Journal of Sustainable Technology in Agriculture

Open Access | www.jostapubs.com | ISSN(online): 3107-6882

Review-Article

November 2026, Volume 2, Issue IV. *doi: 10.65287/josta.202610.de78*

Methodology for Evaluating Artificial Intelligence Chatbots: A Review with Applications to Agricultural Advisory Systems

Farhana L S*

Abstract

Artificial intelligence chatbots increasingly advise farmers, yet their adoption should rest on rigorous, standardised evaluation. This review brings together five standardised instruments for assessing chatbots (the System Usability Scale, the Chatbot Usability Scale, the Checklist for Trust between People and Automation, the AI Chatbots Usage Scale, and the User Engagement Scale) and one assessment method, knowledge testing. It also explains Cronbach's alpha, the reliability criterion most of them rely on. A case study from the literature, in which five large language models answered PhD-level soil science questions, together with deployed systems such as Farmer.Chat and AgroLLM, show how evaluation works in practice and why fluent but unverified advice is risky. Beyond questionnaires, we examine accuracy-based and retrieval-grounded testing. We propose a five-layer protocol that first checks technical accuracy, then usability, trust, engagement, and learning outcomes, to guide the design and validation of chatbots for precision and smallholder farming.

Keywords: Cronbach's alpha, Trust in automation, Usability, User engagement, Large language models, Extension services.

Received: 2026-10-11 | *Accepted:* 2026-11-09 | *Published:* 2026-11-16

© 2026 The Author(s). Licensee: Journal of Sustainable Technology in Agriculture. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

AI is no longer confined to laboratories. It now shows up in precision farming (Liakos *et al.* 2018). As happened with computers and the internet, its real value lies in getting information to people faster and more fairly, rather than in taking over from human experts. This wider use has drawn attention to chatbots, programs that let people interact with a computer by typing or speaking in everyday language. Chatbots first became common in customer service and education. Extension services are the latest place to adopt chatbots (Anusha *et al.* 2025). Surveys trace the field from early rule-based programs to neural conversational agents (Adamopoulou and Moussiades 2020) and show that chatbot research now spans education, health, and service sectors (Okonkwo and Ade-Ibijola 2021; Rapp *et al.* 2021). In agriculture, chatbots have opened new routes for farmers to reach technical information, above all timely, locally relevant advice on crop management, pest and disease identification, market prices, and weather-related risks (Bhuvaneswari *et al.* 2024). This sits within the wider move towards digital and 'smart' farming (Wolfert *et al.* 2017; Liakos *et al.* 2018; Klerkx *et al.* 2019).

A conversational interface, however, says nothing about whether a tool is usable, trustworthy, or helps anyone learn. Showing that a chatbot really serves its users means measuring it systematically. Human-computer interaction (HCI) research has refined such measurement over many years, yet agricultural informatics applies it inconsistently. This review brings together the main scales used to evaluate AI chatbots, explains the statistical reasoning behind their reliability, and looks at how they have been applied in published case studies and real agricultural deployments. Our aim is to provide a reference framework for designing and validating future chatbots in the agricultural sciences.

The objectives of this review are therefore to (i) describe the main standardised instruments for evaluating AI chatbots; (ii) explain Cronbach's alpha as the reliability criterion they share; (iii) examine how these instruments and accuracy-based tests have been used in agricultural case studies and deployments; and (iv) propose a staged, multi-layer protocol for evaluating agricultural chatbots. The remainder of the paper follows this order: Section 2 sets out the review methods, Section 3 presents the instruments, the case studies, and evaluation beyond questionnaires, Section 4 discusses the methodological issues and the proposed protocol, and Section 5 concludes.

1.1. AI chatbots in agriculture: typology and deployment

Agricultural chatbots handle farmers' questions in several different ways. The simplest are rule-based systems that work from predefined keywords, menus, and decision paths. They perform well on questions the developers anticipated, but may be less effective when a farmer asks something that matches none of the keywords, menus, or stored responses. Other systems use Natural Language Processing (NLP) and Machine Learning (ML) to interpret less structured questions and generate responses. Some implementations combine rule-based and AI-based components, although the architecture varies across applications. For example, (Bharti *et al.* 2024) described a rule-based chatbot hosted on Facebook Messenger and evaluated it with 102 farmers in Odisha, India; the study should not be interpreted as evidence of a hybrid rule-based and NLP/ML system.

The type of interaction used by farmers is another factor that can be considered when classifying chatbots. Voice-enabled systems let users speak instead of type, which can matter in rural areas where literacy levels limit the use of text-based services; early work such as FarmChat explored conversational question answering for farmers in India (Jain *et al.* 2018). Context-aware chatbots draw on earlier messages, so a conversation can continue without the user repeating information. Chatbots built into familiar platforms such as WhatsApp, Facebook Messenger, and Telegram have a practical edge too, because farmers can reach the service through apps they already use and need not install anything new.

Several emerging systems show how these general architectures are being carried into agriculture. Farmer.Chat, developed by Digital Green with Microsoft Research, is one example. It uses a retrieval-augmented generation (RAG) framework (Lewis *et al.* 2020), so its answers are built from information retrieved from agricultural knowledge resources. According to (Singh *et al.* 2024), Farmer.Chat has reached more than 15,000 farmers across four countries and has handled over 300,000 queries. AgroLLM adopts a comparable approach by combining a large language model (LLM) with a RAG pipeline linked to an open-source agricultural knowledge base. Information retrieved from this knowledge base is incorporated into the response generation process, helping the system provide more reliable answers and reduce hallucinations (Samuel *et al.* 2025). Comparative work on agricultural question answering likewise found that both retrieval augmentation and fine-tuning improved the answers of a base model (Balaguer *et al.* 2024).

