



ARTICLE

TRDC-RAG for Agricultural Question Answering: A Dual-Channel Retrieval Framework with Iterative Query Rewriting

Xiaoyu Li¹, Xiaoyi Nie^{1,*} and Ke Gu²

¹College of Information and Intelligence, Hunan Agricultural University, Changsha, China

²School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China

*Corresponding Author: Xiaoyi Nie. Email: xynie@hunau.edu.cn

Received: 10 June 2026; Accepted: 27 August 2026; Published: 22 October 2026

ABSTRACT: Retrieval-augmented generation (RAG) effectively mitigates knowledge deficits and hallucination issues in large language models (LLMs) for domain-specific question answering (QA). However, in specialized agricultural domains such as rapeseed cultivation, conventional RAG frameworks suffer from severe retrieval deviations and difficulties in handling complex relational dependencies due to highly colloquial user queries, intricate hierarchical knowledge structures, and implicit logical dependencies. To address these critical limitations, this study introduces TRDC-RAG, a dual-channel retrieval-augmented framework integrating a retrieval-feedback-driven iterative query rewriting mechanism. The proposed framework constructs a multi-layer structured agricultural knowledge base and deploys a retrieval-feedback-driven, two-round query rewriting paradigm. Specifically, a vector block consistency analysis module is developed to dynamically discriminate between strongly and weakly correlated semantic trajectories, executing differentiated weighting strategies to guide precise query reconstruction. During the retrieval phase, a dual-channel architecture operates concurrently, fusing dense vector spaces with structured knowledge graph subgraphs to seamlessly bridge unstructured contextual details and explicit relational logic. Extensive evaluations on a self-constructed rapeseed QA dataset and the benchmark WikiEval dataset via the RAGAS framework demonstrate that TRDC-RAG consistently and significantly out-performs state-of-the-art baselines, including Step-Back, Multi-Query, and RAG-Fusion. On the rapeseed domain corpus, the framework achieves state-of-the-art performance with Context Precision (0.8041), Context Recall (0.8285), Faithfulness (0.9657), and Answer Relevancy (0.9733), markedly suppressing factual hallucinations. These findings confirm that TRDC-RAG provides a robust, scalable, and highly reliable architecture for intelligent agricultural knowledge services without requiring intensive parameter fine-tuning.

KEYWORDS: Retrieval-augmented generation; agricultural question answering; large language models; query rewriting; knowledge graph; intelligent agriculture

1 Introduction

Question answering (QA) systems have become an important component of intelligent information services in both open-domain and domain-specific applications [1]. In knowledge-intensive fields such as agriculture, medicine, and finance, accurate question answering is essential for efficient knowledge acquisition and decision support. In recent years, the increasing complexity of agricultural production systems has generated substantial demand for intelligent agricultural QA services capable of supporting crop management, disease diagnosis, cultivation guidance, and yield prediction [2]. As one of the most important oil crops worldwide, rapeseed plays a critical role in agricultural production and food security.

Consequently, constructing intelligent QA systems for rapeseed-related knowledge has become an important research direction in smart agriculture [3].

Recent advances in large language models (LLMs), including PaLM [4], LLaMA [5], and GPT-4 [6], have significantly improved the capability of natural language understanding and text generation. Despite their strong performance in general-purpose tasks, these models still exhibit substantial limitations in domain-specific agricultural scenarios. In particular, agricultural knowledge often contains specialized terminology, implicit logical relationships, and strong contextual dependencies, making LLMs prone to factual hallucination and semantic misunderstanding when answering professional agricultural questions.

To mitigate these issues, retrieval-augmented generation (RAG) has emerged as an effective framework for enhancing LLMs with external knowledge [7]. By introducing retrieval mechanisms into the generation process, RAG can improve factual consistency and reduce hallucination problems [8]. Several representative frameworks, such as FiD [9] and REALM [10], have demonstrated the effectiveness of retrieval-enhanced generation in open-domain QA tasks. However, existing RAG frameworks still face several challenges in agricultural applications.

First, agricultural user queries are often colloquial and semantically ambiguous, resulting in insufficient retrieval accuracy during the initial retrieval stage. Existing query optimization methods mainly rely on single-round query rewriting or semantic expansion strategies [11], which frequently fail to capture implicit agricultural intent and contextual constraints. Second, agricultural knowledge usually involves complex entity relationships, including crop varieties, diseases, environmental conditions, and cultivation measures. Conventional dense vector retrieval methods primarily focus on semantic similarity while lacking the capability to model explicit logical relationships between agricultural entities. Although knowledge graphs (KGs) provide strong advantages in structured knowledge representation and logical reasoning [12], existing studies have not fully integrated graph reasoning with semantic retrieval in a unified framework. Recent studies such as KAG [13] have attempted to incorporate structured knowledge into LLM-based systems, but the interaction between retrieval feedback, query optimization, and multi-source knowledge fusion remains insufficiently explored.

To address these limitations, this study proposes TRDC-RAG (Two-Round Re-write and Dual-Channel Retrieval-Augmented Generation), a retrieval-enhanced framework for agricultural question answering. The proposed framework introduces a retrieval-feedback-driven iterative query rewriting mechanism based on a multi-layer agricultural knowledge base. Specifically, the framework performs two rounds of query rewriting and utilizes vector block consistency analysis to distinguish strongly related and weakly related knowledge directions, thereby dynamically optimizing query semantics according to retrieval feedback. The proposed retrieval-feedback optimization mechanism is designed to improve semantic alignment and retrieval effectiveness in domain-oriented intelligent question answering systems. Furthermore, a dual-channel retrieval architecture integrating dense vector retrieval and knowledge graph subgraph retrieval is designed to jointly leverage unstructured semantic information and structured relational knowledge. Finally, the retrieved knowledge is integrated into the GLM-4 [14] generation framework to improve reasoning consistency and answer reliability.

The main contributions of this study are summarized as follows:

- (1) A retrieval-feedback-driven two-round query rewriting mechanism is proposed to improve semantic understanding and retrieval accuracy for complex agricultural queries.
- (2) A dual-channel retrieval framework integrating dense vector retrieval and knowledge graph subgraph retrieval is constructed to enhance knowledge coverage, relational association modeling, and contextual completeness.

- (3) A rapeseed-domain knowledge question-answering dataset is developed, and extensive experiments on both domain-specific and public benchmark datasets demonstrate the effectiveness and robustness of the proposed framework.

2 Materials and Methods

The strategic imperative of the proposed TRDC-RAG framework is to systematically eliminate semantic ambiguity and retrieval drift in professional agricultural question answering systems. To achieve this, the architecture transcends conventional static, single-pass retrieval methods by implementing a “retrieval-evaluation-recalibration” closed-loop feedback mechanism, seamlessly coupled with a parallelized dual-channel text-graph fusion network.

2.1 Overall Architecture of TRDC-RAG

The query rewriting enhancement method proposed in this study is not designed as an independent system, but rather as an integrated component within a knowledge graph-based RAG framework. Its primary objective is to improve overall system performance by optimizing the upstream query rewriting module. The overall architecture of the TRDC-RAG system is illustrated in Fig. 1.

