

ARTIFICIAL INTELLIGENCE IN AGRICULTURAL EXTENSION COMMUNICATION: A SYSTEMATIC LITERATURE REVIEW OF 278 STUDIES

Vinay Sonule¹, Dr Moiz Mannan Haque², Chetana Ramesh Bansod³

¹Department of Mass Communication, Rashtrasant Tukdoji Maharaj Nagpur University, vinayresearch0@gmail.com, 0009-0008-1745-4443

²Department of Mass Communication, Rashtrasant Tukdoji Maharaj Nagpur University, moiznews@gmail.com, 0009-0002-5931-1623

³Department of Education, Symbiosis International University, Pune, chetanabansod591@gmail.com, 0009-0006-8756-0546

ABSTRACT

Artificial intelligence (AI) technologies are increasingly deployed to modernise agricultural extension communication, yet the evidence base for their effectiveness remains fragmented and methodologically immature. This systematic review, conducted in accordance with PRISMA 2020 guidelines, synthesises evidence from 278 studies retrieved across five databases, covering the period 2015–2026, to characterise the landscape of AI-enabled extension tools, quantify evidence certainty for key farmer outcomes, and identify critical methodological gaps. Of 1,156 retrieved records, 278 met inclusion criteria following deduplication and two-stage relevance screening; a subset of 30 full-text accessible studies received detailed quality appraisal using the Mixed Methods Appraisal Tool (MMAT) 2018. System development studies predominated (39.6%), with India (24.8%) and Sub-Saharan Africa (18.0%) as the primary geographic contexts; chatbots (32.7%) and mobile/SMS platforms (29.5%) were the most common AI communication modalities. Only one randomised controlled trial (0.4%) was identified, and all 30 MMAT-appraised studies were rated ‘Not assessable’, reflecting the dominance of proof-of-concept designs over evaluative research. Applying a GRADE-adapted narrative grading approach, evidence certainty for knowledge gain and technology adoption outcomes was graded ‘Very Low’, while productivity and economic outcome evidence was insufficient for grading. These findings indicate that, despite rapid technological growth, the field lacks the experimental rigour necessary to inform large-scale policy commitments; future research must prioritise randomised evaluations, longitudinal cohort studies, equity- disaggregated outcome reporting, and offline-capable inclusive system designs to translate AI innovation into verified agricultural development impact.

KEYWORDS: artificial intelligence; agricultural extension; chatbot; systematic literature review; smallholder farmers; PRISMA 2020; MMAT 2018; mobile advisory; knowledge gain; technology adoption

1. INTRODUCTION

Agricultural extension services have historically served as the primary conduit for disseminating agronomic knowledge, improved seeds, and sustainable farming practices to rural smallholder communities worldwide. Despite their acknowledged importance, conventional extension models face persistent structural constraints: low extension-worker-to-farmer ratios, high per-contact costs, geographic remoteness, and an inability to deliver timely, personalised advice at scale (Sahoo and Jena, 2025). These limitations are particularly acute in low- and middle-income countries (LMICs), where the majority of the world's 500 million smallholder farm households operate with limited access to formal advisory services (Korir et al., 2023).

The rapid diffusion of mobile telephony and internet connectivity across rural areas of South Asia, Sub-Saharan Africa, and Southeast Asia has catalysed interest in digital agricultural extension platforms. Within this broader digital revolution, artificial intelligence (AI) technologies — encompassing natural language processing (NLP), machine learning (ML), large language models (LLMs), retrieval-augmented generation (RAG), and computer vision

— hold particular promise. AI-enabled communication tools can theoretically operate continuously, serve millions of users simultaneously, respond in local languages, and personalise advice based on crop, location, and seasonal context (Singh et al., 2024; Collis et al., 2025). Early deployments, ranging from SMS-based advisory systems to multimodal voice chatbots, have demonstrated technical feasibility and initial user acceptance (Bhatia and Kaur, 2024; Mehra and Anitha, 2025).

Yet the proliferation of AI extension tools has outpaced the systematic evaluation of their effectiveness. Fragmented, single-system evaluations predominate in the literature, and the few existing narrative reviews focus on technical architectures or narrow geographic contexts. No prior systematic review has simultaneously mapped the global evidence base, quantified evidence certainty for key farmer outcomes, and applied a rigorous quality appraisal framework such as MMAT 2018 (Hong et al., 2018). This gap is consequential: policymakers, funders, and technology developers lack reliable evidence on which AI communication modalities are most effective, for which populations, and under what conditions.

