Journal of Agricultural Engineering

https://www.agroengineering.org/

Empowering farmers with artificial intelligence: a retrieval-augmented generation based large language model advisory framework

Shreeram Sawant, Rahul Nair, Siddharth Hariharan

Computer Engineering, Terna Engineering College, Navi Mumbai, Maharashtra, India

Corresponding author: Shreeram Sawant, Computer Engineering, Terna Engineering College, C603, Swaraj CHS, Rd no. 13, Tilak Nagar, Mumbai 400089, India. E-mail: sawantshreeram2122@ternaengg.ac.in

Please cite this article as doi: 10.4081/jae.2026.1908

Submitted: 7 July 2025
Accepted: 5 December 2025

Empowering farmers with artificial intelligence: a retrieval-augmented generation based large language model advisory framework

Shreeram Sawant, Rahul Nair, Siddharth Hariharan

Computer Engineering, Terna Engineering College, Navi Mumbai, Maharashtra, India

Corresponding author: Shreeram Sawant, Computer Engineering, Terna Engineering College, C603, Swaraj CHS, Rd no. 13, Tilak Nagar, Mumbai 400089, India. E-mail: sawantshreeram2122@ternaengg.ac.in

Contributions: Shreeram Sawant contributed to conception and design of the RAG-based LLM advisory framework, analysis and interpretation of experimental results, drafting of the original manuscript, and critical revision for important intellectual content. **Rahul Nair** contributed to conception and design of the system architecture, analysis and interpretation of performance data, drafting of methodology sections, and critical revision for important intellectual content. **Siddharth Hariharan** contributed to conception and design of the research approach, analysis and interpretation of results, critical revision of the manuscript for important intellectual content. All authors provided final approval of the version to be published and agreed to be accountable for all aspects of the work.

Conflict of interest: The authors declare no competing interests.

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Dataset: https://doi.org/10.5281/zenodo.15881813

Code: https://doi.org/10.5281/zenodo.17542352

Abstract

This study presents a retrieval augmented generation (RAG) based system designed to provide farmers with expert agricultural advisory services. The framework delivers context aware guidance on critical practices such as crop cultivation, pest and disease management, fertilizer application, and other agronomic practices, and compares the performance of four large language models (LLMs) in generating these recommendations. The system processes package of practices (PoP) documents for five major crops maize,

ragi, sweet potato, cotton, and groundnut through semantic chunking and embedding using Amazon Titan *via* BedrockEmbeddings. Vector representations are indexed in ChromaDB to enable efficient similarity search for query-relevant content retrieval. Upon receiving user queries, the system retrieves the most semantically similar document chunks and incorporates them into structured prompts. Four LLMs such as Llama3.1, Mistral, Phi3, and Qwen2.5 were evaluated for their effectiveness in generating accurate agricultural recommendations. Performance was evaluated across multiple dimensions. Relevance and retrieval were assessed using precision@K, recall@K, mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG). Lexical overlap was measured with the bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE-1, ROUGE-2, ROUGE-L) metrics. Semantic quality was analyzed using Bidirectional Encoder Representations from transformers score (BERTScore) precision, recall, F1, semantic similarity and faithfulness to capture contextual alignment between generated and reference responses. Source attribution was assessed through the attribution score, while efficiency was measured using retrieval time, generation time, and total time. Overall, mistral and Qwen2.5 achieved the highest performance, demonstrating superior relevance, semantic quality, and efficiency. This evaluation highlights which LLMs perform best for the agricultural domain and illustrates the potential of knowledge-grounded AI systems to democratize agricultural expertise, particularly in regions with limited access to traditional advisory services.

Key words: Agricultural advisory systems, large language models (LLMs), question answering, retrieval augmented generation (RAG), semantic retrieval, vector databases.

Introduction

Agriculture plays a vital role in sustaining human life and economic development, especially in regions where farming is the primary occupation. Yet, farmers around the world often face significant challenges in making informed decisions about crop selection, soil health, pest control, irrigation, and market access. These decisions typically require timely, location-specific, and expert-level guidance. In many rural and underserved areas, however, access to agricultural experts, government extension officers, or reliable digital resources remains limited or inconsistent (Dhanabalan and Sathish, 2018).

In recent years, artificial intelligence (AI) has shown great potential in addressing these gaps. Among AI technologies, conversational agents commonly known as chatbots have emerged as accessible tools that can provide real-time responses to user queries (Kar and Haldar, 2016). These systems can simplify complex agricultural knowledge and make it

more accessible to farmers. However, traditional chatbots are often limited in scope. Rule-based systems can only respond to predefined queries, while purely generative models may produce inaccurate or misleading answers due to a lack of grounded information.

To address these shortcomings, the retrieval augmented generation (RAG) architecture has emerged as a robust solution by grounding responses in factual data. This hybrid approach has demonstrated significant success in other high-stakes domains where informational accuracy is non-negotiable. For instance, RAG-based systems have been developed to generate high-quality question-answer pairs for human health risk assessment, provide nuanced sentiment analysis of financial texts, and parse complex international customs documentation. The proven ability of RAG to deliver reliable, context-aware information in specialized fields like medicine and legal services highlights its immense potential for transforming agricultural advisory systems, where precise and trustworthy guidance is equally critical. While traditional rule-based chatbots are inflexible and purely generative models risk producing factually incorrect 'hallucinations', the RAG framework overcomes these limitations. By first retrieving relevant, up-to-date information from a trusted knowledge base before generating a response, RAG ensures that the guidance provided to farmers is both accurate and contextually specific. This grounding in factual documents is crucial for agricultural applications where incorrect advice can have significant real-world consequences.

This paper presents the design, development, and evaluation of a RAG based framework for knowledge-grounded agricultural advisory systems. Focusing on five major crops such as maize, ragi, sweet potato, cotton, and groundnut the system leverages authoritative package of practices (PoP) documents to construct its knowledge base. A key contribution of this work is the systematic evaluation and comparison of four distinct LLMs: Llama3.1, Mistral, Phi3, and Qwen2.5 to determine their effectiveness in generating accurate agricultural recommendations. By integrating document retrieval with generative models, the system effectively addresses the knowledge accessibility gap prevalent in farming communities. This approach aligns with the broader goals of digital agriculture and rural empowerment by making scientific guidance more accessible, context-specific, and scalable, particularly in regions with limited access to traditional agricultural advisory services.

RAG has gained prominence as an effective approach that integrates information retrieval with generative modeling to enhance accuracy and relevance making it particularly valuable in agriculture, where reliable and region-specific knowledge is crucial. Studies such as Zafarmomen and Samadi (2025) on adverse weather reasoning, Wilkho *et al*. (2023) on flash flood detection using Flash Flood BERT (FF-BERT), and Zhou *et al*. (2022) on harvesting social media rescue requests with VictimFinder illustrate the utility of LLMs and BERT-based architectures in handling domain-specific knowledge, reasoning under uncertainty, and delivering contextually accurate outputs. These works provide a strong rationale for applying similar approaches in agriculture to generate precise, context-aware recommendations for farmers. Recent studies have examined the potential of RAG based architectures across domains including agriculture, health, and finance. Khanifar (2025) evaluated models such as Claude 3.5 Sonnet and GPT-4o for soil science queries, reporting 65% accuracy but limited performance for complex contextual questions. Meng *et al*. (2025) proposed a RAG framework for human health risk assessment that improved factual precision through optimized retrieval mechanisms. Xiong *et al*. (2025) developed an agricultural question-answering system using RAG with a localized knowledge base of 7,000 plant protection documents and low rank adaptation of large language models (LoRA) tuned InterLM-20B, demonstrating improved factual consistency and contextual relevance. Similarly, Yin *et al*. (2025) reviewed agricultural foundation models (AFMs) and identified persistent challenges such as dataset quality, training efficiency, and domain variability, emphasizing the need for domain-adapted RAG systems.

In related fields, Hu *et al*. (2025) introduced intelligent customs clearance assistant using retrieval augmented generation (ICCA-RAG) for multimodal customs documentation, achieving higher relevance and factual accuracy, an approach adaptable to heterogeneous agricultural data. Legashev *et al*. (2025) found that graph-based dialogue management surpassed tree-based methods in maintaining conversational coherence based on BLEU and BERTScore metrics. Acharya *et al*. (2025) explored agentic AI, highlighting the potential of autonomous, goal oriented systems for complex decision-making in agriculture. Mathebula *et al*. (2024) proposed language feature extraction and adaptation for reviews (LFEAR), a RAG-enhanced autoregressive fine-tuning model for financial sentiment analysis, which achieved 97% context precision demonstrating the

cross-domain adaptability of RAG techniques for structured reasoning and context retention.

