Development of Generative Chatbot towards Thai Rice Knowledge Using Retrieval-Augmented Generation

Yuttana Jaroenruen[1], Nattapong Kaewboonma[2], Taneth Ruangrajitpakorn[3] and Jirapong Panawong[4,*]

[1]Informatics Innovative Center of Excellence, School of Informatics, Walailak University, Thailand
[2]Faculty of Management Technology, Rajamangala University of Technology Srivijaya, Thailand
[3]Artificial Intelligence Research Group, National Electronics and Computer Technology Center, Thailand
[4]Faculty of Management Science, Nakhon Ratchasima Rajabhat University, Thailand

∗Corresponding Email: jirapong.p@nrru.ac.th

Received January 26, 2025, Revised January 30, 2026, Accepted March 25, 2026, Published July 1, 2026

Abstract. *This paper presents the development of a generative chatbot to support Thai rice farmers by extending rice-relevant knowledge from reliable external sources. The developed chatbot system is based on a lightweight pretrained large language model and is improved through fine-tuning using collected documents related to Thai rice farming. In addition, approved Thai rice information is utilized as an external knowledge base within a retrieval-augmented generation framework to enable responses to domain-specific and technical questions related to Thai rice cultivation. Experimental evaluation compares chatbot systems with and without the application of retrieval-augmented generation. The results indicate that the generative chatbot incorporating retrieval-augmented generation performs better than the baseline system in terms of response relevance and factual reliability. User-driven evaluation shows that the RAG-based chatbot produces 91.43% acceptable responses and 8.57% partially acceptable responses, with no unacceptable responses observed. To further assess system reliability, a grounding-based evaluation was conducted using a controlled query set with document-level reference attribution. The proposed system achieves document-level grounding recall of 0.896 (micro-average) and 0.983 (macro-average), indicating effective utilization of the external knowledge base. These findings suggest that retrieval-augmented generation is an effective approach for developing generative chatbots in domain-specific and localized agricultural advisory contexts, where accuracy, grounding, and trustworthiness of information are essential.*

Keywords: Generative Chatbot, retrieval-augmented generation, Thai rice knowledge, advisory conversational system

1. Introduction

Rice holds an unparalleled significance in Thailand in terms of a staple food and a cornerstone of the country's economy, culture, and rural livelihoods. Economically, rice production is a crucial contributor to Thailand's agricultural sector, which employs a significant proportion of the workforce. The cultivation of rice sustains millions of farming households, particularly in rural regions, where it is often the primary source of income. Despite its importance, rice production in Thailand faces numerous challenges including a lack of personnel to provide suitable production knowledge and access to timely information [1]. Since rice cultivation involves a wide range of rice varieties with unique characteristics that cater to diverse environmental conditions, the complexity of managing these varieties lies in understanding related information in farm management such as suitable farming parameters, water management [2], and difference in invulnerability to pests and diseases in specific regions. However, traditional methods of knowledge distribution face significant barriers, such as limited accessibility, inconsistent outreach, and context-specific information.

In the current information-driven era, chatbots have emerged as sophisticated tools for efficient information retrieval and summarization making them invaluable in numerous fields including agriculture with its ability to provide accessible, accurate, and tailored information to users. Unlike generalized resources, the chatbot uses context-aware algorithms to deliver customized responses based on user input and can be accessed regardless of time or location [3][4]. Chatbots, especially ChatGPT [5], serve as intermediaries between users and vast data repositories to enable efficient and accurate access to relevant information. The main function of the chatbot in this regard is to simplify the process of locating specific data and to summarize extensive and complex datasets into concise and relevant contents. Thus, the issue of rice knowledge distribution involving cultivation practices, pest and disease management, and farming techniques to farmers can be solved using Chatbot. Chatbots hence present a transformative solution for rice knowledge distribution. By offering accessible, real-time, and context-sensitive guidance, a chatbot can empower farmers to adopt best and most suitable practices, solve emerging agricultural issues without needing to search for required information among massive information to read and summarize into compact content by themselves.

In this work, we aim to develop Chatbot using an existing GPT (Generative Pre-trained Transformer) model [5] to process and generate Thai text with RAG (Retrieval-Augmented Generation) [6] to search for relevant information from Thai rice public information as an external knowledge base. The developed chatbot thus can access to external approved knowledge for rice production specific information while maintains the conversational fluency of GPT model. This chatbot will hence response to farmers who need rice production information with accuracy, timeliness, and relevance.

2. Background

2.1 Large Language Models and Retrieval-Augmented Generation

Large Language Models (LLMs) are deep learning–based language models trained on large-scale text corpora to understand and generate human language and have demonstrated strong performance in conversational applications such as chatbots [7].

Several research has highlighted the potential of LLMs across various natural language processing tasks, including specialized applications in fields such as medical and health sciences [8][9] and politics [10]. Built upon the transformer architecture [11], LLMs leverage self-attention mechanisms to model contextual relationships across tokens, enabling fluent and context-aware dialogue generation [12][13][14]. Modern LLMs support multilingual interaction, including Thai, allowing them to generate responses across a wide range of languages using shared representations learned during pretraining. However, despite their generative capabilities, LLMs are fundamentally limited to the knowledge encoded in their training data and lack inherent access to domain-specific, up-to-date, or localized information. As a result, LLM-based chatbots may produce responses that are plausible but inaccurate [15] when applied to specialized domains without external grounding.

