CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.033

A retrieval-augmented large language model for agricultural advisory on crop varieties and cultivation techniques

Tran Thanh Dien[1*], Tran Nguyen Phuc Thinh[2], Le Duy Anh[1], Nguyen Thai-Nghe[2], Phan Khoa Anh[3], and Nguyen Huu Hoa[2]

[1]*Can Tho University Publishing House, Can Tho University, Viet Nam*

[2]*College of Information and Communication Technology, Can Tho University, Viet Nam*

[3]*The Improbability Company, Viet Nam*

**Corresponding author (thanhdien@ctu.edu.vn)*

Article info.

Received 13 Dec 2025
Revised 22 Jan 2026
Accepted 28 May 2026

Keywords

Agricultural advisory systems, hybrid information retrieval, large language models, re-ranking, retrieval-augmented generation, smart agriculture

ABSTRACT

This study presents the design and implementation of an agricultural advisory chatbot to help farmers access reliable information on crop varieties and cultivation techniques, with a focus on rice and mango. The proposed system integrates a large language model (Gemini 2.5 Pro) with a Retrieval-Augmented Generation architecture to mitigate hallucination and improve factual accuracy in domain-specific responses. The knowledge base is constructed from authoritative agricultural technical documents, including crop variety descriptions, cultivation procedures, and pest and disease management guidelines. These materials are systematically preprocessed, segmented, and indexed in the Qdrant vector database to enable efficient semantic retrieval within the Retrieval-Augmented Generation pipeline. To enhance retrieval robustness, the system employs a hybrid search strategy that combines keyword-based retrieval and dense vector search, followed by a Cross-Encoder re-ranking module to optimize contextual relevance before response generation. System performance is evaluated using the Retrieval Augmented Generation Assessment framework. Experimental results demonstrate high reliability, with a Faithfulness score of 91.43% and an Answer Relevancy score of 95.52%. The findings indicate that the proposed approach can deliver accurate, context-aware, and practically applicable agricultural recommendations, highlighting its potential to support digital transformation and AI-driven solutions in smart and sustainable agriculture.

1. INTRODUCTION

Agriculture remains a critical sector for food security and economic stability, particularly in developing countries where smallholder farmers rely heavily on timely and accurate technical information. Decisions related to crop variety selection, cultivation techniques, and pest and disease management require access to reliable knowledge sources. However, traditional agricultural extension services are often constrained by limited human resources, geographical barriers, and delayed information dissemination, making it difficult to meet farmers' practical needs in a timely manner.

Recent advances in large language models (LLMs) have enabled the development of conversational

agents capable of generating fluent and context-aware responses across a wide range of topics. Despite their impressive generative capabilities, LLMs suffer from a fundamental limitation: they rely on static pre-trained knowledge and may produce hallucinated or outdated information when applied to specialized domains such as agriculture (Ji et al., 2023). This limitation poses significant risks in agricultural advisory scenarios, where inaccurate recommendations may lead to economic loss or environmental harm.

Retrieval-Augmented Generation (RAG) has emerged as an effective paradigm for addressing these limitations by grounding language model outputs in external knowledge sources (Lewis et al., 2021). By retrieving relevant documents at inference time and incorporating them into the generation process, RAG-based systems significantly improve factual accuracy and transparency. Recent studies have demonstrated the effectiveness of RAG architectures in knowledge-intensive tasks, including domain-specific question answering and decision support systems (Ji et al., 2023; Yang et al., 2024). Nevertheless, the performance of RAG systems strongly depends on the quality of retrieval mechanisms, particularly in domains characterized by technical terminology and localized expressions.

In agricultural applications, relying solely on dense vector retrieval can yield suboptimal results, as semantically similar documents may not always contain the most relevant technical details. Conversely, keyword-based retrieval methods capture exact term matches but often fail to reflect deeper semantic intent. Hybrid retrieval strategies that combine dense and sparse retrieval methods, together with re-ranking techniques, have been shown to improve retrieval robustness and precision (Cormack et al., 2009; Thakur et al., 2021).

Motivated by these challenges, this study proposes an agricultural advisory chatbot that integrates an LLM with a hybrid RAG architecture. The system combines semantic vector search, keyword-based retrieval, and Cross-Encoder re-ranking to ensure high-quality contextual grounding before response generation. Focusing on rice and mango as representative crops, the chatbot is designed to support technical advisory tasks, comparative reasoning, and multi-turn interactions while minimizing hallucination through prompt-level guardrails.

In terms of methodology, this study contributes a validated RAG framework that systematically integrates hybrid retrieval, Cross-Encoder re-ranking, and grounding-aware prompt design for domain-specific advisory tasks. The framework is empirically evaluated through a combination of quantitative assessment using the Retrieval Augmented Generation Assessment (RAGAs) framework and qualitative scenario-based analysis, providing practical evidence of its effectiveness in improving factual grounding, contextual reliability, and robustness in agricultural advisory applications.

2. RELATED WORK

Recent advances in LLMs have enabled the development of sophisticated conversational agents capable of understanding complex user queries and generating coherent natural language responses. Despite these advances, the application of general-purpose LLMs in knowledge-intensive domains remains challenging due to their reliance on static training data and their tendency to generate hallucinated or unsupported information (Ji et al., 2023). These limitations are particularly critical in domains such as agriculture, where inaccurate recommendations may lead to significant economic losses or adverse impacts on crop production.

To overcome these challenges, RAG has emerged as an effective architectural paradigm that integrates external information retrieval with neural text generation. By conditioning the generation process on documents retrieved from a curated knowledge base, RAG-based systems have demonstrated substantial improvements in factual accuracy and interpretability compared to standalone LLMs (Lewis et al., 2021). However, recent benchmark studies indicate that the effectiveness of RAG systems is highly dependent on the quality of the retrieval component, and that suboptimal retrieval often becomes the primary bottleneck in end-to-end performance (Yang et al., 2024).

Various retrieval strategies have been proposed to enhance the relevance and coverage of retrieved documents. Hybrid retrieval approaches that combine keyword-based methods with dense vector-based semantic search have shown particular promise, as they leverage both exact lexical matching and contextual similarity. Techniques such as Reciprocal Rank Fusion (RRF) have been widely adopted to integrate ranked results from heterogeneous retrievers, often outperforming individual ranking methods in information retrieval tasks (Cormack et al., 2009). Furthermore, re-

ranking models based on Cross-Encoder architectures have been shown to significantly improve retrieval precision by jointly encoding query–document pairs, especially in complex and domain-specific scenarios (Thakur et al., 2021).