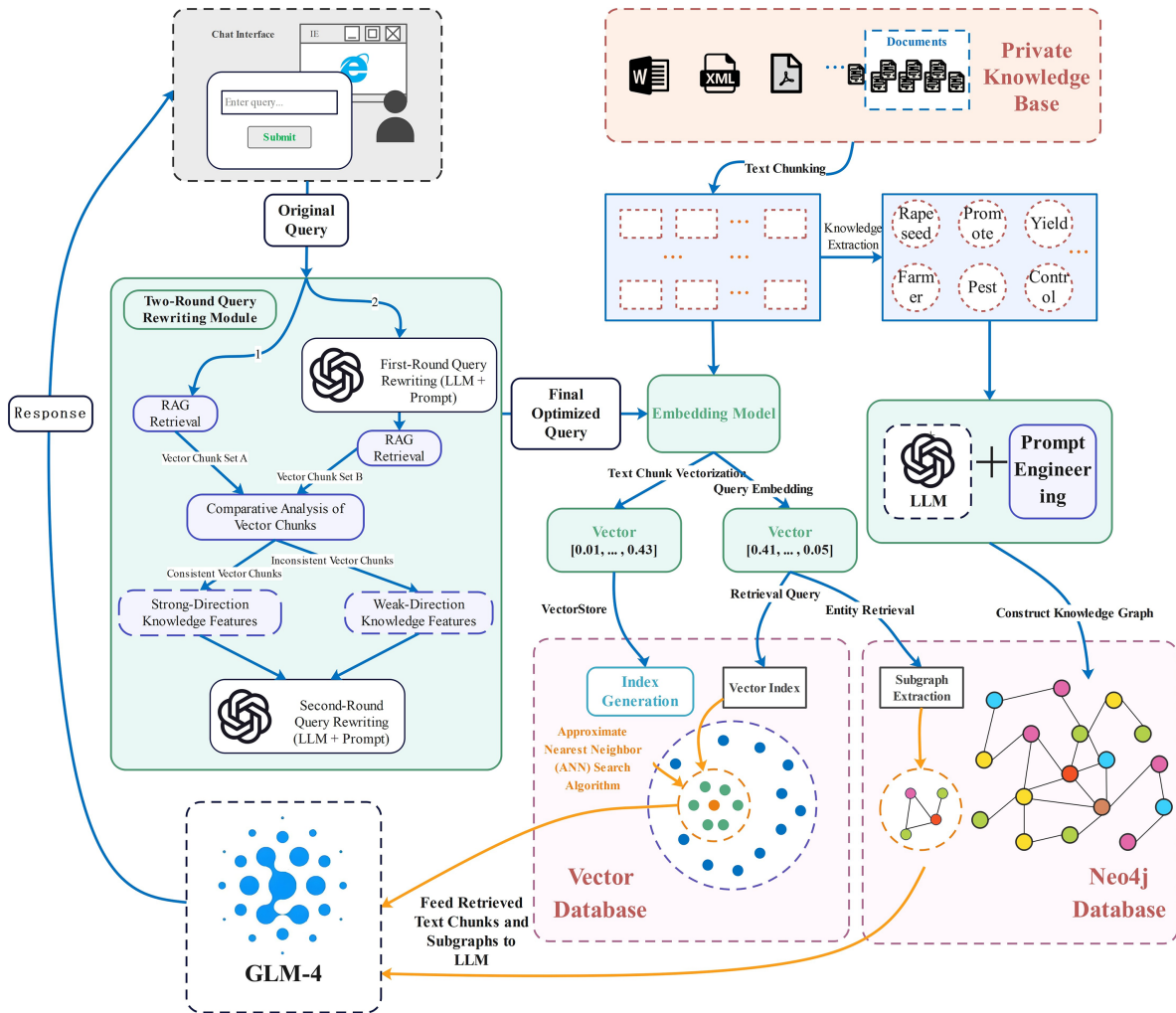


Figure 1: Architectural workflow of the TRDC-RAG framework.

The workflow of the proposed framework follows a three-stage pipeline: query rewriting, knowledge retrieval, and answer generation. Given an original user query, the system first processes it through the core two-round query rewriting module developed in this study. By introducing a retrieval feedback mechanism, this module transforms the conventional single-step semantic rewriting into a multi-step iterative optimization process, aiming to generate an optimal query that is better aligned with the domain-specific knowledge base.

The optimized query is then forwarded to a dual-channel knowledge retrieval layer, which integrates information from both vectorized textual data and knowledge graphs. The retrieved knowledge fragments are subsequently fed into the generation module, where the GLM-4 large language model performs comprehensive reasoning to produce accurate final answers. Furthermore, the incorporation of the knowledge graph enables visualization of the supporting evidence, thereby enhancing the interpretability and transparency of the generated responses.

Let the original user query be denoted as q^0 . The TRDC-RAG framework first performs semantic normalization on q^0 through a two-round query rewriting layer. By integrating the capabilities of the large language model with the constructed multi-layer knowledge base, an optimized query representation, denoted as q^2 , is generated. The overall process can be formally expressed as shown in Eq. (1).

$$q^2 = R(q^0) \quad (1)$$

where $R(\cdot)$ denotes a large language model-based function augmented by a multi-layer knowledge base, which is designed to enhance the clarity and domain relevance of the query.

Subsequently, in the dual-channel knowledge retrieval layer, the system performs textual vector retrieval and knowledge graph subgraph retrieval in parallel. In the textual vector retrieval channel, the optimized query q^2 is encoded into a vector representation v_q . Then, the $Top - k$ most relevant documents are retrieved from the document vector database, where $k = 3$, forming a document set D , as formally defined in Eq. (2).

$$D = Top - k(\cos(v_q, v_d), \forall d \in D) \quad (2)$$

where $\cos(\cdot)$ denotes the cosine similarity function. The symbol d represents a document in the document vector database, and v_d denotes the corresponding vector representation of document d .

In the knowledge graph channel, the system leverages the large language model in conjunction with the LightRAG framework to identify and extract entities (e.g., names, dates, locations, and events) and their corresponding relationships from knowledge documents, thereby constructing structured triples. Meanwhile, a structured metadata annotation scheme is employed to preserve critical information, including the source of the text, contextual positions, and entity categories, ensuring the integrity and traceability of the extracted knowledge.

Based on the extracted triples and associated metadata, a domain-specific knowledge graph is constructed. Subsequently, given the optimized query q^2 , an entity set E_q is identified from the knowledge graph. Taking the entities in E_q as seed nodes, a neighborhood subgraph G_q is retrieved, as formally defined in Eq. (3).

$$G_q = (v_i, r_{ij}, v_j) \in G \mid v_i, v_j \in E_q \quad (3)$$

where v_i and v_j represent the entity nodes in the knowledge graph, r_{ij} represents the semantic relationship between entities v_i and v_j , and G represents the complete domain knowledge graph.

Finally, an attention routing mechanism is designed in the knowledge generation layer to fuse the retrieval results from the two channels and construct a multi-source knowledge base K_q , as shown in Eq. (4).

$$K_q = D \cup G_q \quad (4)$$

Based on the knowledge base K_q , question reasoning is performed. Let $G(\cdot)$ denote the generation module; then, the final answer a is obtained as shown in Eq. (5).

$$a = G(q^2, K_q) \quad (5)$$

where q^2 denotes the final optimized query; K_q represents the multi-source knowledge base formed by integrating the text retrieval results with the knowledge graph subgraphs; and $G(\cdot)$ stands for the instruction-tuned large language model based on GLM-4, which achieves contextual understanding of the text data D and the graph subgraph G_q by fusing the encoding structure and the attention routing mechanism.

Through the synergistic operation of dynamic query optimization, dual-channel knowledge retrieval, and knowledge generation, the TRDC-RAG framework significantly enhances retrieval precision and answer quality for question-answering tasks in the rapeseed domain.

2.2 Overall Design of the Two-Round Query Rewriting Module

The detailed design of the two-round query rewriting module is illustrated in Fig. 2. This module leverages the content of the knowledge base to effectively optimize the original query, transforming the traditional isolated state of query optimization and knowledge retrieval. Its design can be decomposed into three sub-processes:

- (1) Initial rewriting and vector set generation: Facing the user's initial query, the module employs two independent execution paths to obtain retrieval feedback from the knowledge base. First, while maintaining the original semantic features of the query, it directly performs retrieval within the multi-layer knowledge base to recall vector block set A. This is designed to preserve the user's most original intent information and prevent semantic deviation caused by excessive rewriting. Second, it utilizes a large language model (LLM) combined with zero-shot prompt engineering to denoise and standardize the original query, primarily addressing issues such as colloquialisms, ambiguity, and missing basic information. Subsequently, it recalls vector block set B based on this rewritten query.
- (2) Quantitative determination of strong and weak directions: This is the core analysis stage. The system compares vector block set A with vector block set B, identifying "consistent vector blocks" and "inconsistent vector blocks" through semantic similarity calculation. The knowledge base content within the consistent vector blocks is defined as the strong direction, while the content in the inconsistent vector blocks is defined as the weak direction. Weight coefficients are then introduced to integrate the retrieved knowledge base content according to their respective weights, providing a basis for the second round of rewriting.
- (3) Final rewriting: Based on the obtained strong and weak directions and their corresponding weights, specific prompt constraints are formulated, such as "focus on the professional details of the {strong direction}" as the primary basis in the answer, and appropriately refer to the background information of the {weak direction}." Under this targeted guidance, the LLM performs the second round of query rewriting and outputs the final query. This output is no longer a simple natural language sentence, but a semantic instruction that integrates professional standardization and comprehensive knowledge coverage, thereby providing a precise direction for downstream retrieval tasks.

