White Rose Research Online
eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/240241/

Version: Accepted Version

Article:

Liu, L., Li, S., Qi, J. et al. (2026) TARAG: A time-aware retrieval-augmented generation framework for supporting precision crop pest and disease management through large language models. Computers and Electronics in Agriculture, 248. 111786. ISSN: 0168-1699

https://doi.org/10.1016/j.compag.2026.111786

© 2026 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in Computers and Electronics in Agriculture is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown

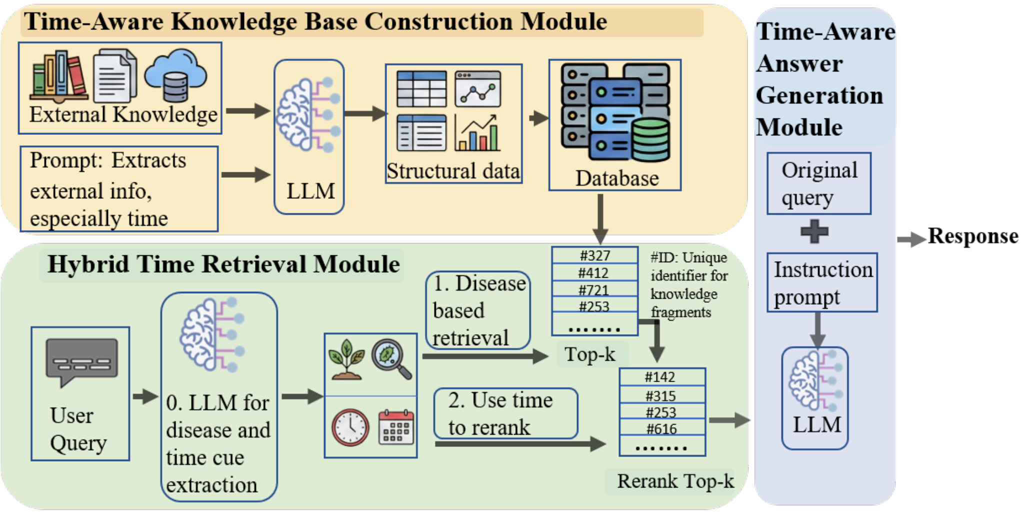
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Graphical Abstract

TARAG: A Time-aware Retrieval-augmented Generation Framework for Supporting Precision Crop Pest and Disease Management through Large Language Models

Lei Liu, Shunbao Li, Jun Qi, Zhipeng Yuan, Po Yang



Highlights

TARAG: A Time-aware Retrieval-augmented Generation Framework for Supporting Precision Crop Pest and Disease Management through Large Language Models

Lei Liu, Shunbao Li, Jun Qi, Zhipeng Yuan, Po Yang

- TARAG introduces a time-aware RAG framework for pest and disease management.
- A time-annotated knowledge base module structures crop stages and pest lifecycles.
- Hybrid retrieval with time-aware re-ranking improves time-aligned evidence retrieval.
- TARAG achieves 99.14% retrieval recall and improves time-consistent decisions.
- TAQA is a new bilingual dataset with 30k+ QA pairs covering 2k+ pest and disease entities.

TARAG: A Time-aware Retrieval-augmented Generation Framework for Supporting Precision Crop Pest and Disease Management through Large Language Models

Lei Liu[a], Shunbao Li[b], Jun Qi[c], Zhipeng Yuan[b,*], Po Yang[b,*]

[a] *School of Software, Yunnan University, Kunming, 650091, China*
[b] *School of Computer Science, University of Sheffield, Sheffield, S10 2TN, United Kingdom*
[c] *Department of Computing, Xian JiaoTong-Liverpool University, Suzhou, 215123, China*

Abstract

Plant diseases and pests cause significant annual crop losses, severely threatening global food security. To mitigate crop losses, large language models (LLMs) demonstrate the ability to alleviate challenges in accessing precise farming support by acting as intelligent assistants that provide farmers with timely and precise decision support services. However, existing efforts still fall short in precision due to the neglect of time-sensitive nature in agricultural practices. They implicitly treat domain knowledge as time-irrelevant, ignoring the impact of crop phenology and pest life stages in their generated recommendations. To address this challenge, we propose a Time-Aware Retrieval-Augmented Generation framework (TARAG), comprising a time-aware knowledge base construction module, a hybrid retrieval module, and a time-based generation module to provide precise pest management suggestions. Firstly, the time-aware knowledge base construction module constructs a time-annotated knowledge base from unstructured documents to provide supplementary agricultural knowledge. Secondly, the hybrid retrieval moudle performs coarse-grained sparse retrieval to ensure relevance, with a

---

[*] Corresponding author.

Email addresses: `leiliu@stu.ynu.edu.cn` (Lei Liu), `shunbao.li@sheffield.ac.uk` (Shunbao Li), `Jun.Qi@xjtlu.edu.cn` (Jun Qi), `zhipeng.yuan1@sheffield.ac.uk` (Zhipeng Yuan), `po.yang@sheffield.ac.uk` (Po Yang)

time-sensitive re-ranking stage that refines the results to achieve both semantic relevance retrieval and time alignment with user queries. Finally, the time-aware generation module leverages the top-k retrieved documents and an instruction prompt to produce the final suggestion. To validate the effectiveness of the proposed framework, we contribute TAQA, the first bilingual, time-annotated agricultural question-answering dataset. Experiments demonstrate that TARAG significantly outperforms state-of-the-art RAG frameworks in retrieval precision and suggestion quality, with 99.14% retrieval recall and an F1 score of 66.85% for generated suggestions. The implementation and datasets for this work are available in https://github.com/leehash1/agri_rag.

Keywords:
Plant disease and pest management, Retrieval-augmented generation, Time-aware agricultural question answering, Time-sensitive information retrieval

---

1. Introduction

The presence of plant diseases and insect pests constitutes a persistent and severe threat to global agricultural production, with a potential annual yield loss of up to 40% of global crops (Junaid and Gokce, 2024). Such biological stresses inflict substantial economic damage across major cropping systems and, more critically, undermine food security for a growing population. Traditional pest and disease management practices, particularly reliance on broad-spectrum chemical pesticides, have led to additional issues, including environmental contamination, pesticide resistance, and non-target ecological impacts (Abate et al., 2000; Deka et al., 2006; Singh et al., 2024). These issues highlight the urgent need for more precise, timely, and context-aware crop protection strategies that can guide farmers toward effective and environmentally responsible interventions. However, obtaining accurate agronomic recommendations remains difficult in practice. Specifically, expert knowledge is unevenly distributed. Extension resources are also limited. Furthermore, textual agricultural guidelines often contain complex, time-dependent instructions that are difficult to interpret without specialized experience (Li et al., 2024). This gap motivates the development of intelligent systems capable of delivering reliable, time-sensitive decision support for crop health management.

Large language models (LLMs), together with retrieval-augmented generation (RAG) frameworks, offer promising capabilities for addressing the challenges outlined above (Yi et al., 2025). The RAG frameworks work by first searching relevant documents in a knowledge base based on user queries, then feeding those retrieved results into an LLM for obtaining a more accurate response. By combining the retrieval-based factual grounding and adaptive reasoning ability of LLMs, LLM-based RAG frameworks mitigate issues arising from incomplete or outdated model priors, enabling more reliable access to domain knowledge and supporting transparent decision-making pathways. These advantages have motivated growing interest in applying LLM-based RAG frameworks to agricultural scenarios, such as crop disease diagnosis, pest management assistance, and knowledge-driven agronomic advisory tools (Kuska et al., 2024)

For example, Pezego (Yuan et al., 2025) implements a RAG framework with a structured expert knowledge base and a chain-of-thought in an Internet of Things system to provide pest management support. AgriGPT further enhances factual grounding by combining dense, sparse, and knowledge-graph retrieval within a Tri-RAG framework (Yang et al., 2025). In addition, other studies emphasize improving retrieval precision through query expansion, hybrid indexing, or reranking strategies, as summarized in the survey (Vizniuk et al., 2025). Collectively, these systems demonstrate the value of domain-aware retrieval and structured knowledge integration for agricultural question answering.

Meanwhile, recent advances in RAG have begun to incorporate time information across the pipeline (Gao et al., 2025). Representative approaches such as TimeR4 extend the standard framework by integrating time processing into query reformulation, retrieval, and downstream reasoning, enabling time-sensitive question answering. These efforts demonstrate the potential of incorporating time signals within a unified RAG pipeline (Zhang et al., 2024b). However, such approaches are primarily designed for general-domain time reasoning, where time information can be normalized into relatively consistent representations (e.g., explicit time points, intervals, or event sequences) and directly utilized across different stages of the pipeline. In contrast, agricultural knowledge exhibits fundamentally different time characteristics, where time is expressed through heterogeneous and context-dependent forms such as phenological stages, seasonal patterns, and biological life cycles. These forms are often implicitly embedded in domain knowledge and cannot be easily normalized into a unified time representation.