The integration of LLMs with retrieval mechanisms has also been explored in several domain-specific conversational systems. For instance, Li et al. (2023) proposed ChatDoctor, a medical chatbot that combines LLMs with external medical knowledge to reduce hallucination and improve response quality. This line of research highlights the potential of retrieval-augmented architectures for delivering reliable expert-level advice. Nevertheless, applications of such approaches in agricultural advisory systems remain limited, and few studies have systematically evaluated their performance using standardized metrics that jointly assess retrieval quality and response generation.

Existing studies demonstrate the effectiveness of RAG-based conversational agents in mitigating hallucinations and improving factual accuracy, while also revealing persistent challenges in retrieval quality and domain adaptation. Building on these findings, the present study proposes a hybrid RAG-based chatbot tailored to agricultural advisory tasks, with a particular focus on crop variety selection and cultivation techniques, and evaluates its performance using a standardized evaluation framework.

3. METHODS

3.1. The proposed model

The proposed system is a domain-specific agricultural advisory model based on an RAG architecture, which integrates a large language model with external agricultural knowledge to deliver accurate, context-aware, and reliable recommendations on crop varieties and cultivation techniques.

The model operates in two main stages: knowledge indexing and query-driven response generation. During indexing, authoritative agricultural documents are preprocessed, segmented into semantically coherent text chunks, encoded into dense vector representations, and stored in a vector database to enable efficient semantic retrieval from heterogeneous knowledge sources.

In the inference stage, user queries are processed through a hybrid retrieval mechanism that combines keyword-based search with semantic vector search. The retrieved candidates are further refined by a Cross-Encoder re-ranking module to select the most relevant contexts. These contexts are injected into a structured prompt, constraining the language model to generate responses strictly grounded in validated agricultural information, thereby reducing hallucinations and improving factual consistency. The model also supports multi-turn interactions by maintaining conversational context across dialogue turns. From an architectural perspective, the system adopts a modular client–server design to ensure scalability and extensibility. The backend manages request routing, session handling, and interactions with retrieval and generation components, while structured data and vectorized knowledge are stored in separate databases for efficiency and maintainability.

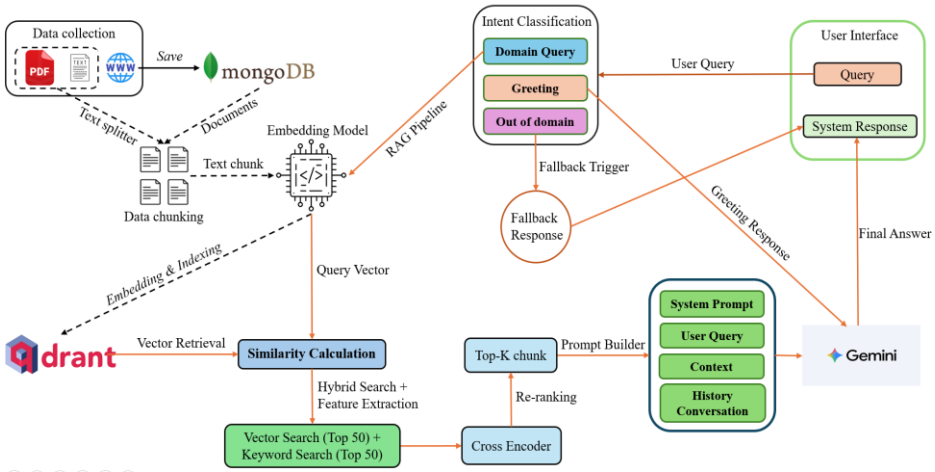

Figure 1. Overview of the proposed RAG-based agricultural advisory model

Overall, the proposed model integrates retrieval, re-ranking, and grounded generation within a unified RAG framework, providing a robust foundation for domain-specific agricultural advisory systems. The overall workflow of the proposed model is illustrated in Figure 1. The individual components of the proposed model are described in detail in the following sections.

3.2. Data collection, knowledge base construction, and preprocessing

The effectiveness of a RAG system largely depends on the quality, coverage, and organization of its knowledge base. In this study, data are collected in a selective and controlled manner, focusing on two major crops in Vietnam: rice and mango. The knowledge base is constructed from authoritative agricultural documents, including textbooks, technical manuals, cultivation guidelines, pest and disease management handbooks, and crop variety descriptions issued by universities, research institutes, and governmental agricultural agencies. The source materials are primarily provided in PDF and DOCX formats to ensure standardization and consistency during processing.

The primary data sources include: (i) academic textbooks and teaching materials from higher education institutions and research centers; (ii) technical handbooks and cultivation protocols published by specialized research institutes such as the Cuu Long Delta Rice Research Institute; and (iii) selected open agricultural knowledge platforms (e.g., WikiCrop), which are considered potential resources for future system expansion. All documents are selected based on three criteria: scientific validity, relevance to Vietnamese agro-ecological conditions, and the presence of clearly defined technical procedures or recommendations.

Before indexing, all documents undergo a structured preprocessing pipeline to ensure data reliability and retrieval efficiency. This process begins with text extraction and cleaning, during which non-informative elements such as headers, footers, page numbers, redundant formatting, and layout artifacts are removed. The cleaned text is then normalized to improve linguistic consistency and prepare the data for segmentation and embedding.

Given the length of agricultural technical documents and the context window limitations of large language models, a text segmentation strategy is applied to balance semantic completeness and retrieval granularity. Documents are divided into smaller, semantically coherent units using a recursive character-based text splitting approach that prioritizes natural textual boundaries such as paragraphs and sentences. Each text chunk is configured with a maximum length of 512 tokens and an overlap of 50 tokens between adjacent chunks. This configuration preserves contextual continuity and mitigates information loss at segmentation boundaries, particularly in procedural descriptions. An illustrative example of the text chunking strategy with character overlap is shown in Figure 2.

After segmentation, each text chunk is enriched with metadata, including document source, page number, crop type, and thematic category. The chunks are then encoded into dense vector representations using a multilingual embedding model and stored in the Qdrant vector database to support semantic retrieval within the RAG architecture. In parallel, the knowledge base is organized in MongoDB according to functional categories, including: (i) crop varieties (biological characteristics, growth duration, yield potential); (ii) plant protection (pests, diseases, symptoms, and treatment methods); (iii) cultivation techniques (standard operating procedures such as VietGAP, fertilization, irrigation, and pruning); and (iv) original reference documents used for citation and evidence grounding.

