JUIT
ADMI Jurnal Ilmiah Teknik
OPEN ACCESS
JUIT Vol 5 No. 2 | May 2026 | ISSN: 2828-6936 (Print), ISSN: 2828-6901 (online), Page. 525-538

Assessing Conversational AI Usability: A System Usability Scale Evaluation of Dola AI

Arimbi Kurniasari[1*], Navizha Velisya AS[2]
Information System, Universitas Gunadarma, Indonesia

Article History
Received : July 17, 2026
Revised : July 25, 2026
Accepted : Aug 17, 2026
Published : Aug 22, 2026

Corresponding author*:
arimbi@staff.gundarma.ac.id

Cite This Article *[APA Style]:*
Kurniasari, A., & Velisya, N. A. (2026). Assessing Conversational AI Usability: A System Usability Scale Evaluation of Dola AI. *Jurnal Ilmiah Teknik*, *5*(2), 525–538.

DOI:
https://doi.org/10.56127/juit.v5i2.1324

Abstract: The rapid development of conversational artificial intelligence has increased the need to evaluate not only system functionality but also the quality of user interaction. Dola AI is an AI-based conversational assistant designed to support users through natural-language interaction; however, widespread use does not necessarily indicate satisfactory usability. **Objective:** This study aims to evaluate the perceived usability of Dola AI using the System Usability Scale (SUS) and to determine its overall level of user acceptance. **Method:** A quantitative descriptive approach was employed involving 50 Dola AI users. Data were collected through an online questionnaire containing the ten standard SUS items measured using a five-point Likert scale. The instrument was evaluated through validity and reliability testing, followed by standardized SUS score calculation and interpretation using adjective-rating and acceptability classifications. **Findings:** All questionnaire items met the validity criterion, while the instrument obtained a Cronbach's alpha of 0.6827**,** indicating adequate internal consistency based on the criterion adopted in the study. The overall SUS score was 69.6**,** placing Dola AI in the "OK" adjective category and the "Marginal" acceptability range. These results indicate that Dola AI provides an adequate level of perceived usability but has not yet reached a clearly high level of user acceptance. **Implications:** The findings provide practical evidence for improving Dola AI, particularly in terms of interaction consistency, perceived complexity, learnability, and user confidence. The study also demonstrates the applicability of SUS as a concise standardized instrument for evaluating conversational AI interfaces. **Originality:** This study provides a standardized empirical usability assessment of Dola AI, for which quantitative usability evidence remains limited. The findings establish an initial usability benchmark that can support future comparative studies and further evaluation of conversational AI applications using broader samples and complementary usability measures.

Keywords: conversational artificial intelligence; Dola AI; System Usability Scale; usability; user experience

INTRODUCTION

Artificial intelligence (AI), particularly systems supported by large language models and natural language processing, has substantially transformed human–computer interaction. One of the most visible developments is the emergence of conversational AI, which enables users to interact with digital systems through natural-language dialogue rather than relying exclusively on conventional graphical user interfaces. Conversational

agents are increasingly used to support information retrieval, customer service, personal assistance, learning, and task automation because they provide relatively direct and intuitive interaction mechanisms (Adamopoulou & Moussiades, 2020; Følstad & Brandtzaeg, 2020; Nicolescu & Tudorache, 2022). As conversational AI becomes embedded in everyday digital activities, usability becomes increasingly important because successful implementation depends not only on technical capability but also on whether users can interact with the system effectively, efficiently, and satisfactorily (Brooke, 1996; ISO, 2018). Therefore, usability evaluation is essential for determining whether the capabilities offered by AI-based systems translate into satisfactory user experiences.