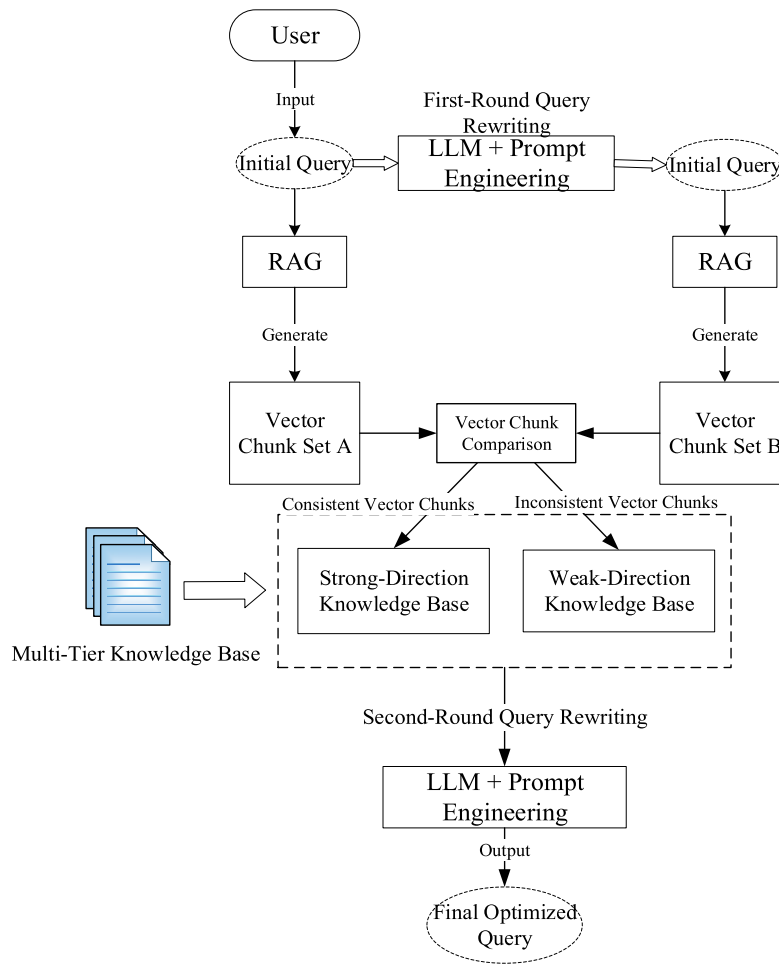


Figure 2: Framework of the two-round query rewriting module.

2.2.1 Design and Construction of the Multi-Layer Knowledge Base

Constructing a multi-layer knowledge base that can clearly reflect the internal organizational structure of domain knowledge is the foundation for the effective operation of the two-round query rewriting mechanism. Traditional retrieval-augmented generation systems typically store knowledge in a flattened vector format. Although this approach facilitates semantic similarity calculations, it neglects the hierarchical affiliations and domain associations among knowledge data. Therefore, based on a systematic analysis of the rapeseed domain knowledge system and the self-built question-answering (QA) dataset, this study constructs a three-layer knowledge base structure to achieve the orderly organization and accurate invocation of domain knowledge.

In terms of overall design, the multi-layer knowledge base adopts a hierarchical division that gradually refines from the macroscopic to the specific. The design of this structure is based on the domain QA dataset constructed in Chapter 3, ensuring a high degree of adaptability between the knowledge organization method and the final QA task. Specifically, the first layer (L1) is defined as the broadest domain boundary, namely “Agriculture”. The second layer (L2) is divided into five major categories based on the core technical links of the rapeseed industry: “Rapeseed Varieties”, “Rapeseed Breeding”, “Cultivation Management”, “Pest and Disease Control”, and “Nutritional Regulation”. These five categories completely cover the key knowledge domains from pre-production to post-production, constituting the pillar of the knowledge system. Finally,

the construction of the third layer (L3) does not use fixed themes for division; instead, it directly utilizes data from the “Remarks” field in the dataset, such as “Weed Damage”, “Rapeseed Cultivation”, and “Cropping System”. This design enables the underlying labels of the knowledge base to remain synchronized with the actual data distribution and user question keywords. For example, for the QA pair “What are the main symptoms of soft rot? (Remark: Weed Damage)” in the dataset, when the system processes the knowledge for storage, it not only extracts its text to generate vectors but also annotates it with structured path information: Agriculture/Pest and Disease Control/Weed Damage. This path accurately reflects the specific position of the knowledge data within the domain system, facilitating the subsequent retrieval of strong and weak direction knowledge bases.

In terms of representation, knowledge at different layers is not stored in isolation, but rather forms inter-layer associations through semantic mapping. For any knowledge fragment d_i , its representation in the knowledge base can be formalized as:

$$d_i \rightarrow \langle L_1, L_2, L_3 \rangle \quad (6)$$

where L_1 denotes the domain constraint, L_2 denotes the professional direction, and L_3 denotes the specific semantic label. During the vectorization process, the text content and its corresponding hierarchical labels are jointly encoded, ensuring that the final vector representation incorporates not only semantic information but also hierarchical structural features.

2.2.2 Design of the Strong and Weak Direction Determination Module

As the core decision-making component within the TRDC-RAG architecture, the strong and weak direction determination module serves the critical function of clarifying query intents and establishing weight constraints. Traditional query rewriting mechanisms rely solely on prompt-based constraints for query optimization, performing only a single round of optimization on the original query without the involvement of external knowledge bases. To ensure that the query rewriting results align more closely with user intents, this study designs a two-round query rewriting process. Concurrently, it leverages external knowledge bases by feeding the retrieved content as context to the large language model (LLM) to ultimately generate the optimal query.

In practical implementation, this mechanism relies on the dual-path retrieval results generated by the two-round query rewriting module. Let the user’s original query be q_0 . The system directly executes retrieval while maintaining the original semantics unchanged to obtain the vector block set $A = \{a_1, a_2, \dots, a_m\}$. Simultaneously, a second retrieval is performed using the query q' , which has undergone the first round of semantic standardization rewriting, yielding the vector block set $B = \{b_1, b_2, \dots, b_n\}$. These two vector block sets reflect the retrieval results under the user’s original intent and the denoised, standardized intent, respectively. The overlap between the two ensures the stability of the query semantics. To quantify this stability, the system determines the consistency relationship of the vector blocks through vector similarity calculations.

For any $a_i \in A$ and $b_j \in B$, cosine similarity is employed as the evaluation metric, as shown in Eq. (7):

$$\text{sim}(a_i, b_j) = \frac{\mathbf{v}_{a_i} \cdot \mathbf{v}_{b_j}}{\|\mathbf{v}_{a_i}\| \|\mathbf{v}_{b_j}\|} = \frac{\sum_{i=1}^d \mathbf{v}_{a_i} \mathbf{v}_{b_j}}{\sqrt{\sum_{i=1}^d (\mathbf{v}_{a_i})^2} \sqrt{\sum_{i=1}^d (\mathbf{v}_{b_j})^2}} \quad (7)$$

where $\text{sim}(a_i, b_j)$ represents the similarity score between vectors a_i and b_j ; d is the dimensionality of the vector space; and $\mathbf{v}_{a_i}$ and $\mathbf{v}_{b_j}$ denote the feature embedding vectors of knowledge fragments a_i and b_j in the high-dimensional vector space, respectively. A similarity threshold τ is established, with τ set to 0.85.

When $\text{sim}(a_i, b_j)$, the vector block is deemed consistent across the two rounds of retrieval. Consequently, the system can partition the retrieval results into two categories: vector blocks that appear or are highly similar in both sets A and B are classified as consistent vector blocks, while those appearing in only one of the retrievals are classified as inconsistent vector blocks. The knowledge base hierarchy and thematic direction corresponding to the consistent vector blocks are defined as the strong direction; conversely, the knowledge base direction corresponding to the inconsistent vector blocks is defined as the weak direction.

After completing the determination of strong and weak directions, the subsequent step involves assigning distinct weights to the different vector block sets, thereby integrating the directional importance into the second round of the query rewriting process. Let K_s denote the knowledge set corresponding to the strong direction, and K_w denote the knowledge set corresponding to the weak direction. When formulating the query rewriting constraints, different weight coefficients are allocated to them, as shown in Eq. (8).

$$w_s : w_w = 4 : 1 \quad (8)$$

where w_s and w_w represent the weights of the strong and weak directions, respectively. The configuration of these parameters is not based on subjective experience, but rather derived from pre-experimental empirical tuning on a validation subset. Specifically, for the similarity threshold, that $\tau < 0.80$ misclassified excessive semantic noise as consistent knowledge, whereas $\tau > 0.90$ led to overly strict filtering that discarded useful context. Thus, $\tau = 0.85$ was adopted to ensure optimal feature separation. Similarly, regarding the weight distribution, testing multiple ratio combinations, including 1:1, 2:1, and 5:1 revealed that when the ratio is lower than 4:1, the background information from the weak direction easily introduces noise, leading to a decrease in Context Precision and semantic drift. Conversely, when the ratio is higher than 4:1, the system tends to overlook necessary background constraints such as regional and climatic conditions, resulting in a reduction in Context Recall. Therefore, the 4:1 setting maximizes the integration of background constraints without diluting core agricultural facts, achieving an optimal balance between precision and recall.

