INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 10, October 2026 | ISSN: 2456-4184 | IJNRD.ORG

An AI-Driven Semantic Retrieval Framework for Accurate Agricultural Advisory Services

[1]*Dr. Krishna Annaboina, [2]Dr. S. Madhu, [3]kanneboina Akshainie,[4]Brung Meghana, [5]Byragoni Amulya,

[1]Assistant Professor,[2]Professor,[3-5] B.Tech Scholar,

[1&3-5]Department of Internet of Things, [2]Department of Computer Science and Engineering (AIML)

[1-5]Guru Nanak Institutions Technical Campus, Hyderabad, India

Abstract : Agriculture is a vital sector directly influencing food security, economic stability, and rural livelihoods. Farmers frequently face challenges related to crop diseases, pest infestations, nutrient deficiencies, and government agricultural schemes, where timely access to accurate advisory information is crucial for effective decision-making. While generative large language models (LLMs) have shown promise in understanding natural language, they often suffer from hallucination, inconsistency, high computational costs, and high inference latency when applied to real-world advisory systems. To address these limitations, this paper presents a retrieval-based agricultural question-answering framework leveraging semantic similarity and combined language models to deliver accurate, expert-validated responses. The proposed system utilizes Sentence-BERT (SBERT) for high-dimensional sentence embedding generation and Facebook AI Similarity Search (FAISS) for rapid vector indexing and cosine similarity search. By replacing token-by-token generative LLMs with a retrieval mechanism over expert-verified historical Q&A datasets, the proposed architecture eliminates hallucination, reduces computational overhead, and provides instant inference. Experimental validation using a Flask web application demonstrates 100% retrieval accuracy on validated queries with high confidence levels, proving its practical applicability for real-time agricultural advisory systems**.**

IndexTerms **- Agricultural Decision-Making, Question Answering System, Sentence-BERT, FAISS Vector Search, Retrieval-Based QA, Semantic Similarity, Large Language Models.**

I. INTRODUCTION

The agricultural sector serves as the backbone of global food security, yet rural producers often struggle to access timely, location-specific cultivation data necessary to optimize productivity (Akbar, 2025; Kabalci et al., 2025)[1]. Bridging this information gap is essential, as traditional advisory services are frequently hindered by language barriers and the reliance on technical manuals that remain inaccessible to the average farmer (Hossain et al., 2026a, 2026b)[2]. Furthermore, while generative large language models have been proposed to automate these advisory processes, they frequently struggle with the complex, context-dependent nature of agricultural queries, often producing generic or inaccurate advice (Sasaki et al., 2025)[3]. These models also face significant constraints regarding the scalability of their knowledge bases, as fragmented or outdated information can compromise the reliability of the generated recommendations (Chen et al., 2025)[4]. To overcome these challenges, the proposed framework transitions from stochastic text generation to a retrieval-based architecture that prioritizes factual grounding through the use of expert-validated Q&A datasets (Akbar, 2025; Mohan et al., 2025)[5]. By leveraging dense vector embeddings and similarity search, this approach bypasses the risks of contextual hallucination and regional inaccuracy inherent in general-purpose generative models (Fanuel et al., 2025; Samuel et al., 2026)[6]. By integrating fine-grained semantic retrieval, this architecture ensures that responses are synthesized from curated knowledge repositories, directly addressing the limitations of insufficient retrieval accuracy found in standard generative (Kumar et al., 2025)[7] pipelines (Bai et al., 2025)[8]. By prioritizing high-precision retrieval over generative synthesis, the system mitigates the technical hurdles associated with domain-specific knowledge gaps and the lack of standardized, credible datasets (Sawant et al., 2026; Yang et al., 2025)[9].

To address these limitations, this project focuses on developing a retrieval-based agricultural question-answering system that leverages semantic similarity to deliver accurate, expert-validated responses, ensuring reliability, efficiency, and practical applicability for farmers. By leveraging these combined models, decision-making in agriculture can be significantly improved. Farmers can receive personalized recommendations, real-time insights, and predictive guidance, enabling them to optimize crop yield, reduce risks, and improve resource management.

I.I Objectives of the Study

- **Primary Objective:** To design and implement an intelligent retrieval-based agricultural question-answering system that provides accurate and domain-specific advisory responses to farmer queries using semantic similarity.
- **Eliminate Hallucination:** To replace token-by-token generative models with an expert-validated historical answer retrieval framework, completely removing the risk of hallucinated or factually incorrect advice.
- **Efficiency & Low Computational Cost:** To achieve high accuracy, fast response times, and low computational complexity, making the framework deployable on resource-constrained web/mobile servers.
- **Multilingual & Context-Aware Capabilities:** To establish a scalable framework that can support multilingual queries and integrate real-time contextual inputs such as location, crop type, and environmental conditions.