Dola AI represents this emerging form of conversational AI by providing AI-assisted services through natural-language interaction. The application is designed to support users in performing various activities while minimizing barriers associated with conventional digital interfaces. Its interaction model includes conversational communication, multimodal input, and integration with communication platforms familiar to users. Such characteristics potentially reduce cognitive and operational effort because users can communicate with the system using interaction patterns that resemble everyday communication. Research on conversational interfaces suggests that natural-language interaction can improve accessibility and convenience when properly designed, although the resulting experience remains dependent on response quality, transparency, interaction consistency, and the system's ability to understand user intentions (Clark et al., 2019; Jain et al., 2018; Følstad et al., 2018). Accordingly, technological sophistication or widespread adoption alone cannot be considered evidence of satisfactory usability.

This issue is especially important because conversational AI differs substantially from conventional point-and-click systems. The quality of interaction may depend on whether the system maintains conversational context, responds appropriately, provides relevant information, and minimizes unnecessary interaction effort. Previous studies indicate that conversational-agent user experience involves multiple dimensions, including conversational quality, perceived usefulness, ease of interaction, accessibility, response performance, trust, and satisfaction (Tubin et al., 2022; Borsci et al., 2022; Nicolescu & Tudorache, 2022). Borsci et al. (2022), for example, demonstrated that conventional usability measures may not fully capture conversational attributes such as response quality and conversational efficiency, leading to the development of the Chatbot Usability Scale. More recent evidence also confirms that chatbot-specific usability measures can reliably

capture interaction quality across different systems and application contexts (Borsci et al., 2023; Borsci & Schmettow, 2024). Therefore, an empirical assessment is necessary to determine whether Dola AI provides an interaction experience that users perceive as usable.

Previous research on usability assessment can be grouped into a first stream focusing on general standardized usability measurement, particularly the System Usability Scale (SUS). SUS was originally introduced by Brooke (1996) as a concise ten-item questionnaire that provides a global assessment of perceived usability. It has subsequently become one of the most widely applied usability instruments because it is relatively quick to administer, technology-independent, and capable of generating a standardized score ranging from 0 to 100 (Bangor et al., 2008; Bangor et al., 2009; Lewis, 2018). Psychometric research has also demonstrated the reliability and practical validity of SUS across different types of interactive systems (Peres et al., 2013; Lewis & Sauro, 2018). Its application has expanded beyond conventional websites and software to contemporary AI-based systems. For example, usability research on ChatGPT has employed SUS to assess perceived effectiveness, efficiency, and satisfaction, demonstrating that the instrument can also provide a standardized usability indicator for generative AI systems (Setiawan et al., 2023).

A second research stream focuses specifically on usability and user experience in chatbots and conversational agents. Unlike conventional software, conversational systems require users to evaluate not only navigation and task completion but also dialogue quality, contextual understanding, conversational coherence, system responsiveness, and accessibility. Systematic reviews indicate that conversational-agent research has used highly heterogeneous evaluation methods, with many studies relying on self-developed questionnaires rather than standardized usability instruments (Tubin et al., 2022). Borsci et al. (2022) therefore developed the Chatbot Usability Scale to capture dimensions that may not be adequately represented by general-purpose usability measures, while subsequent cross-language validation provided further evidence for the reliability of chatbot-specific assessment (Borsci et al., 2023). Research on chatbot conversational design has similarly shown that linguistic, visual, and interactive design practices influence user satisfaction and engagement (Feine et al., 2019; Rapp et al., 2021; Villegas-Ch et al., 2023). Nevertheless, empirical evidence concerning Dola AI remains limited, particularly regarding standardized usability evaluation based on SUS. This creates a specific empirical

gap concerning whether the perceived usability of Dola AI reaches an acceptable level and which aspects of its interaction contribute positively or negatively to the user experience.

Accordingly, this study aims to evaluate the usability of Dola AI using the System Usability Scale (SUS). The study involves 50 users with diverse demographic characteristics and applies the ten SUS items to generate an overall usability score and determine its corresponding usability interpretation. The instrument is also evaluated for validity and reliability before the SUS calculation is performed, consistent with the analytical procedure adopted in the study. SUS was selected because of its brevity, standardized scoring mechanism, broad applicability across interactive technologies, and established psychometric properties (Brooke, 1996; Bangor et al., 2008; Peres et al., 2013; Lewis, 2018). By applying this instrument to Dola AI, the study seeks to provide empirical evidence regarding users' overall perceptions of the application's usability and to identify interaction aspects that may require further improvement.

