Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.074409

REVIEW

A Review of Foundation Models for Multi-Task Agricultural Question Answering

Changxu Zhao[1], Jianping Liu[1,*], Xiaofeng Wang[1], Wei Sun[2], Libo Liu[3], Haiyu Ren[1], Pan Liu[1] and Qiantong Wang[1]

[1]School of Computer Science and Engineering, North Minzu University, Yinchuan, 750021, China
[2]Institute of Plant Protection, Ningxia Academy of Agriculture and Forestry Sciences, Yinchuan, 750002, China
[3]School of Information Engineer, Ningxia University, Yinchuan, 750021, China

*Corresponding Author: Jianping Liu. Email: liujianping01@nmu.edu.cn

Received: 10 October 2025; Accepted: 12 December 2025; Published: 12 March 2026

ABSTRACT: Foundation models are reshaping artificial intelligence, yet their deployment in specialised domains such as agricultural question answering (AQA) still faces challenges including data scarcity and barriers to domain-specific knowledge. To systematically review recent progress in this area, this paper adopts a task–paradigm perspective and examines applications across three major AQA task families. For text-based QA, we analyse the strengths and limitations of retrieval-based, generative, and hybrid approaches built on large language models, revealing a clear trend toward hybrid paradigms that balance precision and flexibility. For visual diagnosis, we discuss techniques such as cross-modal alignment and prompt-driven generation, which are pushing systems beyond simple pest and disease recognition toward deeper causal reasoning. For multimodal reasoning, we show how the fusion of heterogeneous data—including text, images, speech, and sensor streams—enables comprehensive decision-making for diagnosis, monitoring, and yield prediction. To address the lack of unified benchmarks, we further propose a standardised evaluation protocol and a diagnostic taxonomy specifically designed to characterise agriculture-specific errors. Finally, we outline a concrete AQA roadmap that emphasises safety alignment, hallucination control, and lightweight deployment, aiming to guide future systems toward greater efficiency, trustworthiness, and sustainability.

KEYWORDS: Foundation models; agricultural question answering; multimodal learning; large language models; smart agriculture; artificial intelligence

1 Introduction

Agriculture underpins global food security, yet the precise delivery of agronomic knowledge remains a critical bottleneck. While traditional extension services are constrained by expert scarcity and high consultation costs [1], Agricultural Question Answering (AQA) [2,3] has emerged as a transformative solution to democratize access to expert advice [4]. Unlike general Open-Domain QA (ODQA) [5,6], AQA is not merely an information retrieval task but a complex decision-support process designed for specific users—ranging from smallholder farmers seeking diagnosis for crop abnormalities to agricultural extension agents verifying pesticide regulations. The scope of AQA spans the entire production cycle, covering seed selection, pest and disease management (plant protection), and harvest planning, often requiring interaction in multilingual or resource-constrained environments.

The limitations of early AQA systems [7], which relied on static keyword matching or limited structured Knowledge Bases (KB-QA), have been largely overcome by the advent of **Foundation Models**. Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in interpreting unstructured queries across text, images, and sensor data [8]. However, the direct application of general-purpose foundation models to agriculture presents unique, domain-specific challenges that distinctly define AQA as an independent research problem separate from general QA or Visual Question Answering (VQA):

- **Domain Specificity and Terminology:** Agricultural queries involve fine-grained taxonomies of crops and pests (distinguishing *Spodoptera frugiperda* from *Mythimna separata*) and diverse input forms, including local dialects and non-standard descriptions, which are often underrepresented in general web corpora.
- **Spatio-Temporal and Environmental Context:** Unlike factual QA, the "correctness" of an agricultural answer is highly dynamic. A symptom observed in a humid region during the seedling stage typically necessitates a different diagnosis and treatment than the same symptom in an arid region during harvest. General models often fail to incorporate these implicit context variables [9].
- **Safety and Regulatory Compliance:** AQA is safety-critical. Hallucinations in pesticide recommendations or dosage calculations can lead to economic loss, food safety violations, and environmental damage. Systems must strictly adhere to regional regulations (e.g., Maximum Residue Limits) and dynamically updated prohibited substance lists.
- **Resource and Deployment Constraints:** Real-world agricultural scenarios often suffer from unstable network conditions and limited hardware resources. This necessitates the development of lightweight models capable of on-device inference, distinct from the cloud-centric paradigm of general LLMs.

Despite the surge in applying foundation models to these tasks, existing research remains fragmented, often focusing on isolated modalities—such as purely visual disease recognition or text-based policy interpretation—without a unified perspective. Current reviews lack a systematic analysis of how to bridge the gap between probabilistic model generation and the rigorous standards of agronomic science.