The application of conversational AI is now being explored in agricultural advisory programmes at the international level. The Generative AI for Agriculture (GAIA) initiative is led by the International Food Policy Research Institute (IFPRI), in partnership with CABI, SCiO, the University of Florida, and Digital Green, with support from the Gates Foundation. During Phase I (2023–2024), the initiative tested a retrieval-augmented generation (RAG) framework with Digital Green's Farmer.Chat in Kenya and India. Phase II (2025–2027) is building on this work through expanded content aggregation, stronger data governance, integration of real-time and multimodal information, and evaluation and benchmarking protocols (IFPRI 2026). These developments suggest that conversational systems are becoming part of agricultural information and extension services. That makes the quality and reliability of their output a serious concern, because a wrong recommendation on crop management, pests, or diseases can cost farmers money and, in some cases, affect food security (Singh *et al.* 2024).

1.2. Rationale for systematic evaluation

Evaluating AI chatbots serves several linked purposes. It measures task success and usability, exposes design weaknesses that can be fixed in later versions, shows whether systems keep pace with changing user needs, and gives developers, extension agencies, and funders the evidence they need to justify continued investment. Usability research has long cautioned that effectiveness, efficiency, and satisfaction capture different things and are often measured inconsistently (Hornbæk 2006; ISO 2018), and chatbot-specific work has proposed quality attributes such as performance, affect, humanity, and accessibility (Radziwill and Benton 2017). Reviews of dialogue-system evaluation likewise separate automatic metrics from human judgement (Deriu *et al.* 2021), and questionnaire studies show that users' chatbot experience is multi-dimensional (Følstad and Brandtzaeg 2020). In agriculture, the stakes of weak evalu-

ation are higher. In a Perspective informed by real-life testing of GPT-generated technical advice for Nigerian cassava farmers, (Tzachor *et al.* 2023) highlighted that fluent responses could remain broad and insufficiently tailored to crop growth stages and smallholder circumstances. The recommendations discussed included inputs that could exceed farmers' financial capacity. This review describes the resulting fluency–adequacy gap: persuasive language may conceal advice that is not sufficiently specific or practical for farmers. Several broader issues concerning the use of AI in agriculture were discussed by (Tzachor *et al.* 2022), including difficulties in integrating agricultural data, possible socio-ecological effects, and the risk that farmers with limited access to reliable connectivity could be left behind. The lesson is that fluency is not competence. Before chatbots are rolled out in settings where food security is at stake, they should be judged on measured usability, trust, and knowledge outcomes, not on the polish of their conversation.

2. Materials and methods

2.1. Review design

This is a narrative review. The review was designed to produce a methodological reference rather than a pooled estimate, so no meta-analysis was attempted.

2.2. Information sources and search strategy

We searched Google Scholar, ScienceDirect, SpringerLink, PubMed, and arXiv with combinations of the terms 'chatbot evaluation', 'usability scale', 'trust in automation', 'user engagement scale', 'AI chatbot agriculture', and 'large language model advisory'. Searches were extended with the terms 'retrieval-augmented generation agriculture', 'smart farming digital advisory', and 'Cronbach's alpha reliability'. The searches were last run on 14 September 2026. To find further sources, we hand-searched the reference lists of retrieved articles (backward snowballing).

2.3. Eligibility criteria

Only English-language sources published between 1996 (the year the System Usability Scale first appeared) and 2026 were considered, with the exception of foundational psychometric and theoretical works cited for definitions and background (Cronbach 1951; Davis 1989; Lewis 1995). We gave priority to (i) foundational papers on psychometric scale development, (ii) peer-reviewed validation studies reporting reliability coefficients, and (iii) empirical case studies applying these scales to chatbots, especially in agricultural and allied applied-science settings. Preprints (arXiv) were admitted where no peer-reviewed version could be found and are identified as such in the reference list. A scale was generally included if it had been published in a scholarly publication (journal, conference, or edited volume) and reported at least one reliability estimate (typically Cronbach's alpha). The CTPA was retained as an explicit exception because it is a foundational trust instrument central to the review, although the cited validation study did not report a reliability coefficient for the full 12-item scale; reliability evidence for an earlier 10-item revision is reported separately (Spain *et al.* 2008).

Instruments also had to have been used, or be plausibly applicable, to evaluate conversational AI systems.

2.4. Case-study selection and synthesis

No new data were collected. One case study from the existing literature, an assessment of five large language models on PhD entrance questions in soil science (Khanifar 2025), was chosen for detailed discussion because it evaluates chatbot performance in an agricultural domain. It is read alongside five further reported evaluations and deployments of agricultural chatbots (Table 2); sources available only as arXiv preprints are identified as such in the reference list. For each included source, we extracted the instrument or method, the number of items, the constructs measured, the reliability or accuracy evidence, the study population, and the application domain. These were synthesised narratively under five evaluation layers: technical accuracy and grounding, usability, trust, engagement and acceptance, and learning outcomes.

3. Results

3.1. Standardised scales for evaluating AI chatbots

The instruments and methods reviewed here measure different but complementary aspects of chatbot quality: overall usability, interaction quality specific to chatbots, user trust, sustained usage, engagement, and learning outcomes. Each is described below in terms of its structure, psychometric properties, and relevance to evaluating agricultural chatbots.