Based on established usability theory and previous conversational-agent research, this study argues that Dola AI is expected to demonstrate an acceptable level of perceived usability, although user evaluations may vary across individual SUS dimensions. The conversational interaction model may enhance ease of use and learnability by reducing conventional navigation requirements, whereas interaction inconsistency, perceived complexity, and adaptation requirements may reduce users' overall experience (Brooke, 1996; Borsci et al., 2022; Tubin et al., 2022). The study therefore does not treat the total SUS score as the sole source of interpretation but also examines response patterns across individual SUS items as supplementary descriptive evidence. This approach enables the standardized evaluation of overall usability while acknowledging that conversational AI possesses interaction characteristics that may not be entirely captured by a single aggregate score.

RESEARCH METHOD

The unit of analysis in this study was individual users of Dola AI who had prior experience interacting with the application. The study focused on users' perceptions of Dola AI usability, particularly aspects related to ease of use, system complexity, consistency, learnability, confidence, and the need for technical assistance. A total of 50 respondents participated in the study. The individual user was selected as the unit of analysis because usability represents users' subjective evaluation of their interaction with

a system and should therefore be assessed directly from individuals who have experienced the application.

A quantitative descriptive research design was employed to measure and interpret the perceived usability of Dola AI. This approach was selected because the primary objective was to quantify users' usability perceptions and convert them into a standardized usability score rather than to examine causal relationships among multiple variables. Usability was evaluated using the System Usability Scale (SUS)**,** originally developed by Brooke (1996). SUS is a technology-independent instrument consisting of ten standardized statements and has been widely applied because of its simplicity, relatively low respondent burden, and established reliability for assessing interactive systems (Bangor et al., 2008; Peres et al., 2013; Lewis, 2018). Its standardized scoring procedure also enables the resulting usability score to be interpreted against established benchmarks.

The study used primary data obtained directly from Dola AI users. Respondents represented users with different demographic characteristics and levels of experience with digital applications. Data were obtained through a structured questionnaire containing the ten standard SUS statements. Five items (SUS1, SUS3, SUS5, SUS7, and SUS9) were positively worded, whereas the remaining five items (SUS2, SUS4, SUS6, SUS8, and SUS10) were negatively worded. Each statement was measured using a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). This alternating positive–negative structure follows the original SUS instrument and is intended to obtain an overall assessment of perceived system usability (Brooke, 1996; Lewis & Sauro, 2018).

Data collection was conducted by distributing the questionnaire to respondents who had interacted with Dola AI. Before completing the questionnaire, respondents were required to have sufficient experience using the application to evaluate its usability. The questionnaire consisted of respondent characteristics followed by the ten SUS items. The use of a structured questionnaire ensured that all respondents evaluated the same usability dimensions using an identical measurement scale. After collection, responses were screened for completeness before being included in the analysis. The final dataset consisted of 50 complete responses, which were subsequently processed for instrument assessment and SUS scoring.

Data analysis was performed in three stages. First, item validity was evaluated using the Pearson product–moment correlation between each item and the total score, while internal consistency reliability was assessed using Cronbach's alpha. Second, the valid

responses were transformed according to the standard SUS scoring procedure. For positively worded odd-numbered items, the contribution score was calculated as the response value minus 1; for negatively worded even-numbered items, it was calculated as 5 minus the response value. The resulting contribution scores, each ranging from 0 to 4, were summed and multiplied by 2.5 to produce an individual SUS score ranging from 0 to 100 (Brooke, 1996). Finally, the mean SUS score across the 50 respondents was calculated and interpreted using established SUS benchmarks and adjective-rating approaches (Bangor et al., 2009; Lewis, 2018). Item-level responses were additionally examined descriptively to identify usability aspects that received relatively positive or negative evaluations.