To address these gaps, this paper provides a comprehensive, **paradigm-driven review** of foundation models in AQA. We structure our analysis around three core task families: Text-based QA, Visual Diagnostic QA, and Multimodal Reasoning. Moving beyond a descriptive summary, we aim to provide **guidance** for future research. We critically analyze adaptation strategies—including **Prompt Engineering**, Retrieval-Augmented Generation (RAG), and Parameter-Efficient Fine-Tuning (PEFT)—and evaluate their effectiveness in handling the unique challenges of agriculture. Finally, we propose a concrete **AQA Roadmap**, outlining the necessary benchmarks, standardized evaluation protocols, and safety guardrails required to build trustworthy, interpretable, and deployable agricultural AI assistants. As illustrated in Fig. 1, our framework categorizes agricultural QA into three core task families: text-driven, vision-driven, and multimodal-driven tasks, summarizing the applicable models, fine-tuning strategies, and methodologies for each paradigm.

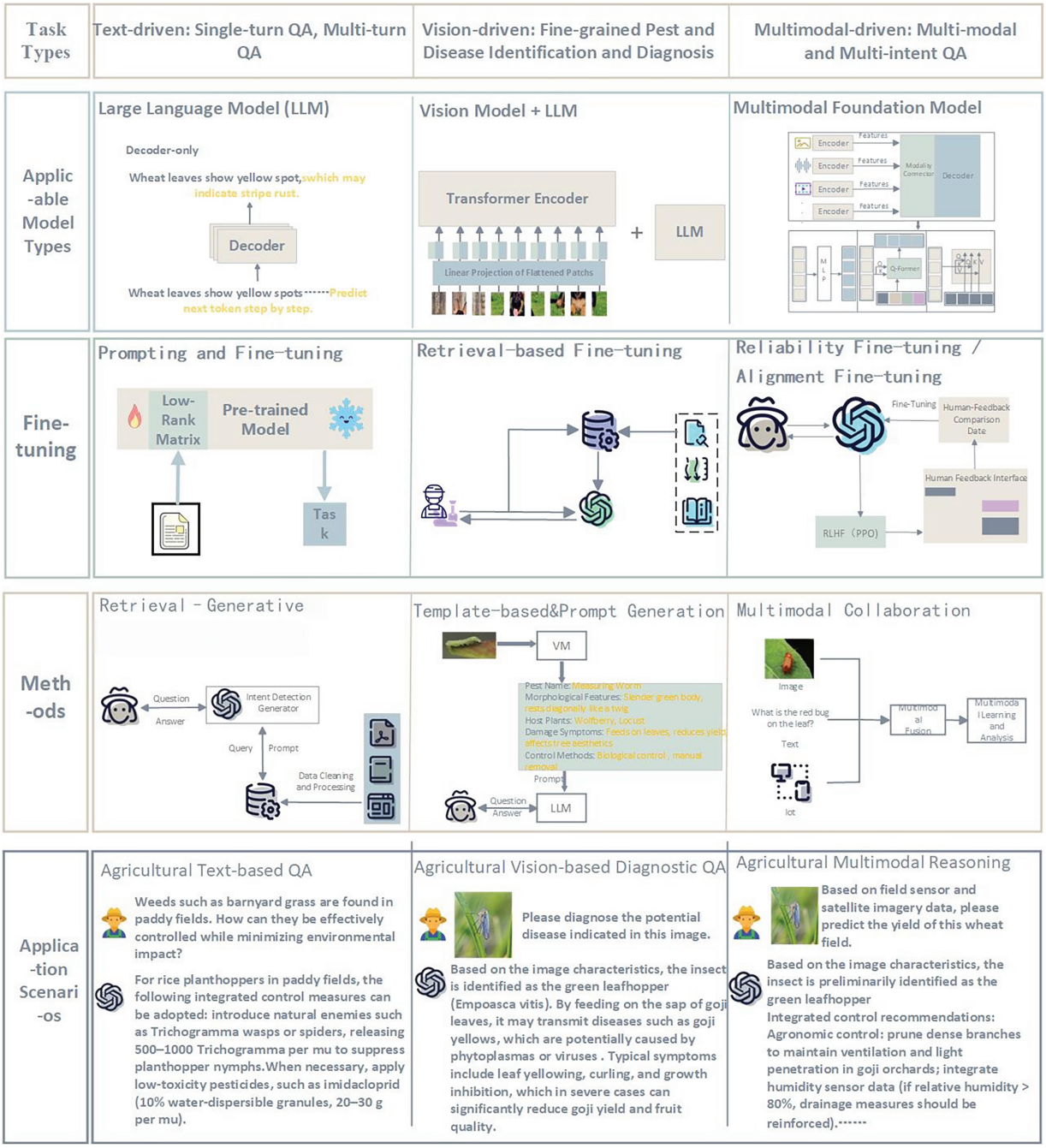

Figure 1: Overview of multitask agricultural question answering scenarios enabled by foundation models, illustrating representative inputs (text, images) and corresponding expert-level outputs (diagnosis, recommendations, compliance checks)