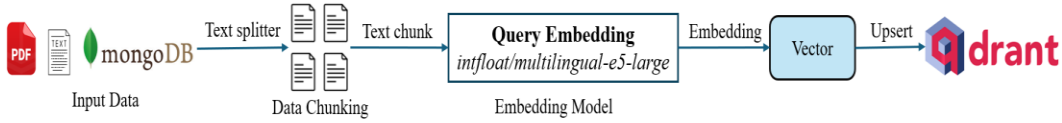

Figure 2. Text chunking strategy with character overlap

To further enhance domain coverage and practical relevance, the system integrates authoritative agricultural resources from national and sectoral institutions. These include the Rice Knowledge Bank of the Vietnam Academy of Agricultural Sciences (VAAS, 2024), sectoral data on agricultural value chains from the (Vietnam Food Association, 2024), and research outputs on rice varieties and cultivation systems from the Cuu Long Delta Rice Research Institute and the Field Crops Research Institute (Cuu Long Delta Rice Research Institute, 2024; Field Crops Research Institute,

2024). For mango cultivation, additional varietal and biological information is collected from specialized crop variety portals such as the "New Crop Varieties" platform maintained by the Institute of Plant Germplasm under VAAS (2024). The overall workflow of data collection, preprocessing, segmentation, embedding, and vector indexing is summarized in Figure 3.

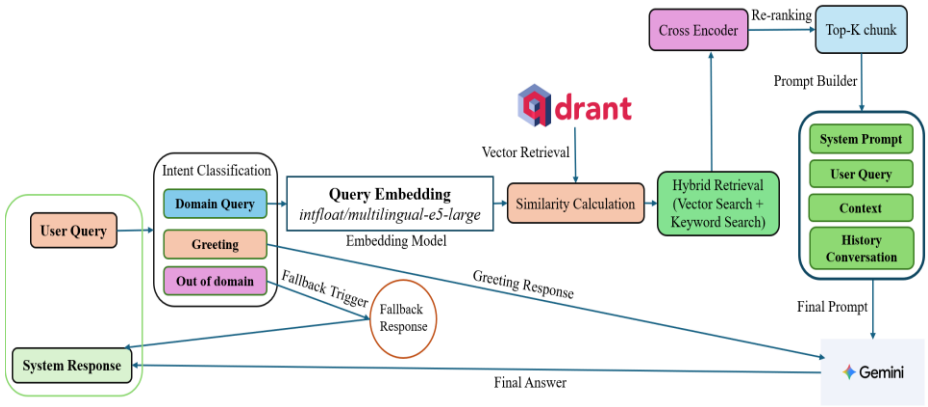

Figure 3. Knowledge base construction and indexing pipeline

Through this systematic pipeline, the proposed system ensures that retrieved contexts are both accurate and contextually coherent for downstream response generation. A statistical summary of the knowledge data processed in the indexing pipeline is presented in Table 1.

Table 1. Statistics of knowledge data volume for the indexing pipeline

Data category	Content description	Quantity	Remarks
Crop varieties	Biological traits, origin, and growth duration of rice varieties (e.g., OM18, ST25) and mango varieties (e.g., Cat Hoa Loc, Cat Chu).	47	Includes 36 rice varieties and 11 mango varieties.
Pests and diseases	Descriptions of symptoms, causal agents (fungi, bacteria, viruses), and visual references of common crop diseases.	20	Covers major diseases such as rice blast, brown planthopper, anthracnose, and black spot.
Disease stages	Disease progression across development stages to support stage-specific advisory recommendations.	20	Decomposed from major diseases for fine-grained advisory support.
Cultivation techniques	Standard operating procedures, fertilization and irrigation practices (e.g., alternate wetting and drying), pruning and canopy management.	14	Based on official guidelines and VietGAP standards.
Reference documents	Original technical manuals and handbooks were used as primary knowledge sources.	10	Retained for citation and evidence grounding.

3.3. Embedding model selection and vector indexing

In a RAG system, embedding quality is a key factor influencing retrieval accuracy, particularly in specialized domains such as agriculture where technical terminology and contextual nuances are prevalent. In this study, text chunks are encoded using the intfloat/multilingual-e5-large embedding model to obtain semantically meaningful vector representations. The multilingual-e5-large model is built on the XLM-RoBERTa-large architecture and optimized for multilingual semantic retrieval. Recent benchmark evaluations report state-of-the-art performance on datasets such as MTEB and MIRACL (Wang et al., 2024), making the model well-suited for Vietnamese agricultural texts that include localized expressions and domain-specific vocabulary.

Each preprocessed text chunk is encoded into a fixed-length 1024-dimensional vector, with the input length capped at 512 tokens to align with the segmentation strategy described in the section above. These vectors represent individual units of

agricultural knowledge and serve as semantic indices for retrieval.

All embeddings are stored in the Qdrant vector database, which supports efficient approximate nearest neighbor search. The system employs the Hierarchical Navigable Small World (HNSW) indexing algorithm with cosine similarity as the distance metric, enabling robust similarity-based retrieval across text segments of varying length and complexity.

The overall workflow of text chunk embedding, vector storage, and similarity-based retrieval is illustrated in Figure 4, showing how segmented agricultural documents are mapped into a unified semantic vector space. By combining a high-quality multilingual embedding model with efficient vector indexing, the system establishes a reliable foundation for downstream hybrid retrieval and re-ranking stages.

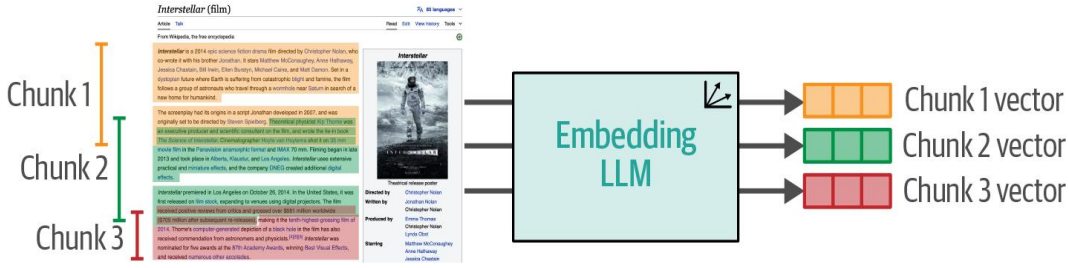

Figure 4. Text chunk embedding and vector indexing in the semantic search space

(Adapted conceptually from Alammar & Grootendorst, 2024)

3.4. RAG techniques and hybrid retrieval search method

LLMs rely on static pre-trained knowledge and are therefore prone to hallucination when applied to domain-specific tasks such as agricultural advisory (Ji et al., 2023). In addition, agricultural queries often contain technical terms, local crop names, or abbreviated disease descriptions, posing challenges for both purely semantic and purely keyword-based retrieval methods. These limitations motivate the adoption of an RAG approach that combines complementary retrieval strategies.