Through the determination of strong and weak directions and the assignment of weights based on vector block consistency, not only is the targetedness of query rewriting enhanced, but a more stable and credible knowledge input is also provided for the subsequent dual-channel retrieval and generation stages. This constitutes a crucial foundation for realizing high-quality intelligent question-answering in the rapeseed domain.

2.2.3 Design of the Optimal Query Generation Module

Following the construction of the multi-layer knowledge base and the determination of strong and weak directions, this paper proposes a two-round optimization mechanism for query rewriting that incorporates an external knowledge base. This design prevents the query rewriting process from relying solely on prompts for semantic constraints, which often fails to fully utilize retrieved external knowledge. By involving the external knowledge base in semantic optimization during the query phase, this approach effectively advances retrieval augmentation to the query rewriting stage.

In the first round of query rewriting, the system primarily addresses the ill-formedness inherent in users' natural language, such as colloquial noise, syntactic disorganization, and the absence of key components. Let q^0 denote the original user query. A zero-shot prompting strategy is employed to generate the initially rewritten query, q^1 . This process is formalized as shown in Eq. (9):

$$q^1 = \mathcal{R}_1(q^0 | \mathcal{P}_1) \quad (9)$$

where $\mathcal{R}_1(\cdot)$ denotes the query rewriting operation executed under the semantic clarification and expression standardization constraint, $\mathcal{P}_1$. This first round of rewriting primarily focuses on maintaining the integrity of the user's original intent. The first-round rewriting prompt is designed to preserve the user's original intent while removing colloquial expressions and completing missing semantic components. The complete prompt template is provided in [Appendix A Table A1](#).

The second round of query rewriting constitutes the most critical generation phase. At this stage, the module must substantially integrate the retrieved sets of strong and weak directions into the rewriting process. Initially, the system executes a vector-to-text decoding operation to extract the original text content corresponding to each vector chunk. Based on their weight attributes, it then constructs a structured knowledge-enhanced context. Serving as the input condition for the second round of query rewriting, this context enables the external knowledge base to participate directly in the semantic reconstruction of the query.

The objective of the second round of query rewriting is to generate a final query optimized for domain-specific retrieval and reasoning. This query not only fully encapsulates the user's intent but also semantically focuses towards highly credible knowledge. Let q^2 denote the final rewritten query. This process is formalized as shown in [Eq. \(10\)](#):

$$q^2 = \mathcal{R}_2(q^0, q^1 \mid K_s, K_w, w_s, w_w, \mathcal{P}_2) \quad (10)$$

where $\mathcal{R}_2(\cdot)$ denotes the Large Language Model (LLM) generation operator functioning under the combined constraints of the external knowledge context and weights. K_s and K_w represent the sets of knowledge texts corresponding to the strong and weak directions, respectively, while w_s and w_w denote their respective weights. The second-round rewriting prompt incorporates strong-direction knowledge as core semantic constraints and weak-direction knowledge as supplementary contextual information. The complete prompt template is provided in [Appendix A Table A2](#).

Through the aforementioned design, the q^2 generated in the second round of rewriting is no longer a simple natural-language query. Instead, it serves as a semantic directive that integrates user intent with domain knowledge. During the subsequent dual-channel knowledge retrieval and generation phase, this directive exhibits superior directional alignment and comprehensiveness, effectively mitigating retrieval deviations caused by semantic ambiguity in complex agricultural problems.

2.3 Design of the Dual-Channel Knowledge Retrieval Module

Given the intricate entity relationships inherently present in rapeseed agricultural texts, a single retrieval mechanism often fails to meet the system's precision requirements for vertical domain knowledge. Therefore, following the preceding normalized query directives, this study constructs a dual-channel retrieval architecture operating in parallel with text vectors and graph subgraphs. This architecture aims to penetrate surface-level semantics, extracting knowledge evidence with semantic or entity associations from both the text corpus and the knowledge graph to form a complementary candidate knowledge set.

As illustrated in [Fig. 3](#), to address the issue where traditional fixed-size chunking methods for long documents tend to disrupt the coherence of agricultural knowledge, a LightRAG-based text segmentation strategy is introduced during the preprocessing stage. By segmenting along semantic boundaries, this strategy not only ensures that the resulting knowledge chunks accommodate the input requirements of the embedding model but also maximizes the preservation of contextual integrity. A tailored LLM-based entity-relation extraction prompt is employed to identify core entities and semantic relations from each text chunk. The complete extraction prompt and output format are provided in [Appendix A Table A3](#). These primarily

include core entities such as crop varieties, cultivation methods, pest and disease types, and nutritional elements, along with relation types like ‘belongs to’, ‘causes’, and ‘acts on’ [15]. To ensure the interpretability of this discrete knowledge, structured tags comprising document IDs and traceability paths are appended to the underlying data. Finally, the labeled data stream undergoes two distinct processing pathways: the extracted triples are constructed into a knowledge graph [16], while the segmented text chunks are mapped into dense vectors via the BAAI/bge-m3 model and ingested into a vector database.

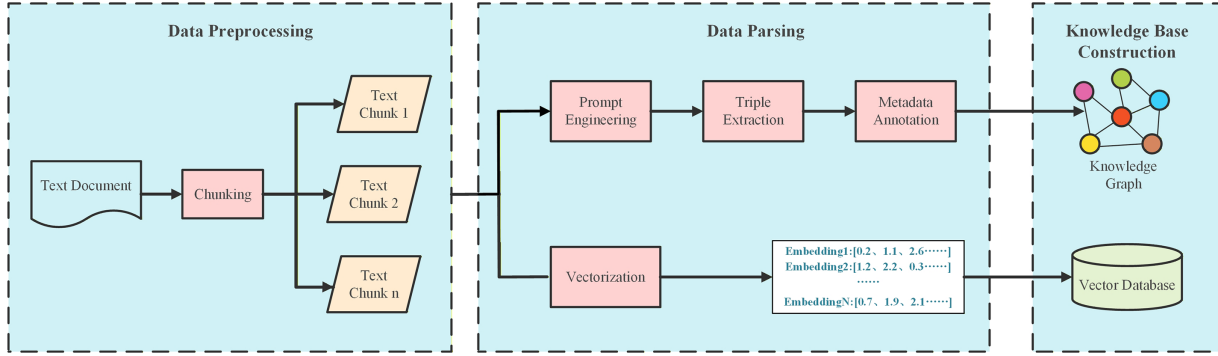


Figure 3: Data processing pipeline for knowledge graph construction and text vectorization.

In the text vector retrieval channel, the system employs cosine similarity as the metric to rank relevance within the semantic space, selecting the *Top – k* most relevant documents to form the candidate set D , as formalized in Eq. (2). This mechanism prioritizes capturing textual information highly aligned with the query semantics, ensuring comprehensive coverage of domain knowledge such as rapeseed cultivation techniques, pest and disease control, and variety breeding, thereby holistically enhancing the representational capacity for the query semantics.

Correspondingly, in the knowledge graph subgraph retrieval channel, an entity set E_q is first extracted from the rewritten query q^2 . Serving as the retrieval starting point, this set is utilized to extract the associated subgraph G_q from the rapeseed domain knowledge graph based on Eq. (3). Compared to pure text retrieval, this channel leverages the advantages of structured representation to more clearly delineate explicit and implicit knowledge—such as variety characteristics, pest and disease associations, and growth stage relationships—thereby compensating for the shortcomings of text-based methods in relation modeling and reasoning capabilities. These two retrieval pathways complement each other in information expression and knowledge organization structures, jointly underpinning a comprehensive knowledge acquisition framework.

Building upon this foundation, the document set D and the subgraph set G_q returned by the dual channels are uniformly integrated according to Eq. (4) to construct a multi-source knowledge base K_q , which serves as the input for the subsequent knowledge generation phase. Through this fusion process, the system’s ability to capture latent semantic information when addressing complex queries is significantly enhanced. The synergistic retrieval of text and graphs not only expands the breadth and depth of knowledge coverage but also provides more reliable and interpretable support for the generation module. In particular, the complementary retrieval pathways provide structured relational constraints and broader contextual evidence, which are beneficial for entity disambiguation, relational association, and logical consistency in complex agricultural QA.