With the emergence of foundation models (FMs), the pre-training–fine-tuning paradigm and cross-modal reasoning capabilities have introduced new technological pathways for agricultural question answering (AQA). As a critical direction integrating artificial intelligence with agricultural knowledge, AQA aims to provide reliable solutions to problems in agricultural production, management, and research through natural language or multimodal inputs. Unlike general-domain QA systems, agricultural scenarios

demand higher levels of domain expertise, scenario diversity, and compliance, encompassing tasks such as policy interpretation, pest and disease diagnosis, agronomic guidance, and yield prediction. Nevertheless, existing studies often concentrate on individual technical dimensions and lack systematic reviews from the perspective of task paradigms, with insufficient discussion of the differences in model adaptation across text-based, vision-based diagnostic, and multimodal reasoning tasks. Li et al. [3] mainly focused on knowledge storage approaches in agricultural QA systems, including corpora, knowledge graphs, and large language models, yet their discussion remained predominantly text-driven and lacked a systematic review of visual and multimodal QA. Wang and Zhao [10] provided a comprehensive summary of large language model applications in agriculture, emphasizing the potential of retrieval-augmented generation (RAG), multimodal fusion, and agent architectures, but offered limited discussion on the independent role of vision models and the complexities of multi-intent QA scenarios. Yang et al. [11] reviewed QA systems in agricultural production and sustainable management, with a strong emphasis on knowledge graphs and text-based methods, while devoting limited attention to the developmental trajectory of foundation models and the emerging direction of agent-based systems. Vizniuk et al. [12] proposed a RAG framework for agricultural decision-making, systematically summarizing the value of retrieval-augmented methods, but their perspective was restricted to text retrieval, with insufficient focus on the integration of multimodal information such as vision, speech, and sensor data, as well as new trends such as lightweight adaptation and causal reasoning. Li et al. [2] adopted a more macro perspective by reviewing the potential of large language models (LLMs), large vision models (LVMs), vision–language models (LVLMs), and multimodal large language models (MLLMs) in agriculture, covering scenarios such as pest and disease diagnosis, crop monitoring, and agricultural machinery automation. However, their work provided limited coverage of QA-specific multi-task scenarios and lacked a systematic synthesis of practical implementations in agricultural QA systems.

This review aims to fill this gap by providing a comprehensive, task-paradigm-driven overview of foundation models for multi-task AQA. Building on the above terminology, we analyze AQA along three major task types: (i) text-based QA over agricultural documents and knowledge resources, (ii) image-based diagnostic QA for plant pests and diseases in a VQA-style setting, and (iii) multimodal AQA that jointly reasons over text, imagery, and other signals. For each task type, we examine how foundation models are adapted via pre-training, fine-tuning, retrieval, and prompt engineering; how AQA datasets are constructed and evaluated; and what practical limitations currently hinder robust deployment in the field.

The main contributions of this survey are four-fold:

- **Framework proposal.** We propose a systematic framework that, for the first time, views AQA through the lens of task paradigms (text-based QA, vision-based diagnostic QA, and multimodal reasoning QA) and explicitly maps these tasks to the capabilities of modern foundation models.
- **Task-type analysis.** We review how large language, vision, and multimodal models are applied across different AQA tasks, highlighting key techniques such as retrieval augmentation, cross-modal alignment, causal and compliant reasoning, and parameter-efficient adaptation in data-scarce agricultural settings.
- **Progress and limitations.** We summarize recent advances in datasets, benchmarks, and empirical results for AQA, and critically analyze current limitations in terms of accuracy, interpretability, transferability across regions and crops, and robustness to domain shift.
- **AQA roadmap.** We distill our findings into a concrete AQA roadmap that includes (i) curated benchmark suites for different task types, (ii) recommended evaluation protocols and reporting practices, (iii) a safety and compliance checklist tailored to agricultural scenarios, and (iv) deployment considerations for integrating AQA systems into real-world agricultural extension and decision-support workflows.