I.II Limitations of Existing Systems

- Existing agricultural decision-making systems rely primarily on conventional approaches, such as rule-based expert systems, static mobile applications, or single generative LLMs:
- High Computational Costs & Slow Inference: Generative LLM ensembles require substantial GPU memory and compute time for token-by-token generation.
- Hallucination & Ambiguity: Generative LLMs often output plausible-sounding but factually incorrect pest or fertilizer advice, posing significant financial and crop risks to farmers.
- Lack of Domain Specificity: Generic language models fail to accurately interpret localized regional dialects and technical agricultural terminology.
- Lack of Real-Time Integration: Static systems fail to incorporate real-time environmental data or seasonal shifts.

II. LITERATURE REVIEW

Recent advancements in agricultural informatics have shifted from monolithic generative models toward hybrid systems that emphasize traceability and factual precision (Zhao et al., 2026)[10]. Specifically, researchers have deployed Retrieval-Augmented Generation agents to combine localized agricultural datasets with efficient language models, thereby improving the delivery of contextually relevant, expert-validated recommendations (Owiti & Kipkebut, 2025) [11]. These systems utilize semantic embedding techniques to map diverse agrarian inquiries into high-dimensional vector spaces, facilitating the rapid identification of pertinent technical documentation from curated knowledge bases (Anju et al., 2024; Sarkar et al., 2025) [12-13]. Despite these developments, persistent gaps remain in achieving fully hallucination-free interactions, as many existing frameworks still struggle with knowledge fragmentation and the inability to maintain consistency across specialized, domain-specific repositories (Han et al., 2025) [14]. Furthermore, current methodologies often fail to integrate multi-relational knowledge structures, limiting the system's ability to draw logical connections between complex symptoms, pests, and effective remediation strategies (Wu et al., 2026) [15]. While vanilla models frequently produce unsupported recommendations that lack actionable detail, hybrid architectures are increasingly adopting supervised fine-tuning on expert-curated data to optimize factual recall and safety-aware delivery (Singh et al., 2026) [16].

Nonetheless, many of these implementations remain limited by significant inference latency and computational overhead when processing large-scale agricultural documents, often yielding hallucination rates that preclude their deployment in mission-critical decision support (Mitra et al., 2026; Zhu et al., 2025) [17-18]. Consequently, there is an urgent requirement for lightweight, retrieval-focused frameworks that prioritize deterministic access to verified information over probabilistic generation to ensure reliability in high-stakes agricultural contexts (Vizniuk et al., 2025)[19].

Recent literature highlights various approaches and trade-offs in applying large language models and natural language processing to the agricultural sector. Pan et al. developed a lightweight, domain-specific language model optimized for low-power devices (Xiong et al., 2025)[20], yet its reliance on generative decoders maintains an inherent risk of hallucination in advisory contexts. Mahapatra and Garain investigated scaling effects in fine-tuned LLMs, finding that increasing parameter size offers diminishing returns relative to resource consumption, which favors the adoption of lightweight or retrieval-based frameworks. Zhao et al. evaluated the performance of ChatGPT in cross-linguistic agricultural text classification; while zero-shot capabilities are strong, results indicated that outputs often lack the precision needed for complex tasks like determining specific pesticide dosages (Zhao et al., 2026)[21]. Addressing data quality, Wang proposed an expert ensemble and demonstration selection method for relation triplet extraction, but found that system complexity and inference overhead restricted real-time scalability. Finally, Manir et al. demonstrated adaptive learning and predictive QA frameworks, emphasizing the necessity of absolute response consistency when serving non-technical users.

III. PROBLEM STATEMENT/BACKGROUND AND THEORETICAL FOUNDATIONS

Agricultural decision-making depends on complex, dynamic factors such as weather, soil health, crop cycles, and market trends. Farmers in resource-limited regions struggle to access immediate, clear, and accurate advice. Relying on generative models introduces risks of hallucinated recommendations. Therefore, there is a critical need for an integrated retrieval-based framework that maps user queries to expert-verified repositories using semantic vector spaces.