The proposed system adopts a RAG architecture in which external knowledge retrieval is tightly integrated into response generation. Instead of generating answers solely from internal model parameters, the system retrieves relevant text chunks from an agricultural knowledge base and injects them into the prompt. This process consists of three stages: retrieval of candidate contexts, prompt augmentation with retrieved evidence, and grounded response generation (Lewis et al., 2021). By decoupling factual retrieval from language generation, the system reduces hallucination and improves response transparency.

To enhance retrieval robustness, a hybrid retrieval mechanism is employed. Semantic vector search is performed using cosine similarity to identify text chunks that are semantically close to the user query, while keyword-based retrieval with the BM25 algorithm captures exact term matches, which are critical for agricultural terminology. The results from both retrieval methods are combined using the RRF technique, which aggregates multiple ranked lists into a single ranking.

The RRF score for a document $d$is computed as:

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)} \qquad (1)$$

As shown in Equation (1), documents that rank highly in any retrieval list receive higher fused scores, where R denotes the set of ranked lists, $\mathrm{rank}_r(d)$ is the rank position of the document d in ranking r, and k is a constant used to control the contribution of lower-ranked documents. Following recommendations from the foundational study by (Cormack et al., 2009), the parameter $k$is set to 60 in this implementation. This configuration balances sensitivity to high-ranked results with the prevention of dominance by any single retrieval method.

3.5. Re-ranking techniques

Although the hybrid retrieval strategy improves document coverage, the initial results may still include noisy or weakly relevant text chunks due to the approximate nature of both semantic and keyword-based retrieval. In agricultural knowledge bases, passages discussing different crop varieties or diseases may use overlapping terminology yet require distinct recommendations. Injecting such

noisy context into the prompt can reduce answer accuracy and increase token usage, particularly within LLM context windows. To address this issue, the system incorporates a Cross-Encoder-based re-ranking stage. Unlike Bi-Encoder models, which encode queries and documents independently, Cross-Encoders jointly process each query–document pair within a single Transformer, enabling fine-grained interaction modeling and more accurate relevance estimation. In this study, the BAAI/bge-reranker-v2-m3 model is employed. After hybrid retrieval, a limited set of top candidate documents is re-scored and re-ordered, and only the highest-ranked text chunks are retained as contextual input for response generation.

The architectural differences between Bi-Encoder and Cross-Encoder models in terms of input processing, computational cost, and semantic precision are summarized in Table 2.

The two-stage retrieval design balances efficiency and accuracy. Bi-Encoder-based retrieval enables fast large-scale search, while Cross-Encoder re-ranking provides high semantic precision on a small candidate set. This combination ensures that the language model receives concise, highly relevant context, thereby improving contextual precision and reducing hallucinations. Prior studies confirm that Cross-Encoder-based re-ranking substantially enhances retrieval quality in knowledge-intensive tasks (Nogueira & Cho, 2019; Thakur et al., 2021).

Overall, the re-ranking module plays a critical role in refining retrieval outputs and ensuring that only high-quality, domain-relevant information is supplied to the generation stage–an essential requirement for reliable agricultural advisory systems.

Table 2. Comparison of the architectural characteristics of Bi-Encoder and Cross-Encoder

Comparison criteria	Bi-Encoder	Cross-Encoder
Input processing mechanism	The query and the document are independently encoded using two separate neural networks (or shared-weight networks), producing two distinct vector representations (Reimers & Gurevych, 2019).	The query and the document are concatenated and jointly encoded in a single Transformer network, enabling direct interaction among all tokens (Nogueira & Cho, 2019).
Scoring mechanism	Similarity is computed by comparing the independently generated vectors, typically using the dot product or cosine similarity.	Relevance is computed directly through full self-attention over the combined query–document input, enabling fine-grained semantic matching.
Retrieval speed	Very fast. Document embeddings can be pre-computed and indexed in advance, making them suitable for large-scale retrieval.	Slower. Relevance scores must be computed from scratch for each query–document pair.
Computational cost	Low. Efficient for searching large collections containing millions of documents.	High. Computationally expensive and suitable only for re-ranking a small set of candidate documents.
Semantic precision	Moderate. May fail to capture complex semantic relationships or subtle contextual distinctions due to independent encoding.	High. Captures deep semantic interactions between query and document through joint attention, leading to more accurate relevance estimation.

3.6. Prompt Engineering and LLM Integration

Prompt engineering

In the proposed system, the LLM functions as the central reasoning component that synthesizes retrieved agricultural knowledge into coherent responses. To ensure factual grounding and prevent unsupported generation, a structured prompt engineering strategy based on in-context learning is employed, in which retrieved document segments are explicitly provided as the sole authoritative source for response generation.

The prompt comprises three main components: a system-level instruction that defines the model's role as an agricultural expert; a context section containing top-ranked document segments obtained after re-ranking; and a set of grounding constraints that restrict the model to the provided context, discourage speculation, and require explicit

acknowledgement when relevant information is unavailable.

To support realistic advisory scenarios, the system handles multi-turn conversations by maintaining and dynamically incorporating dialogue history into the prompt. This enables the model to correctly interpret follow-up queries while preserving contextual continuity across turns.

The system integrates the Gemini 2.5 Pro LLM via the Google Generative AI API, selected for its strong long-context processing capability and proficiency in Vietnamese. During inference, a low temperature setting is used to prioritize factual consistency over creative variation. In addition, a lightweight intent classification module based on a smaller language model distinguishes domain-relevant agricultural queries from non-domain inputs. For out-of-domain queries, a controlled refusal mechanism is triggered, preventing responses beyond the agricultural knowledge base.

By combining structured prompt design with controlled LLM invocation, the system clearly separates knowledge retrieval from language generation, thereby enhancing reliability, transparency, and safety in agricultural advisory applications.

In practice, the system adopts a split-task LLM architecture. Gemini 2.5 Pro is employed as the primary reasoning model for grounded answer generation due to its strong long-context processing and reasoning capabilities. In parallel, Gemini 2.5 Flash is used for lightweight tasks such as intent classification and query routing, where low latency is prioritized. This design improves overall system efficiency while preserving high-quality, context-aware response generation.