Retrieval-Augmented Generation (RAG) is a framework designed to address the knowledge limitations of large language models by integrating information retrieval with text generation. Within the RAG framework, a retriever component first identifies relevant documents from an external knowledge base, and a generator then conditions its response on the retrieved content to produce grounded and contextually informed outputs. The RAG framework was originally introduced to improve performance on knowledge-intensive tasks by enabling language models to access non-parametric memory during inference, thereby improving factual accuracy and reducing hallucination [6]. RAG has since been widely adopted in domain-specific applications where reliability, updateability, and traceability of information are critical, making it particularly suitable for conversational systems operating on curated and authoritative knowledge sources.

Within the framework, the choice of retrieval strategy plays a critical role in determining the relevance and quality of the contextual information provided to the generator. For small, curated, and authoritative document collections, sparse retrieval methods such as term frequency–inverse document frequency (TF-IDF) and Best Matching 25 (BM25) remain effective due to their simplicity, interpretability, and low computational cost [16][17]. These methods rely on lexical matching and do not require large training corpora or specialized hardware, making them suitable for domain-specific knowledge bases with limited scale. While dense retrieval and approximate nearest neighbor search offer advantages for large and semantically diverse corpora, their benefits are less pronounced in small, well-defined datasets, where sparse retrieval can achieve competitive performance with substantially lower complexity [17].

2.2 Conversational AI for Agricultural Advisory Systems

Digital agricultural advisory systems have been widely studied as a means to improve farmers' access to agronomic knowledge, particularly in regions where in-person extension services are limited [18]. Early systems primarily relied on rule-based decision support, frequently asked question (FAQ) interfaces, or SMS and IVR platforms to disseminate farming recommendations [19]. While these approaches improved information accessibility, they typically offered static responses and limited adaptability to diverse user contexts. More recent work has explored conversational and generative chatbot systems to provide interactive and personalized agricultural advice, including deployments targeting smallholder farmers in developing and resource-constrained settings [20]. However, existing agricultural chatbots continue to face challenges related to factual reliability, localization of recommendations, and integration of authoritative domain knowledge, highlighting the need for grounded conversational systems that can support trustworthy agricultural decision-making.

With recent advances in large language models, generative conversational agents have been increasingly explored for knowledge-intensive advisory tasks, including applications in agriculture, healthcare, and education [21]. While such models are capable of producing fluent and contextually appropriate responses, their reliance on parametric knowledge alone raises concerns regarding factual accuracy and hallucination, particularly in high-stakes advisory domains [22]. To address these limitations, retrieval-augmented generation has emerged as a promising approach by grounding responses in external, authoritative knowledge sources, thereby improving reliability and transparency [6]. Prior studies in domain-specific question answering demonstrate that RAG-based systems are better suited for scenarios where correctness, traceability, and updateability of information are essential, making them a natural fit for agricultural advisory applications.

In the Thai agricultural research context, smart agriculture systems have predominantly focused on IoT-based sensing, monitoring, and automation to support precision farming and resource management [23][24]. While these systems play an important role in data collection and operational decision support, they are typically designed as monitoring platforms or rule-based information systems rather than providing ground knowledge [25] or conversational advisory tools. Several chatbot-based applications have been reported in the academic literature, including task-specific systems such as rice disease diagnosis delivered via a chatbot interface [26], LINE-based chatbots integrated with smart agriculture platforms [27], and prototype chatbots designed for automatic responses to narrowly defined agricultural queries such as pest information and price prediction [28]. However, these systems primarily employ rule-based or task-oriented interaction and are not designed as general-purpose conversational advisory agents that integrate and reason over authoritative knowledge sources. Moreover, Thai natural language processing remains challenging due to limited annotated resources and the reliance of multilingual language models on generalized training data, which may not adequately capture local agricultural practices and terminology [29]. These factors highlight the need for localized and knowledge-grounded conversational systems tailored to the Thai agricultural context.

Building upon established research on conversational AI and RAG, this work adapts a globally recognized paradigm to the specific context of Thai rice farming. Unlike prior Thai agricultural information systems that focus primarily on sensing or rule-based interaction, the proposed system integrates a generative chatbot with a curated and authoritative Thai knowledge base to provide grounded and contextually relevant advisory responses. Through this design, the work addresses a gap between global advances in conversational advisory systems and their practical application within localized Thai agricultural settings.

3. Development of Chatbot for Thai Rice-related Knowledge

Our objective in this work is to develop a Thai-based chatbot for Thai rice-related knowledge using RAG. This chatbot is designed for single turn approach which refers to an interaction where the user provides a single query, and the system responds with a single output. There is no back-and-forth dialogue, and the exchange is completed in one turn and will not memorize

information from previous exchanges or maintain context across multiple inputs for sequential conversation. The chatbot is designed to provide information related to rice farming to farmers in Thai natural language based on users' query. The providable information includes rice variety and its specification, rice disease and pest and their symptom/treatment and how to manage rice paddy field. The chatbot system contains 3 main components including external information source, LLM, and RAG as shown in Figure 1.