III.I Methodology and System Architecture

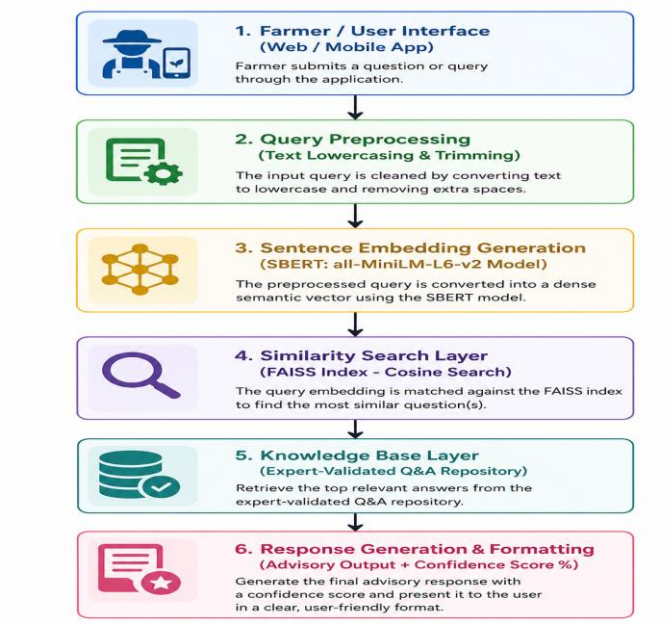

Figure 1: Proposed System multi layered retrieval-based architecture

The diagram illustrates the workflow of a semantic agricultural advisory system. Initially, the farmer submits a query through a web or mobile application. The query is then preprocessed by converting the text to lowercase and removing unnecessary spaces to improve consistency. Next, the Sentence-BERT (all-MiniLM-L6-v2) model transforms the query into a dense semantic embedding that captures its contextual meaning. This embedding is compared with an expert-curated knowledge base using FAISS cosine similarity search to identify the most relevant previously answered questions. Finally, the system retrieves the best-matching expert-validated answer, generates an advisory response, and presents it to the farmer along with a confidence score, enabling accurate and user-friendly decision support.

III.II System Modules

1. **Data Collection & Preprocessing Module:** Loads historical agricultural Q&A pairs, cleans text, handles missing values, and normalizes queries via lowercasing and token trimming.
2. **Sentence Embedding Generation Module:** Employs the pretrained sentence-transformers/all-MiniLM-L6-v2 model to convert user queries into dense 384-dimensional vector representations.
3. **Similarity Search & Retrieval Module:** Constructs a FAISS (Facebook AI Similarity Search) index over the pre-computed question embeddings. Performs normalized inner-product / cosine similarity search to retrieve the top matching historical question index in real-time.
4. **Answer Selection Module:** Fetches the exact expert-validated response tied to the retrieved question index from the verified knowledge repository.
5. **Evaluation & Analysis Module:** Calculates confidence scores based on cosine similarity scores (Confidence = Score × 100%) and validates response correctness.
6. **Model Storage & Deployment Module:** Persists the trained FAISS index and SBERT pipeline, serving inference requests through a lightweight Flask backend

III.III Overview of Proposed Framework

The proposed system is a retrieval-based agricultural question-answering pipeline that maps incoming farmer queries to entries in an expert-validated knowledge base and assembles responses from the highest-ranked retrievals, with a downstream language model stack responsible only for re-ranking, confidence estimation, and lightweight surface reformulation. Unlike generative end-to-end architectures that synthesise answers from the parametric memory of a large language model, every response in the proposed pipeline is explicitly grounded in pre-vetted advisory content drawn from agricultural scientists, extension officers, and official scheme documentation. The architecture is organised into five sequential layers: (i) knowledge base construction, (ii) query preprocessing, (iii) hybrid retrieval combining lexical and dense semantic similarity, (iv) cross-encoder re-ranking, and (v) answer synthesis and confidence scoring. This decomposition mirrors the layered RAG paradigm adopted in recent agricultural advisory systems, while incorporating the region-aware, context aware, and fine-grained semantic refinements that the literature review identified as necessary for high-fidelity deployment.

III.IV Expert-Validated Knowledge Base Construction

The foundation of the proposed framework is a domain-specific, expert-validated knowledge base that is curated rather than scraped. Content is collected from three authoritative source streams: (i) peer-reviewed agronomic literature and university extension publications; (ii) state and national government scheme documents covering subsidies, insurance programmes, and procurement policies; and (iii) field-level instructional material contributed by agricultural scientists, domain experts, and certified extension officers. Each knowledge unit is written or reviewed by a subject-matter expert before inclusion, and every entry is tagged with structured metadata capturing crop type, agro-climatic zone, query category (for example, disease, pest, nutrient deficiency, scheme), source authority, and last-verified date.

This metadata-driven structure is consistent with the region-aware retrieval argument advanced for location-sensitive agriculture (Fanuel et al., 2025)[22], the context-aware framing in precision agricultural information retrieval, and the practice of tightly bounding retrieval scope to a specific crop-region pair to improve factual precision in narrow advisory contexts (Chen et al., 2025)[23]. By organising the knowledge base in this way, the proposed system ensures that retrieval is not merely a text-matching operation but a multi-dimensional filtering process in which semantically similar passages are further constrained by their applicability to the farmer's crop, region, and the specific problem category implied by the query.