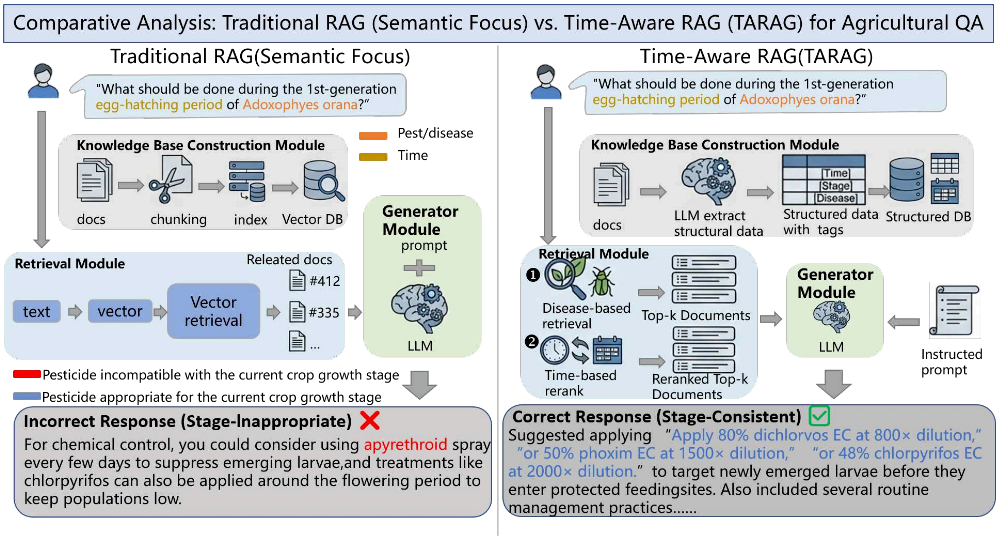


Figure 1: Comparative Analysis of TARAG versus Traditional Retrieval-Augmented Generation. Based on the framework diagram, time sensitivity in agricultural QA involves two aspects. Input time sensitivity means that the user's query often contains explicit or implicit time constraints, such as the 1st-generation egg-hatching, early growth period, or post-harvest management, which directly determine what actions are appropriate. Output time sensitivity requires that the framework's response remains consistent with that specific stage, as the recommended control measures can vary dramatically across different growth or seasonal phases. Traditional RAG frameworks struggle with this because they treat time expressions as ordinary tokens and rely primarily on semantic similarity during retrieval.

This time mismatch can lead to context-insensitive recommendations, as shown in Figure 1. For instance, suggesting pesticides appropriate only for early-instar larvae when the pest population has already reached adulthood (Ye et al., 2025). Therefore, a unified time-aware RAG framework is needed that performs explicit time-based processing from query interpretation and document selection to final response generation, ensuring that recommendations remain consistent with the dynamic biological and ecological processes governing real-world farming.

In response to this demand, we propose TARAG, a unified framework that explicitly enforces time alignment across the entire pipeline for agricultural decision-making. Specifically, TARAG consists of three tightly coupled components: a time-aware knowledge base construction module for time grounding, a hybrid time retrieval module for time alignment, and a time-aware answer generation module for ensuring stage-consistent outputs. This pipeline-level design distinguishes TARAG from existing time-aware RAG approaches that incorporate time signals only in isolated components.

To validate the effectiveness of TARAG, the first time-annotated agricultural question-answering dataset, TAQA, is constructed with approximately 30,196 bilingual QA pairs and time-related metadata, covering over 2,056 types of pests and diseases in agriculture, which serves as a challenge benchmark for evaluating RAG frameworks in agriculture. In the comparison experiments, the TARAG framework outperforms the state-of-the-art RAG framework. In addition, the efficacy of each retrieval stage is demonstrated through ablation experiments.

The main contributions of this paper are as follows:

1) We propose TARAG, a unified retrieval-augmented generation framework designed for time-sensitive agricultural decision support . The framework explicitly integrates time signals, such as crop phenology and pest life cycles, into both the retrieval and generation stages. This framework effectively addresses the chronic issue of time misalignment in traditional RAG systems, ensuring that recommendations are synchronized with real-world biological processes.

2) We develop a hybrid retrieval strategy with a time-aware re-ranking mechanism to improve evidence alignment. It models the time relevance between user queries and candidate documents by fusing sparse lexical matching and dense embeddings. This mechanism significantly mitigates the mismatch between query intent and retrieved knowledge, leading to a marked improvement in retrieval precision for complex farming scenarios.

3) We construct TAQA, a large-scale, bilingual benchmark for time-annotated agricultural question answering . The dataset comprises approximately 30,196 QA pairs across 2,056 pest and disease types, incorporating fine-grained time metadata. Our extensive experiments on this benchmark demonstrate that TARAG outperforms state-of-the-art frameworks across multiple evaluation metrics, providing a robust foundation for future research in the field.

2. Related Work

2.1. Retrieval-Augmented Generation (RAG)

As LLMs demonstrate formidable capabilities in text comprehension and generation, RAG has become a leading paradigm for enhancing the factual reliability and domain adaptability of LLMs. Early RAG frameworks paired sparse lexical retrievers such as BM25 (Robertson et al., 2009) with generative decoders to introduce external knowledge grounding (Lewis et al., 2020), while dense retrieval models, such as DPR (Karpukhin et al., 2020) further strengthened semantic matching through dual-encoder architectures. Subsequent developments introduced hybrid retrievers and advanced re-ranking strategies that combine sparse precision with dense coverage, improving robustness in diverse knowledge domains. More recent variants, including Self-RAG (Asai et al., 2024), ReAct-style retrieval reasoning (Zhang et al., 2024a), and lightweight designs such as LightRAG, target hallucination reduction, chunk optimization, and efficiency under constrained resources (Guo et al., 2024). Methods like HyDE (Gao et al., 2023) or graph-augmented RAG (Edge et al., 2024) extend reasoning by synthesizing hypothetical evidence or leveraging structured knowledge graphs.

While these advancements significantly improve factual retrieval and reasoning, they largely assume that relevance is time-irrelevant, depending only on semantic similarity rather than contextual factors such as time, state, or environment. As a result, existing RAG frameworks often retrieve information that is topically relevant but time-incompatible, leaving them inadequate for domains where the validity of knowledge changes over time, as shown in Figure 1.

To address this gap, time-aware extensions of RAG have emerged in recent years. These methods introduce time cues through time-aware re-ranking, time-adjusted embeddings, or specialized datasets designed for timestamp prediction and time-sensitive question answering. Representative works, including TimeR4 (Qian et al., 2024), TimeQA (Chen et al., 2021), and Ts-Retriever (Wu et al., 2024), integrate time reasoning into retrieval and generation to better handle event ordering and time constraints. However, these methods are primarily developed for general-domain scenarios, where time information can be represented in relatively explicit and consistent forms, such as timestamps, event sequences, or well-structured timelines. Despite their progress, none of these frameworks model the ecological, seasonal, and phenological time structures that define agricultural knowledge. Agronomic recommendations, such as pest control timing, nutrient management schedules, and phenology-dependent disease risks are governed by biological cycles rather than explicit timestamps. Consequently, general-domain time-aware RAG approaches cannot capture the implicit, domain-specific time dependencies that determine whether an agricultural recommendation is appropriate or actionable.

2.2. Pest and Disease Identification

Pest and disease identification has long been a central task in agricultural decision support systems, as accurate diagnosis is a prerequisite for effective intervention. Early approaches relied on rule-based systems and expert knowledge, where symptoms, environmental conditions, and manual observations were encoded into decision rules for disease diagnosis and treatment recommendation (Shafay et al., 2025). While interpretable, these methods are limited by scalability and their dependence on handcrafted knowledge.

With the advancement of machine learning, particularly deep learning, data-driven approaches have become dominant. Convolutional neural networks (CNNs) have been widely applied for crop disease and pest classification from leaf images, achieving high accuracy in controlled settings (Zheng et al., 2025). More recently, Transformer-based architectures and multimodal models have further improved performance by capturing complex visual patterns and integrating textual or contextual information (Liu et al., 2025). These methods significantly enhance automated diagnosis but are primarily focused on perception tasks, often lacking the ability to provide actionable, context-aware recommendations.

To bridge this gap, recent studies have begun exploring large language models (LLMs) for agricultural question answering and decision support (Tzachor et al., 2023). These systems aim to move beyond recognition toward knowledge-driven reasoning, enabling users to query treatment strategies, management practices, and agronomic advice. Nevertheless, most existing approaches treat agricultural knowledge as static and fail to incorporate time constraints that are critical in real-world scenarios.

However, effective pest and disease management is inherently time-sensitive, as appropriate interventions depend on crop growth stages, pest life cycles, and seasonal dynamics. Existing identification and QA-based systems largely ignore such time dependencies, limiting their practical applicability. This highlights the need for frameworks that can jointly model domain knowledge and time-aware reasoning, motivating the integration of time signals into retrieval-augmented generation for agricultural decision support.

2.3. RAG in agriculture

Building upon advances in pest and disease identification and agricultural decision support, LLMs have recently been explored enabling applications such as advisory chatbots, crop disease diagnosis assistants, and agronomic knowledge extraction systems (Gong and Li, 2025) (Yuan et al., 2025). Existing studies often combine LLMs with retrieval components built from agricultural extension manuals or plant protection documents, and some incorporate structured resources such as agricultural knowledge graphs to enhance factual grounding (Lv et al., 2024) (Zhao et al., 2024). These efforts highlight growing interest in applying RAG-style methods to improve the reliability and interpretability of agricultural AI tools.

However, current agricultural RAG applications face two major limitations. First, they treat agronomic knowledge as time-irrelevant and overlook the time dependencies that determine when recommendations are valid. Retrieval systems typically rely on lexical or semantic similarity (Bali et al., 2024) without considering crop growth stages, pest life-cycle phases, or seasonal constraints, leading to time-mismatched evidence, such as retrieving pre-flowering pesticide guidelines for late-season scenarios. Second, the agricultural domain lacks high-quality, time-annotated QA datasets, making it difficult for models to learn or evaluate time-sensitive reasoning. Existing datasets focus on factual content but rarely encode time conditions, preventing systematic assessment of how well RAG frameworks handle time-sensitive queries (Kpodo et al., 2024) (Kasai et al., 2023).