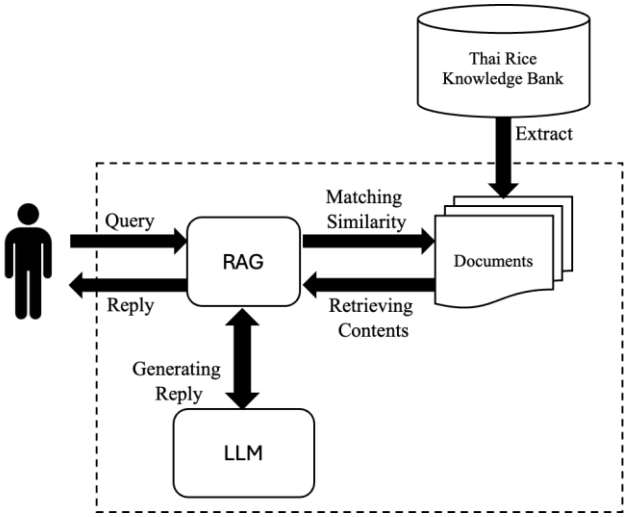


Figure 1: Overall Processes of Chatbot for Thai Rice-related Knowledge

For external documents as extra Thai rice-related knowledge, public data from Department of Rice, Thailand called Rice Knowledge Bank (RKB) [20] are collected. Data in RKB are categorized into groups of topics such as Rice Variety, Rice Pest, Rice Technology, Production and Farming Techniques, and Fertilizer. Each category contains several subcategories, and their leaf pages are separated to a different specific topic containing Thai descriptive content and images as an example for rice variety/pest or infographic for techniques/methods. To use RKB pages as external documents for text-based chatbot, we preprocess the data by removing images and separate each leaf page into a text document. There are 293 documents, and top 3 categories are Rice Variety, Rice Disease and Rice Pest, respectively. These data are used in both finetuning and retrieval phase.

The chosen base LLM in this work is Qwen2.5-7B [30], which uses transformer architectures with RoPE, SwiGLU, RMSNorm, and Attention QKV bias, has 7.61B parameters and supports Thai language. The Qwen 2.5-7B model was chosen as the base LLM due to its strong multilingual capability and competitive performance across non-English benchmarks. According to the official Qwen 2.5 Technical Report [30], Qwen 2.5 models demonstrate competitive multilingual performance on standard evaluation benchmarks such as MMLU, XCOPA, and FLORES-200, achieving scores comparable to or exceeding models of similar size including Llama 3 8B and Gemma 7B. Although Thai-specific results are not separately reported, the benchmark coverage includes Thai as one of the supported languages. These observations together with its open-weight availability and LoRA compatibility motivated its adoption as the base model for this study.

For fine-tuning, the chosen dataset is ThaiQA-SQuAD [31], which is a question-answering dataset in Thai based on Stanford Question Answering dataset structure and the collected RKB data. These datasets help the LLM to acquire greater understanding of Thai jargons as well as a nuance of question-answer patterns of natural Thai. The Qwen2.5-7B model was fine-tuned using the Low-Rank Adaptation (LoRA) [32] approach on a dataset combining ThaiQA-SQuAD and the curated RKB documents. Training was conducted on a single NVIDIA A100 GPU (40 GB VRAM). The model was optimized with the AdamW algorithm and a linear learning-rate scheduler, using a learning rate of approximately 2×10^{-5} and mini-batches of eight samples across three epochs. A dropout rate of about 0.05 was applied to mitigate overfitting, and LoRA parameters were configured with a rank of 8 and $\alpha = 16$ to achieve lightweight adaptation while preserving the base model's stability. During training, the loss consistently decreased across epochs and stabilized near 0.87 after the second epoch, where the validation loss plateaued. This indicates satisfactory convergence for domain adaptation without evidence of overfitting, suitable for a lightweight fine-tuning objective emphasizing domain alignment rather than full model retraining.

LLMs rely on pre-trained knowledge given in a training data. The training data that may not include domain specific knowledge or become outdated over time. RAG thus can help to provides up-to-date information without heavy task on retraining the model and can minimize the risk of generating false or misleading information. In our RAG, the knowledge of the base LLM is augmented with RKB information as an external knowledge base. Since the task in this work is limited to the Thai rice related information with less than 300 documents as a small specific task, we decide to apply the approach of Sparse Vector Representations as it is an efficient approach for small datasets and can cover lexical matching. In the retrieval process, it identifies and retrieves the most relevant pieces of information using BM25 (best matching 25) method [33] to find relevant content from the knowledge base following a given user's query. BM25 ranks the given external documents based on their relevance to the query rather than just matching keywords by computing a similarity score between the query and each document to ensure more context-aware results. Furthermore, BM25 uses a logarithmic scaling of term frequency instead of raw counts to prevent bias toward documents with excessive keyword repetition, unlike TF-IDF which can excessively favor frequent terms.

The prototype system was deployed on a workstation equipped with an NVIDIA A100 GPU (40 GB VRAM), 48 GB system memory, and an AMD EPYC 7313 16-core CPU. Model inference was executed through a Python-based API using the Hugging Face Transformers framework. Average response time measured across 70 test queries was 2.9 seconds per query (standard deviation ± 0.6 s) for the RAG-based chatbot and 2.1 seconds for the baseline without RAG. The extra latency primarily results from retrieval and document-embedding operations. These show that even with the retrieval overhead, the RAG-based chatbot

maintains a practical response time suitable for near-real-time agricultural information services. During inference, the system's GPU memory usage remained below 28 GB, allowing deployment on a single GPU workstation.