3.7. Evaluation framework

Evaluating RAG systems is challenging due to the open-ended nature of generated responses and the strong dependence between retrieval quality and generation reliability. Traditional text similarity metrics such as BLEU or ROUGE are insufficient, as they focus on lexical overlap and fail to assess semantic correctness and factual grounding. Therefore, this study adopts the RAGAs framework, which provides a principled evaluation approach by explicitly modeling the relationships among the user query, retrieved context, and generated answer (Es et al., 2025).

Within the RAGAs framework, four core metrics are employed to assess both retrieval and generation performance. **Faithfulness** measures the extent to which generated answers are supported by the retrieved context and is critical for mitigating hallucination in knowledge-intensive domains (Ji et al., 2023). **Answer Relevancy** evaluates how well the generated response aligns with the user's information need, penalizing irrelevant or excessively verbose answers. **Context Precision** assesses the proportion of retrieved text segments that are relevant to the query, particularly among higher-ranked documents, which is important given the tendency of language models to prioritize early prompt content ("lost in the middle" effect) (Liu et al., 2023). **Context Recall** measures whether the retrieved contexts collectively provide sufficient information to generate a complete and correct answer.

Together, these metrics provide a balanced and comprehensive evaluation of RAG system performance. Context Precision and Context Recall characterize retrieval effectiveness, while Faithfulness and Answer Relevancy assess the reliability and usefulness of generated responses. This evaluation framework enables fine-grained diagnosis of system behavior and serves as a robust foundation for the experimental analysis presented in the subsequent section.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

The experimental setup evaluates the effectiveness of the proposed RAG-based chatbot in agricultural advisory scenarios, focusing on both retrieval quality and response reliability. A curated evaluation dataset is constructed from representative agricultural queries related to crop variety selection and cultivation techniques, with emphasis on rice and mango. The queries reflect realistic farmer information needs, including technical guidance, comparative questions, and procedural inquiries. For each query, a reference answer derived from authoritative agricultural documents serves as the ground truth.

Evaluation follows a standardized protocol aligned with the RAGAs framework. For each query, relevant document segments are retrieved using the hybrid retrieval and re-ranking pipeline and supplied to the language model under the prompt constraints. Both retrieved contexts and generated responses are then assessed using the selected RAGAs metrics. All system parameters, including the embedding configuration, retrieval settings, re-ranking strategy, and prompt structure, are held

constant across evaluation runs to isolate the impact of the retrieval–generation pipeline. Experiments are conducted offline to ensure reproducibility, using the same queries, contexts, and reference answers for all metric computations. Quantitative and qualitative results are presented in the subsequent subsections.

4.2. Quantitative evaluation results

The quantitative results summarized in Table 3 indicate that the proposed system achieves strong overall performance. The Faithfulness score of 0.9143 demonstrates that the generated responses are well grounded in the retrieved source documents, confirming the effectiveness of the RAG architecture in mitigating hallucinations in agricultural advisory tasks. The system also achieves a high Answer Relevancy score of 0.9552, reflecting its ability to accurately capture user intent and produce focused, on-topic responses without unnecessary verbosity. In addition, the combination of hybrid retrieval and re-ranking substantially improves both Context Precision and Context Recall compared with vector-only retrieval. These results highlight the effectiveness of integrating semantic and keyword-based search for technical agricultural documents containing terminology and localized expressions, thereby enhancing the quality of contextual information supplied to the language model.

Table 3. Average evaluation results using the RAGAs scale

Metric	Score	Meaning
Faithfulness	0.9143	Responses are well-grounded in source documents, ensuring high reliability.
Answer Relevancy	0.9552	Answers closely match user queries and satisfy information needs.
Context Precision	0.8415	The re-ranking module effectively filters noisy or irrelevant contexts.
Context Recall	0.8872	Retrieved contexts sufficiently support complete answers.

4.3. Qualitative evaluation results

In addition to quantitative metrics, qualitative evaluation is conducted to examine the system's behavior in realistic agricultural advisory scenarios. The evaluation focuses on the chatbot's ability to perform comparative reasoning, synthesize technical procedures, handle out-of-domain queries safely, and maintain conversational context under noisy input. Four representative scenarios are designed to reflect common information needs in agricultural practice.

Scenario 1: Comparative reasoning queries

This scenario evaluates the system's ability to retrieve information from multiple sources and synthesize it into a structured comparison. Given the query "*What are the main differences between Cat Chu mango and Cat Hoa Loc mango?*", the system correctly identifies comparative intent, retrieves relevant information for both varieties, and generates a coherent comparison across key attributes such as origin, fruit morphology, and flavor characteristics. The response demonstrates effective multi-source integration and logical comparison, highlighting the reasoning capability enabled by RAG architecture. This behavior is illustrated in Figure 5.

Scenario 2: Technical procedures for pest and disease management

This scenario assesses the system's ability to synthesize complex technical procedures. For the query *"How can anthracnose disease be controlled?"*, the chatbot retrieves heterogeneous knowledge related to cultivation practices, nutrient management, and chemical control. The generated response is well structured, emphasizing orchard sanitation and cultivation practices, complemented by precise recommendations on irrigation, nutrition, and appropriate active ingredients for preventive spraying. The response demonstrates high technical accuracy and practical applicability. An example is shown in Figure 6.