Table 1. System Usability Scale Instrument for Dola AI

Code	SUS item	Type
SUS1	I think that I would like to use Dola AI frequently.	Positive
SUS2	I found Dola AI unnecessarily complex.	Negative
SUS3	I thought Dola AI was easy to use.	Positive
SUS4	I think that I would need the support of a technical person to be able to use Dola AI.	Negative
SUS5	I found the various functions in Dola AI were well integrated.	Positive
SUS6	I thought there was too much inconsistency in Dola AI.	Negative
SUS7	I would imagine that most people would learn to use Dola AI very quickly.	Positive
SUS8	I found Dola AI very cumbersome to use.	Negative
SUS9	I felt very confident using Dola AI.	Positive
SUS10	I needed to learn a lot of things before I could get going with Dola AI.	Negative

Note. Items were adapted from the original System Usability Scale (Brooke, 1996) by replacing the generic reference to the evaluated system with "Dola AI." Responses were measured on a five-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

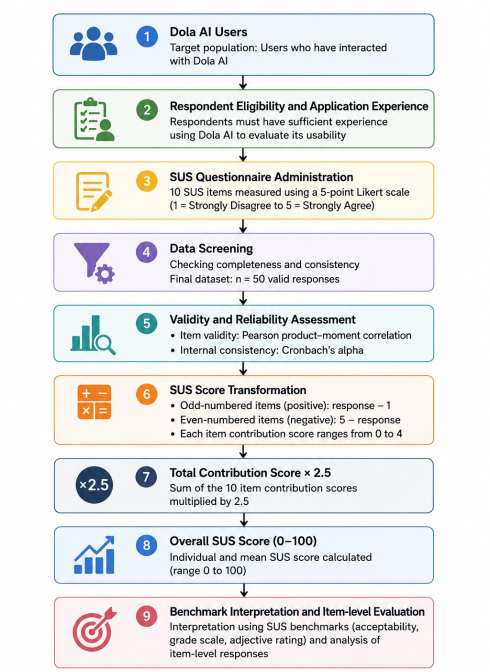

Figure 1. Research procedure for evaluating the usability of Dola AI using the System Usability Scale.

RESULT

Respondent Characteristics

A total of 50 respondents who had experience interacting with Dola AI participated in the usability evaluation. The respondents represented diverse demographic characteristics, including gender, age group, and occupational background. This demographic variation provided a broader representation of user perceptions toward Dola AI usability. The distribution of respondents is presented in Figure 2.

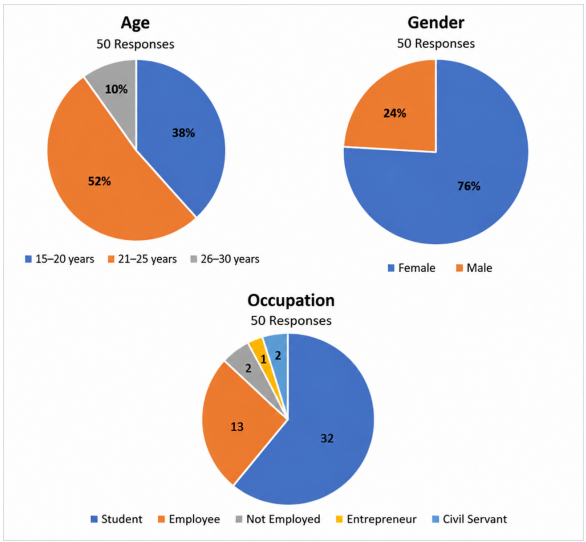

Figure 2. Respondent characteristics

Instrument Validity and Reliability

The validity test showed that all ten SUS items satisfied the predetermined validity criterion. The calculated item–total correlation coefficients ranged from 0.284 to 0.663, and all values exceeded the critical *r*-value of 0.279. Therefore, all SUS items were retained for the subsequent usability analysis. The highest correlation coefficient was observed for Q6 ($r = 0.663$), whereas the lowest was obtained for Q3 ($r = 0.284$).