Several domain-specific studies further demonstrate the real-world impact of RAG in agriculture. Balpande *et al*. (2024) developed an AI-powered chatbot integrated with geographic information system (GIS) and IBM Watson Assistant to deliver localized advice for Kenyan potato farmers. Saha and colleagues (2024) proposed question-to-question inverted index matching (QuIM)-RAG, leveraging question-to-question inverted index matching to enhance semantic accuracy and response relevance. A *et al*. (2024) and V *et al*. (2024) implemented RAG systems that combined knowledge retrieval with sensor-based soil monitoring to improve decision-making and reduce misinformation. Salim *et al*. (2024) designed an open-source platform enabling the deployment of low-resource LLMs for agricultural support, enhancing accessibility and scalability. Arslan *et al*. (2024) also reviewed RAG applications across domains and identified agriculture as an underexplored area compared to medicine and technology, calling for more integration of domain-specific data sources.

Overall, the reviewed literature highlights RAG's transformative potential in enhancing factual accuracy, semantic relevance, and user engagement in AI-driven advisory systems. Despite significant progress, key challenges persist, including contextual understanding, data quality, scalability, and domain adaptation. These findings underline the need for continued research to refine RAG-based frameworks tailored for agriculture, systems capable of democratizing access to reliable, knowledge-grounded, and region-specific farming guidance, thereby advancing sustainable agricultural development.

Materials and Methods

Methodology

This research introduces a RAG based framework for knowledge-grounded agricultural advisory systems, as illustrated in Figure 1, designed to enhance farmer's access to timely and reliable agricultural guidance. The system is built on a RAG framework, which integrates document retrieval with natural language generation to produce responses that are both accurate and context-aware. By retrieving relevant information from a structured agricultural knowledge base and generating tailored responses based on user

input, the system delivers practical, query-specific insights. This section details the system architecture, including the data preparation process, mechanisms for knowledge representation and retrieval, and the end-to-end pipeline used for generating and delivering responses.

The proposed RAG system for agricultural advisory services employs a modular architecture designed to deliver accurate, context-aware responses to farming queries through comparative evaluation of multiple large language models. The system's foundation consists of comprehensive PoP documents covering five strategically selected crops: maize, ragi, sweet potato, cotton and groundnut. These crops were chosen for their regional significance, nutritional value, and economic importance across diverse Indian farming systems. Maize, a highly adaptable cereal crop grown in both Kharif and Rabi seasons, is featured with detailed best practices on hybrid seed selection, precision sowing techniques, fertilizer scheduling, irrigation management, and integrated pest and disease control strategies. Its inclusion reflects its multipurpose role in food, livestock feed, and industrial usage. Ragi (finger millet), recognized for its exceptional nutritional content particularly calcium, fiber, and essential amino acids is cultivated using low-input, eco-friendly methods. The dataset outlines steps for seed priming, organic nutrient management, timely weeding, and biological control measures to ensure productivity and sustainability.

Sweet potato is a climate-resilient root crop valued for its high carbohydrate and vitamin A content. The dataset outlines basic cultivation practices such as selection of healthy vines, ridge planting for better tuber formation, and moisture conservation. It also includes guidance on nutrient application and pest control for improved root quality and yield. Cotton, an important fiber crop, is featured with agronomic recommendations including the use of BT and hybrid varieties, appropriate spacing, and seed treatment. The practices also focus on balanced fertilization and integrated pest management to address bollworms and sucking pests, ensuring healthy crop development. Groundnut (peanut), a leguminous crop known for its protein and oil content, is addressed through key practices such as seed treatment, gypsum application, and proper irrigation during flowering and pegging. The dataset emphasizes disease management strategies for leaf spot and root rot, and highlights the crop's role in soil fertility improvement. Overall, the dataset encompasses crop-specific guidelines including hybrid seed selection, precision

sowing techniques, fertilizer scheduling, irrigation management, integrated pest and disease control strategies, and post-harvest management practices. Area and production statistics for the selected crops are summarized in Table 1 (Government of Kerala, 2020). Document preprocessing begins with structured content extraction from PDF formats to remove formatting inconsistencies and isolate meaningful agricultural information. The cleaned content undergoes semantic chunking, where documents are segmented into coherent, self-contained knowledge units. Each chunk is annotated with relevant metadata including crop type, agricultural activity, seasonal applicability, and regional specificity to enhance retrieval precision. The preprocessing and embedding parameters used in this study, including chunk size, overlap, embedding dimensions, and generation settings such as temperature and top-p sampling, are summarized in Table 2. Each preprocessed text chunk is transformed into dense vector representations using Amazon Titan embeddings *via* BedrockEmbeddings. The embedding process converts textual content into high-dimensional vectors that capture semantic meaning beyond simple keyword matching. Each chunk c_i is transformed into a dense vector representation e_i using an embedding model E,

$$e_i = E(c_i) \quad \text{(eq. 1)}$$

Where E represents the Amazon Titan embedding model e_i and is the resulting vector representation.

These embeddings are indexed and stored in ChromaDB, a vector database optimized for semantic similarity search operations that enables efficient retrieval of contextually relevant information through cosine similarity calculations between query and document vectors. When users submit agricultural queries, the system processes the input through the same embedding model to generate a query vector q . Semantic similarity between the query and stored document chunks is computed using cosine similarity:

$$cosine\_sim(q, ei) = \frac{(q \cdot e_i)}{(||q|| * ||e_i||)} \quad \text{(eq. 2)}$$

The retrieval module identifies the top-k most semantically relevant chunks based on similarity scores, and these retrieved segments are assembled into a coherent context

block that provides comprehensive background information for response generation. All experiments were conducted on a 64-bit operating system desktop workstation equipped with an AMD Ryzen 7 2700 Eight-Core Processor, 16.0 GB of RAM, and an NVIDIA GeForce GTX 1660 SUPER graphics card with 6 GB of VRAM. This hardware configuration was chosen to evaluate the system's performance on accessible, consumer-grade hardware, which is a key consideration for practical deployment in agricultural advisory contexts. The core contribution of this study lies in the systematic comparison of four LLMs accessed through the Ollama framework: Llama3.1, Mistral, Phi3, and Qwen2.5. These models were selected based on a combination of architectural diversity, instruction-following capabilities, open-source availability, and suitability for deployment in resource-constrained agricultural contexts. Mistral is a dense decoder-only transformer known for delivering strong performance relative to its size and excels in instruction-tuned tasks. Llama 3.1, developed by Meta, serves as a widely adopted open-weight baseline demonstrating consistent generalization across domains. Phi-3, a compact, instruction-optimized model developed by Microsoft, was chosen for its impressive performance despite a smaller parameter count, making it a practical candidate for lightweight, in-field deployments. Qwen 2.5, a recent large-scale LLM, is recognized for its robust reasoning capabilities, multi-turn conversation handling, and adaptability across diverse domains, making it particularly suitable for generating accurate, context aware agricultural recommendations. Collectively, these models represent a balanced spectrum of model size, training data diversity, and computational efficiency, enabling a fair assessment of trade-offs between performance and resource consumption.

To ensure an unbiased evaluation, a fixed benchmark dataset comprising 27 agricultural advisory queries across five crops was constructed. Each query was paired with a ground truth answer sourced from authoritative PoP documents published by agricultural extension agencies. The queries reflected real world farmer concerns, including factual, procedural, and temporal questions related to crop varieties, planting schedules, soil and climate conditions, irrigation practices, and nutrient management. This standardized dataset was used uniformly across all model evaluations while a subset of the standardized dataset was used to evaluate faithfulness. The complete list of 27 benchmark queries is provided in Supplementary Material (Section S2). Each model was

independently evaluated using the same retrieval context and prompt format. To ensure factual grounding and attribution, we guided the generator model with a custom, instruction-based prompt. This prompt commands the model to act as an expert, answer using only the provided context, and provide inline citations for every factual claim. The verbatim prompt template is provided in Supplementary Material (Section S1).

This formatted prompt is processed by each LLM to generate natural language responses tailored to the specific agricultural query. The system's performance is assessed using a comprehensive multi-metric evaluation approach that encompasses both retrieval effectiveness and generation quality. To ensure fairness, retrieval relevance was held constant by using a shared embedding model and retrieval pipeline, while ground truth references enabled objective evaluation using both lexical and semantic metrics. Retrieval evaluation employs Precision@K to measure the proportion of relevant documents in top-k results, Recall@K to evaluate the system's ability to retrieve all relevant documents, MRR to assess ranking quality, and NDCG to consider both relevance and ranking position.