These constraints reveal a critical gap: while RAG improves factual grounding, it remains ill-equipped to capture the dynamic, seasonal, and phenological rhythms underlying agricultural decisions. This motivates the development of a time-aware RAG framework, supported by a dedicated time-sensitive agricultural QA dataset, to bridge the disconnect between time-sensitive reasoning and retrieval-augmented generation in agriculture.

3. Methodology

We first provide an overview of the TARAG framework in Subsection 3.1. The framework consists of three principal components: Subsection 3.2 introduces a Time-Aware Knowledge Base Construction module that encodes time cues and phenological metadata; Subsection 3.3 introduces a Hybrid Time Retrieval module that fuses sparse, dense, and time filtering to retrieve time-aligned evidence; and Subsection 3.4 introduces a Time-Based Generation module that synthesizes responses conditioned on retrieved, time-relevant documents.

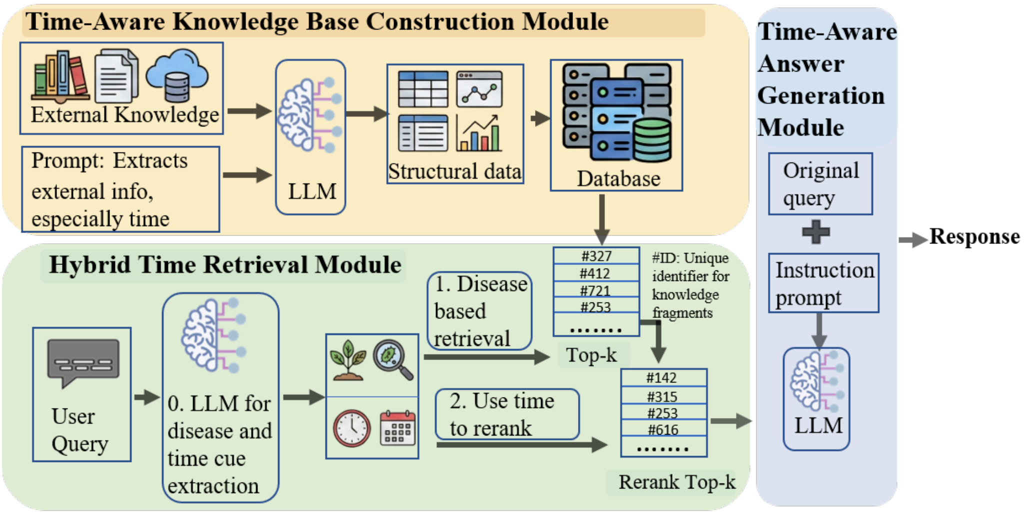


Figure 2: The TARAG framework consists of three main modules: (1) Time-Aware Knowledge Base Construction: utilizes an LLM to structure external knowledge, embedding each entry with associated disease and time information; (2) Hybrid Time Retrieval: processes an original query to perform disease-based retrieval, followed by a time-aware re-ranking of the results to select the top-k most time-relevant documents; (3) Time-Aware Answer Generation: augments the original query with the retrieved contexts and an instruction prompt, guiding an LLM to generate a final, time-sensitive response.

3.1. Overview of TARAG Framework

Unlike conventional question answering settings where document relevance is often approximated by global semantic similarity, agricultural question answering exhibits a more structured relevance mechanism.

In this domain, user queries typically contain two essential components: (1) domain-specific entities, and (2) time conditions. These two components jointly determine whether a piece of knowledge is applicable. Formally, we represent a query as:

$$q = (q_{\text{entity}}, q_{\text{time}}) \tag{1}$$

The entity component q_{entity} includes crop types, pest or disease names, and their variants. These elements are relatively stable and discrete, and primarily determine the topical relevance of retrieved documents.

The time component q_{time} consists of heterogeneous time expressions, which in agricultural scenarios cannot be treated as uniform timestamps. Following the time annotation schema defined in our dataset, we categorize time expressions into four canonical types: (1) phenological stages, (2) seasonal periods, (3) calendar-based time, and (4) relative time expressions. These categories exhibit different semantic characteristics: phenological stages are closely tied to biological processes, seasonal periods provide coarse entity context, calendar-based expressions offer explicit entity references, and relative expressions encode procedural or event-dependent timing. As a result, time in agricultural knowledge is inherently heterogeneous in both granularity and dependency on biological states.

Correspondingly, each retrieved document d is associated with both entity scope and time scope:

$$d = (d_{\text{entity}}, d_{\text{time}}) \tag{2}$$

Based on this decomposition, we define document relevance as a two-stage (nested) process. First, a candidate set is obtained based on entity matching:

$$\mathcal{D}\text{entity} = \text{TopK}\big(Score\text{entity}(q_{\text{entity}}, d_{\text{entity}})\big) \tag{3}$$

where $Score_{\text{entity}}$ measures the degree of entity-level matching (e.g., lexical or semantic similarity), ensuring that retrieved documents are topically relevant. Then, documents within the candidate set are re-ranked according to time alignment:

$$d^{*} = \text{TopK}\big(Score_{\text{time}}(q_{\text{time}}, d)\big), \quad d \in \mathcal{D}_{\text{entity}} \tag{4}$$

Ultimately, d^{*} emerges as the most time-relevant and topically relevant document. This is determined by $\text{Score}_{\text{time}}$, which measures the alignment between the query time condition and the document time scope, ensuring the agronomic validity of the knowledge applied.

This formulation reflects the structured retrieval process in agricultural QA. In practice, entity matching is first used to retrieve a candidate set of documents with high topical relevance, and time alignment is subsequently applied to refine the ranking by enforcing time consistency. Therefore, time does not act as an independent relevance signal, but as a constraint that is applied conditionally on entity-matched candidates.

From a theoretical perspective, this formulation can be viewed as a structured decomposition of relevance, which differs from conventional RAG frameworks that rely on a single similarity signal. By explicitly separating entity grounding and time validity, the model can avoid retrieving documents that are topically correct but time-inappropriate.

This theoretical formulation is consistent with the design of TARAG. The first-stage retrieval focuses on entity-level matching to ensure recall, while the second-stage re-ranking enforces time alignment, ensuring that the final retrieved evidence satisfies both topical relevance and time validity.

3.2. Time-Aware Knowledge Base Construction module

To support fine-grained time retrieval, agricultural documents are transformed from unstructured text into structured JSON records through a controlled prompting procedure. Each record contains standardized fields such as: *disease entity*, *crop stage*, *time validity intervals*, and *recommended treatments.*

Formally, for each raw document x, an LLM produces a structured output:

$$\mathcal{S} = \text{LLMstruct}(x, p_{\text{struct}}) \tag{5}$$

where p_{struct} denotes a schema-constrained prompt. The resulting knowledge base $\mathcal{K} = \mathcal{S}_1, \ldots, \mathcal{S}_n$ thus encodes explicit time metadata, enabling later retrieval modules to perform time-discriminative ranking.

3.3. Hybrid Time Retrieval module

We propose a Three-Stage Hybrid Time Retrieval module that integrates sparse lexical matching and dense time-aware semantic retrieval.

Stage 0: Entity and Time Cue Extraction

Before retrieval, an LLM extracts entity and time cues from the query q:

$$q_{\text{entity}}, q_{\text{time}} = \text{LLM}_{\text{extract}}(q) \tag{6}$$

where q_{entity} contains crop or pest entities, and q_{time} contains time expressions such as crop phenology stages or seasonal information. These extracted cues are then used in the subsequent two retrieval stages.

Stage 1: Sparse Retrieval via BM25 with entity Filtering

The entity cue q_{entity} guides BM25-based sparse retrieval. For a document d, the BM25 score is computed as:

$$\text{BM25}(q_{\text{entity}}, d) = \sum_{t \in q_{\text{entity}} \cap d} \text{IDF}(t) \cdot \frac{f(t,d)\,(k_1 + 1)}{f(t,d) + k_1 \left(1 - b + b\,\frac{|d|}{\text{avgdl}}\right)} \tag{7}$$

where $f(t, d)$ is the term frequency of token t in d, and $k1$, b are standard hyperparameters. Only documents containing at least one disease entity are retained:

$$\mathcal{D}_{\text{BM25}} = \{d \in \mathcal{K} \mid \text{EntityMatch}(q_{\text{entity}}, d) = 1\} \tag{8}$$

Stage 2: Dense Time-Aware Re-Ranking

The extracted time expressions q_{time} and d_{time} are represented as raw text spans (e.g., crop stages or seasonal descriptions). We encode these expressions into dense vectors using a pre-trained sentence embedding model:

$$\mathbf{t} = \text{Enc}_{\text{time}}(t) \tag{9}$$

where Enc_{time} is implemented using a frozen pre-trained encoder (e.g., BERT-based or sentence-transformer). The encoder is not further trained in our framework; instead, it is directly reused to capture semantic similarity between time expressions.

The candidate set $\mathcal{D}_{\text{BM25}}$ is then re-ranked using dense embeddings of time cues extracted from the query and the documents:

$$\mathbf{q}_{\text{time}} = \text{Enc}_{\text{time}}(q_{\text{time}}), \qquad \mathbf{d}_{\text{time}} = \text{Enc}_{\text{time}}(d_{\text{time}}) \tag{10}$$

$$\text{sim}(q, d) = \cos(\mathbf{q}_{\text{time}}, \mathbf{d}_{\text{time}}) \tag{11}$$

where d_{time} denotes time expressions in the document. By leveraging the semantic representation ability of the embedding model, this stage maps time-similar expressions (e.g., related crop stages or seasonal descriptions) into nearby regions in the embedding space, ensuring that retrieved documents are time-aligned with the query and prioritizing contextually appropriate content for downstream generation.