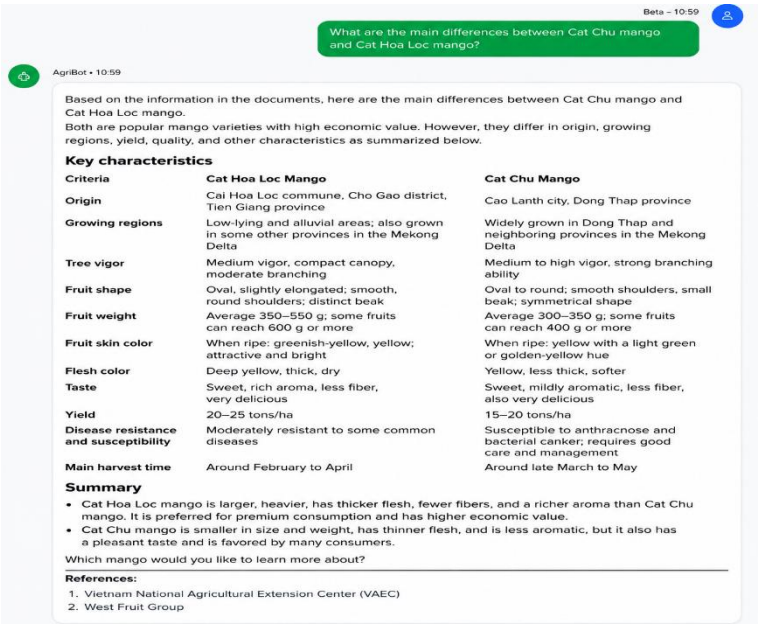

Figure 5. Comparative reasoning using multi-source agricultural information

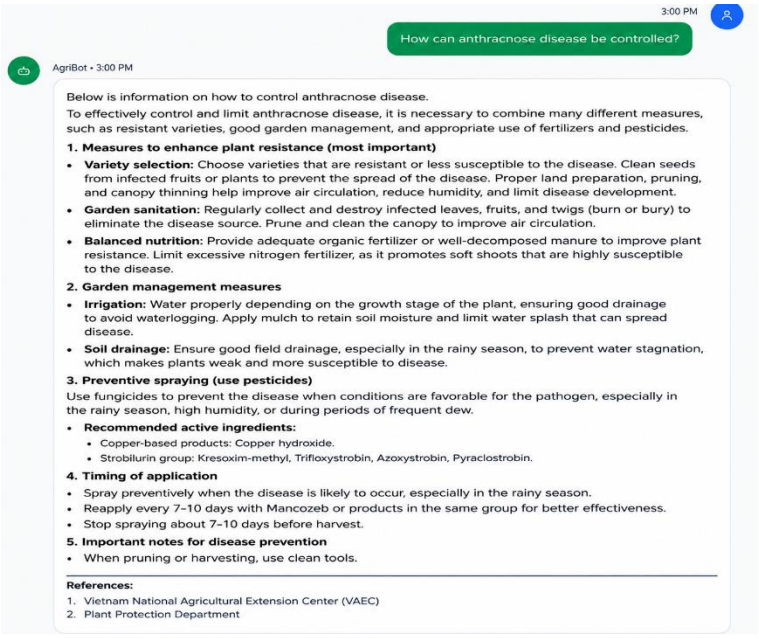

Figure 6. Technical advisory response for pest and disease management procedures

Scenario 3: Handling out-of-domain queries

This scenario evaluates the system's guardrail mechanism. When asked, *"Who is the President of Vietnam?"*, the query classification module correctly identifies the question as out-of-domain and triggers a controlled refusal response. Instead of generating potentially incorrect information, the system politely redirects the user to agricultural topics. This behavior confirms effective domain control and hallucination prevention. The response is illustrated in Figure 7.

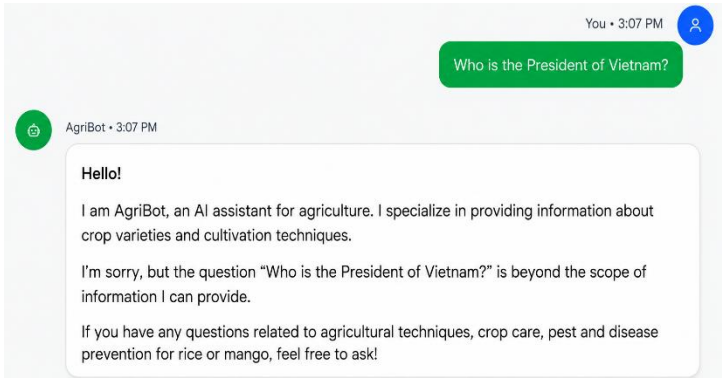

Figure 7. Controlled refusal for out-of-domain queries

Scenario 4: Context preservation and robustness to input errors

This scenario evaluates the system's ability to preserve conversational context across multi-turn interactions and to handle noisy input while maintaining semantic understanding. In a two-turn interaction, the user first describes rice symptoms associated with brown spot disease, allowing the system to correctly identify the disease and retain it as contextual state. In the subsequent turn, the user submits a follow-up query containing severe spelling distortions ("How can it be comtrol"), which is intended to ask about prevention and control measures ("How can it be control"). These results confirm strong context preservation and robustness to input noise, which are essential for real-world agricultural advisory scenarios where users may omit explicit references or introduce typing errors. Representative examples are shown in Figure 8 and Figure 9.

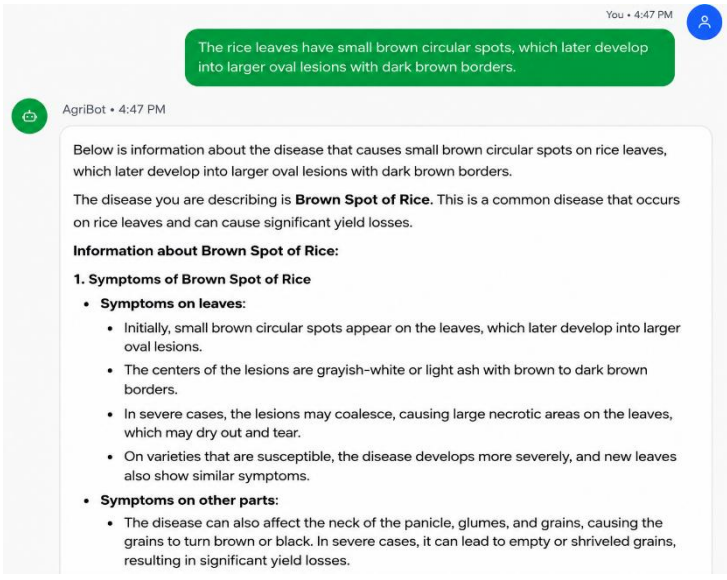

Figure 8. Multi-turn context preservation in agricultural advisory dialogue

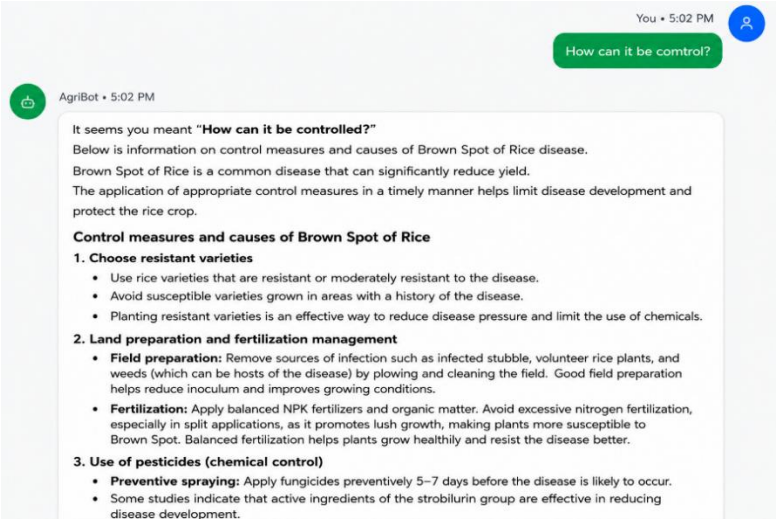

Figure 9. Robust handling of spelling errors in follow-up queries

4.4. Response time analysis

Response time is a critical factor in interactive advisory systems, as it directly affects user experience. To assess the real-time performance of the proposed system, response time is measured on a test set of 50 randomly selected queries, covering both short factual questions and more complex advisory requests.