Generation quality and system performance are assessed across multiple dimensions. Lexical overlap is measured using BLEU for n-gram overlap, and ROUGE-1, ROUGE-2, ROUGE-L for lexical overlap and longest common subsequence assessment. Semantic quality is evaluated using BERTScore precision, recall, and F1, along with semantic similarity to capture meaning beyond surface-level word overlap. Source attribution is measured using an attribution score to determine the reliability of referenced knowledge. Efficiency metrics include retrieval time, generation time, and total time to assess the practical feasibility of the system. Ground truth responses are established for a subset of queries to enable quantitative evaluation, with each of the four language models evaluated against the same benchmark dataset to allow direct performance comparison across different model architectures and capabilities. The complete RAG pipeline operates through a sequential process of query embedding, similarity search, context retrieval, prompt formatting, LLM processing, and response generation. This modular design facilitates easy model comparison by maintaining consistent preprocessing, retrieval, and evaluation components while varying only the generation model, providing robust insights into the relative strengths and limitations of different LLMs for agricultural domain applications.

Evaluation metrics

To evaluate the performance of the proposed RAG-based system for agricultural guidance, a comprehensive set of metrics was employed to assess both the quality of generated responses and the effectiveness of information retrieval. Generation quality was measured using BLEU for n-gram overlap, ROUGE-1, ROUGE-2, ROUGE-L for lexical overlap and longest common subsequence assessment, and BERTScore precision, Recall, and F1 to capture semantic similarity. Retrieval effectiveness was analyzed using Precision@K, Recall@K, MRR, and NDCG. Additionally, source reliability was evaluated through an attribution score, system efficiency was measured *via* retrieval time, generation time, and total time. Faithfulness was calculated manually. Experiments were conducted using four prominent LLMs: Llama 3.1, Mistral, Phi 3, and Qwen 2.5 to systematically compare their capabilities within the RAG framework. Together, these metrics offer a robust framework to examine the system's ability to generate contextually relevant, accurate, and ranked responses, thereby validating its utility in delivering timely and reliable agricultural information to users (Shejuti *et al.*, 2025). The performance of all four LLMs was evaluated using key retrieval metrics such as Precision@K, Recall@K, MRR and NDCG. Since the retrieval pipeline including embeddings, vector database (ChromaDB), and retrieval strategy is fixed and shared across models, the retrieval quality is independent of the specific LLM used for response generation.

Precision@K measures the proportion of relevant items among the top-K retrieved documents and is computed as:

$$\text{Precision@K} = \frac{\text{Number of relevant items in top K recommendations}}{\text{K}} \qquad \text{(eq. 3)}$$

Recall@K evaluates how many relevant documents were retrieved among all possible relevant ones, given by:

$$\text{Recall@K} = \frac{\text{Number of relevant items in top K recommendations}}{\text{Total number of relevant items}} \qquad \text{(eq. 4)}$$

MRR assesses the rank position of the first relevant document in the retrieved list and is expressed as:

$$\mathrm{MRR} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\frac{1}{\mathrm{rank}_i} \quad \text{(eq. 5)}$$

NDCG captures both the relevance and the ranking of retrieved documents. It is calculated as:

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}} \quad \text{(eq. 6)}$$

Where:

$$\mathrm{DCG} = \sum_{\mathrm{i}=0}^{\mathrm{n}}\frac{\mathrm{lst(i)}}{\log_2(\mathrm{i}+1)} \quad \text{(eq. 7)}$$

The next set of metrics provide complementary insights into the quality of generated responses, namely BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore_F1, BERTScore Precision and BERTScore Recall. Starting with BLEU, this metric evaluates the overlap of n-grams between generated and reference texts, offering a quantitative measure of syntactic accuracy (K S *et al.*, 2023). It is particularly useful for assessing word-level and phrase-level precision in generated responses.

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}}\log \mathrm{p_n} \quad \text{(eq. 8)}$$

Where:

$$\mathrm{p_n} = \frac{\text{Number of n gram tokens in system and reference translations}}{\text{Number of n gram tokens in system translation}} \quad \text{(eq. 9)}$$

and

The brevity penalty (BP) = exp(1 - r/c), where c is the length of the hypothesis translation (in tokens), r is the length of the closest reference translation.

In ROUGE-1, which measures the overlap of unigrams (individual words) between the machine-generated response and the reference response (Zhang and Zhang, 2025). It is primarily a recall-based metric but can also be reported with precision and F1-score. ROUGE-1 Precision measures how many unigrams in the generated text are also in the reference text. The F1-score is the harmonic mean of ROUGE-1 Precision and ROUGE-1 Recall.

$$\text{ROUGE} - 1\ (\text{Recall}) = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in reference}} \quad \text{(eq. 10)}$$

$$\text{ROUGE} - 1\ (\text{Precision}) = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in candidate (generated) text}} \quad \text{(eq. 11)}$$

$$\text{ROUGE} - 1\ (\text{F1 Score}) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(eq. 12)}$$

In ROUGE-2 measures the overlap of bigrams (two-word sequences) between the generated text and reference text, providing insight into the system's ability to capture short phrase-level dependencies. The ROUGE-2 score is computed as:

$$\text{ROUGE} - 2\ (\text{Recall}) = \frac{\text{Number of overlapping bigrams}}{\text{Total bigrams in reference}} \quad \text{(eq. 13)}$$

$$\text{ROUGE} - 2\ (\text{Precision}) = \frac{\text{Number of overlapping bigrams}}{\text{Total bigrams in candidate (generated) text}} \quad \text{(eq. 14)}$$

$$\text{ROUGE} - 2\ (\text{F1 Score}) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(eq. 15)}$$

In ROUGE-L, which evaluates the longest common subsequence (LCS) between a generated text and a reference text (P *et al.*, 2025).

LCS(X, Y): Length of the longest common subsequence between two sequences X and Y

$$\text{ROUGE} - \text{L}\{\text{Precision}\} = \frac{\text{LCS}(\text{X}, \text{Y})}{\text{length of candidate (X)}} \quad \text{(eq. 16)}$$

$$\text{ROUGE} - \text{L}\{\text{Recall}\} = \frac{\text{LCS}(\text{X}, \text{Y})}{\text{length of reference (Y)}} \quad \text{(eq. 17)}$$

$$\text{ROUGE} - \text{L}\{\text{F1 Score}\} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(eq. 18)}$$

BERTScore leverages contextual embeddings from pre-trained models such as BERT to evaluate the semantic similarity between generated and reference sentences (Kim *et al*., 2024). Unlike traditional n-gram overlap metrics like BLEU and ROUGE, BERTScore captures meaning-based alignment. The scoring process involves matching each token in the candidate sentence ($\hat{x}$) to the most similar token in the reference sentence (x) to compute precision, and vice versa to compute recall (Irican *et al*., 2024). A greedy matching strategy is employed to maximize similarity between token pairs across sentences. Finally, precision and recall are combined to calculate the F1 score, providing a balanced metric that reflects both semantic relevance and coverage.

BERTScore configuration: we utilize the roberta-large model as the contextual embedding backbone. IDF weighting is disabled (idf=False) to ensure equal weighting of all tokens, and baseline rescaling is not applied (rescale_with_baseline=False).

$$\text{R}_{\text{BERT}} = \frac{1}{|\text{x}|} \sum_{\text{x}_\text{i} \in \text{x}} \max_{\widehat{\text{x}_\text{j}} \in \hat{\text{x}}} \times_{\mathbf{i}}{}^{\text{T}} \widehat{\text{x}_\text{j}} \quad \text{(eq. 19)}$$

$$\text{P}_{\text{BERT}} = \frac{1}{|\hat{\text{x}}|} \sum_{\widehat{\text{x}_\text{j}} \in \hat{\text{x}}} \max_{\text{x}_i \in \text{x}} \times_{\mathbf{i}}{}^{\text{T}} \widehat{\text{x}_\text{j}} \quad \text{(eq. 20)}$$

$$\text{F}_{\text{BERT}} = 2\,\frac{\text{P}_{\text{BERT}} \times \text{R}_{\text{BERT}}}{\text{P}_{\text{BERT}} + \text{R}_{\text{BERT}}} \quad \text{(eq. 21)}$$

Semantic similarity measures the cosine similarity between candidate and reference embeddings:

$$\text{Similarity} = \frac{\mathbf{v}_{\text{candidate}} \cdot \mathbf{v}_{\text{reference}}}{\|\mathbf{v}_{\text{candidate}}\| \|\mathbf{v}_{\text{reference}}\|} \quad \text{(eq. 22)}$$

Source attribution is captured using the attribution score, defined as the ratio of correctly attributed facts to the total facts in the generated response:

$$\text{Attribution} = \frac{\text{Number of cited sources}}{\text{Total retrieved sources}} \quad \text{(eq. 23)}$$

To assess whether generated responses remained factually consistent with retrieved PoP documents, we conducted manual faithfulness evaluation on a representative sample of 20 responses (5 queries × 4 models). For each response, discrete factual claims were extracted, including fertilizer dosages, seasonal timing, soil requirements, and procedural recommendations. Each claim was systematically verified against the corresponding retrieved document chunks used to generate that response.