The internal consistency assessment produced a Cronbach's alpha of 0.6827. Based on the criterion adopted in this study ($\alpha > 0.60$), the questionnaire was considered sufficiently reliable for measuring users' perceptions of Dola AI usability.

Table 2. Validity and Reliability Results of the SUS Instrument

Item	*r*-calculated	*r*-critical	Result
Q1	0.353	0.279	Valid
Q2	0.469	0.279	Valid
Q3	0.284	0.279	Valid
Q4	0.591	0.279	Valid
Q5	0.328	0.279	Valid
Q6	0.663	0.279	Valid
Q7	0.288	0.279	Valid
Q8	0.471	0.279	Valid
Q9	0.289	0.279	Valid
Q10	0.538	0.279	Valid
Cronbach's alpha	**0.6827**	**> 0.60**	**Reliable**

System Usability Scale Score

Following the standard SUS scoring procedure, the transformed scores of the ten items were summed for each respondent and multiplied by 2.5. The cumulative SUS score obtained from the 50 respondents was 3,480, resulting in a mean SUS score of 69.6.

Table 3. Summary of Dola AI Usability Evaluation

Indicator	Result
Number of respondents	50
Number of SUS items	10
Total SUS score	3,480
Mean SUS score	**69.6**
Adjective rating	**OK**
Acceptability range	**Marginal**
Overall interpretation	Moderate/adequate usability

The mean score of 69.6 indicates that Dola AI achieved a usability level close to the commonly used acceptable threshold. According to the interpretation scheme employed in the study, this score was classified as **"**OK**"** on the adjective-rating scale. On the acceptability scale, the score falls within the Marginal category because it lies between 50 and 70.

Overall Usability Classification

The combined interpretation of the SUS results indicates that Dola AI provides a generally usable interaction experience but has not yet reached a clearly high level of usability. The score of 69.6 positions the application near the transition from marginal to acceptable usability. Thus, the results indicate that users were generally able to interact with Dola AI, while some usability aspects still offer opportunities for improvement.

Table 4. Interpretation of the Dola AI SUS Score

Evaluation approach	Dola AI result	Interpretation
SUS score	69.6	Moderate usability
Adjective rating	OK	Usability is adequate
Acceptability	Marginal	Close to the acceptable range
Instrument validity	All 10 items valid	Items retained
Internal consistency	$\alpha = 0.6827$	Reliable according to the study criterion

DISCUSSION

The findings indicate that Dola AI achieved a mean System Usability Scale (SUS) score of 69.6**,** suggesting a moderate level of perceived usability. Based on the interpretation framework adopted in this study, this score corresponds to an **"**OK**"** adjective rating and falls within the **"**Marginal" acceptability range. Importantly, a SUS score should not be interpreted as a percentage; rather, it represents a standardized usability score that enables comparison with established benchmarks (Brooke, 1996; Bangor et al., 2008). The result therefore indicates that Dola AI provides a generally usable interaction experience, although its usability has not yet reached a consistently high level. The proximity of the score to the acceptable range further suggests that relatively targeted improvements in user interaction may increase overall perceived usability.

This finding is particularly relevant because Dola AI operates as a conversational AI system, where usability extends beyond conventional interface navigation. Users interact

with the application through natural-language conversations, making interaction quality dependent on factors such as learnability, conversational clarity, response consistency, perceived complexity, and user confidence. Previous studies have demonstrated that conversational-agent usability involves characteristics such as response quality, accessibility, functionality, and conversational efficiency (Borsci et al., 2022, 2023). Therefore, the moderate SUS score obtained in this study may reflect a combination of positive experiences associated with conversational interaction and remaining difficulties encountered during system use. In practical terms, users may perceive Dola AI as sufficiently straightforward for routine interaction while still encountering aspects that prevent the experience from being evaluated as highly usable.