System Usability Scale (SUS)

(Brooke 1996) developed the System Usability Scale at Digital Equipment Corporation, and it remains one of the most widely used general-purpose usability instruments. It has ten statements, five worded positively and five negatively to counter acquiescence bias, rated on a five-point Likert scale and combined into a score from 0 to 100. Because SUS is not tied to any one technology, it has been applied to hardware, websites, mobile applications, and conversational agents alike. Its reliability is high (approximately $\alpha = 0.91$ across 2,324 questionnaires from 206 studies; (Bangor *et al.* 2008)), and a two-factor structure separating 'usability' from 'learnability' has been reported (Lewis and Sauro 2009). For farmer-assistance chatbots, SUS gives a quick, low-cost benchmark of whether an interface is intuitive and efficient enough to keep using. Its adjective ratings (e.g., 'good', 'excellent', 'best imaginable'; (Bangor *et al.* 2009)) also make scores easy for non-specialists, such as extension officers, to interpret.

Chatbot Usability Scale (BUS)

SUS was designed with general software in mind, not conversational agents. (Borsci *et al.* 2022) therefore developed the Chatbot Usability Scale (BUS) for AI-based conversational

systems. The BUS-15 has 15 items spread over five factors: accessibility of chatbot functions, quality of chatbot functions, quality of conversation and information provided, privacy and security, and response time. In a pilot with 96 volunteers who rated ten customer-service chatbots (385 valid questionnaires), the overall Cronbach's alpha was 0.87, and the scale correlated strongly with the Usability Metric for User Experience–Lite (UMUX-LITE; (Lewis *et al.* 2013)). For agricultural advisory chatbots, the BUS is especially useful because it asks directly whether the system keeps track of the conversation, gives accurate information, protects privacy, and responds promptly. These are the qualities that bear most on farmer trust and continued use.

Checklist for Trust between People and Automation (CTPA)

Trust is a precondition for adopting any automated recommendation system. The most frequently cited instrument for measuring it is the Checklist for Trust between People and Automation developed by (Jian *et al.* 2000), which we abbreviate as CTPA to avoid confusion with other trust questionnaires. It has 12 items covering trust (confidence, dependability, reliability, security, integrity) and distrust (deception, suspicion, wariness, perceived harm). The original authors treated these as opposite poles of one dimension, but later confirmatory factor analyses support a two-factor oblique structure in which trust and distrust are related yet distinguishable (Spain *et al.* 2008). (Spain *et al.* 2008, p. 1336) report that an earlier 10-item modified version used by (Fallon *et al.* 2005) had an internal-consistency reliability of $\alpha = 0.93$; they do not report a reliability coefficient for the full 12-item scale in their own study. The scale responds to changes in trust after system updates or visible errors. Farmers and extension workers will act on a chatbot's advice, which often involves costly inputs such as fertiliser or pesticide, only if they trust it. Trusting a system too much is costly, and so is trusting it too little. The trust literature therefore argues that reliance should match how reliable the system actually is (Lee and See 2004; Hoff and Bashir 2015). Trust in AI systems is also shaped by whether users can see why an answer was given (Shin 2021). The CTPA is therefore well suited to judging whether an agricultural chatbot has earned enough confidence to influence real decisions. It can also flag early signs of distrust that simple satisfaction surveys would miss.

AI Chatbots Usage Scale (ACUS)

More recently, researchers have looked for instruments suited to sustained, real-world chatbot use rather than single-session usability tests. (Ibrahim and Nemt-allah 2026) developed and validated the AI Chatbots Usage Scale (ACUS; the article's title calls it the AI chatbots acceptance and perception scale) with Egyptian university students. It began as a 25-item pool, and the final four-factor model retains 19 items. Separate exploratory (n = 374) and confirmatory (n = 599) factor analyses supported four factors, namely Ease of Use, Perceived Usefulness, Trust, and Accessibility. The total-scale Cronbach's alpha was 0.927 (Ibrahim and Nemt-allah 2026, Tables 1 and 3). These factors closely follow ideas from the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT). That link to established theory is a real strength of the scale. The scale came out of an educational technology setting, but its structure carries over readily to agricultural

chatbots, particularly the Accessibility and Trust dimensions, since connectivity and adoption barriers are recognised in digital agriculture (Klerkx *et al.* 2019; Mohr and Kühl 2021).

User Engagement Scale (UES)

Usability alone does not show whether an interaction is rewarding in itself or likely to be repeated. The User Engagement Scale, developed by (O'Brien and Toms 2010) and later shortened by (O'Brien *et al.* 2018) into a four-factor form, measures Aesthetic Appeal, Focused Attention, Perceived Usability, and Reward. Each subscale uses several Likert-type statements (e.g., 'I lost myself in this experience'; 'This system was attractive'). The short form (UES-SF) has acceptable reliability (McDonald's = 0.81–0.86 with online shoppers and 0.70–0.88 with book-search users; (O'Brien *et al.* 2018)) and places much less burden on respondents. Some chatbots are meant to attract farmers' interest and keep it, rather than be used once and dropped. For these, the UES offers a validated way to quantify the affective and motivational side of the interaction that usability metrics overlook. Engagement matters in practice because satisfaction and perceived usefulness have been shown to drive continued use of AI-powered service agents (Ashfaq *et al.* 2020).

