, such as heightened privacy concerns in healthcare, or personalisation in hospitality. Industry-specific case studies could build on this limitation, potentially adapting the instrument to specific needs.

5.1 Limitations

We acknowledge that there may be additional aspects or constructs not covered by our literature review, despite our efforts to provide a thorough analysis. Nevertheless, from the literature review, we have identified six quality dimensions that contribute to the Continuance of Intention and Satisfaction in the context of AI chatbots. However, the interconnections between these dimensions remain unclear. Additionally, the validity of the proposed instrument has not yet been sufficiently evaluated and further test-retest procedures are necessary in future studies to ensure comprehensive validity testing. This scale was tested solely in the automotive industry, which may limit its generalisability to other contexts. Additional research is required to extend the validity and applicability of this study across various industries.

5.2 Scope of Future Work

This research can further be extended in several directions. Firstly, the evaluation instrument can be tested and retested across multiple companies that have implemented an AI service chatbot, either for internal or external purposes. In that context, we recommend that companies develop an internal satisfaction score range to establish a baseline for user satisfaction within the organisation. Alternatively, a non-parametric test could be employed, to define positive and negative score ranges, archiving the same objectives by ranking data across different score levels ranging from low to high and comparing the non-Gaussian distribution of ranks, as the data

Table 1: Refined 40-item scale after the instrument evaluation process

Constructs	No.	Question
	1	How often do you use chatbots (e.g. ChatGPT) or digital assistants (e.g. Siri,Google Assistant, or Alexa)?
Humanness/ Anthropomorphic Design Cues (ADC)[1, 13]	2	Did the AI chatbot provide its name and/or asked how you are doing?
	3	I perceive that the avatar or the voice of the AI chatbot is similar to a human being.
	4	The AI chatbot remembers what we have been talking about.
	5	The AI chatbot is able to continue the conversation, even if I switch languages.
	6	The AI chatbot can detect/understand my mood based on my tone.
Dialogic Communication[14]	7	The AI chatbot responds to my queries in a timely manner.
	8	I feel that the AI chatbot treats me as a real communication partner, allowing for a two-way dialogue.
	9	The AI chatbot can accurately understand what I mean.
	10	The responses provided by the AI chatbot are relevant to my previous inputs.
	11	The AI chatbot's conversation style is similar to a human being's Information Quality
Information Quality[3, 14]	12	Information provided by the AI chatbot is sufficient.
	13	Information provided by the AI chatbot is clear.
	14	Information provided by the AI chatbot is accurate.
	15	Information provided by the AI chatbot is reliable.
	16	Information provided by the AI chatbot is up-to-date.
	17	The AI chatbot service is available 24/7.
	18	The AI chatbot service is available from any device.
	19	The AI chatbot provides the information I need on time.
Perceived Privacy Risk[8]	20	I trust my interactions with the AI chatbot are confidential.
	21	I trust that the AI chatbot will not misuse my personal information.
	22	I trust the AI chatbot protects my information.
Perceived Usefulness[3]	23	The AI chatbot helps me accomplish things more quickly than a human would.
	24	The AI chatbot is a useful tool in my life.
	25	Using the AI chatbot enables me to reduce the task completion time.
	26	Using the AI chatbot allows me to be more productive.
	27	Using the AI chatbot helps me to perform many tasks more conveniently.
	28	The AI chatbot provides answers in a useful format.
Human-AI Collaboration[7]	29	I can switch between the AI chatbot and a human easily.
	30	If the AI chatbot cannot solve my query, I can easily find a human to help me.
	31	When I switch from the AI chatbot to a human, the human knows what I am asking.
	32	It bothers me to use the AI chatbot when I could talk to a human instead.
	33	Personal attention by a human is very important for me.
Satisfaction[4]	34	I enjoy using the AI chatbot.
	35	It was pleasant to have a conversation with the AI chatbot.
	36	I feel excited when interacting with the AI chatbot.
	37	When I am using the AI chatbot, I feel it's a valuable tool for me.
	38	I think the AI chatbot is an innovative way of providing service.
Continuance Intention[3, 8]	39	I intend to continue using the AI chatbot in the future.
	40	I will strongly recommend others to use the AI chatbot

may exhibit asymmetry [6]. Secondly, an additional direction for extending this research involves an evaluation of functionality prioritisation, which could entail the creation of a second scale based on Kano's or Herzberg's theories. This scale would help identify features that specifically influence user satisfaction or dissatisfaction. Such a scale would be valuable during the chatbot development phase and when assessing potential future enhancements.