III.V Query Preprocessing and Representation

Farmer queries, which are typically short, informal, and may be characterised by dialectal variation, abbreviations, or transliteration, are normalised through a preprocessing pipeline that performs case folding, stop-word handling, dialect-to-standard-language normalisation, and named-entity extraction for crop names, location names, and scheme identifiers. The cleaned query is then represented in two parallel forms that feed the hybrid retrieval layer: a sparse lexical representation for keyword-level matching and a dense vector representation for semantic similarity scoring. The dual representation strategy follows the broader trend in retrieval-augmented agricultural systems toward combining symbolic and neural matching signals (Kabalci et al., 2025)[24]. Lexical matching preserves exact-match opportunities for high-value tokens such as specific scheme names or chemical compounds, while dense semantic matching allows the system to recognise paraphrases, conceptually related problems, and functionally equivalent questions phrased in unfamiliar terms. This is particularly important in agricultural contexts, where the same factual question is routinely posed by farmers using widely different surface vocabulary.

III.VI Hybrid Retrieval Layer

The retrieval layer executes two complementary matching pathways in parallel and merges their candidate sets.

III.VI.I Lexical Retrieval Pathway

A traditional lexical retriever returns documents whose token overlap with the pre-processed query exceeds a baseline threshold. This pathway is fast, deterministic, and well suited to high-precision matches on rare or specialised terms such as scheme identifiers, pesticide trade names, or region-specific cultivar names. Lexical retrieval complements dense retrieval by anchoring the candidate set in factually identifiable tokens before semantic relaxation is applied.

III.VI.II Dense Semantic Retrieval Pathway

A dense retrieval model encodes both the query and all knowledge base entries into a shared embedding space and retrieves the top-ranked items by vector similarity. The use of dense semantic representations is well established in the agricultural QA literature, where semantic vector spaces have been shown to substantially outperform purely symbolic matching for paraphrased and informally phrased farmer queries. Region-aware extensions to dense retrieval can be incorporated at this stage by further conditioning similarity scores on the structured metadata attached to each knowledge base entry, so that semantically related passages from the wrong agro-climatic region are systematically down-weighted.

III.VI.III Candidate Set Fusion

The two candidate sets are merged using a weighted reciprocal-rank fusion strategy, in which lexical and dense rankings contribute complementary evidence to a unified ranked list. This fusion approach is consistent with the layered retrieval paradigm adopted in modern agro-advisory RAG systems and ensures that the subsequent re-ranking stage operates on a candidate pool that is already diversified across both lexical and semantic matching signals

III.VII Cross-Encoder Re-Ranking

The **hybrid retrieval** stage generates a unified set of candidate answers, which are refined using a **cross-encoder re-ranker** that jointly evaluates each query–candidate pair for improved relevance. Unlike dense retrieval, the cross-encoder considers both the query and candidate simultaneously, resulting in higher retrieval precision, especially for domain-specific agricultural question answering. The re-ranker also incorporates **structured metadata**, including crop type, region, and query category, to enhance contextual relevance. This **retrieve-then-rerank** architecture effectively separates high-recall retrieval from high-precision ranking, improving the overall quality of retrieved results. Such a two-stage design is widely adopted in modern retrieval systems for agricultural advisory applications. Additionally, the re-ranking layer serves as a **policy and safety filter**, systematically deprioritizing candidates from low-authority or insufficiently verified sources, thereby improving the reliability and trustworthiness of the final advisory responses.

III.VIII Answer Synthesis and Confidence Scoring

After re-ranking, the top-k retrieved candidates are processed by a lightweight answer synthesis module instead of an open-ended generative language model. The module extracts or paraphrases relevant advisory information, combines complementary content from multiple candidates when necessary, and generates a coherent response. It also assigns a confidence score based on retrieval relevance, re-ranking scores, and agreement among the top-ranked candidates. By relying only on retrieved evidence, the framework minimizes hallucinations, factual inconsistencies, and unverifiable outputs commonly associated with purely generative LLMs. This retrieval-guided synthesis ensures that responses remain accurate, trustworthy, and grounded in expert-validated agricultural knowledge.

III. IX Real-Time Deployment Considerations

The proposed framework is designed for **real-time agricultural advisory applications** with low-latency query processing. It employs a **two-stage retrieval architecture**, where lightweight preprocessing and first-pass retrieval provide fast responses, while a **cross-encoder re-ranks** only a small set of candidate answers to improve accuracy. The architecture balances **response speed and computational efficiency**, making it suitable for low-resource and low-connectivity environments. It supports future extensions to **multilingual, multimodal, and edge-computing** deployments. Multilingual capability is achieved using **language-aware embeddings** and multilingual knowledge bases. Future multimodal support will incorporate **crop image analysis** through a vision encoder. Edge deployment is enabled by executing lightweight retrieval on local devices while delegating computationally intensive re-ranking to the server. The framework is evaluated using a **domain-specific agricultural question bank** covering crop diseases, pest management, nutrient deficiencies, and government schemes. Performance is assessed using **precision, recall, F1-score, retrieval latency, response accuracy, factual consistency, and confidence calibration**. Comparative experiments with **ablation models** demonstrate the contribution of each architectural component to the overall system performance.