In this processing pipeline, different large language models are employed for different tasks. Gemini 2.5 Flash is used for intent classification and query routing due to its low latency, whereas Gemini 2.5 Pro is responsible for response generation, which involves longer inference time but provides higher reasoning quality. This split-task design helps explain the heterogeneous latency observed across processing stages.

Table 4. Decomposition of the average query processing time

Stage	Processing component	Average time	Proportion
1. Intent analysis	Calling the Gemini 2.5 Flash API to classify the query (Agricultural / Greeting / Out-of-domain) and extract keywords	~500 ms	~3.12%
2. Data retrieval	- Query embedding using E5-large: ~800 ms - Vector search with Qdrant: ~600 ms - Metadata query (keyword-based search): ~200 ms	~1500 ms	~9.38%
3. Result refinement	Re-ranking the top 100 retrieved candidates using a Cross-Encoder (BAAI/bge-reranker-v2-m3) and selecting the top 10 most relevant contexts	~2000 ms	~12.5%
4. Answer generation	Time to First Token (TTFT): the time required for Gemini 2.5 Pro to process the retrieved contexts and generate the first output token	~12000 ms	~75%
Total	Time elapsed from clicking "Send" until the first generated token appears	~16000 ms	100%

The processing of each query is decomposed into four main stages, with average processing times and their relative contributions summarized in Table 4. Among these stages, answer generation dominates the total latency, accounting for approximately 75% of the overall response time. This behavior is inherent to large language models such as Gemini 2.5 Pro, which must process substantial contextual input during response generation.

Although the re-ranking module introduces an additional latency of approximately 2000 ms, this overhead represents a favorable trade-off. With re-ranking enabled, Context Precision improves from 0.53 to 0.84, whereas omitting this step reduces the total response time to about 14,000 ms but causes a sharp drop in Context Precision to approximately 0.50–0.55 due to increased contextual noise. These results indicate that the additional latency yields a substantial improvement in response reliability.

The Qdrant vector database demonstrates efficient retrieval performance, with an average vector search time of approximately 600 ms over a collection of several thousand vectors, confirming the scalability of the retrieval component. After the first token is generated (typically after 15–16 seconds), the system achieves an average generation speed of 45–60 tokens per second, allowing complete responses of 300–500 words to be delivered within an additional 8–12 seconds.

4.5. Detailed latency and error analysis

From a latency perspective, the average TTFT ranges from 12 to 16 seconds, with most of the delay attributable to LLM inference, particularly when processing extensive contextual input. In agricultural advisory applications, response speed is secondary to accuracy and reliability. For users such as farmers, a waiting time of approximately 15 seconds is acceptable if it ensures correct and actionable technical guidance. This latency reflects a deliberate design trade-off between speed and precision. In agriculture, informational errors–such as incorrect pesticide application or fertilization timing–can lead to significant economic losses. Therefore, the system prioritizes accuracy and safety over real-time responsiveness. The observed latency is required to execute the full quality assurance pipeline, including hybrid retrieval, Cross-Encoder-based re-ranking, and evidence-grounded response generation, and is considered reasonable for an expert advisory system.

To improve efficiency for simple interactions, the system incorporates a query classification and routing module based on Gemini Flash. This component quickly identifies greetings or out-of-domain queries and handles them using predefined responses or controlled refusals, thereby reducing response time for non-technical requests.

Regarding error analysis, despite achieving a high Faithfulness score (approximately 0.91), several suboptimal cases are observed, primarily due to retrieval limitations rather than fundamental system failures. For complex comparative queries that require information from multiple distant sections of documents, context retrieval may be incomplete due to Top-K constraints, resulting in partial answers. In addition, short or ambiguous keywords may occasionally introduce semantic noise, causing confusion between conceptually similar diseases. While the re-ranking module mitigates this issue, residual ambiguity can still occur under lower filtering thresholds.

Based on these observations, typical error cases identified during evaluation are categorized and summarized in Table 5.

Table 5. Analysis of typical error cases

Error type	Description	Proposed mitigation
Incomplete context retrieval	For complex or comparative queries, the retriever may fail to retrieve all relevant information because essential knowledge is distributed across multiple document segments.	This limitation can be mitigated by increasing the retrieval scope (Top-K) and employing a hybrid retrieval strategy that combines Cross-Encoder re-ranking, enabling broader information coverage while preserving contextual relevance.
Semantic ambiguity	Short or ambiguous keywords (e.g., disease names with overlapping symptoms) may cause confusion between similar agricultural concepts.	This issue is addressed by combining semantic vector search with keyword-based retrieval and applying Cross-Encoder re-ranking to disambiguate semantically similar but contextually distinct document segments.
Context noise	Retrieved contexts may include weakly relevant or redundant text segments, which can distract the language model during response generation.	Contextual noise is reduced by applying a Cross-Encoder-based re-ranking stage that filters and prioritizes highly relevant document segments before prompt construction.
Out-of-domain queries	Users may submit questions outside the agricultural domain, such as political or general knowledge queries.	A lightweight intent classification module is used to detect out-of-domain queries and trigger a controlled refusal mechanism, ensuring the system responds only within its intended domain.
Typographical and input errors	User queries may contain spelling mistakes or character-level distortions that affect retrieval accuracy.	This issue is mitigated by leveraging the robustness of the large language model to noisy input and by relying on semantic retrieval, which reduces sensitivity to surface-level textual errors.

4.6. Evaluation of the re-ranking module

To evaluate the contribution of the re-ranking module, an ablation study is conducted by comparing two system configurations. Configuration A (Baseline) uses vector-based semantic search only, while Configuration B (Proposed) integrates a Cross-Encoder re-ranking model (BAAI/bge-reranker-v2-m3) to rescore retrieved document chunks before response generation. Both configurations are evaluated on the same test set of 30 representative queries using the RAGAs framework. The average results are reported in Table 6.