The Cronbach's alpha analysis assessed internal consistency reliability, which is a necessary condition but in itself insufficient to show evidence of construct validity [12]. To establish construct validity, it is recommended to conduct factor analyses—both Exploratory Factor Analysis (EFA) and subsequently Confirmatory Factor Analysis (CFA). EFA can be used to determine if further item reduction is required or provide confirmation of latent factors

that have been retained by earlier theoretical analyses, through quantitative analysis of item loadings. This can validate if the scale items are reasonably independent of one another. CFA provides further evidence of construct validity through quantitative analysis to confirm the Goodness of Fit of the final factor structure derived from EFA. The use of CFA allows for evaluation of the Goodness of Fit, comparing the chi-square performance of a single common factor model against that of a multifactor model equivalent to the number of constructs present in the evaluation tool [12].

However, both EFA and CFA are susceptible to sample size effects. Further, construct validity testing was unable to be conducted due to the small sample size of the currently administered questionnaire. It is recommended that a minimum sample of 150 observations be collected for EFA, and a minimum sample size of 200 for CFA [12].

Thus, further data collection is required to conduct these analyses to show evidence of convergent and discriminant validity of the constructs included in the scale. In addition, further data collection may also improve Cronbach's alpha in *Humanness* and *Dialogic Communication*, and would also satisfy test-retest reliability, if conducted at two points in time from the same group of respondents. Finally, further testing across different industries would enhance the instrument's practical relevance. Validation across diverse settings could refine user satisfaction measures and improve its overall adaptability.

References

[1] Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* 31, 2 (2021), 427–445. https://doi.org/10.1007/s12525-020-00414-7

[2] Benedikte S Als, Janne J Jensen, and Mikael B Skov. 2005. Comparison of think-aloud and constructive interaction in usability testing with children. In *Proceedings of the 2005 conference on Interaction design and children*. 9–16. https://doi.org/10.1145/1109540.1109542

[3] Muhammad Ashfaq, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. 2020. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics* 54 (2020), 101473. https://doi.org/10.1016/j.tele.2020.101473

[4] David Benyon. 2019. *Designing user experience*. Pearson UK.

[5] Charles Berger, Robert Blauth, and David Boger. 1993. Kano's methods for understanding customer-defined quality. (1993).

[6] Liu Bin and Gao Jianbo. 2019. Understanding the non-Gaussian distribution of revealed comparative advantage index and its alternatives. *International Economics* 158 (2019), 1–11.

[7] Qian Chen, Yeming Gong, Yaobin Lu, and Jing Tang. 2022. Classifying and measuring the service quality of AI chatbot in frontline service. *Journal of Business Research* 145 (2022), 552–568. https://doi.org/10.1016/j.jbusres.2022.02.088

[8] Yang Cheng and Hua Jiang. 2020. How do AI-driven chatbots impact user experience? Examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use. *Journal of Broadcasting & Electronic Media* 64, 4 (2020), 592–614. https://doi.org/10.1080/08838151.2020.1834296

[9] The Interaction Design Foundation. 2024. Scamper: How to use the best ideation Methods. https://www.interactiondesign.org/literature/article/learn-how-to-use-the-best-ideation-methods-scamper [Accessed: (2024)].

[10] The Interaction Design Foundation. 2024. What is Cheatstorming. https://www.interaction-design.org/literature/topics/cheatstorming [Accessed: (2024)].

[11] Frederick Herzberg. 1966. *Work and the nature of man*. World Publishing Company.

[12] Timothy R Hinkin. 1998. A brief tutorial on the development of measures for use in survey questionnaires. *Organizational research methods* 1, 1 (1998), 104–121. https://doi.org/10.1177/109442819800100106

[13] Chin-Lung Hsu and Judy Chuan-Chuan Lin. 2023. Understanding the user satisfaction and loyalty of customer service chatbots. *Journal of Retailing and Consumer Services* 71 (2023), 103211. https://doi.org/10.1016/j.jretconser.2022.103211

[14] Hua Jiang, Yang Cheng, Jeongwon Yang, and Shanbing Gao. 2022. AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior* 134 (2022), 107329. https://doi.org/10.1016/j.chb.2022.107329

[15] Catherine Prentice and Mai Nguyen. 2021. Robotic service quality–Scale development and validation. *Journal of Retailing and Consumer Services* 62 (2021), 102661. https://doi.org/10.1016/j.jretconser.2021.102661

[16] Mark NK Saunders, Philip Lewis, and Adrian Thornhill. 2019. *Research methods for business students*. Pitman.

[17] Shavneet Sharma, Gurmeet Singh, Nazrul Islam, and Amandeep Dhir. 2022. Why do SMEs adopt artificial intelligence-based chatbots? *IEEE Transactions on Engineering Management* 71 (2022), 1773–1786. https://doi.org/10.1109/TEM.2022.3203469

[18] Stanford. 2023. An Introduction to Design Thinking Process Guide. https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf [Accessed: (2023)].

[19] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53. https://doi.org/10.5116/ijme.4dfb.8dfd

[20] Drew Westen and Robert Rosenthal. 2003. Quantifying construct validity: two simple measures. *Journal of personality and social psychology* 84, 3 (2003), 608. https://doi.org/10.1037/0022-3514.84.3.608