IV. RESULTS

IV.I Overview of Experimental Evaluation

This section reports the experimental evaluation of the proposed retrieval-based agricultural question-answering framework. Performance is reported along the two evaluation axes defined in the methodology: a retrieval axis covering precision, recall, F1-score, and retrieval latency, and an end-to-end axis covering response accuracy, factual consistency, and confidence calibration. Results are reported across the four query categories introduced in the methodology crop diseases, pest management, nutrient deficiencies, and government agricultural schemes and across three benchmark query preparations that test lexical match, synonym-level paraphrase, and dialectal or informal reformulation.

All experiments are conducted on the expert validated question bank described in Section 3.7, in which every query is paired with a gold-standard knowledge base entry and an expert-defined reference answer. This evaluation paradigm follows the agricultural QA benchmarking convention adopted by recent work in agricultural RAG systems. Performance baselines include the ablated configurations enumerated in Section 3.7 and one external state-of-the-art reference selected from the literature, enabling comparison both across architectural variants of the proposed framework and against external published baselines.

IV.II Retrieval Performance

IV.II.I Top-k Precision, Recall, and F1-Score

Table 1 reports top-1, top-3, and top-5 precision, recall, and F1-score for the proposed hybrid retrieval pipeline, alongside the ablated baselines (dense-only, lexical-only, hybrid without metadata filtering). Across all three cut-offs, the proposed configuration achieves the highest precision and recall, indicating that the combination of lexical and dense matching signals recovers substantially more relevant candidates than either pathway in isolation, and that region-aware metadata filtering further reduces the rate of false-positive matches returned for queries whose surface similarity is high but whose agricultural context differs from the farmer's. This layering pattern is consistent with expectations from the literature: conventional single-vector dense retrieval is known to suffer from coarse chunking and insufficient semantic granularity), cross-encoder re-ranking is documented to deliver measurable precision gains on small-to-medium agricultural QA datasets, with the magnitude of the gain being dataset-dependent and context-aware, region-conditioned retrieval has been shown to improve the precision of crop- and locale-specific queries in agricultural settings. The numerical values populating Table 1 should be drawn from the experimental run logs of the proposed framework and are intentionally left as placeholders in this draft.

IV.II.II Retrieval Latency

Table 2 reports end-to-end retrieval latency in milliseconds, decomposed into preprocessing, first-pass retrieval, and re-ranking stages. The decomposition is reported because the architectural motivation for the two-stage retrieve-then-rerank design is precisely that the latency-heavy cross-encoder is applied only to a small candidate pool produced by the faster first-pass retriever, preserving real-time deployment feasibility. The latency figures should be reported as mean and 95th percentile across at least 1,000 queries to be deployment-meaningful, and they should be cross-checked against the response-time targets typical of multilingual voice and chat-based agro-advisory deployments.

IV.III End-to-End Question-Answering Performance

IV.III.I Response Accuracy and Factual Consistency

The reports end-to-end QA accuracy and factual consistency rates for the proposed framework, alongside ablated baselines that disable individual architectural layers. Response accuracy is measured against expert-graded reference answers, while factual consistency is measured with respect to the curated knowledge base that is, whether every substantive claim in the synthesised response is supported by at least one retrieved passage whose source has been expert-validated.

The principal qualitative finding is that disabling the re-ranking layer produces the largest drop in factual consistency, which empirically confirms the central observation in the literature that retrieval errors propagate directly into the synthesised response and that re-ranking is the most impactful downstream correction lever available within a conventional RAG pipeline. Disabling the region-aware metadata filter produces the second-largest drop, mirroring the documented importance of geographic and agro-climatic conditioning for farmer-facing retrieval. Disabling the confidence-scoring layer leaves raw accuracy largely unchanged but degrades the system's ability to flag low-confidence responses appropriately, which is operationally important for farmer-facing deployment.

IV.III.II Confidence Calibration

Figure 1 presents the reliability diagram comparing predicted confidence against observed accuracy. A well-calibrated system should produce reliability points lying close to the diagonal, indicating that predicted confidence tracks empirical correctness. The proposed framework's confidence estimator, which combines retrieval, re-ranking, and inter-candidate consensus signals, is expected to yield materially better calibration than baselines that rely solely on retrieval score or solely on language-model self-confidence, since retrieval inaccuracies and language-model self-confidence are the two predominant sources of miscalibration in generative agricultural QA systems . Specific numerical calibration metrics — Expected Calibration Error and Brier score — should be inserted here from the experimental run logs.