Claims were categorized as: i) supported: directly verifiable in retrieved documents with exact or paraphrased matches, ii) partially supported: mostly correct with minor discrepancies in completeness or phrasing, or iii) not supported: contradicting or absent from retrieved documents. Faithfulness was computed as the proportion of supported claims among all verifiable claims (supported + not supported), expressed as a percentage. Generic or non-verifiable statements (e.g., "proper care is needed") were excluded from scoring.

The evaluation was conducted through systematic claim extraction and source document comparison, with ambiguous cases resolved through discussion among all authors. To ensure objectivity, responses were evaluated in random order without advance knowledge of which model generated each response.

Finally, efficiency metrics evaluate system performance in terms of latency. Retrieval time measures the duration to fetch relevant documents ($t_{retrieval\_end} - t_{retrieval\_start}$), generation time captures the duration to produce the response ($t_{generation\_end} - t_{generation\_start}$), and total time is the sum of both.

$$\text{Total Time} = \text{Retrieval Time} + \text{Generation Time} \quad \text{(eq. 24)}$$

Together, these metrics provide a comprehensive framework for assessing both the quality and efficiency of generated responses in retrieval-augmented systems.

Results

This section details the system's performance across retrieval, generation, efficiency metrics, attribution, semantic similarity and faithfulness. For conciseness, aggregate results are presented in the main paper (Tables 3 to 6). The detailed, per-query results for all models and metrics are available in Supplementary Material (Section S3).

Retrieval performance

The retrieval component of the RAG-based agricultural advisory system was evaluated across all four language models using a test set of 27 agricultural queries spanning the five evaluated crops. The system demonstrated strong retrieval effectiveness, achieving a mean Precision@K of 0.6173 (95% CI: 0.5034 - 0.7312) and Recall@K of 0.8704 (95% CI: 0.7664-0.9743). The mean reciprocal rank (MRR) was 0.8889 (95% CI: 0.7792-0.9986), indicating that relevant documents were typically ranked within the top positions. The normalized discounted cumulative gain (NDCG) of 0.8985 (95% CI: 0.8038-0.9932) further confirmed effective ranking quality (Table 3). Figure 2 illustrates the consistency of retrieval metrics across all four LLMs, while Figure 3 shows performance variation across different crops, with cotton and sweet potato achieving the highest retrieval scores.

Generation performance across LLMs

Generation quality was assessed using both lexical overlap metrics (BLEU, ROUGE) and semantic similarity measures (BERTScore). Performance varied considerably across the four evaluated LLMs, as shown in Figure 2.

Llama 3.1 achieved a BLEU score of 0.0454 (95% CI: 0.0185-0.0724), with ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.2913 (95% CI: 0.2384-0.3442), 0.1367 (95% CI: 0.0846-0.1888), and 0.2226 (95% CI: 0.1738-0.2714), respectively (Table 4). BERTScore metrics showed precision of 0.8211 (95% CI: 0.8111-0.8311), recall of 0.8914 (95% CI: 0.8790-0.9038), and F1 of 0.8546 (95% CI: 0.8448-0.8643).

Mistral demonstrated improved performance over Llama 3.1, with BLEU = 0.0570 (95% CI: 0.0354-0.0786), ROUGE-1 = 0.3737 (95% CI: 0.3233-0.4242), ROUGE-2 = 0.1872 (95% CI: 0.1238-0.2505), and ROUGE-L = 0.2916 (95% CI: 0.2347-0.3484) (Table 4). BERTScore precision, recall, and F1 were 0.8473 (95% CI: 0.8366-0.8579), 0.8976 (95% CI: 0.8855-0.9097), and 0.8715 (95% CI: 0.8618-0.8812).

Phi-3 exhibited the lowest generation quality among the evaluated models, recording BLEU = 0.0227 (95% CI: 0.0133-0.0322), ROUGE-1 = 0.2540 (95% CI: 0.2124-0.2955), ROUGE-2 = 0.0886 (95% CI: 0.0617-0.1155), and ROUGE-L = 0.1815 (95% CI: 0.1476-0.2154) (Table 4). BERTScore values were precision = 0.8187 (95% CI: 0.8088-0.8285), recall = 0.8894 (95% CI: 0.8788-0.9000), and F1 = 0.8523 (95% CI: 0.8441-0.8606).

Qwen 2.5 achieved the highest performance across most metrics, with BLEU = 0.0824 (95% CI: 0.0219-0.1429), ROUGE-1 = 0.3712 (95% CI: 0.3085-0.4339), ROUGE-2 = 0.1910 (95% CI: 0.1170-0.2651), and ROUGE-L = 0.2899 (95% CI: 0.2210-0.3589) (Table 4). Its BERTScore precision, recall, and F1 scores were 0.8435 (95% CI: 0.8313-0.8558), 0.9034 (95% CI: 0.8893-0.9174), and 0.8721 (95% CI: 0.8609-0.8834), respectively. Qwen 2.5 outperformed Phi-3 in BLEU score and ROUGE-1 score.

Figure 3 reveals notable performance variation across crops, with maize and cotton generally yielding higher generation quality scores compared to ragi and sweet potato across most models.

Response time analysis

Figure 4 presents the time performance characteristics of the system. Retrieval times remained relatively consistent across LLMs, with Llama 3.1 recording 1.0548 seconds (95% CI: 1.0407 - 1.0690), Mistral 1.1089 seconds (95% CI: 1.0476-1.1702), Phi-3 1.0589 s (95% CI: 1.0500-1.0678), and Qwen 2.5 1.2174 seconds (95% CI: 1.1397-1.2951) (Table 5). Llama 3.1 and Phi-3 demonstrated the fastest retrieval performance at approximately 1.05-1.06 s.

Generation times varied more substantially across models. Phi-3 demonstrated the most efficient generation at 9.4881 s (95% CI: 6.6947-12.2816), followed by Mistral at 12.4030 s (95% CI: 10.3250-14.4809), Qwen 2.5 at 14.2463 s (95% CI: 12.1128-16.3798), and Llama 3.1 at 16.8863 s (95% CI: 14.5324-19.2402).

Total response times ranged from 10.5470 s (95% CI: 7.7539-13.3402) for Phi-3 to 17.9411 s (95% CI: 15.5842-20.2980) for Llama 3.1, with Mistral and Qwen 2.5 recording 13.5119 s (95% CI: 11.4061-15.6176) and 15.4637 s (95% CI: 13.2951-17.6323), respectively. These response times indicate acceptable latency for practical agricultural advisory applications. Across crops, the system maintained consistent performance with total response times falling within similar ranges, demonstrating scalability across different agricultural domains.

Attribution and semantic similarity

Attribution scores, measuring the system's ability to ground responses in retrieved documents, remained consistently high across all models. Phi-3 achieved the highest attribution score at 0.6914 (95% CI: 0.5951-0.7876), followed by Llama 3.1 at 0.6420 (95% CI: 0.5457-0.7382), Qwen 2.5 at 0.6296 (95% CI: 0.5306-0.7287), and Mistral at 0.5802 (95% CI: 0.4795-0.6810), as shown in Table 5. Across crops, attribution scores ranged from 0.58 to 0.72, indicating reliable source-grounding behavior across different agricultural domains.

Semantic similarity scores demonstrated strong consistency across LLMs (Figure 6). Qwen 2.5 achieved the highest score at 0.7613 (95% CI: 0.6965-0.8260), followed by Mistral at 0.7456 (95% CI: 0.6919-0.7992), Llama 3.1 at 0.7328 (95% CI: 0.6739-0.7918), and Phi-3 at 0.7325 (95% CI: 0.6702-0.7949). Across crops, semantic similarity ranged from 0.67 to 0.81, with cotton and sweet potato showing the highest semantic alignment between generated and reference responses.

Faithfulness and factual accuracy

Manual evaluation of 20 representative responses revealed high factual consistency across all evaluated models (Table 6). A total of 157 discrete factual claims were extracted and verified against source documents.