The present systematic review addresses this gap with three objectives: (1) to characterise the global landscape of AI applications in agricultural extension communication through structured PICOS extraction across 278 included studies; (2) to synthesise evidence on the effectiveness of AI communication tools for farmer knowledge gain and technology adoption, with GRADE- adapted evidence certainty grading; and (3) to appraise methodological quality across 30 full- text studies using MMAT 2018 and identify research priorities for advancing the field. By providing the most comprehensive evidence synthesis to date, this review equips research communities, national agricultural extension systems, and international development organisations with a rigorous foundation for informed investment and future research design.

2.METHODS

2.1Protocol and Registration

The review protocol was designed in accordance with PRISMA 2020 reporting guidelines (Page et al., 2021). The review question was formulated using the PICOS framework: Population (smallholder and subsistence farmers in any country); Intervention (AI-based communication tools including chatbots, voice assistants, mobile applications, LLMs, and SMS/IVR systems); Comparison (traditional extension, alternative digital tools, or no intervention); Outcome (knowledge gain, technology adoption, productivity, economic outcomes, user satisfaction, and system performance); and Study design (any empirical or system-development study published in a peer-reviewed venue or as a credible preprint).

2.2Search Strategy and Database Coverage

Systematic searches were conducted across five sources: SciSpace Deep Search, SciSpace Full Text Index, SciSpace Library, Google Scholar, and PubMed, covering publications from January 2015 to June 2026. The review period was selected to capture the contemporary era of deep-learning-based NLP while excluding very early, pre-smartphone-era extension literature that is not comparable to current AI deployments. Search terms combined subject-domain concepts (agricultural extension, farmer advisory, agri-advisory, rural knowledge dissemination) with technology descriptors (artificial intelligence, chatbot, large language model, voice assistant, NLP, mobile application, SMS advisory) and outcome keywords (knowledge gain, technology adoption, farmer behaviour).

2.3Eligibility Criteria

Inclusion criteria comprised: peer-reviewed journal articles and conference proceedings; preprints from recognised repositories (arXiv, Zenodo); studies examining AI-based communication tools in any agricultural extension context; any geographic setting; English- language publications.

Exclusion criteria comprised: purely agronomic studies without an AI communication component; hardware IoT studies without advisory or communication dimensions; non-English publications; records below the relevance scoring thresholds applied at each screening stage.

2.4Screening Process

Screening proceeded in two stages. In Stage 1 (abstract screening), records were evaluated against a relevance threshold of $\geq 4.0/5.0$ on a validated five-point scale. Records meeting this threshold advanced to Stage 2 (full-text screening), where a higher threshold of $\geq 4.5/5.0$ was applied to confirm relevance and satisfy the more stringent full-text inclusion criteria.

2.5Data Extraction

Data extraction was conducted at two levels. For all 278 included studies, structured extraction was performed using original SLR columns covering study methodology, AI technology types, farmer engagement factors, and geographic context. A subset of 30 studies with accessible full- text PDFs received detailed PICOS extraction. Where study design information was insufficient for complete PICOS coding, the 'Not specified' category was applied.

2.6Quality Appraisal

Methodological quality was assessed using the Mixed Methods Appraisal Tool (MMAT) 2018 (Hong et al., 2018) for the 30 full-text accessible studies. MMAT 2018 comprises two universal screening questions (S1: clear research question; S2: data adequately address the research question) and five design-specific quality criteria covering qualitative, quantitative randomised, quantitative non-randomised, quantitative descriptive, and mixed-methods designs. For studies classified as system development or simulation — which lack a dedicated MMAT category — adapted criteria were applied: clear objectives; appropriate technical methods; specified evaluation methods; reported performance metrics; discussion of limitations and generalisability.

2.7Evidence Certainty Grading

Evidence certainty for key outcome domains (knowledge gain, technology adoption, productivity/yield, economic outcomes, user satisfaction, and system performance) was assessed using a narrative approach adapted from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework. Certainty was

appraised across five domains: study design limitations, risk of bias, inconsistency, indirectness, and imprecision. Certainty levels are classified as High, Moderate, Low, or Very Low, with evidence rated Insufficient when outcome data were too sparse for grading.

3.RESULTS

3.1PRISMA Flow and Study Selection

Database searches retrieved 1,156 records in total: SciSpace Deep Search (n = 746), SciSpace Full Text Index (n = 200), SciSpace Library (n = 2), Google Scholar (n = 197), and PubMed (n = 11). After automated deduplication, 808 unique records were retained (–348 duplicates).