Knowledge testing

The instruments above capture what users think, whereas knowledge testing measures what they have actually learned. It is an assessment method rather than a standardised scale. Participants sit a test before and after using the chatbot, and the gain attributable to it is usually compared with paired t-tests or analysis of variance (ANOVA). Reliability can still be checked. Cronbach's alpha is a widely used reliability estimate for knowledge tests as well as questionnaires (Tavakol and Dennick 2011), and for right-or-wrong items its dichotomous form, the Kuder–Richardson formula 20 (KR-20), is commonly used (Kuder and Richardson 1937; Cronbach 1951). In agriculture, this approach shows whether a chatbot genuinely improves farmers' grasp of pest identification, integrated crop management, or climate-risk mitigation. Usability and engagement scores cannot provide that evidence.

Complementary and related instruments

Several related tools beyond the five core instruments appear often in the wider usability literature and deserve a brief mention. The Post-Study System Usability Questionnaire (PSSUQ) is described in (Lewis 1995), whereas the Usability Metric for User Experience (UMUX) was developed by (Finstad 2010). The two-item UMUX-LITE was later introduced by (Lewis *et al.* 2013). These shorter usability measures can be used alongside SUS or in its place. The Technology Acceptance Model (TAM) and UTAUT supply the theoretical scaffolding from which several newer chatbot-specific scales, the ACUS among them, were derived (Davis 1989; Venkatesh *et al.* 2003). Both stress perceived usefulness and ease of use as the closest determinants of adoption intention. None of these tools was developed specifically for agricultural chatbots, but their constructs recur throughout the scales reviewed above and help in interpreting results across studies.

3.2. Cronbach's alpha as a reliability benchmark

Nearly every scale discussed above reports Cronbach's alpha (Cronbach 1951; Streiner 2003) as evidence of internal consistency, the degree to which items within a questionnaire measure the same underlying construct. Formally, for a scale of k items,

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{yi}^2}{\sigma_x^2}\right)$$

where k is the number of items in the questionnaire, σ_{yi}^2 is the variance of each item, and σ_x^2 is the variance of the total score.

Values of α normally fall between 0 and 1. (Tavakol and Dennick 2011) note that reports of acceptable values range from 0.70 to 0.95, and that a very high value (above about 0.90) may suggest redundant items. The reliability values reported for the scales reviewed here (Table 1) generally fall in this range. Researchers evaluating agricultural chatbots should treat α as necessary but not sufficient evidence of scale quality (Sijtsma 2009). High internal consistency does not guarantee that a scale measures the intended construct validly, or that it will behave the same way across farmer populations that differ in language and culture.

3.3. Comparative synthesis of evaluation scales

Table 1 brings together the structure, reliability estimates, and main constructs of the five scales reviewed and the knowledge-testing method, as a single reference for researchers choosing instruments for agricultural chatbot studies.

Scale	Items	Primary construct(s)	Reliability	Source
System Usability Scale (SUS)	10	General usability, learnability	$\alpha \approx 0.91$ (2,324 questionnaires) (Bangor *et al.* 2008)	(Brooke 1996)
Chatbot Usability Scale (BUS)	15	Accessibility of functions, quality of functions, quality of conversation and information, privacy and security, response time	$\alpha = 0.87$ overall (96 volunteers, 385 questionnaires) (Borsci *et al.* 2022)	(Borsci *et al.* 2022)

Scale	Items	Primary construct(s)	Reliability	Source
Checklist for Trust between People and Automation (CTPA)	12	Trust and distrust in automated systems	No αreported for the 12 item scale; α= 0.93 for an earlier 10-item modified scale (Spain *et al.* 2008)	(Jian *et al.* 2000)
AI Chatbots Usage Scale (ACUS)	19	Ease of use, perceived usefulness, trust, accessibility	α= 0.927 total (Egyptian university students, n = 599) (Ibrahim and Nemt-allah 2026)	(Ibrahim and Nemt-allah 2026)
User Engagement Scale, Short Form (UES-SF)	12	Aesthetic appeal, focused attention, perceived usability, reward	ω = 0.70–0.88 by subscale (online shoppers; book-search users) (O'Brien *et al.* 2018)	(O'Brien *et al.* 2018)
Knowledge testing (assessment method)	Variable	Objective knowledge gain (pre–post)	α(or KR-20), computed for each test (Tavakol and Dennick 2011)	(Tavakol and Dennick 2011)

Table 1: Comparative summary of standardised scales for AI chatbot evaluation

The table shows a clear division of labour among the instruments. SUS and BUS address usability at the level of the interface and the interaction. CTPA addresses the cognitive and affective judgement of reliability that comes before behavioural reliance. ACUS combines usability with predictors of sustained adoption, UES-SF captures the motivational and hedonic quality of the interaction, and knowledge testing is unique in measuring downstream educational impact rather than perception. No single scale is enough. A chatbot that is pleasant to use but not trusted may go unused, while one that is trusted but dull may be respected yet rarely consulted of the farmer's own accord. Wherever resources allow, a battery of several instruments is therefore advisable.

3.4. Case studies from the literature

Evaluating large language models on soil science questions

(Khanifar 2025) evaluated five generative AI chatbots (GPT-4o, GPT-4o Mini, Claude 3.5 Sonnet, Gemini 1.5 Pro, and Gemini 1.5 Flash) on 105 PhD entrance examination questions in soil science. Questions were entered in Persian, and GPT-4o was also tested on English translations to check for a language effect. Responses were scored against the official answer key. GPT-4o and Claude 3.5 Sonnet performed best, answering about 65% of questions correctly, while Gemini 1.5 Flash and GPT-4o Mini scored lowest. GPT-4o's Persian and English results did not differ significantly. Both leading models did better on the fundamental-level first section, and the author reports that none of the chatbots handled questions needing multifaceted analysis or practical experience. This suggests caution about substituting chatbot output for expert agronomic advice, although one examination-based study cannot settle the question.