Model-specific faithfulness scores demonstrated strong performance across all LLMs: Mistral achieved the highest faithfulness at 100.0% (95% CI: 100.0-100.0%), followed by Qwen 2.5 at 94.5% (95% CI: 85.08-100.0%), Llama 3.1 at 90.6% (95% CI: 77.8-100.0%), and Phi-3 at 91.6% (95% CI: 80.72-100.0%). The consistency of high

faithfulness scores across models confirms the effectiveness of the RAG architecture in grounding generated responses in authoritative agricultural knowledge.

Analysis by claim type revealed that soil and climate requirements achieved highest faithfulness, followed by numerical dosages such as fertilizer rates and spacing, timing recommendations, and procedural steps. This pattern suggests that models excel at extracting and reproducing structured factual information, with minor challenges in synthesizing multi-step procedures.

The most common faithfulness issues were: i) minor omissions of alternative options when multiple valid approaches exist (2 instances) for example, mentioning only one planting season when documents specify two options; ii) logical interpretation errors, such as misreading "or" as "and" when describing seasonal alternatives (1 instance); and iii) source confusion when multiple documents contained similar but contextually distinct recommendations (2 instances). Notably, formatting and phrasing ambiguities accounted for most partially supported claims, rather than substantive factual errors.

Collectively, these findings suggest that while lexical overlap metrics such as BLEU remain modest reflecting the variability of natural language generation, semantic-oriented measures such as ROUGE and BERTScore demonstrate strong contextual and semantic fidelity. Mistral and Qwen 2.5 consistently delivered superior overall performance across both lexical and semantic dimensions, underscoring the importance of model selection and domain-specific fine-tuning for optimizing RAG-based agricultural advisory systems. Although the system demonstrates robust performance across five evaluated crops (maize, ragi, sweet potato, cotton, and groundnut) using PoP documents from Indian extension sources, its generalization to other crops, regions, or updated PoPs remains untested.

Discussion

Interpretation of retrieval performance

The high Recall@K of 0.8704 indicates that the system successfully identifies relevant agricultural information for the vast majority of queries, which is critical for ensuring farmers receive comprehensive guidance. The MRR of 0.8889 suggests that relevant documents are typically positioned within the top two results, reducing the need for users to sift through multiple irrelevant entries. This retrieval effectiveness likely stems from

the domain-specific nature of the PoP documents and the effectiveness of the embedding-based retrieval mechanism (384-dimensional embeddings with Top-K=3) employed in the RAG architecture.

However, the moderate Precision@K (0.6173) suggests that approximately 38% of retrieved documents may not be directly relevant to the query. This could indicate either overgeneralization in the retrieval mechanism or ambiguity in agricultural queries that legitimately connect to multiple topics. The consistency of retrieval metrics across all four LLMs suggests that retrieval quality is primarily determined by the embedding and retrieval strategy rather than the downstream language model, which is expected given the shared retrieval architecture (Figure 2).

The crop-level variation observed in Figure 3, with cotton and sweet potato achieving higher retrieval scores, may reflect differences in document structure, terminology consistency, or query complexity across crops. Future refinement of query understanding or the implementation of re-ranking mechanisms may improve precision without sacrificing recall.

Analysis of generation quality

The modest BLEU scores (ranging from 0.02 to 0.08) across all models are consistent with performance patterns observed in other open-ended generation tasks, particularly in domains requiring specialized knowledge. Unlike machine translation tasks where BLEU scores above 0.3 are common, agricultural advisory involves substantial paraphrasing and contextual adaptation of retrieved information, which naturally results in lower lexical overlap with reference responses. Given the sample size of 27 queries, these BLEU scores reflect typical variation in agricultural advisory phrasing rather than model deficiencies.

In contrast, the higher ROUGE scores (ROUGE-1 ranging from 0.25 to 0.37) and particularly the strong BERTScore F1 scores (all above 0.85) indicate that models successfully capture semantic content despite surface-level variation in phrasing. The consistently high BERTScore Recall (>0.89 across all models) suggests that generated responses comprehensively cover the information present in reference answers, which is more relevant to practical utility than exact word matching. The semantic similarity

analysis corroborates these findings, showing consistent alignment (0.73-0.76) between generated and reference responses across all models (Figure 6).

Faithfulness and safety considerations

While the system demonstrated high faithfulness, several limitations warrant acknowledgment. First, faithfulness evaluation was conducted on only 20 responses (18.5% of the 108 total model-query combinations), which may not capture all potential error modes. However, the stratified sampling approach covering all models, all crops, and diverse query types provides reasonable confidence in the generalizability of findings.

Faithfulness assessment focused on factual claim verification rather than contextual appropriateness. A response may be factually accurate according to retrieved documents yet inappropriate for specific field conditions not captured in the query (e.g., region-specific pest pressures, soil amendments for extreme pH). Production deployment would require additional context-gathering mechanisms to ensure recommendations match farmer circumstances.

While no high-risk errors were observed in the evaluated sample, the statistical possibility of rare but severe errors cannot be eliminated. Agricultural advisory systems deployed in practice should include: i) prominent disclaimers that AI-generated advice requires validation by local extension officers, particularly for pest management and chemical applications; ii) dosage verification mechanisms that flag recommendations outside normal ranges; and iii) regular human review of system outputs to identify and correct emerging error patterns.

Model comparison and performance patterns

Qwen 2.5 and Mistral emerged as the top-performing models across both lexical and semantic metrics. Qwen 2.5 achieved the highest BLEU (0.0824), BERTScore recall (0.9034), and semantic similarity (0.7613), while Mistral demonstrated strong balanced performance with the highest ROUGE-1 (0.3737) and competitive BERTScore F1 (0.8715, semantic similarity 0.7456). These models appear better suited to agricultural domain language, possibly due to their training data composition or architectural refinements that enhance instruction-following and factual grounding.

The underperformance of Phi-3 across generation quality metrics (BLEU: 0.0227, ROUGE-1: 0.2540, BERTScore F1: 0.8523) presents an interesting trade-off with its computational efficiency. Despite achieving the lowest generation quality scores, Phi-3 demonstrated the fastest total response time (10.55 s) faster than Qwen 2.5 (15.46 s) and faster than Llama 3.1 (17.94 s). This efficiency stems primarily from its generation speed (9.49 seconds), which is faster than its nearest competitor. Notably, Phi-3 also achieved the highest attribution score (0.6914), suggesting strong source-grounding behavior despite lower overall generation quality. This indicates that Phi-3's limitations lie primarily in language generation fluency and completeness rather than in its ability to utilize retrieved information appropriately.

Llama 3.1 exhibited the slowest total response time (17.94 s) despite moderate generation quality (ROUGE-1: 0.2913, BERTScore F1: 0.8546), suggesting architectural inefficiencies that make it less attractive for deployment. Its generation time (16.89 s) was longer than Phi-3's and longer than Qwen 2.5's, without commensurate improvements in output quality.

Retrieval times were remarkably consistent across models (1.05-1.22 seconds). This consistency confirms that retrieval performance is determined by the shared embedding and vector search architecture rather than the downstream language model, validating the design decision to separate retrieval and generation components in the RAG architecture.

Limitations and scope

Several limitations constrain the generalizability of these findings. First, the evaluation was conducted on a relatively small test set of 27 queries covering five crops (Maize, ragi, sweet potato, cotton, and groundnut) using PoP documents from Indian agricultural extension sources. While this sample provides initial evidence of system effectiveness, larger-scale evaluation across more diverse queries, additional crops, alternative agricultural systems (e.g., organic farming, precision agriculture), and PoPs from different geographical regions would strengthen confidence in the findings.

All evaluation was conducted in English, whereas many Indian farmers prefer regional languages such as Hindi, Tamil, Malayalam, or Kannada. The system's effectiveness in

multilingual scenarios either through translation or native multilingual models, requires investigation.

The static nature of the knowledge base may limit applicability as agricultural recommendations evolve with climate change, new pest varieties, or updated research findings. The system currently lacks mechanisms to flag outdated information or integrate real-time updates.

The observed crop-level performance variation suggests that system effectiveness may depend on document quality and domain characteristics. This variation has not been systematically investigated to understand whether it reflects inherent crop complexity, data quality issues, or other factors.