The results show that re-ranking leads to substantial improvements across all quality-related metrics. Context Precision increases by 58.4%, indicating that the Cross-Encoder effectively filters noisy or weakly relevant contexts that vector search alone fails to distinguish. Context Recall also improves significantly (+87.5%), reflecting the effectiveness of the "retrieve broadly, then filter precisely" strategy.

Improvements in retrieval quality translate directly into better generation performance. Faithfulness increases from 0.76 to 0.91, indicating reduced hallucination, while Answer Relevancy improves from 0.83 to 0.95, demonstrating stronger alignment with user intent. These findings are consistent with prior studies highlighting the superior semantic discrimination of Cross-Encoder architectures (Nogueira & Cho, 2019).

From a performance perspective, re-ranking introduces an additional latency of approximately 0.8 seconds per query. However, this overhead represents an acceptable trade-off given the substantial gains in retrieval accuracy and response reliability. Overall, the ablation study confirms that the re-ranking module is a critical component of the proposed system, significantly enhancing the trustworthiness and practical usefulness of agricultural advisory responses.

Table 6. Compare the performance of configurations with and without re-ranking

Metric	Baseline (Vector Search only)	Proposed (Vector Search + Re-ranking)	Improvement
Context Precision	0.53	0.84	+58.4%
Context Recall	0.32	0.60	+87.5%
Faithfulness	0.76	0.91	+19.7%
Answer Relevancy	0.83	0.95	+14.5%
Response time	1.2 s	2.0 s	−0.8 s

4.7. Evaluation of system stability with augmented test datasets

To assess the robustness and statistical reliability of the initial evaluation results (N = 30), the test dataset is progressively expanded to include 40 and 50 queries. The additional queries increase both diversity and difficulty, covering exceptional agricultural scenarios and multi-step reasoning tasks. The objective is to examine whether performance metrics remain stable as the evaluation scope grows. Metric variations below approximately 5% are considered indicative of stable and representative results.

The quantitative results across different dataset sizes are summarized in Table 7.

Table 7. Performance comparison as the test suite scales

Metric	N = 30	N = 40	N = 50	Standard deviation
Faithfulness	0.9143	0.9199	0.9138	0.0034
Answer Relevancy	0.9552	0.9453	0.9527	0.0051
Context Precision	0.8400	0.8372	0.8395	0.0015
Context Recall	0.5960	0.5921	0.5889	0.0026

To quantify metric variability across the three datasets (N = 30, 40, and 50), standard deviation (SD) is used as a measure of stability. Standard deviation reflects the dispersion of values around their mean, with lower values indicating greater robustness (Walpole et al., 2012). For each metric, SD is computed as:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N - 1}} \quad (2)$$

The standard deviation defined in Equation (2) is used to assess the stability of evaluation metrics across different test set sizes, where x_i denotes the metric value for each dataset, μ is the mean value, and $N = 3$.

The results indicate strong system stability. Faithfulness shows minimal variation (SD = 0.0034), demonstrating consistent grounding behavior even as query diversity increases. Context Precision remains stable around 0.84 (SD = 0.0015), confirming the robustness of the hybrid retrieval and re-ranking architecture. Answer Relevancy exhibits minor fluctuations due to the inclusion of more challenging queries but maintains a low overall variation (SD = 0.0051), indicating consistent alignment with user intent.

Overall, metric variations across expanded test sets are negligible (below 1%), suggesting that performance has converged. These findings confirm that the initial evaluation conducted with 30 queries is statistically reliable and representative of the system's real-world performance.

5. CONCLUSION

This study presents a validated Retrieval-Augmented Generation (RAG) framework for domain-specific agricultural advisory that integrates hybrid retrieval, Cross-Encoder re-ranking, and grounding-aware prompt design. Methodologically, the results show that combining semantic vector search with keyword-based retrieval and re-ranking significantly improves retrieval precision, contextual coverage, and factual grounding, effectively mitigating hallucination in knowledge-intensive question-answering systems.

The framework is instantiated as an agricultural advisory chatbot for rice and mango cultivation. Quantitative evaluation using the RAGAs framework and qualitative scenario-based analysis confirm its ability to support complex advisory tasks, including comparative reasoning, procedural synthesis, multi-turn dialogue handling, and safe rejection of out-of-domain queries, with acceptable response latency.

Although the current implementation is limited in scope, the proposed framework is generic and scalable. Future work will extend the knowledge base to additional crops, incorporate multilingual and multimodal inputs, and integrate real-time agricultural data to further support smart and sustainable agriculture.

ACKNOWLEDGMENT

This study is funded by Can Tho University (Code: CTCS2025-01-04).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. The authors have no financial, professional, or personal relationships that could have influenced the research results or the interpretation of the findings.

REFERENCES

Alammar, J., & Grootendorst, M. (2024). *Hands-On Large Language Models: Language Understanding and Generation*. O'Reilly Media.

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009* (pp. 758–759). ACM. https://doi.org/10.1145/1571941.1572114

Cuu Long Delta Rice Research Institute. (2024). *Cuu Long Delta Rice Research Institute (CLRRI) – Official Website*. https://clrri.org/ (in Vietnamese).

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024, March). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations* (pp. 150-158).

Field Crops Research Institute. (2024). *Field Crops Research Institute (FCRI) – Official Website*. Field Crops Research Institute. https://fcri.com.vn/ (in Vietnamese).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, December). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 9459-9474)

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). *ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge* (No. arXiv:2303.14070). arXiv. https://doi.org/10.48550/arXiv.2303.14070

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, *12*, 157-173.

Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. *arXiv Preprint*, *abs/1901.04085*. https://arxiv.org/abs/1901.04085

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP 2019)* (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv Preprint*. https://arxiv.org/abs/2104.08663

VAAS. (2024). *Lessons on Rice Cultivation*. https://vaas.vn/kienthuc/Caylua/10/ (in Vietnamese).

Vietnam Academy of Agricultural Sciences, Institute of Plant Germplasm. (2024). *Mango Seedlings and Varieties*. https://giongcaytrongmoi.com/9/giong-cay-an-qua/giong-cay-xoai (in Vietnamese).

Vietnam Food Association. (2024). *Vietnam Food Association (Vietfood) – Official Website*. https://vietfood.org.vn/ (in Vietnamese).

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report* (No. arXiv:2402.05672). arXiv. https://doi.org/10.48550/arXiv.2402.05672

Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., ... Dong, X. L. (2024). CRAG: Comprehensive RAG benchmark. *Advances in Neural Information Processing Systems*, 37, 10470-10490.