Abstract screening at the $\geq 4.0/5.0$ threshold yielded 281 records advancing to full-text review. Full-text screening at $\geq 4.5/5.0$ excluded three records, resulting in 278 included studies. Of these, 30 had accessible full-text PDFs that received detailed PICOS extraction and MMAT quality appraisal; the remaining 248 studies were characterised through original SLR column extraction. The high retention rate from abstract to full-text screening (281 to 278; 99.0%) reflects stringent abstract-stage filtering and the focused inclusion criteria.

3.2Study Characteristics and Temporal Distribution

The 278 included studies exhibit a pronounced temporal trend: 31.7% of all included records were published in 2025 alone, reflecting the rapid acceleration of AI agricultural extension research coinciding with widespread LLM availability and renewed development-sector interest in digital extension. Journal articles constituted the majority of the evidence base (56.1%), with conference proceedings, preprints, and technical reports comprising the remainder.

3.3Geographic Distribution

As shown in Table 1, geographic coverage was heavily concentrated in South Asia and Sub-Saharan Africa. India dominated with 24.8% of studies (n = 69), consistent with substantial government investment in digital agriculture under programmes such as the Digital Agriculture Mission. Sub-Saharan Africa accounted for 18.0% (n = 50), with Kenya (n = 8; 2.9%), Nigeria (n = 7; 2.5%), and Tanzania (n = 5; 1.8%) as the most frequently represented countries. China contributed 4.0% (n = 11) of the evidence base. A large proportion of studies (46.0%) fell into the 'Other/Mixed/Not Specified' category, encompassing multi-country investigations, global technical reviews, and system-development papers without a defined field deployment geography.

Table 1. Geographic distribution of included studies (n = 278).

Geography	Count	Percentage
India	69	24.8%
Sub-Saharan Africa	50	18.0%
China	11	4.0%
Kenya	8	2.9%
Nigeria	7	2.5%
Tanzania	5	1.8%
Other / Mixed / Not Specified	128	46.0%

Note. Percentages do not sum to 100 due to rounding.

3.4Study Design and Methodology

Table 2 presents the distribution of study designs. System development and prototype studies comprised the single largest category (39.6%), reflecting the developmental stage of AI agricultural extension as a research field. Secondary analyses and reviews (17.3%) indicate growing efforts to synthesise emerging findings. Survey and descriptive studies (13.3%) and mixed-methods designs (8.6%) provided user-centred and contextual data. Critically, only one completed randomised controlled trial was identified (0.4%), and experimental or quasi-experimental designs accounted for just 4.0% of the corpus. This near- total absence of causal inference capacity is the single most important methodological characteristic of the field and severely constrains the certainty of all effectiveness conclusions drawn from this review.

Table 2. Study design distribution (n = 278).

Study Design	Count	Percentage
System Development / Prototype	110	39.6%
Review / Secondary Analysis	48	17.3%
Survey / Descriptive	37	13.3%
Mixed Methods	24	8.6%
Case Study	18	6.5%

Qualitative	15	5.4%
Experimental / Quasi-Experimental	11	4.0%
Observational	8	2.9%
Cross-Sectional	6	2.2%
Randomised Controlled Trial	1	0.4%

3.5AI Technology Types

Table 3 summarises the AI technology categories identified across included studies (multiple technologies per study permitted). Chatbot systems (32.7%) and mobile/SMS platforms (29.5%) predominate, consistent with the widespread mobile penetration in target geographies and the relative technical accessibility of rule-based and retrieval-based conversational agents. Classical machine learning (23.7%) and NLP (18.7%) underpin the majority of system implementations. The emergence of LLM/RAG architectures (16.5%) reflects post-2022 adoption of generative AI in agricultural contexts, though these remain a minority of the evidence base. Voice/IVR systems (14.7%) represent an inclusive modality relevant to low-literacy and feature-phone users (Rajanala et al., 2025; Mehra and Anitha, 2025). Computer vision (13.7%) is primarily applied to crop disease diagnosis and yield estimation.

Table 3. AI technology types (n = 278; multiple per study permitted).

AI Technology	Count	Percentage
Chatbot / Conversational Agent	91	32.7%
Mobile Application / SMS	82	29.5%
Classical Machine Learning	66	23.7%
Natural Language Processing (NLP)	52	18.7%
Large Language Model / RAG	46	16.5%
Voice / IVR System	41	14.7%
Computer Vision	38	13.7%
Deep Learning	35	12.6%
Recommendation System	28	10.1%
Expert System	22	7.9%

Note. Percentages sum to >100% due to multiple technologies per study.