Conclusions

This study introduced a RAG-based framework for knowledge-grounded agricultural advisory systems, designed to deliver timely and accurate guidance to farmers. By leveraging a structured, domain-specific dataset and integrating it with advanced natural language processing techniques, the system bridges the gap between farmer queries and authoritative agricultural knowledge. The modular architecture facilitates efficient document retrieval and context-aware response generation, ensuring clarity and relevance in outputs. A comprehensive evaluation was conducted using linguistic metrics (BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore precision, BERTScore recall, BERTScore F1, semantic similarity), source attribution metrics (attribution score), retrieval-based and efficiency metrics (Precision@K, Recall@K, MRR, NDCG, Retrieval Time, Generation Time, Total Time) and Faithfulness to compare the performance of four large language models: Llama3.1, Mistral, Phi3, and Qwen2.5. The evaluation across five crop domains: maize, ragi, sweet potato, cotton, and groundnut revealed performance variability, highlighting the importance of domain-specific tuning and dataset enrichment. This research underscores the practical viability of RAG based systems in real-world agricultural settings, offering scalable, intelligent tools to support informed decision-making.

Future enhancements could significantly expand the system's capabilities and real-world impact. The framework could evolve into a more comprehensive precision crop management tool by integrating real-time data sources such as local weather forecasts,

soil conditions, and historical farming records. A promising direction includes developing intelligent pest and disease management features, where the system could process multimodal inputs, allowing farmers to upload images of affected crops for instant diagnosis and treatment advice. Furthermore, the system could provide climate-smart sustainability guidance by incorporating climate models and research on adaptive farming practices to help farmers maintain productivity amid climate change. To ensure broad accessibility, developing multilingual support and deploying the framework on mobile platforms will be crucial for empowering smallholder farmers across diverse agro-ecological zones and truly democratizing agricultural expertise.

Online Supplementary Material
Table S1. Benchmark queries for evaluation of the RAG-based agricultural advisory system.
Table S2. Query-wise retrieval and generation times for all LLMs.
Table S3. Performance metrics for LLM responses across crops and queries.
Table S4. Source attribution and retrieval effectiveness metrics for LLM responses across crops and queries.
Table S5. Semantic similarity and source attribution scores for LLM responses across crops and queries.

References

A, S., Krishnan, A.G., V, G. 2024. Leveraging technology to empower millet farmers a retrieval-augmented generation approach with large language models. Proc. 5th IEEE Global Conf. Advancement in Technology (GCAT), Bangalore; pp. 1-7.

Acharya, D.B., Kuppan, K., Divya, B. 2025. Agentic AI: autonomous intelligence for complex goals - A comprehensive survey. IEEE Access 13:18912-18936.

Arslan, M., Ghanema, H., Munawarb, S., Cruza, C. 2024. A survey on RAG with LLMs. Procedia Comput. Sci. 246:3781-3790.

Balpande, M., Mahajan, K., Bhandarkar, J., Borse, G., Badjat, S. 2024. AI powered agriculture optimization chatbot using RAG and GenAI. Proc. IEEE Silchar Subsection Conf. (SILCON 2024), Agartala; pp. 1-6.

Dhanabalan, T., Sathish, A. 2018. Transforming Indian industries through artificial intelligence and robotics in industry 4.0. Int. J. Mech. Eng. Technol. 9:835-845.

Government of Kerala, Directorate of Economics and Statistics, EARAS Division. 2020. Agricultural Statistics 2018-19. Available from: https://ecostat.kerala.gov.in/storage/publications/239.pdf

Hu, R., Liu, S., Qi, P., Liu, J., Li, F. 2025. ICCA-RAG: intelligent customs clearance assistant using retrieval-augmented generation (RAG). IEEE Access 13:39711-39726.

Irican, B.B., Sivri, M., Kokach, V., Kocacinar, B., Akbulut, F.P. 2024. QBot: domain-specific chatbots with retrieval-augmented generation and vector embedding for

complex documentation queries. Proc. Innovations in Intelligent Systems and Applications Conf. (ASYU), Ankara; pp. 1-6.

K S, N.P., S, S., T N, T., Yuvraaj, Y., D A, V. 2023. Conversational chatbot builder – smarter virtual assistance with domain specific AI. Proc. 4th Int. Conf. Emerging Technology (INCET), Belgaum; pp. 1-4.

Kar, R., Haldar, R. 2016. Applying chatbots to the internet of things: opportunities and architectural elements. arXiv:1611.03799.

Khanifar, J. 2025. Evaluating AI-generated responses from different chatbots to soil science-related questions. Soil Adv. 3:100034.

Kim, M., Kim, D., Park, Y., Jeong, D. 2024. Development of an expert chatbot for digital forensics using RAG model implementation. Proc. Int. Conf. Platform Technology and Service (PlatCon), Jeju; pp. 182-187.

Legashev, L., Shukhman, A., Badikov, V., Kurynov, V. 2025. Using large language models for goal-oriented dialogue systems. Appl. Sci. 15:4687.

Mathebula, M., Modupe, A., Marivate, V. 2024. Fine-tuning retrieval-augmented generation with an auto-regressive language model for sentiment analysis in financial reviews. Appl. Sci. 14:10782.

Meng, W., Li, Y., Chen, L., Dong, Z. 2025. Using the retrieval-augmented generation to improve the question-answering system in human health risk assessment: the development and application. Electronics 14:386.

P, K., M, H., Hayagreevan, V. 2025. Development of interactive assistance for academic preparation using large language models. Proc. Int. Conf. Computational, Communication and Information Technology (ICCCIT), Indore; pp. 265-269.

Saha, B., Saha, U., Zubair Malik, M. 2024. QuIM-RAG: advancing retrieval-augmented generation with inverted question matching for enhanced QA performance. IEEE Access 12:185401-185410.

V, N., G. A, S., S, G., M, K., A, M., S, T. 2024. AgriBot: An integrated chatbot platform for precision agriculture and farmer support using deep learning techniques. Proc. Int. Conf. Power, Energy, Control and Transmission Systems (ICPECTS), Chennai; pp. 1-6.

Wilkho, R.S., Chang, S., Gharaibeh, N.G. 2023. FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. Adv. Eng. Inform. 59:102293.

Zhou, B., Zou, L., Mostafavi, A., Lin, A., Yang, M., Gharaibeh, N., et al. 2022. VictimFinder: harvesting rescue requests in disaster response from social media with BERT. Comput. Environ. Urban Syst. 95:101824.

Xiong, J., Pan, L., Liu, Y., Zhu, L., Zhang, L., Tan, S. 2025. Enhancing plant protection knowledge with large language models: a fine-tuned question-answering system using LoRA. Appl. Sci. 15:3850.

Yin, S., Xi, Y., Zhang, X., Sun, C., Mao, Q. 2025. Foundation models in agriculture: a comprehensive review. Agriculture 15:847.

Zafarmomen, N., Samadi, V. 2025. Can large language models effectively reason about adverse weather conditions? Environ. Model. Softw. 188:106421.

Zhang, W., Zhang, J. 2025. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics 13:856.