2.4 Design of the Knowledge Generation Layer

Following semantic normalization and multi-source knowledge retrieval, the TRDC-RAG framework integrates knowledge evidence through the knowledge generation layer to produce the final answer for the user. To accommodate the semantic complexity and knowledge density inherent in question-answering within the Chinese rapeseed domain, this study incorporates GLM-4, part of the ChatGLM series developed by Zhipu AI, as the primary generation engine. Equipped with a deep comprehension of Chinese semantics and the capability to effectively retain long-context information, GLM-4 significantly enhances the logical consistency and content completeness during the generation process.

In terms of specific implementation, based on the knowledge generation process delineated in Eq. (5), the generation function $G(\cdot)$ is invoked using the ultimately optimized query q^2 and the fused knowledge base K_q as inputs, yielding the final answer a . Specifically, the document set D and the graph subgraph G_q , obtained via the dual-channel retrieval, are first encoded into a context embedding sequence $\{h_D, h_G\}$. Subsequently, the generation module calculates the attention weights for both types of information based on the optimal query q^2 , as formulated in Eq. (11).

$$\alpha_D = \text{softmax}(q'^T K_D^T), \alpha_G = \text{softmax}(q^2 K_G^T) \quad (11)$$

where K_D and K_G denote the key vectors of the document and the subgraph, respectively. By utilizing the attention weights α_D and α_G , the model can adaptively select information sources based on the current generation context. The final fused representation is defined as formulated in Eq. (12).

$$K_q = \alpha_D V_D + \alpha_G V_G \quad (12)$$

where V_D and V_G denote the value vectors of the document and the subgraph, respectively. This architectural design ensures that the answer generation process exploits both the granular details inherent in unstructured text and the structural constraints provided by the knowledge graph, thereby achieving semantic complementarity and enhancement.

The knowledge generation layer serves not only as the core output module of the TRDC-RAG framework but also as a pivotal bridge integrating semantic understanding, knowledge selection, and language generation. Capitalizing on the advantages of ChatGLM (GLM-4) in Chinese semantic modeling, and synergizing with the upstream dynamic queries and dual-channel knowledge construction, this layer facilitates the generation of controllable, high-quality answers for complex domain-specific problems. Consequently, it significantly bolsters the overall robustness and practical efficacy of the system.

3 Results

3.1 Experimental Setup

This section details the experimental configuration used to evaluate the impact of the proposed two-round query rewriting mechanism on the performance of the RAG system, primarily encompassing the experimental datasets, baseline methods, and evaluation metrics.

(1) Implementation Details

To ensure computational efficiency and system stability during large-scale corpus processing and model inference, all experiments were conducted on a robust hardware platform. The system was equipped with an Intel(R) Xeon(R) Gold 6338 CPU, 32 GB of RAM, and dual NVIDIA T4 GPUs (16 GB VRAM each), operating on Ubuntu 18.04. The software environment was constructed using Python 3.10 and the PyTorch 1.10.2 deep learning framework. For the core RAG components, the BAAI/bge-m3 model was employed

for text embedding generation, FAISS was used as the vector indexing library, and GLM-4 served as the foundational large language model for query rewriting and answer generation. Neo4j Community Edition 5.23.0 was utilized as the graph database to support knowledge graph construction, storage, and subgraph retrieval. Automated evaluation was conducted using the RAGAS framework (version 0.1.18). To ensure consistency, the chunk size is set to 1200 across all datasets, and the random seed was fixed at 42 across all computational processes.

(2) Datasets

To validate the practical effectiveness and domain adaptability of the TRDC-RAG framework in question answering tasks, this study focuses on rapeseed as the target domain and constructs two datasets from different sources. The first is a domain-specific QA corpus built from professional rapeseed knowledge, and the second is the publicly available WikiEval dataset used within the RAGAS framework, which serves to evaluate model performance across different scenarios.

Specifically, the open-source dataset generation tool Easy Dataset (<https://github.com/ConardLi/easy-dataset>) is employed to process rapeseed-related textual data, resulting in 1319 QA pairs. The dataset was collected from two authoritative sources to ensure the reliability, professionalism, and practical relevance of the knowledge. First, the academic knowledge corpus was derived from publicly available scientific literature and agricultural technical documents. A total of 231 core research articles were collected from authoritative databases, including CNKI and Web of Science, together with industry standards, technical specifications, and guideline documents issued by the Ministry of Agriculture and Rural Affairs of China. These resources provide a systematic and well-validated theoretical foundation for the knowledge base. Second, the practical knowledge corpus was compiled from agricultural production experiences summarized by domain experts and further verified through peer review. This subset contains approximately 70,000 words of textual content and covers five major knowledge domains: rapeseed varieties, breeding, cultivation management, pest and disease control, and nutritional regulation, as shown in Table 1. The corpus is constructed from domain literature and expert knowledge, and is manually annotated to ensure semantic clarity, standardized terminology, and factual correctness [17]. This dataset serves as the primary training and evaluation corpus for the TRDC-RAG framework in the agricultural domain, forming the basis for assessing its capability in professional semantic understanding and knowledge localization.

Table 1: Rapeseed knowledge question and answer pair.

Question	Answer	Remark
What are the main symptoms of soft rot disease?	The main symptoms of soft rot disease (<i>Pectobacterium carotovorum</i>) include water-soaked soft decay at the basal stem of rapeseed plants, often accompanied by a foul odor. The disease is primarily caused by bacterial invasion through insect wounds or natural plant openings, or via insect-borne transmission. Control measures include chemical control of insect vectors, ridge cultivation to reduce field humidity, and the application of well-decomposed fertilizers to prevent pathogen introduction.	Weed/ Pest-Related

(Continued)

Table 1 (continued)

Question	Answer	Remark
What is the primary transmission pathway of rapeseed viral diseases?	The primary transmission pathway of rapeseed viral diseases is through aphids. Large-scale outbreaks of aphids often lead to the spread of viruses. In particular, Turnip mosaic virus (TuMV) is the dominant virus affecting rapeseed in China, and its transmission mainly depends on aphid vectors. Therefore, aphids are the key factor in the spread of rapeseed viral diseases.	Weed/Pest-Related
What roles does rapeseed play in crop rotation systems?	Rapeseed plays several important roles in crop rotation systems: (1) enhancing soil fertility by improving nutrient content; (2) improving soil microbial community structure by promoting the growth and diversity of beneficial microorganisms. These functions contribute to both ecological and economic benefits in agricultural systems.	Rapeseed Cultivation
What are the main reasons for the northward expansion of winter rapeseed cultivation in recent years?	The primary reason for the northward expansion of winter rapeseed cultivation is the successful breeding of cold-resistant Brassica-type winter rapeseed varieties. This advancement enables rapeseed to adapt to colder climates, extending its cultivation range to regions as far north as 48° N latitude, thereby overcoming previous climatic limitations.	Cropping System

To further evaluate the generalization and robustness of the proposed method, the publicly available benchmark dataset WikiEval from the RAGAS framework is introduced. This dataset is derived from 50 Wikipedia pages and covers diverse domain knowledge, including corresponding questions, answers, and contextual information. It enables the assessment of TRDC-RAG in heterogeneous and complex document environments, particularly in terms of robustness and retrieval-augmented generation capability.

(3) Baseline Methods

To objectively evaluate the effectiveness of the proposed two-round query rewriting mechanism in improving generation quality, the following baseline methods are adopted for comparison. All methods share the same underlying retrieval pipeline to obtain initial candidate contexts and utilize the same GLM-4 model as the answer generator, ensuring a fair comparison.

Standard RAG: The standard RAG approach segments documents into fixed-length text chunks and stores them in a vector database. During retrieval, semantic similarity between the user query and the stored text chunks is computed to identify the most relevant passages. The retrieved content is then provided as contextual information to the large language model for answer generation.

Decomposition [18]: This method decomposes a complex query into multiple sub-questions. Each sub-question is independently retrieved and answered, and the final response is constructed by aggregating the intermediate results.

Step-Back Prompting [19]: This approach rewrites a specific query into a more general form. Relevant information is then retrieved based on both the original and the generalized queries, and the final answer is generated by integrating the retrieved knowledge.

Multi-Query: Multiple reformulated versions of the user query are generated and used to query the vector database, thereby increasing the likelihood of retrieving relevant documents.

RAG-Fusion [20]: This method generates multiple queries and re-ranks the retrieved results by combining reciprocal rank fusion with customized vector similarity weighting, leading to more comprehensive and accurate retrieval outcomes.