3.4. Time-Aware Answer Generation Module

In the final stage, TARAG employs an LLM to generate answers conditioned on both the retrieved evidence and a time-aware grounding prompt. Let $\mathcal{D}_k = d_1, \ldots, d_k$ denote the selected documents. The generator produces:

$$a = \text{LLM}(q, \mathcal{D}k, p_{\text{time}}) \tag{12}$$

where p_{time} explicitly instructs the model to respect time constraints, restrict answers to the correct crop stage, and avoid treatments irrelevant to the specified period.

This module ensures that the final answer reflects both domain correctness (e.g., correct pesticide type and dosage) and time appropriateness (e.g., avoiding high-toxicity chemicals during pre-harvest safety interval).

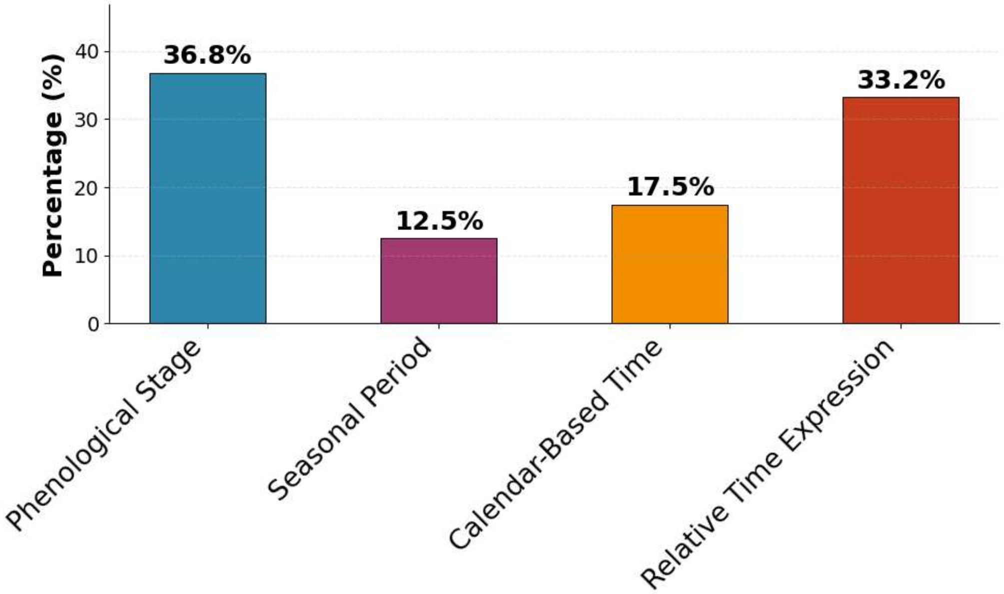


Figure 3: Distribution of time expression categories in the TAQA dataset. Phenological stages (e.g., "during flowering") constitute the majority (36.8%), followed by seasonal periods (12.5%), calendar-based times (17.5%), and relative expressions (33.2%). This indicates a strong domain-specific bias towards ecological time references.

4. Dataset Construction: TAQA

To support systematic evaluation of time-aware reasoning in agricultural question answering, we construct **TAQA**, a bilingual (Chinese–English) dataset centered on fine-grained time-sensitive understanding of crop diseases and pests. This section details the design motivation, data collection pipeline, time annotation schema, and statistical characteristics of the dataset.

4.1. Motivation and Design Principles

Existing agricultural QA datasets primarily focus on factual correctness while overlooking the time dimension that is essential for field decision-making. Disease outbreaks, pest population dynamics, and recommended control strategies all depend strongly on crop phenology and seasonal factors. Consequently, models lacking time situational awareness often fail to produce stage-specific or time-valid answers.

The construction of TAQA is guided by three principles:

Table 1: Overview of TAQA Dataset Statistics

Dimension	**Statistics**
Number of QA pairs	30,000+
Crop categories	100+ distinct crop types
Unique pest/disease entities	2,000+ unique entities, such as "Apple Leaf Rust Blight","Banana stem borer weevil"
Time coverage	Full seasonal cycle and long-term agronomic stages
Average query length	∼20 words
Example of query diversity	“After flowering, do grapes become susceptible to black rot in 10–14 days? How should it be prevented?” “How to control grape black rot about two weeks after flowering?”

- **Time Explicitness**: each instance must contain at least one explicit or implicit time cue (e.g., “Heading stage”, “after harvest”, “late August”).
- **Agronomic Faithfulness**: questions and answers must align with region-specific cultivation cycles and validated plant protection guidelines.
- **Retrieval Compatibility**: documents are segmented and annotated to support time-aware retrieval modules such as time filtering and hybrid sparse–dense ranking.

4.2. Data Sources and Annotation Process

- **Data Sources:** The dataset is compiled from authoritative agricultural resources, including: (1) national and provincial plant protection manuals; (2) digital agricultural extension archives; (3) peer-reviewed agronomy literature (Yao et al., 2024; Yan et al., 2025); (4) farmer consultation logs and expert Q&A forums.

  All raw texts are processed following a two-stage pipeline:

- **Question Generation and Verification**

  We adopt a hybrid human-in-the-loop generation workflow. LLMs are prompted to produce time-based questions based on each document

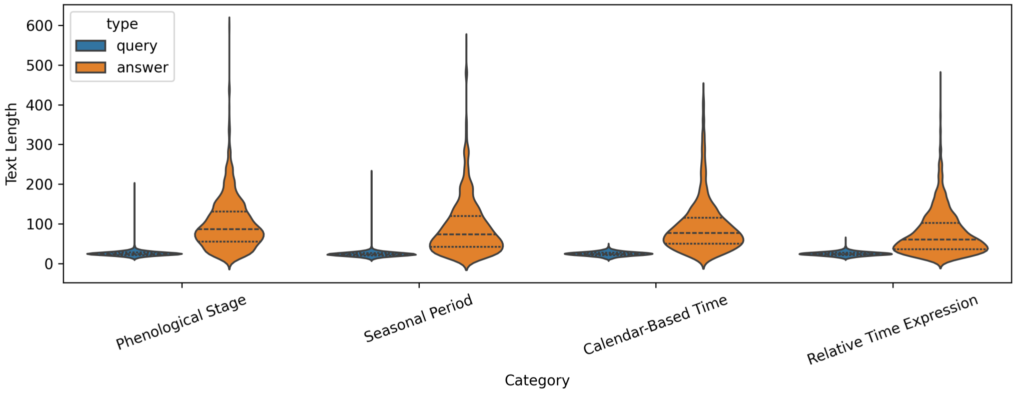


Figure 4: Summary statistics of text lengths for both queries and answers across four time-related categories in the agricultural QA dataset. Across all categories, queries remain uniformly concise (median about 24 to 25 words, IQR about 6 to 7) with low variance, whereas answers are substantially longer and more dispersed. The Phenological Stage category contains the largest sample size and yields the longest answers on average (mean about 101 words, IQR about 76), followed by Calendar-Based Time and Seasonal Period, whose answers also span a wide range (maximum up to 552 words). The Relative Time Expression category exhibits relatively shorter answers (mean about 77 words) but still shows considerable variability. Overall, the consistent gap between query and answer length highlights that time-related agricultural questions are brief and well defined, while their corresponding explanations require richer and category-dependent contextual information.

segment. Domain experts subsequently verify correctness, remove hallucinations, and ensure agronomic validity.

- **Golden Answer and Evidence Alignment**

  Each question in TAQA is paired with a gold-standard answer and its corresponding evidence is defined at the document level. Instead of selecting a minimal continuous text span, annotators identify the specific source document(s), represented by unique document identifiers, that contain sufficient information to support both the factual and time components of the answer. This design provides an explicit grounding signal for retrieval evaluation, enabling rigorous assessment of whether a framework can retrieve the correct evidence document(s) necessary for answering time-sensitive agricultural questions.

Table 2: Examples of time-aware agricultural question answering. The time condition is explicitly specified in the query and categorized into different time types. Red represents pests and diseases, blue represents time, and related documents indicate which document the question was extracted from.

Query	Time (Category)	Answer	Relative Docs
How should maize northern leaf blight be controlled at the early infection stage?	Early infection stage (Relative Time Expression)	Apply fungicides such as 5% chlorothalonil WP (300× dilution) or 80% Sukejing WP (1000× dilution). Spray once every 10 days for 2–3 consecutive applications.	#387
What should be done after maize northern leaf blight has occurred after maize harvest?	Post-harvest (Relative Time Expression)	Remove and destroy diseased plant residues both inside and outside the field.	#388
How should maize Fusarium seedling blight be treated at the seed stage?	Seed stage (Phenological stage)	Treat seeds with 50% carbendazim WP at 0.2%–0.3% of seed weight, or soak seeds in 70% thiophanate-methyl WP (500× dilution).	#1524
How should Monema flavescens be controlled in winter?	Winter (Seasonal Period)	During winter, remove overwintering cocoons by pruning branches. Collected cocoons should be destroyed to eliminate pest sources.	#63
How should Monema flavescens be controlled around June?	Around June (Calendar-Based Time)	During the egg hatching period, spray 20% insecticide suspension (1500× dilution) or 5% KASIKE suspension (1500× dilution) at the early larval stage.	#64

4.3. Time Annotation Schema

To explicitly capture the time semantics underlying agricultural phenomena and disease dynamics, TAQA adopts a structured time annotation schema focused on the categorization of time expressions. Although time information in agricultural texts can span multiple dimensions, the present work annotates the *Time Category* dimension, which establishes a unified taxonomy for normalizing heterogeneous time mentions across sources.