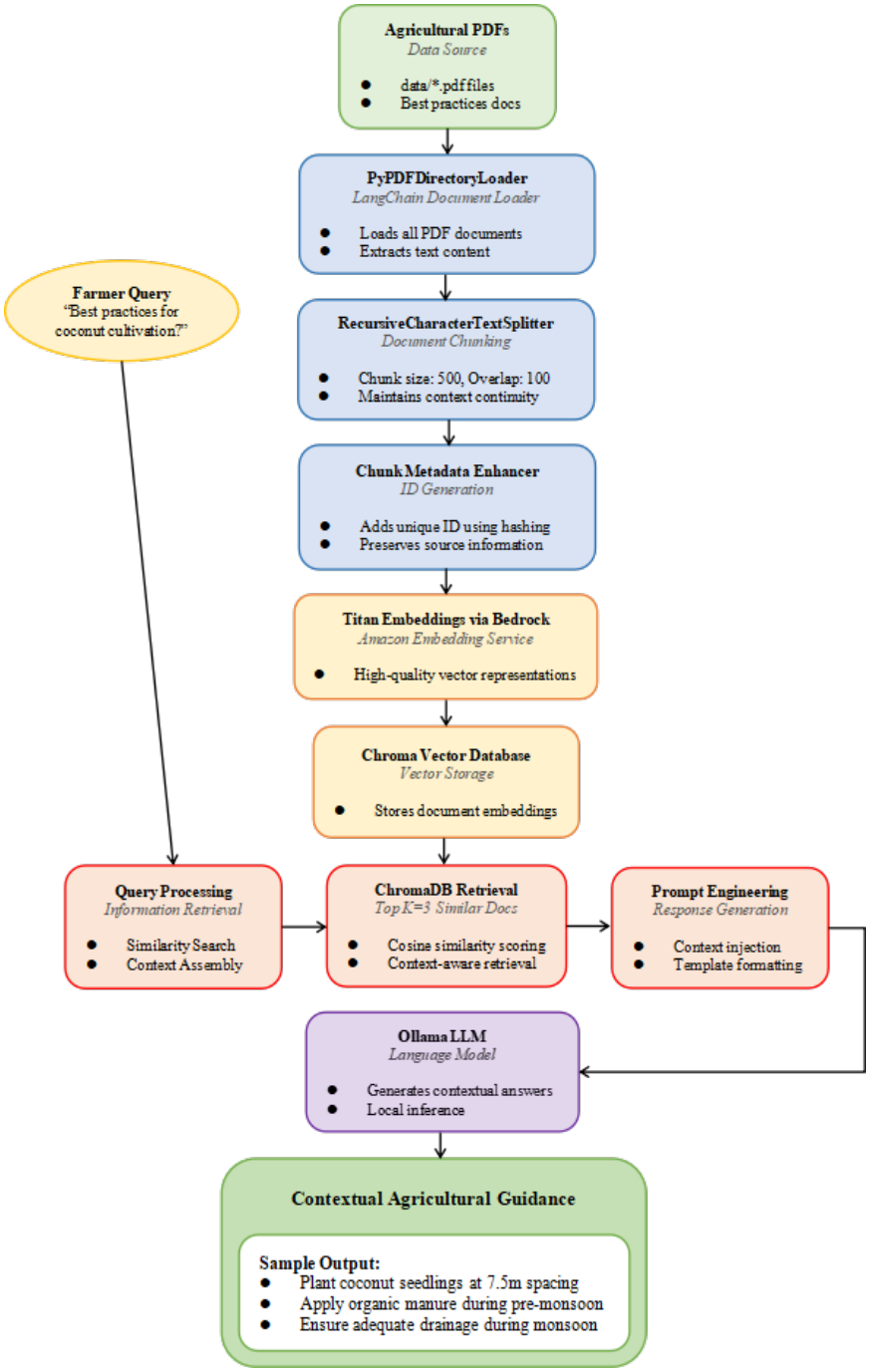

Figure 1. System architecture of the retrieval-augmented generation (RAG) framework for agricultural guidance.

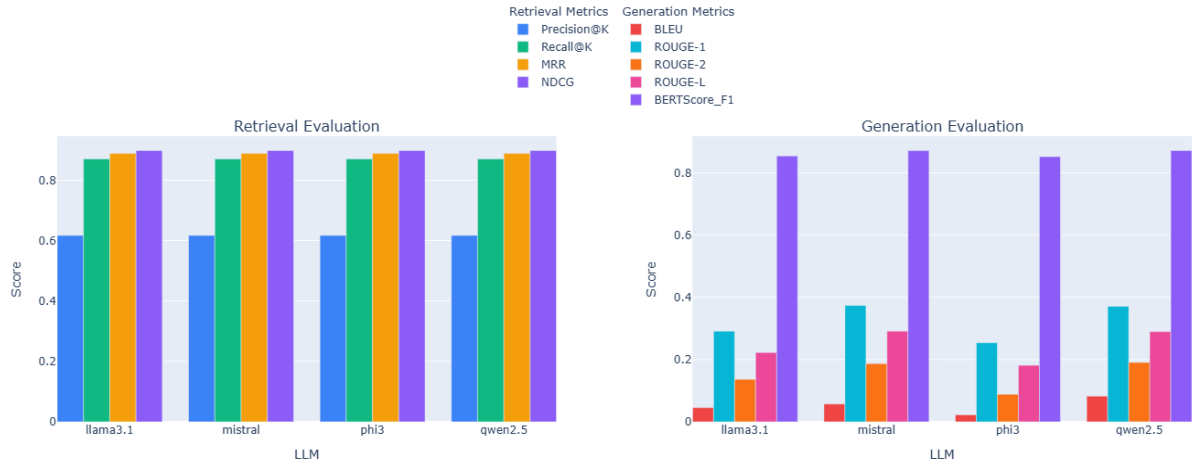

Figure 2. LLM performance based on retrieval and generation metrics.

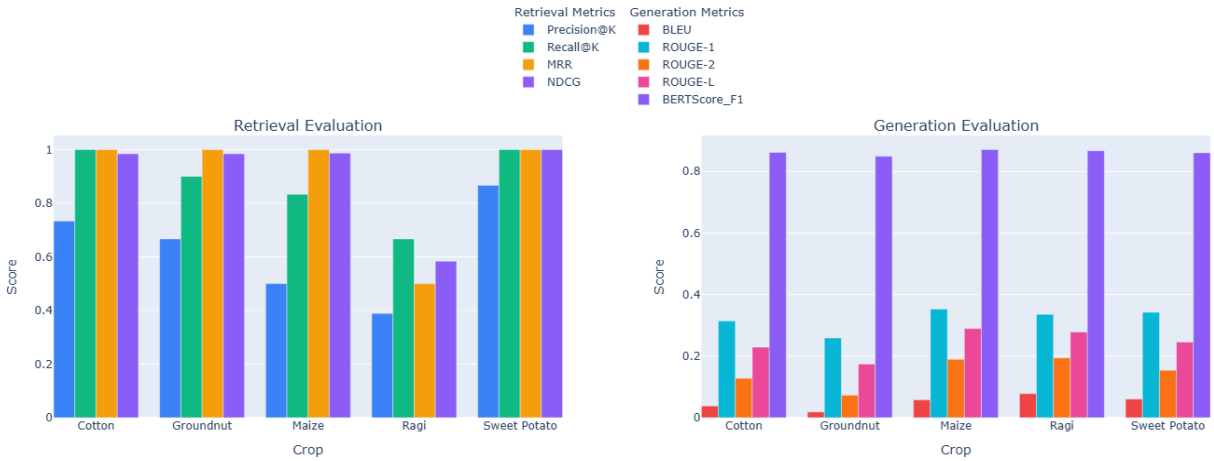

Figure 3. Crop domain performance based on retrieval and generation metrics.

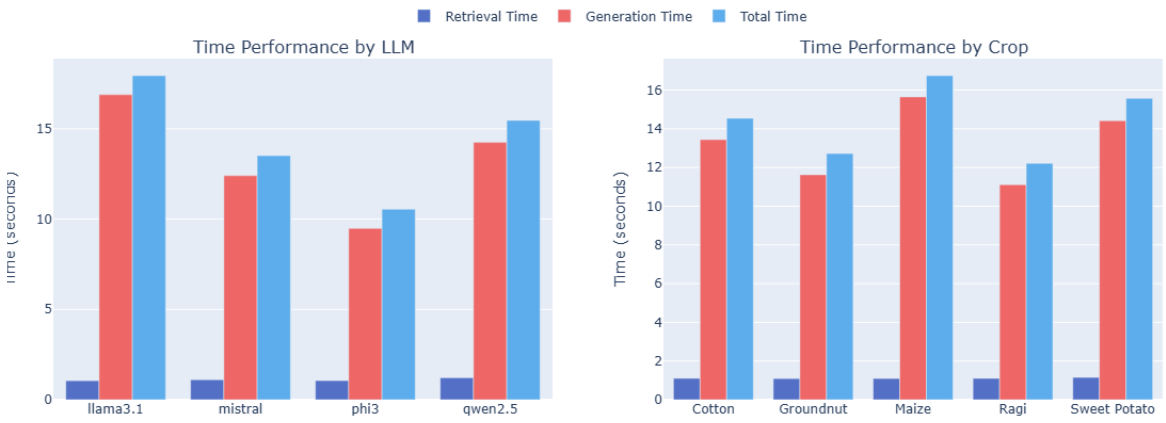

Figure 4. Time performance based on LLMs and crop domain.

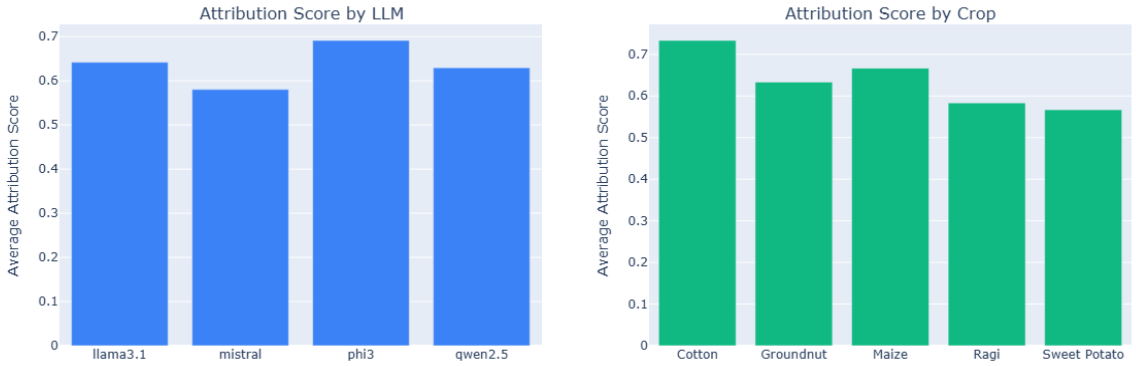

Figure 5. Attribution score based on LLMs and crop domain.