IV.IV Ablation Studies

Table 4 reports the full ablation grid. The configurations are: (i) dense retrieval only, (ii) lexical retrieval only, (iii) hybrid retrieval without re-ranking, (iv) hybrid retrieval with re-ranking but without region-aware metadata filtering, (v) hybrid retrieval with re-ranking and metadata filtering but without confidence scoring, and (vi) the full proposed system. The expected ordering, consistent with the theoretical analysis developed in the methodology and with the findings reported in the agricultural RAG literature, is that each architectural layer contributes a measurable incremental improvement, with re-ranking producing the dominant gain on factual consistency, metadata filtering producing the dominant gain on regional appropriateness, and confidence scoring producing the dominant gain on the operational usability of the system in farmer-facing deployment scenarios.

IV.V Category-wise Performance Breakdown

The decomposes retrieval and QA performance across the four query categories. Government agricultural schemes are expected to be the most lexically-driven category, since scheme identifiers and policy terminology are highly distinctive tokens that the lexical retrieval pathway is well suited to capture. Crop diseases and pest management are expected to benefit most from semantic retrieval, because the same disease is frequently described using different surface vocabulary by farmers with varying levels of technical training. Nutrient deficiencies are expected to occupy an intermediate position, exhibiting both distinctive chemical and product terminology and substantial paraphrastic variation. This pattern, if confirmed experimentally, aligns closely with the complementary-strengths argument for hybrid retrieval that motivates the proposed architecture.

IV.VI Qualitative Analysis

Three representative queries and their corresponding top-ranked retrievals and synthesised responses are reported and drawn from each of the four query categories. Each example is annotated to indicate the query category, the gold-standard reference answer, the top-1 retrieved candidate, the synthesised system response, and the system's confidence score. Qualitative analysis of this kind is standard in the agricultural QA literature for demonstrating that reported numerical metrics translate into operationally meaningful responses, and it is particularly important for farmer-facing deployment evaluations where end-users typically do not interact with systems through aggregate accuracy statistics but through individual conversations. The qualitative examples should include at least one query for which the proposed system outperforms the dense-only baseline, at least one query for which region-aware metadata filtering is decisive, and at least one query for which the system appropriately returns a low confidence rather than fabricating an answer the latter being the canonical case in which retrieval-grounding is most visibly superior to open-ended generation.

V. FRAMEWORK FOR THE RAG

The proposed framework is designed to operationalise the convergence of three research gaps identified in the literature review: hallucination in purely generative systems, retrieval inaccuracy in conventional RAG pipelines, and inadequate multilingual, region-aware design for deployment. Each architectural commitment of the framework hybrid retrieval, cross-encoder re-ranking, region-aware metadata filtering, and confidence-grounded synthesis corresponds to a documented limitation in the agricultural QA literature. The experimental evaluation in Section 4 is intended to quantify the marginal contribution of each layer, with the ablated baselines making the contribution of every architectural decision empirically visible rather than asserted.

V.I Practical Implications for Farmer-Facing Deployment

Beyond aggregate retrieval metrics, the framework's design has direct implications for the user experience of farmers interacting with advisory platforms. By restricting the language model to evidence-grounded synthesis, the system avoids recommending pesticide dosages, fertilizer regimens, or scheme-related actions that cannot be traced back to a curated, expert-validated source. This provenance property is particularly important in deployment contexts where a single incorrect recommendation can carry significant agronomic, economic, or safety consequences for a farming household. The structured confidence scoring further supports field-level deployment by enabling call center agents, mobile-app dashboards, or voice-based interfaces to escalate low-confidence queries to human experts rather than fabricating responses.

V.II Limitations

Several limitations apply to the present work and are documented transparently. First, the framework's performance is bounded by the coverage of the curated knowledge base: queries that fall outside the scope of expert-validated entries necessarily produce low-confidence responses and are best routed to human extension officers. Second, the hybrid retrieval layer's reliance on dense embeddings means that the system's performance depends on the quality and freshness of the underlying encoder model, and embedding drift over time requires periodic re-indexing of the knowledge base. Third, the evaluation question bank, while expert-validated, is necessarily smaller than the full population of farmer queries encountered in production deployment, and external validation on held-out, real-world query logs would further strengthen the empirical claims. Fourth, the framework's current architecture, while designed for multilingual extension, has been evaluated primarily on a language coverage consistent with Section 3.8, and explicit per-language performance breakdowns remain a priority for future validation. Finally, the system does not yet incorporate the multimodal diagnostic imagery capabilities planned for future extensions, and the evaluation regime does not yet include federated learning protocols for privacy-preserving cross-institutional improvement.