Time Category: Each time expression in the dataset is assigned to one of four canonical categories that frequently appear in agronomic decision-making:

- **Phenological Stage:** Expressions describing crop growth or pest development phases, such as "seedling stage", "fruit set stage", or "overwintering period". These stages are strongly associated with risk variation and treatment efficacy.
- **Seasonal Period:** Coarse-grained references to agricultural seasons or sub-seasonal intervals, including phrases like "early spring", "mid-summer", or "late autumn". Such expressions provide broad time context but often require integration with phenological cues for actionable interpretation.
- **Calendar-Based Time:** Mentions expressed using explicit dates or ranges, e.g., "August", "late July", or "during September". These expressions facilitate alignment with real-world time but may vary in precision across sources.
- **Relative Time Expression:** Context-dependent references such as "before harvest", "after flowering", or "during pesticide application". These expressions often encode procedural relations rather than absolute time and are essential for modeling agronomic workflows.

This categorization supports downstream time-sensitive reasoning by mapping heterogeneous time expressions, each with distinct granularity and linguistic form, into their corresponding normalized time categories. This structured normalization enables the framework to align queries and documents based on consistent category-level time semantics rather than relying on surface-level textual similarity. Furthermore, an analysis of the TAQA dataset, as shown in Table 1 and 2, and Figure 3 and 4, demonstrates that these

Table 3: A comprehensive evaluation of the retrieval module

Model	**Recall**					**nDCG**				
	R@1	R@3	R@5	R@10	R@20	nD@1	nD@3	nD@5	nD@10	nD@20
Contriever (Izacard et al., 2021)	5.07%	9.84%	12.55%	17.02%	22.62%	5.07%	7.83%	8.94%	10.37%	11.78%
TS-Contriever (Wu et al., 2024)	7.07%	12.41%	15.36%	20.29%	25.97%	7.07%	10.16%	11.38%	12.97%	14.40%
E5 (Wang et al., 2024)	78.33%	92.29%	95.03%	96.71%	97.51%	78.33%	86.59%	87.72%	88.28%	88.48%
Agri-sentence-transformer (Rezayi et al., 2025)	1.42%	2.90%	3.68%	5.40%	7.73%	1.42%	2.27%	2.59%	3.14%	3.73%
BCE (Youdao, 2023)	69.77%	84.65%	88.58%	91.94%	94.32%	69.77%	78.56%	80.19%	81.28%	81.89%
Qwen3-Embedding-0.6B (Zhang et al., 2025)	71.16%	84.99%	88.53%	91.48%	93.57%	71.16%	79.35%	80.81%	81.77%	82.31%
Qwen3-Embedding-4B (Zhang et al., 2025)	76.28%	89.19%	92.07%	94.49%	96.21%	76.28%	83.91%	85.11%	85.89%	86.33%
Our	**80.52%**	**93.45%**	**96.31%**	**98.44%**	**99.14%**	**80.52%**	**86.62%**	**87.75%**	**88.28%**	**88.49%**

time categories are widely distributed across the corpus, highlighting their central role in structuring time-based retrieval and reasoning within agricultural question-answering tasks.

5. Experiments

5.1. Experimental environment

The experimental setup was conducted on a high–performance hardware platform comprising a 12th Generation Intel(R) Core(TM) i7-12700H CPU operating at 2.70GHz, an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM (driver version 525.125.06), and 24 GB of system memory. The implementation of the experimental framework was carried out using Python 3.9.23 and the PyTorch 2.5.1 deep learning library, ensuring compatibility and efficiency in handling computationally intensive tasks.

5.2. Experimental Design

Our experimental evaluation is conducted from two complementary perspectives: (1) a retrieval-focused evaluation that isolates the retriever to assess the time alignment and relevance of retrieved documents, and (2) a full RAG evaluation that examines how time information is preserved and utilized throughout the entire pipeline to produce contextually appropriate agronomic recommendations.

5.2.1. Evaluation of the Retrieval Module

To assess retrieval quality, we adopt two widely used ranking-based metrics: Recall@k(%) and nDCG@k(%). Recall@k evaluates whether the retriever successfully surfaces the relevant time-aligned documents within the top-k results:

$$\text{Recall@}k = \frac{|\text{Relevant} \cap \text{Retrieved@}k|}{|\text{Relevant}|} \tag{13}$$

where, Relevant denotes the set of all documents that are truly relevant and time-aligned with the query, while Retrieved@k denotes the set of the top-k documents returned by the retrieval method.

This metric is particularly suitable for time-sensitive agricultural QA because relevant evidence is often sparse and time-dependent; thus failing to retrieve such passages directly undermines downstream generation accuracy.

To complement recall, we use normalized Discounted Cumulative Gain (nDCG@k), which considers both correctness and ranking position:

$$\text{nDCG@}k = \frac{1}{\text{IDCG@}k} \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{14}$$

where rel_i denotes the relevance of the i-th retrieved document.

Definition of IDCG@k: IDCG@k (Ideal DCG at rank k) is the maximum possible DCG value for the top-k documents, obtained by sorting the candidate set in descending order of their true relevance scores:

$$\text{IDCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i^*} - 1}{\log_2(i+1)} \tag{15}$$

where rel_i^* is the relevance score of the i-th document in the ideally ranked list.

Computation of rel_i: In our time-sensitive agricultural retrieval setting, rel_i is defined as a binary or graded score reflecting both semantic relevance and time correctness of document d_i with respect to the query q:

$$rel_i = \begin{cases} 1, & \text{if } d_i \text{ is relevant and time aligned with } q \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

nDCG@k therefore measures how well the top-k retrieved documents are both relevant and time-consistent, which is critical because downstream generation quality deteriorates when irrelevant or time-mismatched passages dominate the top ranks.

We focus on the top-20 retrieved documents (i.e., Recall@20 and nDCG@20). Retrieving beyond this depth provides diminishing gains: introducing excessive context into the LLM not only increases computational cost but also risks overwhelming the generator with noisy or time-inconsistent information.

5.2.2. Evaluation of the End-to-End RAG Framework

For full pipeline evaluation, we adopt standard QA metrics, namely Recall(%), Precision(%), and F1(%), computed over the top-10 retrieved documents that support control recommendations, rather than over generated answer candidates. This threshold balances coverage with reliability, ensuring that metrics reflect the practical usefulness of the recommendations in real-world agricultural decision-making.

To account for the strong dependency between treatment validity and time conditions in agricultural pest management, we further introduce a task-specific interpretation of TP/FP/FN/TN.

Instead of using their generic definitions, we operationalize them based on whether the generated recommendation is both agronomically correct and time-appropriate:

- TP: the model outputs a pesticide and dosage that are agronomically correct and consistent with the required time condition.
- FP: the model suggests a pesticide that is either not supported by the reference answer or violates time constraints (i.e., time-inappropriate treatment).
- FN: the model fails to include a required pesticide that is specified in the reference answer under the given time condition.
- TN: the model correctly avoids recommending pesticide application when treatment is unnecessary or time-prohibited.

Based on these definitions, the evaluation metrics are computed as:

$$\text{Precision@10} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall@10} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$\text{F1@10} = \frac{2 \times \text{Precision@10} \times \text{Recall@10}}{\text{Precision@10} + \text{Recall@10}} \tag{18}$$

This design allows us to quantify not only factual correctness but also the time sensitivity of generated recommendations, providing a unified measure of how well a RAG framework aligns agronomic actions with the appropriate phenological or seasonal stage.

5.3. Baselines

To comprehensively evaluate both the retrieval and end-to-end capabilities of our proposed TARAG framework, we compare it against a diverse set of strong baselines. These baselines cover (1) standalone retrievers widely used in domain-general and time-aware retrieval research, and (2) representative RAG frameworks that reflect different prompt-based, retrieval-augmented, and structure-enhanced paradigms.

5.3.1. Retrieval Baselines

We compare TARAG's hybrid time-aware retriever with a diverse set of dense and sparse retrieval baselines, including both general-purpose and time-sensitive models: **Contriever** (Izacard et al., 2021), **TS-Contriever** (Wu et al., 2024),**BCE** (Youdao, 2023), **E5** (Wang et al., 2024), **Qwen-Embedding** (Zhang et al., 2025), **Agri-Sentence-Transformer** (Rezayi et al., 2025), **BM25** (Robertson et al., 2009), and **BGE-M3** (Chen et al., 2024).

5.3.2. RAG Framework Baselines

For the end-to-end comparison, we evaluate TARAG against several representative RAG and prompting frameworks that reflect different philosophies of reasoning, retrieval integration, and knowledge organization, and we employ DeepSeek's 1.5B, 7B, and 14B models as the underlying generation backbone (Guo et al., 2025).

- **Direct Prompting**: The LLM answers queries without any retrieval augmentation, providing a lower bound for model performance.
- **Chain-of-Thought (CoT) Prompting (Wei et al., 2022)**: A reasoning-enhanced prompting strategy that encourages step-by-step generation, serving as a non-retrieval reasoning baseline.
- **Naive RAG (Lewis et al., 2020)**: The standard retrieve-then-generate pipeline where retrieved passages are directly concatenated with the query.
- **TimeR4 (Qian et al., 2024)**: A time-enhanced RAG framework incorporating rule-based time re-ranking for improved time-sensitive retrieval.