3.6Communication Channels

Mobile applications dominated delivery modalities (35.3%; Table 4), capitalising on smartphone proliferation in rural South Asia and parts of Sub-Saharan Africa. Web-based platforms (16.2%) and SMS/text messaging (13.7%) offered lower-device-specification alternatives, while voice and IVR systems (11.5%) addressed low-literacy and non-smartphone users. The limited use of WhatsApp and other messaging apps (6.5%) is noteworthy given their widespread adoption in many target regions and the potential for low-friction deployment through existing social networks.

Table 4. Primary communication channels (n = 278).

Communication Channel	Count	Percentage
Mobile Application	98	35.3%
Web-Based Platform	45	16.2%
SMS / Text Messaging	38	13.7%
Voice Call / IVR	32	11.5%
Multiple Channels	27	9.7%
WhatsApp / Messaging Apps	18	6.5%
Not Specified	20	7.2%

3.7Reported Outcomes

System performance metrics — including accuracy, intent recognition, and response latency — were the most frequently reported outcomes (40.3%; Table 5), consistent with the dominance of system development study designs. User satisfaction and acceptance (28.1%) were commonly assessed via self-administered surveys. Knowledge gain (16.2%) and technology adoption or behaviour change (13.7%) represent substantively important outcomes but are less frequently evaluated. Productivity/yield impacts were reported by only 7.9% of studies, and economic outcomes by 6.5%. Among the 30 studies with detailed full-text extraction, sample sizes were predominantly small: 60.0% enrolled fewer than 100 participants, 26.7% enrolled 100–500, and only 13.3% exceeded 500 participants (median approximately 100).

Table 5. Primary outcome categories (n = 278; multiple per study permitted).

Outcome Category	Count	Percentage
System Performance Metrics	112	40.3%
User Satisfaction / Acceptance	78	28.1%
Knowledge Gain / Learning	45	16.2%
Adoption / Behaviour Change	38	13.7%
Not Specified	35	12.6%
Productivity / Yield Impact	22	7.9%
Economic Outcomes	18	6.5%

Note. Percentages sum to >100% due to multiple outcomes per study.

4.Evidence Synthesis

Evidence synthesis focused on the 13 studies with detailed full-text evidence relevant to the two primary effectiveness questions: (1) Do AI communication tools improve farmer knowledge? and (2) Do AI communication tools promote technology adoption?

4.1Knowledge Gain Evidence

Multiple studies report improvements in farmer decision-making capacity, information access, and problem-resolution efficiency following AI tool use. Among the 50-farmer pilot of the Vayazh conversational AI system, participants reported significant improvements in decision-making and agricultural productivity following regular use of the tool (Shriya et al., 2025). Field testing of the KrishiDisha chatbot with 150 farmers confirmed practical impact through enhanced decision confidence and reduced time-to-information (Bhatia and Kaur, 2024).

Evidence for problem-solving efficiency is provided by Aenugu and Ramaraju (2025), who conducted a pilot study with 83 registered users and 62 documented query interactions, finding that chatbot-assisted queries achieved a resolution success rate increase of over 40% compared with non-chatbot baseline interactions. This metric suggests that AI tools can meaningfully accelerate information access and problem resolution, though the pilot scale and absence of a concurrent control group preclude causal attribution.

Indirect evidence for knowledge gain potential derives from system correctness metrics. The AgriLLM prototype achieved 89.6% accuracy on a battery of 500 standardised farmer queries with substantially fewer incorrect recommendations than baseline models (AgriLLM Research Team, 2025). The Dhenu2 voice-enabled chatbot by Mehra and Anitha (2025) demonstrated 97.53% response accuracy and 91.35% contextual relevance across multilingual evaluation sets. High technical accuracy is a necessary prerequisite for knowledge transmission, but the pathway from system correctness to internalised farmer knowledge remains poorly documented in the literature.

The most direct adoption-oriented knowledge evidence comes from Prabhu and Saravanakumar (2025), who report that 72% of 50 Tamil-speaking farmers adopted one or more recommended agronomic practices following voice chatbot interaction over a four-month period. In a substantially larger deployment, the Ama Krushi chatbot in Odisha, India, engaged 1,200 users over its evaluation period; Sahoo and Jena (2025) report that 52% of users perceived improvement in input-use decisions and crop planning following chatbot use. While these figures are encouraging, both rely on self-reported perceptions of change rather than pre-post knowledge assessments or independent behavioural verification.

4.2Technology Adoption Evidence

Technology adoption evidence is characterised by encouraging but methodologically limited findings. Self-reported adoption figures range from 52% (Sahoo and Jena, 2025) to 72% (Prabhu and Saravanakumar, 2025), and engagement metrics suggest feasibility at scale. Korir et al. (2023) evaluated a chatbot providing expert agronomic advice to Kenyan potato farmers and found 80% expressed high likelihood of future use alongside 60% overall user satisfaction, indicating positive attitudinal precursors to sustained adoption. However, attitudes and intentions are imperfect proxies for actual behaviour change.