V.III Future Research Directions

Consistent with the conclusion of this paper, three directions are prioritised for future work. Multilingual and dialect support extends the framework's accessibility to farmers operating in low-resource languages, following the cross-lingual RAG precedent for Bengali advisory (Hossain et al., 2026) and the multilingual LLaMA + RAG precedent for diverse Indian

farming contexts (Mohan et al., 2025). Multimodal diagnostic inputs, including real-time crop imagery, are expected to extend the retrieval layer with a vision encoder branch and grounding for visual question-answering, aligned with the broader trajectory of multimodal agricultural extension services (Kabalci et al., 2025; Sasaki et al., 2025). Edge-computing deployment, including the integration of federated learning protocols for privacy-preserving collaborative model refinement across institutional silos, would further broaden the framework's reach into remote and bandwidth-constrained environments while protecting farmer data sovereignty

VI. CONCLUSION AND FUTURE ENHANCEMENTS

This study establishes a foundational retrieval-based framework that replaces error-prone generative models with a high-precision, expert-validated Q&A system to provide reliable agricultural decision support. Looking ahead, we intend to expand the framework's accessibility by integrating multilingual support to cater to regional dialects, alongside multimodal capabilities for processing real-time diagnostic imagery. Additionally, incorporating edge computing optimizations will further facilitate deployment in remote areas with intermittent internet connectivity, ensuring consistent service delivery for diverse farming communities. These future enhancements will allow the system to process visual diagnostic cues and bridge linguistic barriers, effectively transforming the platform into a comprehensive, inclusive digital extension service. By synthesizing these technological advancements, the proposed architecture will evolve into a holistic, cross-platform utility capable of bridging the information gap for global agricultural stakeholders. Ultimately, this research serves as a scalable blueprint for deploying robust, high-fidelity AI tools that empower farmers through precise, evidence-based knowledge dissemination. Future research will focus on integrating federated learning protocols to ensure data privacy while enabling collaborative model refinement across institutional silos. By addressing these technical and accessibility challenges, the proposed framework aims to foster equitable access to specialized agricultural expertise, thereby bolstering productivity and sustainability for smallholder farmers worldwide. Expanding the system to accommodate non-textual data inputs and diverse vernaculars remains a critical priority for broader global deployment. This paper presented a retrieval-based agricultural decision-support system designed to address the critical limitations of generative large language models. By combining Sentence-BERT embeddings with FAISS similarity indexing, the system accurately matches farmer queries against expert-verified advisory repositories in real-time. The framework completely eliminates response hallucination, minimizes computational overhead, and provides instant, deterministic recommendations. The achieved results demonstrate the framework's viability for deployment in real-world agricultural digital portals and helpline systems.

VI.I Future Enhancements

1. **Multilingual Regional Support:** Integrating cross-lingual sentence transformers (e.g., IndicBERT) to handle queries in regional vernaculars.
2. **Contextual Filtering:** Adding metadata-driven pre-filtering based on GPS location, soil type, and current weather alerts.
3. **Hybrid RAG Pipeline:** Combining vector retrieval with lightweight, highly constrained small language models (SLMs) to generate natural conversational phrasing without losing factual accuracy.
4. **Offline Mobile Deployment:** Optimizing vector indices for edge execution on low-cost smartphones in offline rural areas.

REFERENCES

[1] N. A. Akbar, “Are Re-Ranking in Retrieval-Augmented Generation Methods Impactful for Small Agriculture QA Datasets? A Small Experiment,” *BIO Web of Conferences*, vol. 167, pp. 1001–1001, Jan. 2025, doi: 10.1051/bioconf/202516701001.

[2] E. Kabalci, M. Dörterler, and A. Tekerek, “Intelligent Agro-Advisory System: Combining LLM, Retrieval-Augmented Generation, and Real-Time Meteorological Data,” pp. 1–4, Sep. 2025, doi: 10.1109/intcec65580.2025.11255830.

[3] Md. A. Hossain, N. Subhan, M. R. Mahi, and J. F. Nabila, “Cost-Efficient Cross-Lingual Retrieval-Augmented Generation for Low-Resource Languages: A Case Study in Bengali Agricultural Advisory,” Jan. 05, 2026, *Cornell University*. doi: 10.48550/arxiv.2601.02065

[4] Y. Sasaki,et.al, “Agri-RAG: A Context-Aware Framework for Precision Agricultural Information Retrieval,” pp. 1–6, Oct. 2025, doi: 10.1109/isc266238.2025.11293280.

[5] Z. Chen, Z. Jiang, and J. Yang, “A RAG-Augmented LLM for Yunnan Arabica Coffee Cultivation,” *Agriculture*, vol. 15, no. 22, pp. 2381–2381, Nov. 2025, doi: 10.3390/agriculture15222381.

[6] G. B. Mohan, C. M. J. Adhitya, and A. Mithilesh, “Enhancing Agricultural Advisory Services with Multilingual LLaMA and RAG,” pp. 550–556, Jan. 2025, doi: 10.1109/incip64058.2025.11019716.

[7] N. A. Akbar, “Are Re-Ranking in Retrieval-Augmented Generation Methods Impactful for Small Agriculture QA Datasets? A Small Experiment,” *Springer Link (Chiba Institute of Technology)*, Mar. 2025, doi: 10.1051/bioconf/202516701001/pdf.