- **LightRAG (Guo et al., 2024)**: A lightweight and modular RAG framework supporting multiple retrieval paradigms and efficient ranking mechanisms.

This collection of baselines ensures a rigorous evaluation across both retrieval quality and time-aligned generation, enabling a comprehensive assessment of TARAG's contributions.

5.4. Main Results

5.4.1. Overall Performance

Across both retrieval-only and full RAG evaluations, TARAG achieves the best performance among all compared frameworks. In the retrieval task, TARAG's hybrid time-aware retriever consistently outperforms all dense and sparse baselines, achieving the highest Recall@20 (99.14%) and nDCG@20 (88.49%), as shown in Table 3. Similarly, in end-to-end evaluation, TARAG delivers the strongest Precision@10 (up to 54.41%), Recall@10 (up to 86.67%), and F1@10 (up to 66.85%), demonstrating superior time grounding and factual consistency compared to direct prompting, CoT prompting, and existing RAG frameworks such as TimeR4 and LightRAG, as shown in Table 4. Notably, the gains in Precision@10 and F1@10 are particularly important, as these metrics directly reflect the ability to filter time-irrelevant evidence in time-sensitive agricultural decision-making.

Notably, while several dense retrievers and RAG baselines have shown strong performance in general-purpose question answering benchmarks, their effectiveness drops markedly in our time-sensitive agricultural setting. This performance gap highlights the unique challenge of time grounding in agricultural decision support, where correct recommendations depend not only on semantic relevance but also on strict alignment with crop growth stages and seasonal constraints.

5.4.2. Why TARAG Performs Better

The superiority of TARAG can be attributed to two key design choices that directly address the time characteristics of agricultural knowledge.

First, the time-aware knowledge base construction enriches each passage with explicit time metadata, enabling the retriever to filter out seasonally or phenologically incompatible information at an early stage. This design effectively mitigates a common failure mode observed in dense retrievers, which often retrieve passages with high semantic similarity but invalid time

Table 4: Comparison of Recall@10, Precision@10, and F1@10 across models and methods.

Model	**Method**	**Recall@10**	**Precision@10**	**F1@10**
DeepSeek-R1-1.5B	Direct Prompt	0.0%	0.0%	0.0%
	CoT	1.12%	1.08%	1.10%
	Naive RAG	33.86%	31.92%	32.86%
	TimeR4	47.80%	38.53%	42.67%
	Light RAG	14.38%	14.33%	14.38%
	Our	**63.56%**	**46.98%**	**54.03%**
DeepSeek-R1-7B	Direct Prompt	3.68%	3.56%	3.62%
	CoT	9.31%	8.80%	9.00%
	Naive RAG	40.27%	35.45%	37.70%
	TimeR4	59.68%	43.80%	50.31%
	Light RAG	22.67%	21.28%	21.95%
	Our	**76.36%**	**53.11%**	**62.50%**
DeepSeek-R1-14B	Direct Prompt	9.38%	9.28%	9.33%
	CoT	23.01%	21.68%	22.32%
	Naive RAG	57.56%	46.13%	51.22%
	TimeR4	78.64%	51.56%	62.28%
	Light RAG	39.94%	33.77%	36.60%
	Our	**86.67%**	**54.41%**	**66.85%**

applicability. Prior studies on time reasoning have similarly emphasized that explicit time constraints are essential in decision-critical domains, and our results further confirm this observation in the agricultural context.

Second, the hybrid time-aware retrieval mechanism integrates sparse lexical matching with time-weighted dense embeddings, improving ranking quality for queries containing agricultural time expressions. By jointly modeling lexical cues and time relevance, TARAG retrieves evidence that is not only semantically relevant but also time-aligned with the query intent. This directly benefits the generation stage, as the LLM receives cleaner and more time-compatible evidence rather than noisy or time-conflicting documents, leading to higher end-to-end Precision and F1 scores.

5.4.3. Why Existing Baselines Underperform

Dense retrievers such as Contriever (Izacard et al., 2021), BGE (Chen et al., 2024), and E5 (Wang et al., 2024) excel at semantic matching and have demonstrated strong performance in a wide range of open-domain retrieval tasks. However, they treat time information as unstructured text, making them insensitive to phenological constraints that are critical in agricultural decision-making. As a result, these models frequently retrieve passages that describe correct diseases but inappropriate time periods, such as recommending chemicals effective during overwintering when the query concerns a summer growth stage. This behavior explains why dense retrievers achieve high Recall@20 but substantially lower nDCG@20 and end-to-end F1@10, as shown in Tables 3 and 4.

Sparse retrievers such as BM25 rely on lexical overlap and can partially capture explicit time expressions. However, they fail to recognize semantically equivalent but lexically divergent time phrases (e.g., “pre-bloom” vs “bud break”), resulting in poor recall for time-sensitive queries. Similarly, baseline RAG frameworks, including Direct Prompting, CoT-Prompting (Wei et al., 2022), Naive RAG (Lewis et al., 2020), and LightRAG (Guo et al., 2024), lack explicit time grounding mechanisms. These methods depend on general-purpose generative reasoning, which often overlooks implicit time cues and leads to seasonally or phenologically inappropriate recommendations.

It is worth noting that these baseline frameworks were not originally designed for fine-grained agricultural time reasoning, which partially explains their limited performance in our setting. In contrast, TARAG explicitly incorporates time constraints during retrieval, reducing both false positives caused by time-irrelevant evidence and false negatives caused by missing time-appropriate measures. This targeted modeling of time relevance ultimately leads to significantly improved F1@10 in time-sensitive agricultural decision support.

5.5. Ablation Studies

To better understand the contribution of each component in our time-aware retrieval-augmented framework, we conduct a series of ablation studies focusing on two aspects: (1) disentangling the hybrid retriever into its dense and sparse components, and (2) validating the generality of our hybrid retrieval strategy by pairing BM25 with alternative dense retrievers.

Table 5: Ablation study on Retrieval Module

Method	**Recall@K**			**nDCG@K**		
	1	**3**	**10**	**1**	**3**	**10**
BM25	58.22%	79.14%	91.25%	58.22%	69.69%	73.95%
BGE	78.19%	89.19%	95.24%	78.19%	84.65%	86.90%
Our	**80.52%**	**93.45%**	**98.44%**	**80.52%**	**86.62%**	**88.28%**

5.5.1. Effect of Hybrid Retrieval

Our hybrid retriever combines a dense encoder with a sparse BM25 retriever to jointly capture semantic relevance and lexical-time cues. To isolate the effect of this design, we compare the full hybrid retriever with its individual components (Dense-only and BM25-only).

Table 5 shows that the hybrid retriever consistently outperforms both standalone dense and standalone sparse retrieval. This confirms that dense retrieval excels at capturing semantic similarity, while BM25 contributes complementary lexical grounding essential for agricultural terminology and time expressions. The combination yields superior recall and nDCG, demonstrating the necessity of integrating both signals for time-sensitive agricultural QA.

The clear performance drop when removing either component indicates that both semantic matching and lexical relevance are crucial for time-sensitive agricultural retrieval. BM25 often retrieves passages containing explicit time markers (e.g., "pre-bloom", "early fruiting"), while dense retrievers capture higher-level biological transitions not explicitly stated in text. Their complementary strengths explain the observed performance gains.

5.5.2. Generalization Across Dense Retrievers

To examine whether our hybrid retrieval design benefits only a particular dense encoder or is generally applicable, we substitute the dense encoder with several widely used models (BCE, E5, and Qwen Embedding). For each dense model, we compare dense-only retrieval with its hybrid variant (Dense Model + BM25).

As shown in Figure 5, adding BM25 consistently improves both Recall@20 and nDCG@20 across all dense encoders. This demonstrates that our hybrid retrieval strategy is not tied to a specific dense model but is a general enhancement applicable to a wide range of retrievers.

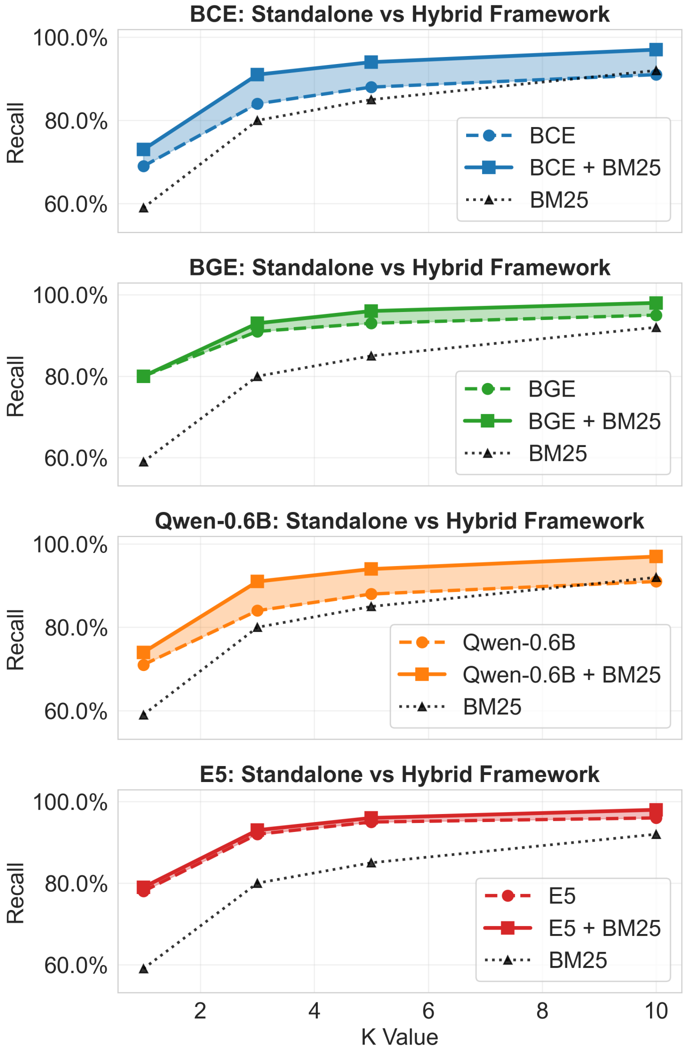


Figure 5: Performance evaluation of the hybrid retrieval framework. This figure uses four subplots (arranged in a 4×1 pattern) to compare the recall (Recall@K) of different dense retrieval models under the hybrid framework with the BM25 baseline. The horizontal axis represents the K value (1, 3, 5, 10), and the vertical axis represents the recall. The curves show that integrating each model (such as BCE and Qwen-0.6B) into the proposed hybrid framework significantly outperforms their individual use or the BM25 baseline. The average improvement percentage marked above further validates the effectiveness of the framework.