At a larger scale, the Farmer.Chat platform reported engagement with over 15,000 farmers and 300,000 queries across multiple countries, with mixed-methods evaluation indicating positive signals for farming practice improvement (Singh et al., 2024). While the scale of engagement demonstrates logistical feasibility of AI extension at reach, the summary reporting of outcomes in the available preprint limits assessment of adoption depth and quality.

Adoption barriers documented across the evidence base illuminate the limitations of headline adoption figures. The AIEP initiative identifies voice lag, low-resource automatic speech recognition limitations in local languages, and labour-intensive agricultural corpus curation as persistent quality constraints that directly affect user experience and willingness to continue use (Collis et al., 2025). Barriers related to digital divides — particularly for women farmers, tribal communities, and users with low digital literacy — are documented in the Ama Krushi evaluation (Sahoo and Jena, 2025) and in the Kenyan context (Korir et al., 2023). These barriers suggest that headline adoption rates, typically measured among early adopters with existing digital connectivity, likely overestimate population-level uptake. The Climate Agronomy Learning Commons initiative in Kenya's Western

Highlands further illustrates how knowledge system design must account for connectivity and literacy constraints to achieve equitable reach (Climate Agronomy Working Group, 2025).

4.3 Evaluation Designs Employed

Evaluation designs across the synthesised studies reflect four dominant approaches: small-scale pilots, user satisfaction surveys, system performance benchmarks, and at-scale observational engagement analyses. Small pilots enrolling between 50 and 150 participants represent the most common design, as evidenced by the Vayazh evaluation with 50 farmers (Shriya et al., 2025), the AgriLLM field trial with approximately 112 participants (AgriLLM Research Team, 2025), and the KrishiDisha validation with 150 farmers (Bhatia and Kaur, 2024). While small pilots enable rapid iteration and support proof-of-concept demonstrations, they are statistically underpowered to detect modest but policy-relevant effect sizes and cannot support generalisable conclusions.

User satisfaction surveys employing self-administered Likert scales and structured interviews dominate the evaluation repertoire. These instruments efficiently capture user perceptions but are vulnerable to social desirability bias and cannot distinguish between transient satisfaction with a novel technology and sustained utility in agricultural decision-making. System performance benchmarks — reporting intent recognition rates, response accuracy, and query resolution times — are technically objective but constitute at best indirect evidence of farmer-level impact: a chatbot that answers 91% of queries correctly (Prabhu and Saravanakumar, 2025) will only translate that accuracy into improved farmer outcomes if queries are well-formed, responses are comprehensible, and recommended practices are subsequently applied. The disconnect between technical performance and field impact represents a pervasive but rarely theorised gap in the literature.

At the larger scale, the Farmer.Chat evaluation (Singh et al., 2024) employed a mixed quantitative-qualitative approach capturing engagement volumes, query content, and farmer narrative accounts of behaviour change. This richer multi-method design is a valuable model

for the field, though the aggregated reporting of key outcomes in the available preprint version prevents detailed quality appraisal. No study in the synthesis evidence base reported a completed randomised controlled trial, and only the AIEP initiative (Collis et al., 2025) described a prospectively designed comparison arm, framing it as a planned — rather than completed — evaluation.

4.4 Evidence Certainty Summary

Table 6 presents the GRADE-adapted evidence certainty assessment for each outcome domain. Knowledge gain and technology adoption are both graded Very Low certainty, reflecting combined downgrading across study design limitations (predominantly small pilots without randomisation), high risk of bias from self-reported outcomes, indirectness of system performance proxies, and imprecision from small sample sizes. Productivity and economic outcomes are graded Insufficient due to the rarity of such measurements: only 7.9% of studies reported any productivity indicator, and no study provided verified yield or income effect estimates. User satisfaction was graded Low, based on consistent but subjective self-report data across multiple studies. System performance (accuracy, intent recognition) was rated Moderate, reflecting objective measurement but the absence of farmer-level outcome linkage.

Table 6. GRADE-adapted evidence certainty by outcome domain.