4. Evaluation

4.1 Evaluation Overview and Design Rationale

This section introduces the evaluation design adopted in this study. The proposed chatbot is a retrieval-augmented generation (RAG) system intended to provide reliable and practical rice-farming knowledge to Thai farmers by grounding its responses in an authoritative document corpus. As such, a single evaluation method is insufficient to capture both real-world usefulness and system-level grounding behavior. To address this, we adopt a two-tier evaluation protocol, where each tier serves a distinct purpose and answers a different evaluation question.

The first tier focuses on practical validity under real usage conditions. This evaluation is conducted using real queries collected from rice farm workers and farm owners, reflecting authentic information needs encountered in practice. Because such queries are open-ended, context-dependent, and may not have a single canonical textual answer, this tier relies on expert judgment to assess whether the chatbot's responses are acceptable and useful in real agricultural settings. The goal of this evaluation is not to benchmark linguistic similarity, but to determine whether the system can deliver responses that are considered correct and appropriate by domain experts.

The second tier focuses on grounding robustness and retrieval correctness. While expert evaluation provides insight into real-world usability, it does not directly reveal whether the chatbot consistently grounds its responses in the correct authoritative sources. To examine this aspect, we conduct a controlled evaluation using a larger set of curated queries that are explicitly designed to be answerable from the Rice Knowledge Bank. This evaluation emphasizes document attribution and grounding behavior, analyzing whether the system retrieves and utilizes the correct knowledge documents when generating responses. This tier therefore assesses the reliability of the RAG mechanism itself, independent of surface-level language variation.

By separating evaluation into these two complementary tiers, we avoid over-reliance on a single metric or perspective. Expert judgment is used where no strict ground truth exists, while grounding-based analysis is applied where authoritative references can be clearly defined. This design ensures that the evaluation remains aligned with the practical goals of an agricultural advisory chatbot while still providing objective evidence of system behavior at scale.

4.2 User-Driven Evaluation

This evaluation focuses on assessing the chatbot's performance under realistic usage conditions using queries derived from actual rice farming practices. The objective of this evaluation is to examine whether the chatbot can generate responses that are considered accurate, relevant, and practically useful by domain experts when answering real-world questions posed by rice farm workers and farm owners.

A total of 70 queries related to Thai rice farming were collected from rice farm workers and farm owners. These queries reflect common information needs encountered in practice, including rice variety characteristics, pest and disease management, and field management techniques. Unlike synthetic or researcher-generated benchmarks, these queries were not constructed to target specific documents but rather represent authentic user inquiries.

The queries were categorized into two levels of complexity. Simple queries consist of questions focusing on a single topic or a specific factual aspect that can typically be answered using one knowledge source, such as "How many amyloses does Khao Khan Chan rice variety have?" or "What pests should be monitored when growing RD15 rice?". Complex queries involve multiple aspects or require synthesis across several topics, for example "What rice varieties are suitable for cultivation in the Northeastern region during the rainy season?" or "What integrated pest management strategies are appropriate for rice blast disease in organic farming?". In total, 50 queries were classified as simple and 20 as complex. The classification was jointly reviewed by two agricultural experts to ensure consistency.

Since these real-world queries are open-ended and may not have a single canonical textual answer, this evaluation does not rely on automatic text-matching metrics. Instead, the assessment is based on expert judgment, reflecting how such responses would be evaluated in real agricultural advisory contexts.

The chatbot responses were evaluated by three academic experts, comprising two senior officers from the Department of Rice, Thailand, and one university lecturer specializing in agricultural science. All evaluators have at least ten years of experience in rice production, pest management, and agricultural knowledge transfer. To reduce potential bias, all responses were anonymized before evaluation, and the experts were unaware of whether a response was generated by the baseline chatbot or the RAG-enhanced system. Each response was rated using the following criteria:

- **Agree**: The response is accurate, relevant, and sufficiently addresses the query
- **Partially agree**: The response is generally accurate but contains minor irrelevant information or incomplete details
- **Disagree**: The response is inaccurate or does not appropriately address the query

For each query, the final rating was determined using a majority voting mechanism among the three experts. In cases where no majority decision was reached, the response was assigned a "Partially agree" rating as a middle ground. This approach ensures that the evaluation outcome is not overly influenced by a single evaluator's judgment.

Inter-rater reliability was assessed using Fleiss' κ coefficient, yielding a value of approximately 0.78, which indicates substantial agreement among the evaluators.

4.2.1 Evaluation Results

Table 1 presents the evaluation results comparing the baseline chatbot and the RAG-based chatbot across simple and complex query types. The results show that the chatbot incorporating retrieval-augmented generation consistently outperforms the baseline system.

For simple queries, the RAG-based chatbot achieved 96% "Agree" ratings, compared to 88% for the baseline. For complex queries, the improvement is more pronounced, with the RAG-based system achieving 80% "Agree" ratings compared to 60% for the baseline. Notably, the RAG-based chatbot produced no "Disagree" responses for either query type, whereas the baseline system generated several responses rated as inaccurate.