(4) Evaluation Metrics

To comprehensively evaluate the performance of the proposed TRDC-RAG framework in retrieval-augmented question answering tasks, this study adopts the automated evaluation framework RAGAS (Retrieval-Augmented Generation Assessment) proposed by Shahul et al. [21]. RAGAS is a reference-free evaluation method that does not require manually annotated data and is specifically designed for assessing RAG-based systems.

The evaluation process relies on four key components: the user query, the retrieved relevant contexts, the answer generated by the RAG system, and the reference (ground-truth) answer. Based on these inputs, RAGAS defines multiple evaluation metrics to assess system performance from different perspectives.

Among these, Context Precision measures the degree of relevance between the retrieved documents and the generated answer, reflecting whether the system accurately retrieves knowledge fragments that are helpful for answering the query, as formally defined in Eq. (13).

$$\text{Context_precision} = \frac{|TP|}{K} \quad (13)$$

where K denotes the total number of retrieved documents, and $|TP|$ represents the number of documents evaluated by the LLM as helpful to the user query. This metric calculates the ratio of helpful documents to the total number of retrieved documents.

Context recall measures the extent to which the retrieved documents contain the necessary information to deduce the ground truth answer, as shown in Eq. (14).

$$\text{Context_recall} = \frac{|V_t|}{|S_t|} \quad (14)$$

where $|S_t|$ denotes the total number of key points extracted from the ground truth answer using an LLM, and $|V_t|$ represents the number of these key points for which corresponding information can be identified within the context, as evaluated by the LLM.

Faithfulness evaluates the extent to which the generated answer is faithful to the retrieved context. Given the propensity of LLMs to hallucinate, in an ideal RAG system, the answer should be entirely derived from the relevant context, as formulated in Eq. (15).

$$F = \frac{|V_a|}{|S_a|} \quad (15)$$

where $|S_a|$ denotes the total number of key points extracted from the generated answer using an LLM, and $|V_a|$ represents the number of these key points that the LLM verifies as being inferable from the relevant context.

Answer_relevancy reflects the degree of alignment between the generated answer and the user query, as shown in Eq. (16)

$$\text{Answer_relevancy} = \frac{1}{n} \sum_{i=1}^n \text{sim}(q, q_i) \quad (16)$$

where n denotes the number of generated questions, q_i represents the i generated question, and q is the original user query.

3.2 Comparative Experiments

To comprehensively evaluate the performance differences of the models in both the vertical domain and general scenarios, this section conducts comparative experiments on the self-constructed rapeseed QA dataset and the public WikiEval dataset, respectively. The evaluation metrics strictly employ the four core metrics of the RAGAS framework: Context_precision, Context_recall, Faithfulness, and Answer_relevancy. The experimental results are presented in [Tables 2 and 3](#).

Table 2: Comparison of experimental results on the rapeseed knowledge question.

Methods	Context_precision	Context_recall	Faithfulness	Answer_relevancy
Standard RAG	0.6433	0.7125	0.7000	0.7240
Step back	0.6500	0.5667	0.8500	0.9604
Decomposition	0.8000	0.7000	0.9417	0.7740
Multi query	0.7900	0.8000	0.8550	0.9681
RAG-fusion	0.8000	0.6333	0.8143	0.9708
TRDC-RAG	0.8041	0.8285	0.9657	0.9733

Table 3: Comparison of experimental results on the WikiEval dataset.

Methods	Context_precision	Context_recall	Faithfulness	Answer_relevancy
Standard RAG	0.7233	0.6982	0.7125	0.7520
Step back	0.7500	0.5000	0.7633	0.8625
RAG-fusion	0.8021	0.7833	0.9150	0.8902
Decomposition	0.8105	0.7750	0.8000	0.9025
Multi query	0.8000	0.7823	0.9256	0.9186
TRDC-RAG	0.8250	0.8075	0.9633	0.9263

In the rapeseed domain dataset, as shown in [Table 2](#), the TRDC-RAG method achieves optimal performance across all four metrics. Particularly regarding the Context_recall metric, which targets retrieval completeness, TRDC-RAG reaches 0.8285, significantly outperforming RAG-Fusion (0.6333) and Step-Back (0.5667). This indicates that when confronting terminology-dense and complex agricultural queries, the question abstraction approach of Step-Back is prone to losing key features. In contrast, guided by dual-channel retrieval and multi-layer knowledge base constraints, TRDC-RAG effectively combines the precise retrieval of core keywords with comprehensive coverage of background information. Regarding generation quality, TRDC-RAG's high Faithfulness of 0.9657 and Answer_relevancy of 0.9733 directly validate the efficacy of the second-round knowledge-injected rewriting. This compels the LLM to strictly adhere to professional facts for reasoning, rather than engaging in divergent associations. Conversely, although the Decomposition method exhibits a Context_precision advantage (0.8000) when dismantling complex questions, its excessive segmentation disrupts semantic coherence, causing its Answer_relevancy to drop to 0.7740.

In the general WikiEval dataset, as shown in [Table 3](#), TRDC-RAG consistently maintains a comprehensive leading advantage, with both its Context_precision (0.8250) and Faithfulness (0.9633) reaching the

highest values. Notably, compared to the domain dataset, the overall performance of all baseline methods is more balanced on WikiEval, reflecting the relatively lower semantic complexity and knowledge barriers of the general dataset. However, TRDC-RAG exhibits no performance degradation in this scenario; rather, it further widens the gap with the baseline models. This proves that the framework's mechanism for determining strong and weak directions—based on vector chunk consistency—possesses excellent robustness and cross-domain generalization capabilities, without overfitting to the specific agricultural corpus.

3.3 Ablation Study

To further verify the actual contributions of the key modules within the TRDC-RAG framework to the overall performance enhancement, an ablation study was designed and conducted in this section. By systematically removing core constituent modules under uniform datasets and evaluation metrics, the performance variations across different configurations were compared. This allows for an analysis of the roles and necessity of the strong/weak direction determination and the two-round query rewriting mechanisms in the rapeseed knowledge QA task.

Regarding the experimental setup, the complete TRDC-RAG model served as the control baseline to construct three ablated variants. The first is TRDC-RAG (Unweighted), a variant with the strong/weak direction determination module removed. Although it retains the two-round rewriting architecture, it merely concatenates the initial retrieval results in the second round without applying differential weights, aiming to verify the effectiveness of the feedback weighting mechanism. The second is TRDC-RAG (One-round), a variant omitting the second-round rewriting phase. This model solely executes the initial standardized rewriting and a single retrieval, essentially functioning as a basic RAG model equipped with hierarchical storage. It is utilized to assess the core contribution of the iterative mechanism in resolving semantic deviation issues. The third is TRDC-RAG (Text-only), which removes the knowledge graph retrieval channel while preserving the complete two-round query rewriting and dense vector retrieval mechanisms. This variant relies solely on vector-based retrieval and is employed to assess the contribution of graph-enhanced retrieval to semantic association modeling and contextual completeness in agricultural question answering.

As presented in [Table 4](#), the comparative analysis between TRDC-RAG (Full) and TRDC-RAG (One-round) reveals that the omission of the second-round query rewriting mechanism exerts the most pronounced negative impact on overall model performance. Specifically, Context_recall experiences a substantial decline of 18.55 percentage points, a relative drop of approximately 22.4%, and Faithfulness decreases by 12.36 percentage points. This outcome indicates that in complex agricultural knowledge QA scenarios, relying solely on a single round of query normalization and denoising is insufficient to fully uncover the latent complex semantics and relevant background constraints embedded in the queries. Given that agricultural inquiries frequently involve multidimensional associations—such as pathogens, environmental factors, and management practices—a single semantic optimization struggles to achieve comprehensive coverage. By introducing the two-round iterative rewriting, the system is empowered to recalibrate query intent based on initial retrieval feedback. This broadens the retrieval perspective from localized semantic matching to a more comprehensive semantic space, thereby significantly enhancing both the breadth of knowledge recall and the completeness of the generated content.

A further comparison between the TRDC-RAG (Full) and TRDC-RAG (Unweighted) configurations reveals that, while retaining the two-round rewriting pipeline, the model lacking strong/weak direction weight modeling maintains a relatively high Context_recall (decreasing by merely 1.76 percentage points). However, its Context_precision and Faithfulness suffer reductions of 5.19 and 7.23 percentage points, respectively. This phenomenon demonstrates that in the absence of strong/weak direction discrimination, a substantial volume of weakly relevant knowledge from the initial retrieval is indiscriminately incorporated

into the rewriting process. This inevitably dilutes critical information with background noise, consequently causing the generation model to overlook core factual details.