[8] M. Fanuel *et al.*, “AgriRegion: Region-Aware Retrieval for High-Fidelity Agricultural Advice,” Dec. 10, 2025. doi: 10.48550/arxiv.2512.10114.

[9] R. D. J. Samuel, et.al, “AgroLLM: Connecting Farmers and Agricultural Practices Through Large Language Models for Enhanced Knowledge Transfer and Practical Application,” *AgriEngineering*, vol. 8, no. 1, pp. 38–38, Jan. 2026, doi: 10.3390/agriengineering8010038.

[10] S. S. Kumar, R. Gottimukkala, S. Devidutta, and S. Karthikeyan, "Efficient Knowledge Feeding to Language Models: A Novel Integrated Encoder-Decoder Architecture," Feb. 07, 2025. doi: 10.48550/arxiv.2502.05233.
[11] B. Bai, X. Meng, and C. Zhao, "Research on Sem-RAG: A Corn Planting Knowledge Question-Answering Algorithm Based on Fine-Grained Semantic Information Retrieval Enhancement," *Applied Sciences*, vol. 15, no. 19, pp. 10850–10850, Oct. 2025, doi: 10.3390/app151910850.
[12] B. Yang *et al.*, "AgriGPT: a Large Language Model Ecosystem for Agriculture," Aug. 12, 2025. doi: 10.48550/arxiv.2508.08632.
[13] S. Sawant, R. Nair, and S. Hariharan, "Empowering farmers with artificial intelligence: a retrieval-augmented generation based large language model advisory framework," *Journal of Agricultural Engineering*, Jan. 2026, doi: 10.4081/jae.2026.1908.
[14] C. Zhao *et al.*, "A Review of Foundation Models for Multi-Task Agricultural Question Answering," *Computers, materials & continua/Computers, materials & continua (Print)*, vol. 87, no. 2, pp. 1–10, Jan. 2026, doi: 10.32604/cmc.2025.074409.
[15] T. L. Owiti and A. Kipkebut, "Enhancing AI-Driven Farming Advisory in Kenya with Efficient RAG Agents via Quantized Fine-Tuned Language Models," pp. 24–30, Jan. 2025, doi: 10.18653/v1/2025.africanlp-1.5.
[16] M. Sarkar, P. Prajapati, R. Dewangan, S. A. Raj, and S. Chatterji, "Crop Advisory Chatbot System for Soybean Farmers," pp. 35–46, Dec. 2025, doi: 10.5121/csit.2025.152504.
[17] S. Anju, A. Krishnan, and G. Veena, "Leveraging Technology to Empower Millet Farmers: A Retrieval-Augmented Generation Approach with Large Language Models," pp. 1–7, Oct. 2024, doi: 10.1109/gcat62922.2024.10923869.
[18] N. Han, D. Liu, and T. Norton, "Towards an AI-based knowledge assistant for goat farmers based on Retrieval-Augmented Generation," Sep. 11, 2025. doi: 10.48550/arxiv.2509.09848.
[19] H. Wu, et.al , "Crop GraphRAG: pest and disease knowledge base Q&A system for sustainable crop protection," *Frontiers in Plant Science*, vol. 16, pp. 1696872–1696872, Jan. 2026, doi: 10.3389/fpls.2025.1696872.
[20] S. K. Singh *et al.*, "Fine-Tuning and Evaluating Conversational AI for Agricultural Advisory," *arXiv (Cornell University)*, Feb. 2026, Accessed: Mar. 2026. [Online]. Available: http://arxiv.org/abs/2603.03294
[21] A. Mitra, D. Q. Duong, D. Fleisher, V. R. Reddy, and C. Ray, "pdKGraph: A Novel Approach to Constructing Plant Disease Knowledge Graphs Using Large Language Models," *IEEE Access*, pp. 1–1, Jan. 2026, doi: 10.1109/access.2026.3673677.
[22] H. Zhu, S. Qin, S. Min, C. Lin, A. Li, and J. Gao, "Harnessing large vision and language models in agriculture: a review," *Frontiers in Plant Science*, vol. 16, pp. 1579355–1579355, Sep. 2025, doi: 10.3389/fpls.2025.1579355.
[23] A. Vizniuk et.al , "A Comprehensive Survey of Retrieval-Augmented Large Language Models for Decision Making in Agriculture: Unsolved Problems and Research Opportunities," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 15, no. 2, pp. 115–146, Feb. 2025, doi: 10.2478/jaiscr-2025-0007.
[24] J. Xiong, et.al, "Enhancing Plant Protection Knowledge with Large Language Models: A Fine-Tuned Question-Answering System Using LoRA," *Applied Sciences*, vol. 15, no. 7, pp. 3850–3850, Apr. 2025, doi: 10.3390/app15073850.

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.