Overall, the RAG-based chatbot obtained 64 "Agree" responses (91.43%) and 6 "Partially agree" responses (8.57%), with no responses rated as "Disagree". In contrast, the baseline system yielded 56 "Agree" responses (80.00%), 11 "Partially agree" responses (15.71%), and 3 "Disagree" responses (4.29%). These results indicate that incorporating retrieval-augmented generation improves the practical usefulness and reliability of the chatbot when responding to real-world user queries.

Table 1: Evaluation results of the chatbot systems

System	Query type	Agree	Partially agree	Disagree
Without RAG (baseline)	Simple Query (50)	44 (88%)	5 (10%)	1 (2%)
	Complex Query (20)	12 (60%)	6 (30%)	2 (10%)
With RAG	Simple Query (50)	48 (96%)	2 (4%)	0
	Complex Query (20)	16 (80%)	4 (20%)	0

4.2.2 Qualitative Analysis and Example Responses

To further illustrate the difference between the two systems, Table 2 presents an example query along with responses generated by the baseline chatbot and the RAG-based chatbot. While both systems identify relevant pests associated with the queried rice variety, the RAG-based chatbot provides more detailed and actionable information, including management strategies grounded in authoritative rice knowledge sources.

A closer examination of responses rated as "Disagree" in the baseline system reveals that many errors stem from confusion between similar rice varieties or pests with closely related names. In several cases, factual attributes such as amylose content or pest characteristics were incorrectly mixed. These issues are largely mitigated in the RAG-based system, which grounds its responses in retrieved documents from the Rice Knowledge Bank.

This evaluation demonstrates that, under realistic usage conditions, the RAG-enhanced chatbot is better able to generate responses that are considered acceptable and useful by agricultural experts. However, this user-driven evaluation does not directly assess whether the system consistently retrieves the correct underlying documents. This aspect is examined separately in the grounding-based evaluation presented in the following section

4.3 Grounding-Based Evaluation

While the user-driven evaluation assesses the chatbot's practical usefulness under real-world conditions, it does not directly examine whether the system consistently grounds its responses in the correct authoritative knowledge sources. To analyze this aspect, we conduct a grounding-based evaluation focusing on document attribution behavior within the retrieval-augmented generation process.

The objective of this evaluation is to examine whether the chatbot retrieves and utilizes the appropriate documents from the Rice Knowledge Bank when answering domain-specific queries. Unlike the previous evaluation, which relies on expert judgment of response quality, this evaluation emphasizes retrieval correctness and grounding reliability at the document level.

4.3.1 Evaluation Setting and Design

To evaluate grounding behavior under controlled conditions, a controlled query set was constructed based on realistic agricultural information needs while ensuring explicit correspondence to the Rice Knowledge Bank. Rather than generating queries arbitrarily, the construction process was anchored in real user queries described in Section 4.2 to preserve practical relevance.

First, 25 queries were randomly selected from the original set of 70 real-world queries and used as seed queries for controlled augmentation. These seed queries were expanded through a semi-automatic query augmentation process using a large language model. The model was instructed to generate realistic variations of each seed query while preserving the original intent and ensuring that all queries remained answerable using the Rice Knowledge Bank. This process produced 250 candidate queries by varying attributes such as rice variety, cultivation region, season, pest or disease type, and management aspect. During augmentation, the Rice Knowledge Bank document identifiers used as the basis for each generated query were recorded to establish intended reference provenance.

All generated queries were manually reviewed, and queries that were duplicated, unclear, or could not be grounded in required Rice Knowledge Bank documents were excluded. After filtering, 217 queries remained and were used as the final controlled query set for grounding-based evaluation. Unlike the real-world query set used in Section 4.2, all queries in this controlled set were explicitly designed to correspond to identifiable documents in the Rice Knowledge Bank, enabling systematic evaluation of document grounding behavior.

Table 2: Example of Query and Difference of Responses from the two

Type	Thai (Original Output)	English (Literary-translated)
Query	ศัตรูพืชอะไรบ้างที่ต้องระวังในการปลูกข้าว กข15	What pests should be watched out for when growing rice variety RD15?