Outcome	Certainty	Principal Reasons for Downgrading	Priority
Knowledge Gain	Very Low	Small pilots; self-report bias; indirect proxies; imprecision	High
Technology Adoption	Very Low	Observational designs; self-report; short follow-up; no controls	High
Productivity / Yield	Insufficient	<8% of studies measure this outcome; no verified effects	Critical

5. Quality Appraisal

5.1 MMAT 2018 Screening Questions

MMAT 2018 screening was applied to all 30 full-text accessible studies (Hong et al., 2018). Question S1 ('Are there clear research questions or objectives?') was answered Yes for all 30 studies (100%), reflecting the practice of articulating explicit system development or evaluation objectives across the literature. Question S2 ('Do the collected data allow addressing the research questions?') was answered Yes for only 15 studies (50.0%), with the remaining 15 (50.0%) rated Can't Tell. The high Can't Tell rate arises from a fundamental gap between ambitious objectives — typically expressed in terms of farmer behaviour change or knowledge improvement — and the

predominantly technical evidence actually presented, which often comprises accuracy benchmarks and usability test results from small convenience samples.

5.2 Design-Specific Quality Criteria

Among the 30 full-text studies, 26 (86.7%) were classified as system development or simulation studies, one (3.3%) as quantitative descriptive, one (3.3%) as qualitative, and two (6.7%) as mixed methods. No quantitative RCTs or quantitative non-randomised studies were present, confirming the study design distribution observed across the full 278-study corpus. For the 26 system development studies, adapted MMAT criteria were applied: all 26 (100%) met the criterion for clear objectives; 24 (92.3%) demonstrated appropriate technical methods; only 12 (46.2%) specified an evaluation method beyond technical benchmarking; 18 (69.2%) reported performance metrics; and merely 6 (23.1%) discussed limitations and generalisability. The declining criterion-fulfilment across the chain from objective-setting to limitations discussion reflects a widespread 'proof-of-concept' culture that prioritises demonstration over rigorous evaluation.

5.3 Overall Quality Distribution

Following MMAT 2018 protocol, no composite quality score was computed. Overall quality was characterised descriptively. All 30 assessed studies were rated 'Not assessable' for overall quality. This classification reflects the fundamental mismatch between the system development designs employed and the evaluation frameworks required for standard quality appraisal. 'Not assessable' does not indicate worthless research: system development studies serve a critical role in establishing technical feasibility, surfacing implementation constraints, and generating hypotheses for subsequent evaluation. However, they cannot provide robust evidence of effectiveness or scalability, and the uniform 'Not assessable' rating across the entire corpus signals an immature field requiring deliberate investment in rigorous evaluation infrastructure.

6. DISCUSSION

6.1 Principal Findings and Interpretation

This systematic review synthesised 278 studies on AI-enabled agricultural extension communication and found a rapidly growing, technically innovative, but methodologically immature field. The evidence base is characterised by four dominant features: (1) a heavy concentration of system development studies that demonstrate technical feasibility without rigorous impact evaluation; (2) geographic concentration in India and Sub-Saharan Africa, with major gaps in Latin America, Southeast Asia, and the Middle East; (3) outcome measurement skewed toward system performance metrics and user satisfaction surveys, with critical farmer outcomes (productivity, income) rarely measured; and (4) Very Low certainty evidence for the outcomes that matter most for agricultural development — knowledge gain and technology adoption — owing to small samples, self-report bias, and the near-total absence of controlled experimental designs.

These findings are consistent with — and extend — earlier narrative accounts of digital extension limitations (Korir et al., 2023; Collis et al., 2025). The evidence pattern also mirrors the trajectory of mHealth research in the early 2010s, where enthusiasm for technology-enabled health communication preceded rigorous impact evaluation by a decade. Agricultural AI extension faces a similar risk of premature scaling without sufficient evidence of effectiveness.

6.2 The Evaluation Gap and Its Implications

The identification of only one RCT (0.4%) across 278 studies is the single most consequential finding of this review. In the absence of randomised or well-controlled quasi-experimental evidence, causal claims about AI extension effectiveness cannot be substantiated, and cost-effectiveness comparisons with traditional extension cannot be conducted. This gap is not merely academic: national governments and international development organisations are currently making large investments in AI extension platforms on the basis of feasibility pilots and user satisfaction surveys that are structurally incapable of answering the question 'does this work better than what we already do?'

Priority research designs to address this gap include: multi-site RCTs with pre-specified power calculations (adequate sample sizes to detect modest adoption effects require $n > 500$ per arm), pre-registered analysis plans, and minimum follow-up periods of two agricultural seasons to assess sustained adoption; quasi-experimental designs exploiting geographic roll-out variation or discontinuities; and longitudinal cohort studies tracking technology use, dis-adoption, and downstream productivity outcomes over multiple years. The AIEP initiative's proposed triple-arm evaluation — comparing AI-only, hybrid AI-plus-human, and human-only advisory conditions — exemplifies the rigorous design the field now requires (Collis et al., 2025).