The result can also be interpreted in relation to previous evaluations of AI-based conversational systems. SUS has increasingly been applied to generative AI and chatbot interfaces because it provides a standardized measure that facilitates comparison across interactive technologies. Previous evaluation of ChatGPT, for example, reported a SUS score of approximately 67.4**,** indicating that generative conversational systems may achieve moderate usability despite their advanced technological capabilities (Setiawan et al., 2023). Against this context, Dola AI's score of 69.6 suggests a broadly comparable usability profile. However, direct comparisons should be interpreted cautiously because differences in respondent characteristics, application context, task complexity, sampling procedures, and versions of AI systems may influence SUS scores. More importantly, the present finding reinforces the argument that sophisticated AI functionality does not automatically translate into superior user experience; usability remains dependent on how effectively technological capabilities are presented and experienced through the interaction interface.

The instrument assessment provides additional evidence regarding the measurement quality of the study. All ten SUS items exceeded the critical item–total correlation value of 0.279 and were therefore retained as valid indicators under the criterion applied in the study. The Cronbach's alpha coefficient of 0.6827 also exceeded the minimum reliability criterion of 0.60 adopted in the original analysis. Nevertheless, this reliability coefficient should be interpreted conservatively. Although it indicates adequate internal consistency according to the study criterion, it remains below the frequently used threshold of .70. Thus, the result supports the use of the instrument for the present exploratory usability assessment but also indicates that the consistency of responses across SUS items was not particularly

strong. This may partly reflect the heterogeneous nature of usability perceptions or differences in how respondents interpreted positive and negatively worded SUS statements.

The demographic composition of the sample also provides important context for interpreting the results. Most respondents were aged 21–25 years (52%)**,** followed by those aged 15–20 years (38%) and 26–30 years (10%). Female respondents constituted 76% of the sample, while male respondents accounted for 24%. Occupationally, students represented the largest group, comprising 32 of the 50 respondents, followed by employees with 13 respondents; the remaining respondents were distributed among smaller occupational categories. Consequently, the observed SUS score primarily represents the perceptions of relatively young and predominantly student users rather than the usability experience of the broader population. Because younger users may have greater familiarity with digital applications and conversational technologies, the usability score obtained in this study should not automatically be generalized to older or less digitally experienced populations.

From a practical perspective, the findings indicate that Dola AI should prioritize interaction refinement rather than merely expanding functionality. A marginal-to-acceptable usability position suggests that improvements should focus on reducing unnecessary complexity, increasing consistency, improving learnability, strengthening users' confidence during interaction, and minimizing situations in which users require additional assistance. These priorities correspond closely to the dimensions assessed by the SUS instrument. For conversational AI specifically, developers should also consider response clarity, conversational coherence, error recovery, feedback mechanisms, and transparency when the system cannot accurately interpret a request (Borsci et al., 2022). Improving these elements could reduce interaction friction and potentially shift users' evaluations from the current “OK” level toward “Good” or higher usability categories.

The study also contributes methodologically by demonstrating the applicability of SUS for providing a rapid global evaluation of Dola AI. However, SUS is a general-purpose usability instrument rather than a conversational-AI-specific measure. Consequently, the score of 69.6 summarizes overall perceived usability but does not directly identify conversational problems such as contextual misunderstanding, inappropriate responses, conversational breakdowns, perceived intelligence, or response relevance. Instruments specifically designed for conversational agents, such as the Chatbot Usability Scale, incorporate dimensions more closely associated with dialogue-based interaction (Borsci et

al., 2022, 2023). Future research could therefore combine SUS with chatbot-specific instruments, task-based usability testing, qualitative interviews, or interaction-log analysis to determine why users assign particular usability evaluations and which conversational mechanisms require improvement.