Response from baseline	เพลี้ยกระโดดสีน้ำตาลดูดกินน้ำเลี้ยงทำให้ต้นข้าวเหลืองเฉาและตายเป็นหย่อมๆ หนอนกอข้าวเจาะเข้าไปในลำต้นทำให้ใบยอดเหี่ยวและรวงข้าวลีบ และหนอนม้วนใบข้าวกัดกินเนื้อใบ	Brown planthoppers suck sap from rice plants, causing yellowing, wilting, and localized death. Rice stem borers burrow into the stem, causing wilting leaves and shriveled rice ears. Rice leaf rollers feed on the leaf tissue.
Response from system with RAG	ศัตรูพืชที่ควรระวัง 1. เพลี้ยกระโดดสีน้ำตาล ลักษณะการทำลาย: ดูดกินน้ำเลี้ยงจากต้นข้าว ทำให้ต้นเหลืองเฉาและตายเป็นหย่อม (อาการไหม้) การจัดการ: ใช้ศัตรูธรรมชาติ เช่น แมลงเต่าทองหรือแตนเบียน หลีกเลี่ยงการใช้ปุ๋ยไนโตรเจนมากเกินไป ใช้สารเคมีเมื่อระบาดหนัก 2. หนอนกอข้าว ลักษณะการทำลาย: เจาะเข้าไปในลำต้น ทำให้ใบยอดเหี่ยว (อาการยอดเหี่ยว) และรวงลีบ การจัดการ: ไถกลบฟางข้าวหลังเก็บเกี่ยว เพื่อลดการสะสมตัวอ่อน ใช้สารชีวภัณฑ์ เช่น เชื้อบีที 3. หนอนม้วนใบข้าว ลักษณะการทำลาย: ม้วนใบข้าวเป็นหลอดและกัดกินเนื้อใบ การจัดการ: ใช้กับดักแสงไฟเพื่อล่อผีเสื้อ ฉีดพ่นสารชีวภาพหรือสารเคมีตามคำแนะนำ	Pests to watch out for 1. Brown planthopper Damage: Sucks sap from rice plants, causing yellowing, wilting, and localized death (burnt appearance) Management: Use natural predators such as ladybugs or parasitic wasps. Avoid excessive nitrogen fertilizer. Apply chemicals when infestation is severe. 2. Rice stem borer Damage: Burrows into the stem, causing wilting of the top leaves (wilting symptoms) and shriveled rice ears.) Management: Plow rice straw after harvest to reduce larvae buildup. Use biopesticides like Bacillus thuringiensis (BT). 3. Rice leaf roller Damage: Rolls rice leaves into tubes and feeds on the leaf tissue. Management: Use light traps to attract moths. Spray biological or chemical pesticides as recommended.)

For each query, a reference document set was defined to represent the required knowledge necessary to answer the query correctly. Candidate reference documents were derived from the document identifiers recorded during query augmentation and were subsequently verified by the authors to confirm that the referenced documents directly support the informational intent of the query. Queries associated with incorrect, ambiguous, or insufficient references were excluded during this verification process. The resulting reference sets may consist of a single document or multiple documents depending on the nature of the query. All references are defined at the document level, and no passage-level annotation is applied.

Grounding correctness is evaluated by comparing the reference document set for each query with the documents retrieved by the chatbot during response generation. Evaluation is performed using recall-based document identity matching, focusing on whether the system retrieves the required documents. Retrieved documents that are not part of the reference set are not penalized, as the evaluation emphasizes grounding sufficiency rather than retrieval exclusivity.

Grounding performance is reported using two complementary aggregations. A macro-averaged grounding score is computed by averaging the per-query grounding scores, reflecting typical query-level grounding behavior. In addition, a micro-averaged document-level grounding recall is computed over all reference documents to reflect overall grounding coverage across the corpus. Formally, grounding recall is defined as:

$$GroundingRecall = \frac{\sum_{q \in Q} |G(q) \cap R(q)|}{\sum_{q \in Q} |G(q)|} \quad (1)$$

where $G(q)$ denotes the set of required reference documents for query q, and $R(q)$ denotes the set of documents retrieved by the system, with evaluation focusing on coverage of required grounding documents rather than linguistic similarity.

4.3.2 Grounding Evaluation Results and Analysis

Grounding evaluation was conducted on the controlled query set consisting of 217 queries constructed as described in Section 4.3.1. The evaluation examines the system's ability to retrieve the required Rice Knowledge Bank documents necessary to ground each response.

Across the controlled query set, queries vary in the number of required reference documents. A substantial proportion of queries require only a single reference document, while the remaining queries involve multiple documents. Queries with two or more required reference documents represent more complex grounding cases, as correct grounding depends on retrieving all relevant documents rather than a single source. The distribution shown in Table 3 provides context for interpreting the grounding performance. Grounding performance results are summarized in Table 4.

Table 3: Distribution of reference documents per query

Reference Docs	Queries	Proportion (%)
1 document	145	66.80%
2 documents	38	17.50%
≥3 documents	34	15.70%
Total	217	100%

Table 4: Grounding evaluation results

Metric	Aggregation	Value
Grounding recall	Macro (per query)	0.983
Grounding recall	Micro (per document)	0.896
Queries	—	217

The macro-averaged grounding score, computed by averaging per-query grounding recall, is 0.983. This indicates that, for a typical query, the system retrieves most or all of the required reference documents. In contrast, the micro-averaged grounding recall, computed over all reference documents across the query set, is 0.896. This value reflects overall grounding coverage at the document level and is more sensitive to partial grounding in queries requiring multiple reference documents.

The difference between macro- and micro-averaged scores can be attributed to the composition of the controlled query set. Queries requiring a single reference document contribute equally to both aggregations and tend to achieve full grounding. In contrast, queries requiring multiple reference documents are more sensitive to partial retrieval, where one required document may be omitted while others are successfully retrieved. These partial grounding cases contribute disproportionately to the reduction in micro-averaged recall.

Qualitative inspection of grounding outcomes indicates that most partial grounding cases involve the omission of a single supplementary document rather than complete grounding failure. In addition, when additional non-reference documents are retrieved, they are typically conceptually related to the required references and belong to the same knowledge category within the Rice Knowledge Bank. Such over-retrieval does not contradict the required grounding and does not prevent retrieval of the core reference documents.

Overall, the grounding evaluation shows that the retrieval-augmented chatbot consistently retrieves the documents required to ground its responses, with remaining errors primarily arising from incomplete coverage in multi-document queries rather than retrieval of unrelated or contradictory information.