6.3 Technology Design Implications

The dominance of mobile applications (35.3%) and chatbot interfaces (32.7%) reflects pragmatic deployment choices aligned with existing farmer technology access patterns. The evidence from voice/IVR systems (Mehra and Anitha, 2025; Prabhu and Saravanakumar, 2025) and from LLM-based implementations (AgriLLM Research Team, 2025; Singh et al., 2024) suggests that these modalities hold particular promise for inclusive delivery. However, key design challenges persist. Language coverage for low-resource agricultural languages remains

limited, with automatic speech recognition and machine translation quality constraining usability for non-dominant language speakers (Collis et al., 2025). Explainability and farmer trust in AI recommendations are rarely addressed in system design or evaluation, despite their importance for adoption among sceptical user groups (Sahoo and Jena, 2025).

The near-absence of offline-capable and hybrid digital-human models is a significant coverage gap. In contexts with intermittent connectivity — common across Sub-Saharan Africa and parts of South Asia — systems that require constant internet access exclude the most marginalised farmers. Future system designs should prioritise edge-computing deployments, WhatsApp- integrated bots capitalising on existing messaging app use, and models that seamlessly route complex queries to human extension agents.

6.4 Equity and Inclusion Considerations

The evidence base reveals systematic inattention to equity dimensions of AI extension. Women farmers, tribal communities, and low-literacy users are rarely the focus of dedicated evaluation, and outcome data are seldom disaggregated by gender, age, land tenure, or connectivity status. Given that women farmers constitute 60–80% of the agricultural workforce in many Sub-Saharan African contexts and face distinctive information access barriers, this omission has major policy significance. Equity-focused evaluations should be embedded as mandatory components of future AI extension research designs and funding conditions.

6.5 Comparison with Existing Reviews

Previous reviews of digital agricultural extension have noted the dominance of feasibility studies, geographic concentration in South Asia, and the limited evidence base for behaviour change. The present review strengthens these observations with a substantially larger and more systematically screened corpus (278 studies vs. fewer than 50 in most prior reviews), formally applies MMAT 2018 quality appraisal, and provides the first GRADE-adapted certainty grading for AI extension outcomes. The finding that all 30 full-text quality-appraised studies receive 'Not assessable' ratings is more definitive than prior qualitative assessments of evidence quality and should serve as a benchmark against which future reviews measure progress.

7. LIMITATIONS

Several limitations should be considered when interpreting the findings of this review.

1. Full-text access constraint. Detailed PICOS extraction and MMAT quality appraisal were completed for 30 of 278 included studies (10.8%), limited by full-text PDF accessibility. While the uniformity of quality ratings across the assessed subset and the consistency of design patterns observable in abstract-level data suggest that the broader corpus exhibits similar methodological characteristics, the 248 abstract-only studies were not individually quality-appraised.

2. Publication and language bias. The review was restricted to English-language publications, excluding potentially relevant research in Hindi, Chinese, Swahili, Portuguese, and other languages with active agricultural AI research communities. Publication bias toward positive technical results is likely given the predominance of system development and feasibility studies, which are rarely published if systems perform poorly.

3. Observational causal inference. The near-total absence of experimental designs across included studies prevents causal claims about AI intervention effects at either the study level or the synthesis level. Associations between AI tool use and reported knowledge or adoption improvements may reflect selection effects, confounding by farmer motivation and digital literacy, or social desirability bias in self-report data.

8. CONCLUSION

This systematic review of 278 studies presents the most comprehensive evidence synthesis to date on AI-enabled agricultural extension communication. The evidence base confirms technical feasibility and demonstrates promising signals for farmer knowledge improvement and technology adoption. However, across all farmer-centred effectiveness outcomes, evidence certainty is Very Low or Insufficient, driven by the dominance of small system-development pilots, heavy reliance on self-reported outcomes, and the near-total absence of randomised or controlled experimental designs.

The field now faces a critical inflection point. Continued investment in proof-of-concept system development, without proportionate investment in rigorous impact evaluation, risks creating a literature that justifies — but cannot empirically substantiate — large-scale deployment of AI extension tools. We call urgently for multi-site RCTs, longitudinal cohort studies, equity- disaggregated evaluations, and independent verification of adoption and productivity outcomes as the research priorities that will determine whether AI extension can deliver on its transformative potential for smallholder agriculture globally.

ACKNOWLEDGEMENTS

[To be completed by the authors.]

FUNDING

[To be completed by the authors.]

Declaration of Competing Interests

The authors declare no competing interests.