Table 4: Performance comparison of different TRDC-RAG variants in the ablation study.

Methods	Context_precision	Context_recall	Faithfulness	Answer_relevancy
TRDC-RAG (Text-Only)	0.7385	0.7642	0.8416	0.8874
TRDC-RAG (Unweighted)	0.7522	0.8109	0.8934	0.9105
TRDC-RAG (One-Round)	0.6854	0.6430	0.8421	0.8566
TRDC-RAG (Full)	0.8041	0.8285	0.9657	0.9733

A further comparison between the TRDC-RAG (Full) and TRDC-RAG (Text-only) configurations highlights the indispensable role of the knowledge graph retrieval channel in agricultural question answering. While preserving the complete two-round query rewriting mechanism, the text-only variant removes graph-based retrieval and relies solely on dense vector retrieval for contextual knowledge acquisition. Experimental results show that Context_precision, Context_recall, Faithfulness, and Answer_relevancy decrease by 8.16%, 7.76%, 12.85%, and 8.83%, respectively. Among these metrics, the most pronounced degradation is observed in Faithfulness, indicating that vector-based retrieval alone is insufficient for maintaining factual consistency in knowledge-intensive agricultural QA tasks.

The aforementioned experimental results collectively illustrate that the performance superiority of TRDC-RAG stems from the holistic design of multi-mechanism synergy, rather than localized improvements to isolated modules. Specifically, the two-round query rewriting mechanism establishes a robust foundation for knowledge coverage, while the strong/weak direction discrimination—underpinned by the multi-tier knowledge base—imposes effective constraints on semantic precision. The absence of either mechanism leaves the model vulnerable to insufficient coverage or semantic drift during complex rapeseed agricultural QA tasks. Only through the effective integration of both can the system strike a stable balance between knowledge breadth and professional depth, thereby sustaining the generation of high-quality and trustworthy domain knowledge.

4 Discussion

To address the prevalent issues of ambiguous query intents, semantic deviation, and insufficient utilization of multi-source knowledge in domain-specific question answering (QA) within complex agricultural scenarios, the proposed TRDC-RAG framework demonstrates superior performance across multiple core evaluation metrics. Specifically, the framework achieves a context precision of 0.8041, a context recall of 0.8285, alongside answer faithfulness and relevance scores of 0.9657 and 0.9733, respectively. These results indicate that traditional Retrieval-Augmented Generation (RAG) methods, which rely on static, single-round semantic rewriting, struggle to effectively accommodate terminology-intensive and semantically complex query tasks in the agricultural vertical domain.

Mechanistically, TRDC-RAG enables the dynamic identification of both strongly and weakly correlated knowledge trajectories by constructing a three-tier rapeseed knowledge base and employing a semantic analysis mechanism based on vector chunk consistency. Unlike conventional single-round query optimization methods, the proposed framework establishes a “retrieval-evaluation-recalibration” closed-loop feedback mechanism [22]. Following the initial retrieval, the system filters out low-relevance noise and integrates high-confidence knowledge as semantic constraints into subsequent query rewriting stages. This significantly enhances semantic alignment capabilities and retrieval accuracy for complex agricultural

queries. Ablation experiments further validate the critical role of this mechanism. The removal of the two-round query rewriting mechanism led to a marked decline in both context precision and answer relevance, demonstrating the profound significance of retrieval feedback-driven dynamic semantic optimization in complex agricultural QA tasks.

Building upon existing research regarding the logical reasoning limitations of Large Language Models (LLMs) [23], the present results further elucidate the strong complementarity between unstructured text vector retrieval and knowledge graph (KG) relational reasoning. In traditional LLM applications based on plain-text semantic matching, dense vector retrieval effectively captures implicit semantic similarities; however, it remains susceptible to relational ambiguity and semantic drift when processing queries with explicit agricultural logic constraints, such as “disease transmission pathways” or “cultivar attribution.” To overcome these challenges, the proposed dual-channel collaborative retrieval mechanism effectively compensates for the deficiencies inherent in single-mode retrieval paradigms. Specifically, KG subgraph retrieval supplies explicit entity relation constraints and structured reasoning pathways, while an attention-based fusion mechanism dynamically integrates multi-source knowledge. This facilitates the synergistic modeling of structured relational knowledge and unstructured semantic information. Experimental findings confirm that this mechanism not only enhances the logical consistency of complex agricultural reasoning tasks but also substantially mitigates factual hallucinations (factual deviations) in domain-specific QA.

Within the context of smart agriculture, the TRDC-RAG framework delineates a viable pathway for extending LLMs from general-purpose QA to specialized, vertical-domain knowledge services. Agricultural knowledge is inherently characterized by strong regionality, empirical dependence, and dynamic evolution, rendering traditional general LLMs ill-equipped to stably meet the demands of complex agricultural scenarios. This study demonstrates that, without the necessity for large-scale parameter fine-tuning of the foundational LLM, the dynamic enhancement via multi-tier knowledge bases and the dual-channel retrieval mechanism can effectively augment the model’s knowledge utilization capacity and generation stability within the agricultural domain [24]. These outcomes not only offer robust support for rapeseed cultivation management, pest and disease control, and agricultural knowledge services, but also establish a reference paradigm for the deployment of artificial intelligence technologies in other specialized vertical domains.

Despite the promising experimental results achieved by the proposed framework in the rapeseed domain QA task, certain limitations remain. First, the current system is predominantly oriented towards textual corpora and structured KGs, lacking the integration of multimodal agricultural information such as agricultural imagery, lesion recognition photographs, and environmental perception data. Given that agricultural diagnostic tasks inherently rely heavily on visual cues, future research must incorporate Multimodal Large Language Models (MLLMs) to facilitate cross-modal agricultural knowledge retrieval and reasoning [25]. Second, agricultural knowledge exhibits a highly dynamic nature; however, the current construction processes for the multi-tier knowledge bases and KGs still necessitate partial manual intervention, resulting in relatively high construction and maintenance costs. Consequently, further investigation into LLM-driven automated KG updating and dynamic verification mechanisms is warranted to enhance the system’s long-term adaptability and cross-domain generalization capabilities within open agricultural environments. Finally, the dual-channel retrieval architecture and two-round query rewriting mechanism inevitably introduce additional computational latency compared to Standard RAG. Specifically, empirical measurements show that the average inference latency per query increases from approximately 7.3 s (Standard RAG) to 13.2 s in TRDC-RAG. This latency overhead stems primarily from the sequential LLM query rewriting and graph traversal operations. In addition, although the key parameters, including the similarity threshold, strong–weak directional weighting ratio, and Top-k value, were determined through preliminary empirical tuning, a systematic sensitivity analysis across a wider parameter space was not conducted

in the current study. Future work will investigate the effects of these parameters on retrieval precision, context recall, and generation quality through controlled sensitivity experiments, thereby providing a more comprehensive assessment of the robustness of TRDC-RAG.

5 Conclusions

This study proposes TRDC-RAG, a dual-channel retrieval-augmented framework with iterative query rewriting for agricultural question answering. To address the limitations of conventional RAG frameworks in semantic understanding, retrieval accuracy, and relational knowledge integration within complex agricultural scenarios, the proposed framework introduces a retrieval-feedback-driven two-round query rewriting mechanism and integrates dense vector retrieval with knowledge graph subgraph retrieval.

Experimental results on both the self-constructed rapeseed knowledge question-answering dataset and the public WikiEval benchmark demonstrate that TRDC-RAG consistently outperforms several representative baseline methods in context precision, context recall, faithfulness, and answer relevancy. The ablation studies further verify the effectiveness of the iterative rewriting strategy and the strong–weak directional weighting mechanism in improving semantic alignment and reducing irrelevant retrieval information.

Compared with conventional retrieval-enhanced generation frameworks, TRDC-RAG provides stronger semantic understanding capability and more reliable knowledge integration for agricultural question answering tasks. The proposed framework demonstrates the potential of combining retrieval feedback optimization with structured relational reasoning in domain-oriented intelligent QA systems.