5. Discussion

This work develops a generative chatbot for providing rice farming information to Thai rice farmers. We compare a baseline system using a base large language model (LLM) with a retrieval-augmented generation (RAG)–based system that incorporates an external domain-specific knowledge base. The results indicate that the generative chatbot using RAG outperforms the baseline in terms of providing more relevant information and including more significant domain-specific details in its responses. All responses generated by the RAG-based chatbot were accepted by the evaluators, whereas some responses from the baseline system were deemed factually incorrect or fabricated, reflecting extrinsic hallucination behavior commonly observed in generative models without external grounding.

Beyond overall response quality, grounding behavior provides an important signal of system reliability in agricultural advisory settings. The grounding-based evaluation shows that most remaining errors arise from incomplete document coverage in queries requiring multiple reference documents, rather than from retrieval of unrelated or contradictory information. This distinction is particularly important in agricultural decision support, where partially grounded but contextually relevant information is preferable to ungrounded or fabricated responses. The observed grounding behavior suggests that document-level grounding evaluation is an appropriate mechanism for assessing reliability in curated agricultural knowledge systems.

The RAG-based generative chatbot thus demonstrates the capability to answer questions that exceed the model's internal training data, which is particularly beneficial for specialized domains where accuracy and domain-specific knowledge are critical. By grounding generated content in external documents, RAG reduces the risk of hallucination and helps maintain user trust in the system. These results reinforce the role of RAG as a key component in improving the usefulness and trustworthiness of generative chatbots for agricultural information services.

Although the Rice Knowledge Bank contains only 293 documents, the corpus was deliberately selected for its reliability and Thai-language coverage of rice production topics issued by the Department of Rice. In addition to serving as an authoritative knowledge source, these documents encode localized agricultural knowledge that reflects Thailand-specific rice varieties, cultivation practices, climate conditions, and pest management strategies. Without such localized grounding, a base large language model may rely on generalized patterns learned from large-scale multilingual or global data, which may not be compatible with Thai agricultural contexts and could lead to inappropriate or misleading recommendations. The intent of this study was therefore to demonstrate a generative chatbot integrating verified and localized external agricultural knowledge, rather than to construct a large-scale information retrieval system. The relatively small corpus enabled manual quality checking and ensured that domain terminology, language usage, and contextual relevance were appropriate for both fine-tuning and retrieval.

At the same time, this design choice introduces limitations related to scalability and dependency on trustworthy external documents. While manual verification is feasible for a small, curated corpus such as the Rice Knowledge Bank, scaling to larger datasets may require additional mechanisms such as automated fact-checking, reranking, or contextual validation to ensure information quality. Reranking strategies, for example, can reorder retrieved documents based on relevance or authority before

response generation, but implementing such techniques in Thai-language settings may require additional resources, including sufficiently large Thai-language corpora for model training or fine-tuning. Moreover, although the chosen base LLM supports Thai as part of its multilingual capability, Thai is not its primary training focus, which may result in responses that are understandable but occasionally lack naturalness. Addressing these challenges and improving Thai-language generation quality remain important directions for future work.

6. Conclusion

This paper presents the development of a generative chatbot to support Thai rice farmers by integrating retrieval-augmented generation with verified external agricultural knowledge. The proposed system is built upon a lightweight pretrained large language model, Qwen2.5-7B, and incorporates domain-specific Thai rice farming documents both for fine-tuning and as a retrieval corpus. By grounding generated responses in localized external knowledge, the chatbot is able to provide contextually relevant and domain-specific information beyond the model's internal training data.

The chatbot was evaluated using both real user queries and a controlled grounding-based evaluation. In user-driven evaluation, the RAG-based system produced acceptable responses for all test queries, whereas the baseline model exhibited factual errors and hallucinated content in several cases. The grounding-based evaluation further shows that the system retrieves the required reference documents for the majority of queries, achieving high grounding recall at both the query and document levels, with remaining errors primarily arising from incomplete document coverage in multi-document queries rather than ungrounded or contradictory generation. These findings highlight the importance of document-level grounding as a reliability signal for generative chatbots in agricultural advisory contexts.

The use of a curated and localized knowledge base is shown to be particularly important for Thai agricultural applications, where generalized knowledge from large-scale multilingual data may not align with local farming practices, crop varieties, or environmental conditions. This study therefore supports the adoption of retrieval-augmented generation with localized and trustworthy external knowledge as a practical approach for deploying generative AI systems in domain-specific and region-sensitive settings.

For future work, we plan to explore graph-based retrieval-augmented generation to represent structured agricultural knowledge more explicitly and to further improve grounding across complex queries. In addition, the use of Thai-specific pretrained language models or large-scale Thai-language corpora for further model adaptation remains an important direction for improving the naturalness and effectiveness of Thai-language response generation.

Disclaimers

This research does not require human research ethics approval for the following reasons: 1. The chatbot system development uses publicly available data from the Rice Knowledge Bank of the Department of Rice, without directly collecting data from human subjects. 2. The system evaluation was conducted by three academic experts, assessing the quality of system-generated responses, without experimentation or data collection from actual farmers or end users. 3. This research focuses on system development and testing, specifically improving the chatbot's question-answering capabilities using RAG technique compared to the baseline system, without studying human impact. 4. There is no collection of personal data, no risks to human rights or welfare, and no experiments that could potentially harm research participants. Therefore, this research does not fall under the scope requiring human research ethics approval, as it is a technical research project developing and testing an AI system using existing public data.