Data Availability

The PRISMA flow data, PICOS extraction tables, and MMAT appraisal matrices supporting this review are available from the corresponding author upon reasonable request.

REFERENCES

1. AgriLLM Research Team (2025). Empowering Smallholder Farmers with AgriLLM: A Low-Latency, High-Accuracy AI Assistant for Context-Relevant Farming Advice. Zenodo. <https://doi.org/10.5281/zenodo.15649531>
2. Aenugu, M. R. and Ramaraju, R. N. (2025). Design and Evaluation of a Natural Language-Based Chatbot for Smart Farming Advisory Services. *Future of Mechanical Engineering and Practices*, 1(3), pp. 53–65. <https://doi.org/10.69996/fmep.2025018>
3. Bhatia, R. and Kaur, S. (2024). Development and Evaluation of a Smart Agricultural Chatbot. In *Proceedings of the 2024 IEEE UEMCON (Ubiquitous and Smart Computing)*, pp. 158–163. IEEE. <https://doi.org/10.1109/uemcon62879.2024.10754742>
4. Climate Agronomy Working Group (2025). From Data Points to Designers: Climate Agronomy Learning Commons for Kenya's Western Highlands. Zenodo. <https://doi.org/10.5281/zenodo.15791048>
5. Collis, S., Kinyua, F., Kumar, V., Lakounga, H., Merz, C., Pandey, K. and Resch, C. (2025). Building AI-Based Advisory Services for Smallholder Farmers: Technical Learnings from the AIEP Initiative. *arXiv preprint arXiv:2601.11537*. <https://doi.org/10.48550/arxiv.2601.11537>
6. Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, M., Cargo, R., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. and Vedel, I. (2018). The Mixed Methods Appraisal Tool (MMAT) Version 2018 for Information Professionals and Researchers. *Education for Information*, 34(4), pp. 285–291. <https://doi.org/10.3233/EFI-180221>
7. Korir, M. K., Mwangi, W. and Kimwele, M. W. (2023). Artificial Intelligence-Based Chatbot Model Providing Expert Advice to Potato Farmers in Kenya. In *Proceedings of the 2023 IEEE AFRICON*, pp. 1–6. IEEE. <https://doi.org/10.1109/africon55910.2023.10293557>
8. Mehra, B. and Anitha, J. (2025). Empowering Indian Farmers with Multilingual AI: A Voice-Enabled ChatBot Using Dhenu2. In *Proceedings of the 2025 International Conference on Communication, Control, Robotics and Intelligent Systems (ICC-ROBINS)*, pp. 556–560. IEEE. <https://doi.org/10.1109/icc-robins64345.2025.11086147>
9. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P. and Moher, D. (2021). PRISMA 2020 Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews. *BMJ*, 372, n160. <https://doi.org/10.1136/bmj.n160>
10. Prabhu, S. V. and Saravanakumar, S. (2025). AI Powered Voice Chatbot for Agri Advisory in Tamil. *International Scientific Journal of Engineering and Management*, 4(7), pp. 1–9. <https://doi.org/10.55041/isjem04793>
11. Rajanala, M. S., Sivakiran, M., Sairevanth, M., Subbulakshmi, M. and Peralta, M. D. S. (2025). Agricultural Chatbot Voice Assistant Using NLP Techniques. *International Journal of Scientific and Academic Technology*, 16(1). <https://doi.org/10.71097/ijst.v16.i1.2892>
12. Sahoo, S. L. and Jena, A. (2025). Harnessing Artificial Intelligence for Agricultural Advisory Services: A Critical Review of Farmer-Led Experiences with the 'Ama Krushi' Chatbot in Odisha. *Archives of Current Research International*, 25(8), pp. 600–608. <https://doi.org/10.9734/acri/2025/v25i81443>
13. Shriya, M., Anusuya, V., A., S. and B., S. (2025). Vayazh — Leveraging AI and NLP to Empower Farmers with Real-Time Agricultural Insights. In *Proceedings of the 2025 International Conference on Computing, Communication, and Engineering Systems (ICCES)*, pp. 2130–2136. IEEE. <https://doi.org/10.1109/icc67310.2025.11337213>
14. Singh, N., Wang'ombe, J., Okanga, N., Zelenska, T., Repishti, J., JayasankarG, K., Mishra, S., Manokaran, R., Singh, V., Rafiq, M. I. and Gandhi, R. (2024). Farmer.Chat: Scaling AI- Powered Agricultural Services for Smallholder Farmers. *arXiv preprint arXiv:2409.08916*. <https://doi.org/10.48550/arxiv.2409.08916>