Despite these promising results, several limitations remain. First, the current framework is primarily designed for text-based agricultural knowledge and does not yet incorporate systematic multimodal or multi-hop reasoning benchmarks. Second, although key parameters were determined through preliminary empirical tuning, comprehensive sensitivity analysis was not conducted. Third, the current experiments were performed under a fixed random seed and do not include statistical significance testing across multiple independent runs. Future research will therefore investigate multimodal agricultural QA, dedicated multi-hop reasoning benchmarks, systematic parameter sensitivity analysis, and repeated-run statistical significance testing. In parallel, lightweight retrieval strategies and automated knowledge graph updating will be explored to reduce inference overhead and improve scalability in practical agricultural service applications.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Hainan Provincial Sanya Yazhou Bay Science and Technology Innovation Joint Project, grant number ZDYF2025GXJS154.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Xiaoyu Li and Xiaoyi Nie; methodology, Xiaoyu Li; software, Xiaoyu Li; validation, Xiaoyu Li, Xiaoyi Nie and Ke Gu; formal analysis, Xiaoyu Li; investigation, Ke Gu; resources, Xiaoyi Nie and Ke Gu; data curation, Xiaoyu Li; writing—original draft preparation, Xiaoyu Li; writing—review and editing, Xiaoyi Nie and Ke Gu; visualization, Xiaoyu Li; supervision, Xiaoyi Nie; project administration, Xiaoyi Nie; funding acquisition, Xiaoyi Nie. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Xiaoyu Li, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his/her role as [EB] of this journal, [Gu Ke] had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

Abbreviations

LLM Large Language Model

RAG Retrieval-Augmented Generation

Appendix A

Table A1: Prompt design for the first-round query rewriting.

Role

You are an agricultural information expert proficient in rapeseed cultivation technology, responsible for transforming colloquial queries into standard written search commands.

Skill: Query Rewriting

Please perform denoising and standardization on the following user query. Requirements:

1. Preserve the original intent; adding unmentioned information is strictly prohibited.
2. Replace colloquialisms with standard agronomic terminology.
3. Complete any omitted subjects.

Output Requirements:

Output only the rewritten sentence; do not include any explanations.

Table A2: Prompt design for the second-round query rewriting.

##Role:

You are a query optimization assistant based on an external knowledge base. The system has retrieved relevant reference materials. Your task is to reconstruct the most precise query based on the provided information.

Skill: Query Rewriting

Instruction:

Based on the core facts (strong guidance) and reference background (weak guidance) provided by the system, reconstruct the initially rewritten user query.

Input Context:

1. User Query: {first_round_query}
2. Core Facts–Mandatory Inclusion: {decoded_strong_text_list}
3. Reference Background–Supplementary: {decoded_weak_text_list}

Rewriting Requirements:

- 1. Core Anchoring:** The generated query must explicitly include domain-specific terms from the *core facts* to ensure high retrieval precision.
- 2. Context Enrichment:** Incorporate relevant information from the *reference background*, such as temporal constraints, climatic conditions, or growth stages, to improve semantic completeness.
- 3. Logical Reconstruction:** Integrate the above information into a coherent and logically structured complex query, rather than a simple aggregation of keywords.

Output Requirement:

Only output the rewritten query sentence. Do not include any explanations.

Table A3: Design of prompt words for entity–relation extraction in rapeseed domain.**## Role**

You are a Knowledge Graph Expert capable of extracting knowledge triples based on user input.

Skill: Triple Extraction

According to the user's input requirements, and in conjunction with the textual information in the knowledge base {Rapeseed Breeding and Cultivation}, comprehensively extract all triple information present in the text. The triple information includes entities (e.g., *Brassica rapa* type rapeseed, rapeseed oil, flea beetles, etc.) and relations (e.g., belongs to, originates from, method, etc.), and must be output strictly following the specified format.

Output Structure:

{“Entity”: “Rapeseed”, “Relation”: “suffers from disease”, “Disease Name”: “Clubroot”}
 {“Entity”: “Rapeseed Clubroot”, “Relation”: “control method”, “Measure”: “Crop rotation”}
 {“Entity”: “Rapeseed”, “Relation”: “yield influencing factor”, “Factor”: “Soil pH”}

Constraints

1. Extract triples strictly in accordance with the defined ontology structure.
2. You are solely responsible for information extraction; do not answer any questions or provide additional explanations.

References

1. Wang T, Wang N, Cui YP, Liu J. Agricultural technology knowledge intelligent question-answering system based on large language model. *Smart Agric.* 2023;5(4):105–16. doi:10.12133/j.smartag.SA202311005.
2. Usigbe MJ, Asem-Hiablie S, Uyeh DD, Iyiola O, Park T, Mallipeddi R. Enhancing resilience in agricultural production systems with AI-based technologies. *Environ Dev Sustain.* 2024;26(9):21955–83. doi:10.1007/s10668-023-03588-0.
3. Tzachor A, Devare M, Richards C, Pypers P, Ghosh A, Koo J, et al. Large language models and agricultural extension services. *Nat Food.* 2023;4(11):941–8. doi:10.1038/s43016-023-00867-x.
4. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. *J Mach Learn Res.* 2023;24:1–113.
5. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. *arXiv:2302.13971.* 2023. doi:10.48550/arXiv.2302.13971.
6. Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, et al. Peer review of GPT-4 technical report and systems card. *PLoS Digit Health.* 2024;3(1):e0000417. doi:10.1371/journal.pdig.0000417.
7. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst.* 2020;33:9459–74.
8. Wen S, Qian L, Hu M, Chang Z. Review of research progress on question-answering techniques based on large language models. *Data Anal Knowl Discov.* 2024;8(6):16–29.
9. Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Kerrville, TX, USA: Association for Computational Linguistics; 2021. p. 874–80. doi:10.18653/v1/2021.eacl-main.74.
10. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: *Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria.* p. 3929–38.
11. Li M, Lv X, Zou J, Chen T, Zhang C, An S, et al. Query expansion in the age of pre-trained and large language models: a comprehensive survey. *ACM Trans Inf Syst.* 2026;44(5):1–42. doi:10.1145/3816248.

12. Zhang H, Wang X, Han L, Li Z, Chen Z, Chen Z. Research on question answering system on joint of knowledge graph and large language models. *J Front Comput Sci Technol.* 2023;17:2377–88. doi:10.3778/j.issn.1673-9418.2308070.
13. Liang L, Bo Z, Gui Z, Zhu Z, Zhong L, Zhao P, et al. KAG: boosting LLMs in professional domains via knowledge augmented generation. In: *Proceedings of the ACM on Web Conference; 2025 Apr 28–May 2; New York, NY, USA.* p. 334–43. (In Chinese). doi:10.1145/3701716.3715240.
14. Glm T, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. *arXiv:2406.12793.* 2024. doi:10.48550/arXiv.2406.12793.
15. Li DM, Luo SS, Zhang XP, Xu F. Review on named entity recognition. *J Front Comput Sci Technol.* 2022;16(9):1954–68. (In Chinese).
16. Guo Z, Xia L, Yu Y, Ao T, Huang C. LightRAG: simple and fast retrieval-augmented generation. In: *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025.* Kerrville, TX, USA: Association for Computational Linguistics; 2025. p. 10746–61. doi:10.18653/v1/2025.findings-emnlp.568.
17. Jiang J, Yan L, Liu H, Xia Z, Wang H, Yang Y, et al. Knowledge assimilation: implementing knowledge-guided agricultural large language model. *Knowl Based Syst.* 2025;314:113197. doi:10.1016/j.knosys.2025.113197.
18. Zhou D, Schärli N, Hou L, Wei J, Scales N, Wang X, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv:2205.10625.* 2022. doi:10.48550/arXiv.2205.10625.
19. Zheng HS, Mishra S, Chen X, Cheng HT, Chi EH, Le QV, et al. Take a step back: evoking reasoning via abstraction in large language models. In: *Proceedings of the International Conference on Learning Representations 2024; 2024 May 7–11; Vienna, Austria.* p. 20279–316.
20. Rackauckas Z. Rag-fusion: a new take on retrieval-augmented generation. *arXiv:2402.03367.* 2024. doi:10.48550/arXiv.2402.03367.
21. Shahul ES, James J, Espinosa Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; 2024 Mar 17–22; St Julians, Malta.* p. 150–8. doi:10.18653/v1/2024.eacl-demo.16.
22. Jiang Z, Xu F, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, et al. Active retrieval augmented generation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore.* p. 7969–92. doi:10.18653/v1/2023.emnlp-main.495.
23. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. *IEEE Trans Knowl Data Eng.* 2024;36(7):3580–99. doi:10.1109/tkde.2024.3352100.
24. Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, et al. Retrieval-augmented generation for AI-generated content: a survey. *Data Sci Eng.* 2026;11(1):1–29. doi:10.1007/s41019-025-00335-5.
25. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, et al. A survey on multimodal large language models. *Natl Sci Rev.* 2024;11(12):nwae403. doi:10.1093/nsr/nwae403.