Real-world agricultural deployments and the limits of fluency

Deployed systems beyond this case study show both the promise and the risks of agricultural chatbots at scale. The Farmer.Chat developers report that RAG-based generative AI can meaningfully extend agricultural service delivery, citing quantitative usage metrics and qualitative farmer feedback gathered across four countries (Singh *et al.* 2024). In the AgroLLM study, the authors report that a GPT-4o Mini configuration paired with RAG reached 93% accuracy on agricultural knowledge queries, ahead of baseline models without retrieval (Samuel *et al.* 2025). By contrast, (Tzachor *et al.* 2023) examined ChatGPT's weed-control advice for Nigerian cassava farmers through an appraisal of the generated recommendations and found that, although the responses were technically coherent, they had limited practical use. It provided no specific information on the growth stage and recommended inputs beyond what smallholders could afford. This 'plausibility trap' illustrates how fluent language can conceal agronomic inadequacy. This supports the central argument of this review: standardised, multi-dimensional evaluation with instruments such as those in Table 1 is a practical safeguard before chatbots are trusted with consequential agricultural decisions.

Study	System and domain	Evaluation approach	Main finding (as reported)
(Khanifar 2025)	Five LLM chatbots; PhD-entrance soil science questions (Persian; English for GPT-4o only)	Accuracy against the official answer key	GPT-4o and Claude 3.5 Sonnet best (about 65% correct); weaker on the advanced sections

Study	System and domain	Evaluation approach	Main finding (as reported)
(Tzachor *et al.* 2023)	ChatGPT; weed-control advice for cassava farmers, Nigeria	Appraisal of the content of generated advice	Fluent but generic advice; no growth-stage specificity; inputs beyond smallholder means ('plausibility trap')
(Singh *et al.* 2024)	Farmer.Chat (RAG); smallholders in four countries	Usage metrics and qualitative farmer feedback (developer-reported)	More than 15,000 farmers reached and over 300,000 queries handled
(Samuel *et al.* 2025)	AgroLLM (RAG with GPT-4o Mini)	Accuracy on agricultural knowledge queries (developer-reported)	93% accuracy, ahead of unaugmented baseline models
(Silva *et al.* 2023)	GPT-3.5 and GPT-4; Brazilian agricultural exam questions	Accuracy on exam-style questions, compared across prompting strategies	GPT-4 scored higher than GPT-3.5, and the prompting strategy changed the scores
(Balaguer *et al.* 2024)	Retrieval augmentation vs fine-tuning; agricultural question answering	Generated answers compared with reference answers for quality	Each approach did better than the base model

Table 2: Selected evaluations and deployments of AI chatbots and large language models in agriculture, with the evaluation approach used

3.5. Evaluation beyond questionnaires: accuracy, grounding, and risk

Human-like conversational cues can also raise how humanlike and credible a chatbot appears (Go and Sundar 2019). A polished, friendly bot may therefore score well on usability, trust, and engagement scales without being more correct. This is another expression of the fluency–adequacy gap discussed in this review, and it is the main reason subjective scales should be paired with objective accuracy testing rather than used alone.

Agricultural advice has further requirements. It has to be valid for the local agro-ecology and crop stage, affordable for smallholders, and delivered in the languages and channels farmers use. Digital tools are changing who produces agricultural advice and who farmers trust to give it. Judging a chatbot on its answers alone would miss something. Its place within the advisory networks that farmers already rely on matters too (Rose and Chilvers 2018; Fielke *et al.* 2020; Tzachor *et al.* 2022).

4. Discussion

4.1. Critical issues and methodological considerations

Several methodological caveats limit the confidence we can place in existing scale-based evaluations. First, a high Cronbach's alpha confirms internal consistency but not construct or predictive validity. A scale can be highly reliable and still miss the dimension it claims to measure, yet few agricultural chatbot studies report confirmatory factor analysis or criterion validity alongside reliability coefficients. Second, the original CTPA has been shown to be skewed toward positive ratings, which may inflate reported trust scores (Gutzwiller *et al.* 2019). This finding should not be generalised to the other instruments reviewed here. Third, most of the scales reviewed here were validated with urban, higher-education, or industrial-technology samples. Their psychometric properties have rarely been re-validated among smallholder farmers who differ in literacy, language, and prior digital exposure, which raises fair questions about measurement invariance across such groups. Fourth, sample sizes in applied agricultural case studies are often small, and too small for stable factor-analytic conclusions, even when the point estimate of reliability looks strong. Finally, subjective measures (usability, trust, engagement) and objective ones (knowledge tests) coexist, and it remains unresolved how they should be weighted or combined into one judgement of chatbot 'quality'. That is an open problem for future psychometric work in this area.

4.2. A staged protocol for evaluating agricultural chatbots

Bringing the scales and the accuracy-based methods together, we propose five evaluation layers that can be applied in sequence, informed by the evaluation domains discussed by (Ding *et al.* (2024)). The five-layer sequence is the framework proposed in this review. The order matters. Technical accuracy and safety are checked first, because it is not worth asking farmers to rate a tool that gives wrong advice. Usability, trust, engagement, and learning outcomes are then measured using accuracy testing combined with the instruments in Table 1 in a pilot with representative users, and the knowledge test links these perceptions to what farmers actually learn or do.