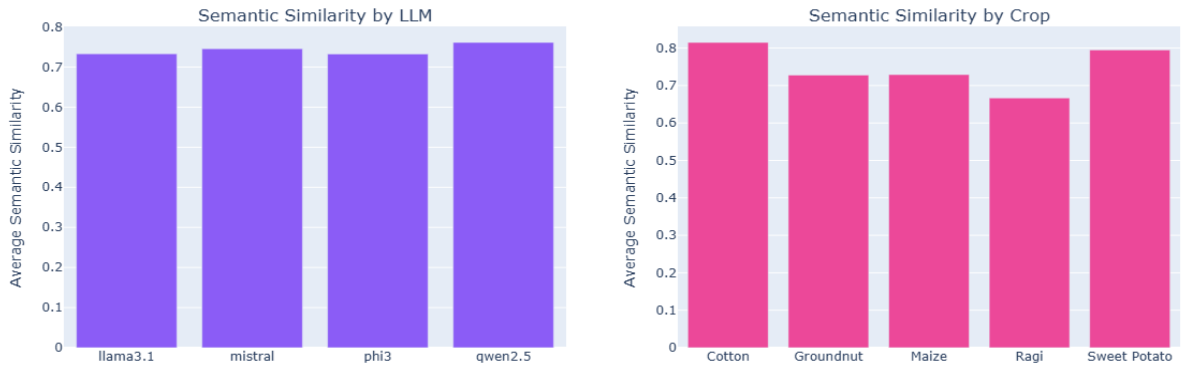

Figure 6. Semantic similarity based on LLMs and crop domain.

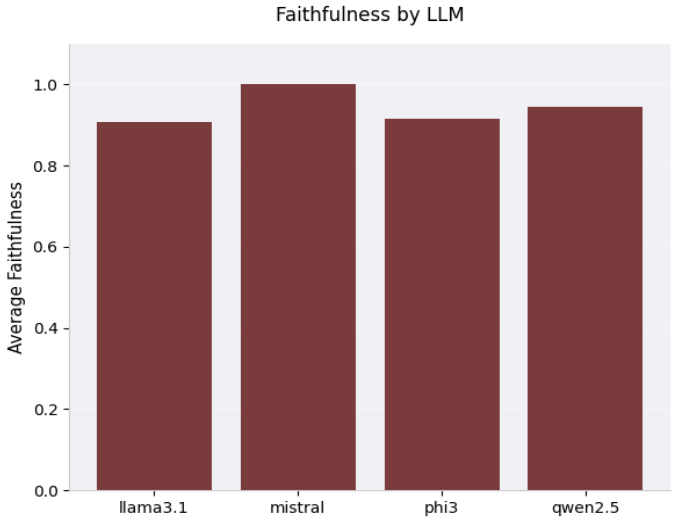

Figure 7. Faithfulness based on LLMs.

Table 1. Area and production statistics for selected crops of Kerala, India (2018-19).

Crop	Area (in hectare)	Production (in metric tons)
Maize	104	144
Ragi	225	271
Sweet potato	210	3060
Cotton	59	90
Groundnut	187	239

Source: Government of Kerala, 2020.

Table 2. Hyperparameters and settings.

Parameter	Value
Chunk size	500 characters
Chunk overlap	100 characters
Embedding dimension	384
Top-K	3 documents
Similarity threshold	0.5
Temperature	0.7
Top_P	0.9
Max tokens	512

Table 3. Uniform performance metrics for all four LLMs.

Metrics	Mean ± [CI_lower, CI_upper]
Precision@K	0.6173 ± [0.5034, 0.7312]
Recall@K	0.8704 ± [0.7664, 0.9743]
MRR	0.8889 ± [0.7792, 0.9986]
NDCG	0.8985 ± [0.8038, 0.9932]

Table 4. Generation performance metrics for all four LLMs.

LLM	Metrics	Mean ± [CI_lower, CI_upper]
Llama3.1	BLEU	0.0454 ± [0.0185, 0.0724]
	ROUGE-1	0.2913 ± [0.2384, 0.3442]
	ROUGE-2	0.1367 ± [0.0846, 0.1888]
	ROUGE-L	0.2226 ± [0.1738, 0.2714]
	BERTScore_P	0.8211 ± [0.8111, 0.8311]
	BERTScore_R	0.8914 ± [0.8790, 0.9038]
	BERTScore_F1	0.8546 ± [0.8448, 0.8643]
Mistral	BLEU	0.0570 ± [0.0354, 0.0786]
	ROUGE-1	0.3737 ± [0.3233, 0.4242]
	ROUGE-2	0.1872 ± [0.1238, 0.2505]
	ROUGE-L	0.2916 ± [0.2347, 0.3484]
	BERTScore_P	0.8473 ± [0.8366, 0.8579]
	BERTScore_R	0.8976 ± [0.8855, 0.9097]
	BERTScore_F1	0.8715 ± [0.8618, 0.8812]
Phi3	BLEU	0.0227 ± [0.0133, 0.0322]
	ROUGE-1	0.2540 ± [0.2124, 0.2955]
	ROUGE-2	0.0886 ± [0.0617, 0.1155]
	ROUGE-L	0.1815 ± [0.1476, 0.2154]
	BERTScore_P	0.8187 ± [0.8088, 0.8285]
	BERTScore_R	0.8894 ± [0.8788, 0.9000]
	BERTScore_F1	0.8523 ± [0.8441, 0.8606]
Qwen2.5	BLEU	0.0824 ± [0.0219, 0.1429]
	ROUGE-1	0.3712 ± [0.3085, 0.4339]
	ROUGE-2	0.1910 ± [0.1170, 0.2651]
	ROUGE-L	0.2899 ± [0.2210, 0.3589]
	BERTScore_P	0.8435 ± [0.8313, 0.8558]
	BERTScore_R	0.9034 ± [0.8893, 0.9174]
	BERTScore_F1	0.8721 ± [0.8609, 0.8834]

Table 5. System performance characteristics for all four LLMs.

LLM	Metrics	Mean ± [CI_lower, CI_upper]
Llama3.1	Semantic similarity	0.7328 ± [0.6739, 0.7918]
	Attribution score	0.6420 ± [0.5457, 0.7382]
	Retrieval time	1.0548 ± [1.0407, 1.0690]
	Generation time	16.8863 ± [14.5324, 19.2402]
	Total time	17.9411 ± [15.5842, 20.2980]
Mistral	Semantic similarity	0.7456 ± [0.6919, 0.7992]
	Attribution score	0.5802 ± [0.4795, 0.6810]
	Retrieval time	1.1089 ± [1.0476, 1.1702]
	Generation time	12.4030 ± [10.3250, 14.4809]
	Total time	13.5119 ± [11.4061, 15.6176]
Phi3	Semantic similarity	0.7325 ± [0.6702, 0.7949]
	Attribution score	0.6914 ± [0.5951, 0.7876]
	Retrieval time	1.0589 ± [1.0500, 1.0678]
	Generation time	9.4881 ± [6.6947, 12.2816]
	Total time	10.5470 ± [7.7539, 13.3402]
Qwen2.5	Semantic similarity	0.7613 ± [0.6965, 0.8260]
	Attribution score	0.6296 ± [0.5306, 0.7287]
	Retrieval time	1.2174 ± [1.1397, 1.2951]
	Generation time	14.2463 ± [12.1128, 16.3798]
	Total time	15.4637 ± [13.2951, 17.6323]

Table 6. Faithfulness evaluation results by model and query.

Query	Llama3.1	Mistral	Phi3	Qwen2.5
What are the main planting seasons for maize? (planting seasons)	0.83	1	1	1
What is the recommended manuring and fertilizer application schedule and dosage for sweet potato cultivation? (fertilizer)	0.93	1	0.86	1
What are the optimal environmental and soil conditions required for ragi cultivation? (soil/climate)	1	1	1	1
What is the recommended manuring and fertilizer application schedule and dosage for cotton cultivation? (fertilizer)	0.77	1	0.92	0.85
What are the different growing seasons for groundnut cultivation? (seasons)	1	1	0.8	0.875
Mean ± [CI_lower, CI_upper]	0.906 ± [0.778, 1]	1 ± [1, 1]	0.916 ± [0.8072, 1]	0.945 ± [0.8508, 1]