The uniform improvement across all models indicates that hybrid retrieval effectively offsets the time insensitivity of dense retrievers by injecting explicit lexical signals, which are particularly important for agricultural phenology, stage-specific management actions, and seasonal pest behaviors.

These ablations collectively validate that hybrid time-aware retrieval is an essential component of our TARAG framework, providing a stronger foundation for downstream time-based answer generation.

5.6. Qualitative Case Study of Existing RAG Limitations

To further illustrate the practical advantages of our time-aware RAG framework, we present a case study comparing model outputs on a time-sensitive agricultural query. Across different baseline configurations, we observe distinct error patterns reflecting the limitations of each method. Direct Prompting often fails to produce actionable recommendations, as the model tends to provide vague or generic descriptions without specifying concrete chemical control measures or growth-stage aligned treatments. Naive RAG shows partial improvement by retrieving relevant segments; however, the presence of substantial irrelevant context, combined with the model's susceptibility to hallucination, leads to inaccurate or inconsistent control recommendations. LightRAG, despite leveraging a knowledge-graph-based structure, suffers from significant information loss during graph construction and frequently retrieves semantically misaligned fragments, resulting in erroneous or incomplete answers. For more details, please see Table 6. To systematically characterize these recurring vulnerabilities, Table 7 categorizes the baseline failures into three principal modes: time misidentification, time-misaligned retrieval, and time inconsistency. As illustrated, while traditional pipelines frequently output time disjointed or generic advice, TARAG effectively anchors the generation to the exact phenological stage, consistently delivering precise and actionable agronomic treatments.

5.7. Failure Mode Analysis

While the proposed TARAG framework effectively mitigates the aforementioned baseline errors through explicit time modeling and hybrid retrieval, it still exhibits specific failure modes when processing highly ambiguous queries. Our analysis indicates that performance degradation primarily occurs when time cues are weakly specified or implicitly bounded. For example, as shown in Table 8, user expressions such as "before peak emergence" introduce partial time ambiguity. Lacking a hard chronological boundary,

Table 6: Case study comparing four methods on a time-sensitive agricultural question. Incorrect outputs are marked in red; correct chemical recommendations from our method are marked in blue.

Method	**Question: "What should be done during the 1st-generation egg-hatching period of *Adoxophyes orana*?"**
Direct Prompt	During the first-generation egg-hatching period of Adoxophyes orana, timing is especially important because the young larvae are most vulnerable before they roll leaves or enter fruit clusters. At this stage, it helps to intensify field monitoring and confirm hatch timing through degree-day accumulation or direct scouting.
Naive RAG	Some steps you can take include removing the badly affected leaves to reduce early infestations and keeping an eye on the orchard for newly hatched larvae. It may also help to adjust field management to keep the canopy dry and improve airflow. For chemical control, you could consider using a pyrethroid spray every few days to suppress emerging larvae, and treatments like chlorpyrifos can also be applied around the flowering period to keep populations low.
LightRAG	When dealing with the first-generation egg-hatching period of Adoxophyes orana, it is important to manage the larvae before they enter leaves or shoots. According to the retrieved information, growers can consider applying dormant-season treatments such as lime-sulfur or even high-concentration oil emulsions, which help suppress overwintering pest populations. Chemical control: Spray with 50% phoxim EC at a dilution of 1,000 times, or 10% imidacloprid EC at a dilution of 2,000 times, or 10% imidacloprid WP at a dilution of 2,000 times, or 40% cypermethrin EC at a dilution of 600-800 times.
Our (TARAG)	Identified that the pest was in the 1st–generation egg-hatching stage and provided control measures suitable for this timing. Suggested applying "80% dichlorvos EC at 800 × dilution," "or 50% phoxim EC at 1500 × dilution," "or 48% chlorpyrifos EC at 2000× dilution."to target newly emerged larvae before they enter protected feeding sites. Also included several routine management practices, such as improving orchard cleanliness, maintaining regular monitoring during hatch, and checking young larval activity to ensure treatments are applied at the optimal moment.

Table 7: Representative failure modes of existing RAG frameworks compared with TARAG. In the user queries, red represents pests and blue represents time constraints. For the model outputs, traditional RAG failures across the pipeline are highlighted in red, with the specific error type denoted in parentheses, while TARAG provides precise, stage-aware recommendations in blue.

User Query	**Traditional RAG Output**	**TARAG Output (Ours)**
What control measures should be taken during the early infestation stage of wheat aphids?	"I recommend focusing on general field sanitation and ensuring proper winter irrigation. Keep the field clean to reduce future outbreaks." (**Time Misidentification**)	"For the early infestation stage, rapid suppression is critical. Apply a targeted foliar spray of 10% imidacloprid WP (1500× dilution) to effectively control the initial population before it spreads."
What should be done during the egg-hatching period of rice stem borers?	"It is highly recommended to perform deep soil plowing and practice stubble burning after the rice harvest." (**Time-Misaligned Retrieval**)	"During the egg-hatching period, treatments must target the larvae before stem penetration. Immediately apply chlorantraniliprole SC to achieve optimal control of the newly hatched rice stem borers."
How to control cotton bollworm during larval emergence?	"For controlling cotton bollworm during this stage, a comprehensive approach is best. Set up adult pheromone traps to monitor moth flights, and simultaneously apply late-stage chemical defoliants to expose the bolls." (**Time Inconsistency**)	"To control cotton bollworm during larval emergence, stage-consistent biological treatments are required. Apply Bacillus thuringiensis (Bt) directly to the foliage to effectively target the actively feeding young larvae."

Table 8: Analysis of TARAG's residual error when handling queries with ambiguous time boundaries. In the user query, red represents pests and blue represents time constraints. For the model outputs, incorrectly hallucinated or out-of-stage extensions are marked in red.

User Query	Traditional RAG Output	TARAG Output (Residual Error)
What actions should be taken before peak emergence of leaf miners?	To effectively manage leaf miners, a comprehensive approach is recommended. You should perform winter orchard pruning to eliminate overwintering pupae. Additionally, apply systemic larval pesticides to control feeding larvae, and set up adult pheromone traps.	Before the peak emergence of leaf miners, prioritize setting up pheromone monitoring networks and applying light preventive sprays. To ensure complete control, also apply a high-concentration abamectin (avermectins) treatment to eradicate larvae already mining inside the leaves.

the framework retrieves a broader time scope, which subsequently leads the generation module to output a mixture of both preventive and reactive measures, rather than a single pinpointed solution. These observations highlight a current limitation of our time-aware design: it relies heavily on extractable time anchors. Consequently, further improvements are required to enhance the model's complex time-reasoning capabilities during the generation phase, specifically for resolving vague, continuous, or highly implicit time expressions.

6. Conclusion and Future Work

In this work, we present TARAG, a framework designed to address the time misalignment that commonly arises in agricultural question answering. By integrating a time-indexed knowledge base, hybrid retrieval mechanisms, and time-based answer generation, TARAG substantially improves both retrieval quality and end-to-end QA performance. Our experiments demonstrate that time metadata plays a pivotal role in aligning model outputs with crop phenology, pest development stages, and season-dependent management constraints. The proposed TAQA dataset further enables systematic evaluation of time-sensitive reasoning, filling a critical gap in agricultural QA research.

While TARAG marks a significant step toward reliable and context-aware agricultural assistance, several avenues remain open for exploration. First, time metadata in agricultural texts is often implicit; future work could incorporate automated time inference models to enhance coverage and granularity. Second, integrating real-time environmental signals, such as weather trajectories or pest monitoring data, would allow the framework to adapt to dynamically changing field conditions. Finally, extending TARAG to multimodal settings, including remote sensing imagery and sensor data, may further enhance its utility in precision crop management.

Overall, we hope this work provides a foundation for developing time-robust RAG frameworks that support safer, more accurate, and more sustainable agricultural decision-making.

References

Abate, T., van Huis, A., Ampofo, J.K.O., 2000. Pest management strategies in traditional agriculture: an african perspective. Annual review of entomology 45, 631–659. `https://doi.org/10.1146/annurev.ento.45.1.631`.

Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H., 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection, in: The Twelfth International Conference on Learning Representations. `https://openreview.net/forum?id=hSyW5go0v8`.