Layer	Core question	Candidate measures	Data source
1. Technical accuracy and grounding	Is the advice correct, specific, safe, and affordable?	Expert-scored accuracy against reference answers; groundedness and hallucination checks; stage- and location-specific rubric	Question bank; expert panel

Layer	Core question	Candidate measures	Data source
2. Usability	Can farmers use it easily?	SUS (Brooke 1996); chatbot-specific usability scale (BUS) (Borsci *et al.* 2022); task success	Representative users
3. Trust and reliance	Do users trust it, and rely on it appropriately?	Checklist for Trust between People and Automation (CTPA) (Jian *et al.* 2000); observed reliance on advice	Representative users
4. Engagement and acceptance	Will they keep using it?	UES-SF (O'Brien *et al.* 2018); ACUS (TAM/UTAUT constructs) (Ibrahim and Nemt-allah 2026); usage logs	Users; platform logs
5. Learning and outcomes	Did knowledge or practice improve?	Pre- and post-test knowledge test scores; follow-up on practice adoption	Pre/post cohort

Table 3: Proposed five-layer protocol for evaluating AI chatbots in agriculture

Whichever layers are used, reliability should be reported for the study's own sample and language, not assumed from the original validation, and sample sizes should be large enough for the factor-analytic checks that justify using a scale in a new population. Instruments translated for farmers should be piloted for comprehension, and positive-response bias should be considered when interpreting scores.

4.3. Future perspectives

Chatbot evaluation is likely to move toward advanced analytics, real-time feedback loops, and adaptive learning algorithms that keep recalibrating advisory content from aggregated farmer interactions. Linking chatbot platforms with Internet of Things (IoT) sensors, drone-based imaging, and predictive crop models could make generated advice considerably more relevant to local context in precision agriculture (Kamilaris and Prenafeta-Boldú 2018). Building standardised usability, trust, and engagement instruments into deployment pipelines, rather than running them once as a post-hoc study, would also allow quality to be monitored continuously at scale. Culturally adaptive, multilingual, and voice-first features will stay essential for widening access among rural and underserved groups, particularly women and low-literacy users who have historically been left out of intermediary-dependent extension services. Future work on scale development should prioritise validation directly with smallholder farmers, de-

velop hybrid instruments that capture usability, trust, and factual accuracy in one validated framework, and report both reliability and validity evidence transparently so that studies can be compared.

4.4. Limitations of this review

This review is a narrative synthesis and not a systematic review with a registered protocol, so some relevant studies may have been missed. Only English-language sources were considered, and several agricultural chatbot evaluations are available only as preprints or conference papers that have not been peer-reviewed. Reported figures for deployed systems (for example, usage and accuracy) come from the developers' own reports and were not independently verified. The proposed protocol is conceptual and has not yet been tested against a single chatbot in the field.

5. Conclusion

AI chatbots are a genuine advance for agricultural information dissemination and decision support, but their value is realised only when usability, trustworthiness, engagement, and educational impact are measured rigorously and transparently. This review has synthesised five standardised instruments (the System Usability Scale, Chatbot Usability Scale, Checklist for Trust between People and Automation, AI Chatbots Usage Scale, and User Engagement Scale) and the knowledge-testing method together with the statistical reasoning of Cronbach's alpha behind their reliability claims. The case study from the literature on soil science question answering, together with real deployments such as Farmer.Chat and AgroLLM, show that standardised evaluation is both feasible and necessary. We propose a five-layer protocol that places accuracy checks before perception measures and ends with learning outcomes. As generative AI chatbots become part of extension and advisory systems, researchers, developers, and policymakers must make sure that adoption decisions rest on validated, context-appropriate measurement and not on the persuasive fluency of chatbot output alone.

Declaration of any AI tool

Artificial intelligence (AI) tools were used solely for grammatical correction and language improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

References

Adamopoulou E, Moussiades L (2020). "Chatbots: History, Technology, and Applications." *Machine Learning with Applications*, **2**, 100006. doi:10.1016/j.mlwa.2020.100006.

Anusha K, Lakshmi KN, Pavan VV, Ganesh G, Lakshmi SV (2025). "Smart Chatbot for Farmers' Agricultural Guidance and Assistance." *International Journal of Communication Networks and Information Security*, **17**(3), 182–186.

Ashfaq M, Yun J, Yu S, Loureiro SMC (2020). "I, Chatbot: Modelling the Determinants of Users' Satisfaction and Continuance Intention of AI-Powered Service Agents." *Telematics and Informatics*, **54**, 101473. doi:10.1016/j.tele.2020.101473.

Balaguer A, Benara V, de Freitas Cunha RL, *et al.* (2024). "RAG Vs Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture." *arXiv preprint arXiv:2401.08406*. URL https://arxiv.org/abs/2401.08406.

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.

Bangor A, Kortum PT, Miller JT (2008). "An Empirical Evaluation of the System Usability Scale." *International Journal of Human-Computer Interaction*, **24**(6), 574–594. doi:10.1080/10447310802205776.

Bharti P, Cabrera E, Miro B, Sharma S, Florey C, Sahoo MR, *et al.* (2024). "Chatbot for Promoting Best Crop Management Practices to Rice Farmers in Odisha, India." *Asian Journal of Agricultural Extension, Economics & Sociology*, **42**(10), 257–264. doi:10.9734/ajaees/2024/v42i102581.