References

[1] Siripong A. Analysis of government policies on rice production in Thailand: a policy evaluation study. International Journal of Agriculture. 2024;9(1):9–21.

[2] Chandolikar N, Dale C, Koli T, Singh M, Narkhede T. Agriculture assistant chatbot using artificial neural network. In: Proceedings of the 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA). IEEE; 2022. p. 1–5. doi:10.1109/ICACTA54488.2022.9753433

[3] Gupta S, Chen Y. Supporting inclusive learning using chatbots: A chatbot-led interview study. Journal of Information Systems Education. 2022;33(1):98–108.

[4] [4] Suebsombut P, Sureephong P, Sekhari A, Chernbumroong S, Bouras A. Chatbot application to support smart agriculture in Thailand. In: Proceedings of the 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE; 2022. p. 364–367. doi:10.1109/ECTIDAMTNCON54251.2022.9720318

[5] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. arXiv preprint arXiv:1801.06146. 2018.

[6] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401. 2020.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

[8] Ding J, Li B, Xu C, Qiao Y, Zhang L. Diagnosing crop diseases based on domain-adaptive pre-training BERT of electronic medical records. Applied Intelligence. 2022. doi:10.1007/s10489-022-04346-x

[9] Khare Y, Bagal V, Mathew M, Devi A, Deva Priyakumar U, Jawahar CV. MMBERT: Multimodal BERT pretraining for improved medical VQA. arXiv preprint arXiv:2104.01394. 2021.

[10] Liu R, Jia C, Wei J, Xu G, Wang L, Vosoughi S. Mitigating political bias in language models through reinforced calibration. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(17):14857–14866. doi:10.1609/aaai.v35i17.17602

[11] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2019.

[12] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33.

[13] Hanslo R. Deep learning transformer architecture for named entity recognition on low-resourced languages. arXiv preprint arXiv:2111.00830. 2021.

[14] Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, et al. Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787. 2019.

[15] Maryamah M, Irfani MM, Raharjo EBT, Rahmi NA, Ghani M, Raharjana IK. Chatbots in academia: A retrieval-augmented generation approach for improved efficient information access. In: Proceedings of the 16th International Conference on Knowledge and Smart Technology (KST). IEEE; 2024. p. 259–264. doi:10.1109/KST60798.2024.10499652

[16] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval. 2009;3(4):333–389. doi:10.1561/1500000019

[17] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008. ISBN: 9780521865715

[18] Aker JC. Dial "A" for agriculture: Using ICTs for agricultural extension. Agricultural Economics. 2011;42(6):631–647. doi:10.1111/j.1574-0862.2011.00545.x

[19] Qiang C, Kuek SC, Dymond A, Esselaar S. Mobile applications for agricultural extension services. World Bank; 2012. (Report — DOI not assigned)

[20] Farmer.Chat. Scaling AI-powered agricultural services for smallholder farmers. arXiv preprint. 2024.

[21] Ibrahim M, et al. Evaluating ChatGPT responses to farmers' questions on rice cultivation. Scientific Reports. 2024.

[22] Arreerard R, Mander S, Piao S. A survey on Thai language resources and tools. In: Proceedings of the Language Resources and Evaluation Conference (LREC). 2022.

[23] Boonying S. A smart water management in a paddy field using IoT technology and machine learning. Information Technology Journal KMUTNB. 2021;17(2):1–10.

[24] Somsuphaprungyos S. Centralizing real-time data using remote sensing towards smart farming applications in a public area: A case study of Ayutthaya. Information Technology Journal KMUTNB. 2022;18(2):1–12.

[25] Ruangrajitpakorn T, Prombut C, Supnithi T. Development of an ontology-based personalized web from a rice knowledge website. In: Proceedings of the 13th International Conference on Knowledge, Information and Creativity Support Systems (KICSS). IEEE; 2018. doi:10.1109/KICSS.2018.8950556

[26] Temniranrat P, Kiratiratanapruk K, Kitvimonrat A, Sinthupinyo W, Patarapuwadol S. A system for automatic rice disease detection from rice paddy images serviced via a chatbot. Computers and Electronics in Agriculture. 2021;185:106156. doi:10.1016/j.compag.2021.106156

[27] Suebsombut P, Sureephong P, Sekhari A, Chernbumroong S, Bouras A. Chatbot application to support smart agriculture in Thailand. arXiv preprint arXiv:2308.02524. 2023.

[28] Hongtong N. Prototype chatbot for automatic responses on coconut pest and price prediction. 2023.

[29] Rice Knowledge Bank. Thai Rice Knowledge Base. Department of Rice, Thailand. Available from: https://newwebs2.ricethailand.go.th/webmain/rkb3/

[30] Team Q. Qwen2.5: A Party of Foundation Models!. Qwen Blog. 2024. Available from: https://qwenlm.github.io/blog/qwen2.5/

[31] PyThaiNLP. ThaiQA SQuAD Dataset. GitHub repository. 2020. Available from: https://github.com/PyThaiNLP/thaiqa_squad

[32] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.