Bali, A., Bhagwat, A., Bhise, A., Joshi, S., 2024. Semantic similarity detection and analysis for text documents, in: 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), IEEE. pp. 1–9. `http://dx.doi.org/10.1109/ic-ETITE58242.2024.10493834`.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z., 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216 4. `https://doi.org/10.48550/arXiv.2402.03216`.

Chen, W., Wang, X., Wang, W.Y., 2021. A dataset for answering time-sensitive questions. arXiv preprint arXiv:2108.06314. `https://doi.org/10.48550/arXiv.2108.06314`.

Deka, M., Bhuyan, M., Hazarika, L., 2006. Traditional pest management practices of assam. Indian Journal of Traditional Knowledge 5, 75–78. `https://www.academia.edu/99492232/Traditional_pest_management_practices_of_Assam?sm=b`.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R.O., Larson, J., 2024. From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130. `https://doi.org/10.48550/arXiv.2404.16130`.

Gao, L., Ma, X., Lin, J., Callan, J., 2023. Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1762–1777. `https://doi.org/10.18653/v1/2023.acl-long.99`.

Gao, Y., Xiong, Y., Zhong, Y., Bi, Y., Xue, M., Wang, H., 2025. Synergizing rag and reasoning: A systematic review. arXiv preprint arXiv:2504.15909 `https://doi.org/10.48550/arXiv.2504.15909`.

Gong, R., Li, X., 2025. The application progress and research trends of knowledge graphs and large language models in agriculture. Computers and electronics in agriculture 235, 110396. `https://doi.org/10.1016/j.compag.2025.110396`.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 `https://arxiv.org/abs/2501.12948`.

Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C., 2024. Lightrag: Simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779. `https://doi.org/10.48550/arXiv.2410.05779`.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E., 2021. Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118. `https://doi.org/10.48550/arXiv.2112.09118`.

Junaid, M., Gokce, A., 2024. Global agricultural losses and their causes. Bulletin of Biological and Allied Sciences Research 2024, 66–66. `https://doi.org/10.54112/bbasr.v2024i1.66`.

Karpukhin, V., Oguz, B., Min, S., Lewis, P.S., Wu, L., Edunov, S., Chen, D., Yih, W.t., 2020. Dense passage retrieval for open-domain question answering., in: EMNLP (1), pp. 6769–6781. `https://doi.org/10.18653/v1/2020.emnlp-main.550`.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N.A., Choi, Y., Inui, K., et al., 2023. Realtime qa: What's the answer right now? Advances in neural information processing systems 36, 49025–49043. `https://doi.org/10.48550/arXiv.2207.13332`.

Kpodo, J., Kordjamshidi, P., Nejadhashemi, A.P., 2024. Agxqa: A benchmark for advanced agricultural extension question answering. Computers and Electronics in Agriculture 225, 109349. `https://doi.org/10.1016/j.compag.2024.109349`.

Kuska, M.T., Wahabzada, M., Paulus, S., 2024. Ai for crop production–where can large language models (llms) provide substantial value? Computers and electronics in agriculture 221, 108924. `https://doi.org/10.1016/j.compag.2024.108924`.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33, 9459–9474. `https://doi.org/10.48550/arXiv.2005.11401`.

Li, S., Yuan, Z., Peng, R., Leybourne, D., Xue, Q., Li, Y., Yang, P., 2024. An effective farmer-centred mobile intelligence solution using lightweight deep learning for integrated wheat pest management. Journal of Industrial Information Integration 42, 100705. `https://doi.org/10.1016/j.jii.2024.100705`.

Liu, J., Zhou, C., Zhu, Y., Yang, B., Liu, G., Xiong, Y., 2025. Ricepest-detr: A transformer-based model for accurately identifying small rice pest by end-to-end detection mechanism. Computers and Electronics in Agriculture 235, 110373. `https://doi.org/10.1016/j.compag.2025.110373`.

Lv, B., Wu, H., Chen, W., Chen, C., Miao, Y., Zhao, C., 2024. Veg-mmkg: Multimodal knowledge graph construction for vegetables based on pre-trained model extraction. Computers and electronics in agriculture 226, 109398. `https://doi.org/10.1016/j.compag.2024.109398`.

Qian, X., Zhang, Y., Zhao, Y., Zhou, B., Sui, X., Zhang, L., Song, K., 2024. Timer4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 6942–6952. `https://doi.org/10.18653/v1/2024.emnlp-main.394`.

Rezayi, S., Liu, Z., Wu, Z., Dhakal, C., Ge, B., Dai, H., Mai, G., Liu, N., Zhen, C., Liu, T., Li, S., 2025. Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications. IEEE Transactions on Big Data 11, 1235–1246. `https://doi.org/10.1109/TBDATA.2024.3442542`.

Robertson, S., Zaragoza, H., et al., 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval 3, 333–389. `http://dx.doi.org/10.1561/1500000019`.

Shafay, M., Hassan, T., Owais, M., Hussain, I., Khawaja, S.G., Seneviratne, L., Werghi, N., 2025. Recent advances in plant disease detection: challenges and opportunities. Plant Methods 21, 140. `https://doi.org/10.1186/s13007-025-01450-0`.

Singh, A., Shraogi, N., Verma, R., Saji, J., Kar, A.K., Tehlan, S., Ghosh, D., Patnaik, S., 2024. Challenges in current pest management practices: Navigating problems and a way forward by integrating controlled release system approach. Chemical Engineering Journal 498, 154989. `https://doi.org/10.1016/j.cej.2024.154989`.

Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., Johal, S., King, B., 2023. Large language models and agricultural extension services. Nature food 4, 941–948. `https://doi.org/10.1038/s43016-023-00867-x`.

Vizniuk, A., Diachenko, G., Laktionov, I., Siwocha, A., Xiao, M., Smolg, J., 2025. A comprehensive survey of retrieval-augmented large language models for decision making in agriculture: Unsolved problems and research opportunities. Journal of Artificial Intelligence and Soft Computing Research 15, 115–146. `https://doi.org/10.2478/jaiscr-2025-0007`.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F., 2024. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672. `https://doi.org/10.48550/arXiv.2402.05672`.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35, 24824–24837. `https://doi.org/10.48550/arXiv.2201.11903`.

Wu, F., Liu, L., He, W., Liu, Z., Zhang, Z., Wang, H., Wang, M., 2024. Time-sensitve retrieval-augmented generation for question answering, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 2544–2553. `https://doi.org/10.1145/3627673.3679800`.

Yan, R., An, P., Meng, X., Li, Y., Li, D., Xu, F., Dang, D., 2025. A knowledge graph for crop diseases and pests in china. Scientific Data 12, 222. `https://doi.org/10.1038/s41597-025-04492-0`.

Yang, B., Zhang, Y., Feng, L., Chen, Y., Zhang, J., Xu, X., Aierken, N., Li, Y., Chen, Y., Yang, G., et al., 2025. Agrigpt: A large language model ecosystem for agriculture. arXiv preprint arXiv:2508.08632. `https://doi.org/10.48550/arXiv.2508.08632`.

Yao, X., Hao, X., Liu, R., Li, L., Guo, X., 2024. Agcner, the first large-scale chinese named entity recognition dataset for agricultural diseases and pests. Scientific Data 11, 769. `https://doi.org/10.1038/s41597-024-03578-5`.

Ye, K., Hu, G., Tong, Z., Xu, Y., Zheng, J., 2025. Key intelligent pesticide prescription spraying technologies for the control of pests, diseases, and weeds: A review. Agriculture 15, 81. `https://doi.org/10.3390/agriculture15010081`.

Yi, W., Zhang, L., Kuzmin, S., Gerasimov, I., Liu, M., 2025. Agricultural large language model for standardized production of distinctive agricultural products. Computers and Electronics in Agriculture 234, 110218. `https://doi.org/10.1016/j.compag.2025.110218`.

Youdao, N., 2023. Bcembedding: Bilingual and crosslingual embedding for rag. `https://github.com/netease-youdao/BCEmbedding/` (Accessed: 26 November 2025).

Yuan, Z., Liu, K., Li, S., Peng, R., Leybourne, D., Musa, N., Yang, P., 2025. Pezego: A precision agriculture system based on large language models and internet of things for pest management. IEEE Internet of Things Journal 12, 49120–49130. `https://doi.org/10.1109/JIOT.2025.3586374`.

Zhang, L., Zhang, H., Wang, C., Liang, P., 2024a. Rag-enhanced commit message generation. arXiv preprint arXiv:2406.05514. `https://doi.org/10.48550/arXiv.2406.05514`.

Zhang, S., Xue, Y., Zhang, Y., Wu, X., Luu, A.T., Zhao, C., 2024b. Mrag: A modular retrieval framework for time-sensitive question answering. arXiv preprint arXiv:2412.15540 `https://doi.org/10.48550/arXiv.2412.15540`.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J., 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176. `https://doi.org/10.48550/arXiv.2506.05176`.

Zhao, X., Chen, B., Ji, M., Wang, X., Yan, Y., Zhang, J., Liu, S., Ye, M., Lv, C., 2024. Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection. Agriculture 14, 1359. `https://doi.org/10.3390/agriculture14081359`.

Zheng, Y., Qi, J., Yang, Y., Yang, P., Yuan, Z., 2025. Aphid-yolo: A lightweight detection model for real-time identification and counting of aphids in complex field environments. IEEE Transactions on AgriFood Electronics 3, 605–614. `https://doi.org/10.1109/TAFE.2025.3600008`.