Bhuvaneswari C, Pokhariya HS, Yarde P, Vekariya V, Patil H, Natrayan L (2024). "Implementing AI-Powered Chatbots in Agriculture for Optimisation and Efficiency." In *Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2024)*. IEEE, Bengaluru, India. doi:10.1109/IDCIoT59759.2024.10467644.

Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, *et al.* (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**(1), 95–119. doi:10.1007/s00779-021-01582-9.

Brooke J (1996). "SUS: A Quick and Dirty Usability Scale." In PW Jordan, B Thomas, BA Weerdmeester, IL McClelland (eds.), *Usability Evaluation in Industry*, pp. 189–194. Taylor & Francis, London.

Cronbach LJ (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, **16**(3), 297–334. doi:10.1007/BF02310555.

Davis FD (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly*, **13**(3), 319–340. doi:10.2307/249008.

Deriu J, Rodrigo A, Otegi A, Echegoyen G, Rosset S, Agirre E, *et al.* (2021). "Survey on Evaluation Methods for Dialogue Systems." *Artificial Intelligence Review*, **54**, 755–810. doi:10.1007/s10462-020-09866-x.

Ding H, Simmich J, Vaezipour A, Andrews N, Russell T (2024). "Evaluation Framework for Conversational Agents with Artificial Intelligence in Health Interventions: A Systematic Scoping Review." *Journal of the American Medical Informatics Association*, **31**(3), 746–761. doi:10.1093/jamia/ocad222.

Fallon CK, Bustamante EA, Ely KM, Bliss JP (2005). "Improving User Trust with a Likelihood Alarm Display." In *Proceedings of the 11th International Conference on Human-Computer Interaction*. Las Vegas, NV.

Fielke S, Taylor B, Jakku E (2020). "Digitalisation of Agricultural Knowledge and Advice Networks: A State-of-the-Art Review." *Agricultural Systems*, **180**, 102763. doi:10.1016/j.agsy.2020.102763.

Finstad K (2010). "The Usability Metric for User Experience." *Interacting with Computers*, **22**(5), 323–327. doi:10.1016/j.intcom.2010.04.004.

Følstad A, Brandtzaeg PB (2020). "Users' Experiences with Chatbots: Findings from a Questionnaire Study." *Quality and User Experience*, **5**, 3. doi:10.1007/s41233-020-00033-2.

Go E, Sundar SS (2019). "Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions." *Computers in Human Behavior*, **97**, 304–316. doi:10.1016/j.chb.2019.01.020.

Gutzwiller RS, Chiou EK, Craig SD, Lewis CM, Lematta GJ, Hsiung CP (2019). "Positive Bias in the "Trust in Automated Systems Survey"? An Examination of the Jian et al. (2000) Scale." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pp. 217–221. doi:10.1177/1071181319631201.

Hoff KA, Bashir M (2015). "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors*, **57**(3), 407–434. doi:10.1177/0018720814547570.

Hornbæk K (2006). "Current Practice in Measuring Usability: Challenges to Usability Studies and Research." *International Journal of Human-Computer Studies*, **64**(2), 79–102. doi:10.1016/j.ijhcs.2005.06.002.

Ibrahim AR, Nemt-allah MA (2026). "Development and Psychometric Validation of the AI Chatbots Acceptance and Perception Scale for Higher Education Students." *Scientific Reports*, **16**, 22941. doi:10.1038/s41598-026-62798-4.

ISO (2018). "ISO 9241-11:2018 Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts." International Organization for Standardization. Accessed: 7 August 2026, URL https://www.iso.org/standard/63500.html.

Jain M, Kumar P, Bhansali I, Liao QV, Truong K, Patel S (2018). "FarmChat: A Conversational Agent to Answer Farmer Queries." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **2**(4), 170:1–170:22. doi:10.1145/3287048.

Jian JY, Bisantz AM, Drury CG (2000). "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics*, **4**(1), 53–71. doi:10.1207/S15327566IJCE0401_04.

Kamilaris A, Prenafeta-Boldú FX (2018). "Deep Learning in Agriculture: A Survey." *Computers and Electronics in Agriculture*, **147**, 70–90. doi:10.1016/j.compag.2018.02.016.

Khanifar J (2025). "Evaluating AI-Generated Responses from Different Chatbots to Soil Science-Related Questions." *Soil Advances*, **3**, 100034. doi:10.1016/j.soilad.2025.100034.

Klerkx L, Jakku E, Labarthe P (2019). "A Review of Social Science on Digital Agriculture, Smart Farming and Agriculture 4.0: New Contributions and a Future Research Agenda." *NJAS – Wageningen Journal of Life Sciences*, **90–91**, 100315. doi:10.1016/j.njas.2019.100315.

Kuder GF, Richardson MW (1937). "The Theory of the Estimation of Test Reliability." *Psychometrika*, **2**(3), 151–160. doi:10.1007/BF02288391.

Lee JD, See KA (2004). "Trust in Automation: Designing for Appropriate Reliance." *Human Factors*, **46**(1), 50–80. doi:10.1518/hfes.46.1.50_30392.

Lewis JR (1995). "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use." *International Journal of Human-Computer Interaction*, **7**(1), 57–78. doi:10.1080/10447319509526110.

Lewis JR, Sauro J (2009). "The Factor Structure of the System Usability Scale." In M Kurosu (ed.), *Human Centered Design, HCD 2009*, volume 5619 of *Lecture Notes in Computer Science*, pp. 94–103. Springer, Berlin. doi:10.1007/978-3-642-02806-9_12.