Finally, the findings should be interpreted in light of several limitations. First, the study involved only 50 respondents, limiting the generalizability of the results. Second, the demographic distribution was relatively concentrated among young users, women, and students, which may introduce sampling bias. Third, the study relied primarily on self-reported perceptions rather than objective usability indicators such as task-completion time, task-success rate, error frequency, or interaction efficiency. Fourth, the cross-sectional evaluation captures usability perceptions at a particular point in time, whereas conversational AI systems can change rapidly as their interfaces and underlying models are updated. Future studies should therefore employ larger and more heterogeneous samples, combine subjective and objective usability metrics, and conduct longitudinal or comparative evaluations across conversational AI platforms. Such extensions would provide a more comprehensive understanding of Dola AI usability and clarify whether improvements in system design produce measurable changes in user experience.

CONCLUSION

This study demonstrates that Dola AI provides a generally adequate level of usability based on the System Usability Scale (SUS). The mean SUS score of 69.6 places the application in the "OK" adjective category and within the "Marginal" acceptability range. These results indicate that users are generally able to interact with Dola AI effectively, but the application has not yet reached a consistently high level of usability. The validity assessment also showed that all ten SUS items were valid, while the Cronbach's alpha value of 0.6827 indicated adequate internal consistency based on the criterion adopted in this study.

The main scientific contribution of this study lies in providing empirical usability evidence for Dola AI, a conversational AI application that has received limited standardized usability evaluation. By applying SUS to Dola AI, the study contributes quantitative data that can be used as a baseline for future evaluations of conversational AI usability. The findings also demonstrate that a general usability instrument such as SUS can provide a concise overall assessment of user perceptions in conversational-AI environments, while highlighting the need to interpret the aggregate score together with

specific aspects such as learnability, interaction consistency, perceived complexity, and user confidence.

This study is nevertheless subject to several limitations. The evaluation involved only 50 respondents, with a demographic composition dominated by younger users and students, which limits the generalizability of the findings to broader user populations. In addition, the study relied primarily on self-reported perceptions and did not include objective usability indicators such as task completion time, error rate, task success, or interaction logs. Future research should therefore involve larger and more diverse samples and combine SUS with objective usability measures, qualitative user feedback, or conversational-AI-specific instruments to provide a more comprehensive evaluation of Dola AI usability.

REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations* (pp. 373–383). Springer. https://doi.org/10.1007/978-3-030-49186-4_31

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123.

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing, 26*, 95–119. https://doi.org/10.1007/s00779-021-01582-9

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2023). A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation. *Personal and Ubiquitous Computing, 27*, 317–330. https://doi.org/10.1007/s00779-022-01690-0

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., & Cowan, B. R. (2019). The state of speech in HCI:

Trends, themes and challenges. *Interacting with Computers, 31*(4), 349–371. https://doi.org/10.1093/iwc/iwz016

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies, 132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: Findings from a questionnaire study. *Quality and User Experience, 5*, Article 3. https://doi.org/10.1007/s41233-020-00033-2

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In S. S. Bodrunova (Ed.), *Internet Science* (pp. 194–208). Springer. https://doi.org/10.1007/978-3-030-01437-7_16

International Organization for Standardization. (2018). *Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts (ISO 9241-11:2018).* International Organization for Standardization.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (pp. 895–906). Association for Computing Machinery. https://doi.org/10.1145/3196709.3196735

Lewis, J. R. (2018). The System Usability Scale: Past, present, and future. *International Journal of Human-Computer Interaction, 34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R., & Sauro, J. (2018). Item benchmarks for the System Usability Scale. *Journal of Usability Studies, 13*(3), 158–167.

Nicolescu, L., & Tudorache, M. T. (2022). Human-computer interaction in customer service: The experience with AI chatbots A systematic literature review. *Electronics, 11*(10), 1579. https://doi.org/10.3390/electronics11101579

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the System Usability Scale (SUS). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57*(1), 192–196. https://doi.org/10.1177/1541931213571043

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies, 151*, 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

Setiawan, A., Kurniawan, E., & others. (2023). *Procedia Computer Science*.