Lewis JR, Utesch BS, Maher DE (2013). "UMUX-LITE: When There's No Time for the SUS." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2099–2102. doi:10.1145/2470654.2481287.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, *et al.* (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 9459–9474. Curran Associates, Red Hook, NY. URL https://arxiv.org/abs/2005.11401.

Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D (2018). "Machine Learning in Agriculture: A Review." *Sensors*, **18**(8), 2674. doi:10.3390/s18082674.

Mohr S, Kühl R (2021). "Acceptance of Artificial Intelligence in German Agriculture: An Application of the Technology Acceptance Model and the Theory of Planned Behaviour." *Precision Agriculture*, **22**, 1816–1844. doi:10.1007/s11119-021-09814-x.

O'Brien HL, Cairns P, Hall M (2018). "A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form." *International Journal of Human-Computer Studies*, **112**, 28–39. doi:10.1016/j.ijhcs.2018.01.004.

O'Brien HL, Toms EG (2010). "The Development and Evaluation of a Survey to Measure User Engagement." *Journal of the American Society for Information Science and Technology*, **61**(1), 50–69. doi:10.1002/asi.21229.

Okonkwo CW, Ade-Ibijola A (2021). "Chatbot Applications in Education: A Systematic Review." *Computers and Education: Artificial Intelligence*, **2**, 100033. doi:10.1016/j.caeai.2021.100033.

Radziwill NM, Benton MC (2017). "Evaluating Quality of Chatbots and Intelligent Conversational Agents." 1704.04579, URL https://arxiv.org/abs/1704.04579.

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. doi:10.1016/j.ijhcs.2021.102630.

Rose DC, Chilvers J (2018). "Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming." *Frontiers in Sustainable Food Systems*, **2**, 87. doi:10.3389/fsufs.2018.00087.

Samuel DJ, Skarga-Bandurova I, Sikolia D, Awais M (2025). "AgroLLM: Connecting Farmers and Agricultural Practices Through Large Language Models for Enhanced Knowledge Transfer and Practical Application." *arXiv preprint arXiv:2503.04788*. URL https://arxiv.org/abs/2503.04788.

Shin D (2021). "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI." *International Journal of Human-Computer Studies*, **146**, 102551. doi:10.1016/j.ijhcs.2020.102551.

Sijtsma K (2009). "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha." *Psychometrika*, **74**(1), 107–120. doi:10.1007/s11336-008-9101-0.

Silva B, Nunes L, Estevão R, Aski V, Chandra R (2023). "GPT-4 as an Agronomist Assistant? Answering Agriculture Exams Using Large Language Models." *arXiv preprint arXiv:2310.06225*. URL https://arxiv.org/abs/2310.06225.

Singh N, Wang'ombe J, Okanga N, *et al.* (2024). "Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers." *arXiv preprint arXiv:2409.08916*. URL https://arxiv.org/abs/2409.08916.

Spain RD, Bustamante EA, Bliss JP (2008). "Towards an Empirically Developed Scale for System Trust: Take Two." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pp. 1335–1339. doi:10.1177/154193120805201907.

Streiner DL (2003). "Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency." *Journal of Personality Assessment*, **80**(1), 99–103. doi:10.1207/S15327752JPA8001_18.

Tavakol M, Dennick R (2011). "Making Sense of Cronbach's Alpha." *International Journal of Medical Education*, **2**, 53–55. doi:10.5116/ijme.4dfb.8dfd.

Tzachor A, Devare M, King B, Avin S, Ó hÉigeartaigh S (2022). "Responsible Artificial Intelligence in Agriculture Requires Systemic Understanding of Risks and Externalities." *Nature Machine Intelligence*, **4**, 104–109. doi:10.1038/s42256-022-00440-4.

Tzachor A, Devare M, Richards C, *et al.* (2023). "Author Correction: Large Language Models and Agricultural Extension Services." *Nature Food*, **4**, 1112. doi:10.1038/s43016-023-00904-9.

Venkatesh V, Morris MG, Davis GB, Davis FD (2003). "User Acceptance of Information Technology: Toward a Unified View." *MIS Quarterly*, **27**(3), 425–478. doi:10.2307/30036540.

Wolfert S, Ge L, Verdouw C, Bogaardt MJ (2017). "Big Data in Smart Farming – A Review." *Agricultural Systems*, **153**, 69–80. doi:10.1016/j.agsy.2017.01.023.

! Publication & Reviewer Details

Publication Information

- **Submitted:** *11 October 2026*
- **Accepted:** *09 November 2026*
- **Published (Online):** *16 November 2026*

Reviewer Information

- **Reviewer 1:**
  Anonymous
- **Reviewer 2:**
  Anonymous

Disclaimer/Publisher's Note

The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2025): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits non-commercial use, sharing, and reproduction in any medium, provided the original work is properly cited and no modifications or adaptations are made.

Affiliation:

Farhana L S*
Department of Agricultural Statistics
College of Agriculture, Vellayani
Thiruvananthapuram India
E-mail: farhanals786@gmail.com

Journal of Sustainable Technology in Agriculture https://www.jostapubs.com/
PAPAYA Academic Press, Statoberry LLP, https://www.statoberry.com/papaya
November 2026, Volume 2, Issue IV *Submitted:* 2026-10-11
doi:10.65287/josta.202610.de78 *Accepted:* 2026-11-09