The remainder of this paper is organized as follows. Sections 2 through 4 systematically review the application of foundation models across three core AQA paradigms: text-based QA, visual diagnostic QA,

and multimodal reasoning. For each paradigm, we examine the foundational architectures, adaptation strategies (such as fine-tuning and RAG), and representative benchmarks. Section 5 addresses the lack of unified standards by proposing a standardized evaluation protocol and a rigorous diagnostic taxonomy for error analysis. Section 6 critically analyzes the challenges of real-world deployment, focusing on safety alignment, hallucination control, and ethical considerations, and outlines a comprehensive roadmap for future research. Finally, Section 7 concludes the paper.

To ensure a comprehensive and representative review, a systematic literature search was conducted following standardized academic retrieval principles. The search covered multiple major databases, including Web of Science, Engineering Village, Springer Link, ScienceDirect, CNKI, arXiv, Google Scholar, Baidu Scholar, and AMiner.

The search terms combined domain-related and methodological keywords such as: "Foundation Model," "Large Language Model (LLM)," "Multimodal Large Language Model (MLLM)," "Retrieval-Augmented Generation (RAG)," "Agricultural Question Answering," "Plant Diseases and Pests," "Knowledge Graph," and "Agent." Boolean operators ("AND," "OR") were used to refine the results and to capture intersections across AI, agriculture, and multimodal reasoning research.

The temporal scope was limited to 2022–2025, ensuring the inclusion of the most recent studies following the emergence of foundation models. The initial retrieval yielded 121 publications, from which duplicates, non-peer-reviewed papers, and studies not directly related to agricultural QA were excluded through manual screening. Ultimately, 34 high-quality studies were retained for in-depth analysis, covering representative works on model architecture, dataset construction, and domain adaptation in agriculture.

The overall workflow of the literature search and screening is illustrated in Fig. 2, which outlines the stages of database retrieval, filtering, and final inclusion.

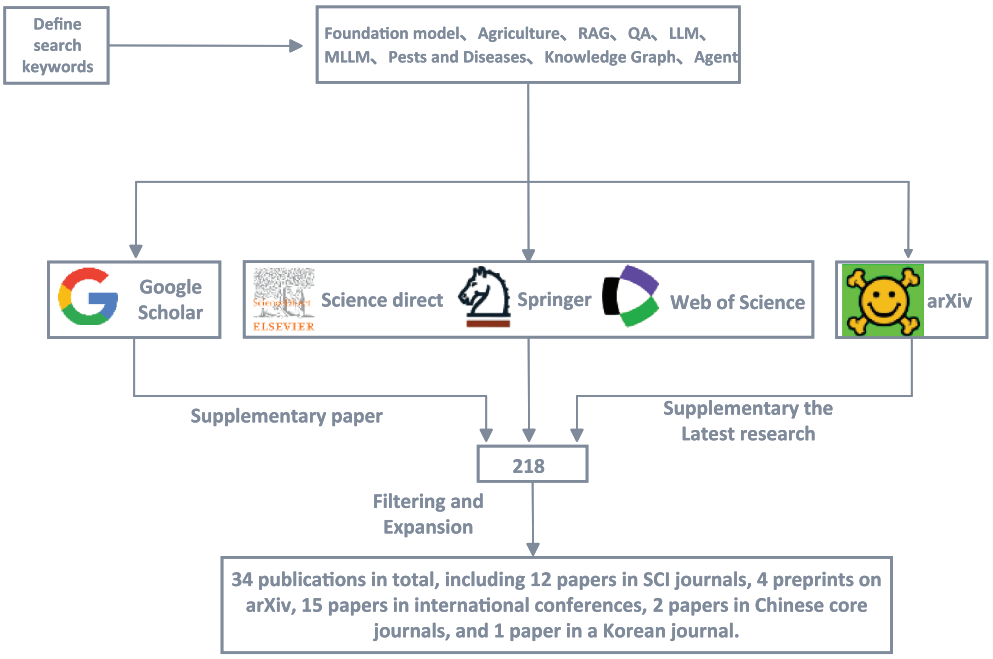

Figure 2: Literature review workflow, showing database selection, keyword formulation, screening criteria, and sequential filtering steps used to identify relevant AQA studies

2 Language Models for Text-Based QA

2.1 Foundations of Language Models

Language models are the core foundation models in natural language processing, primarily based on the Transformer [13] architecture. By leveraging the self-attention mechanism, Transformers capture global word dependencies, overcoming the limitations of traditional RNNs and CNNs in long-range semantic modeling. Their typical architectures fall into three categories (Fig. 3):

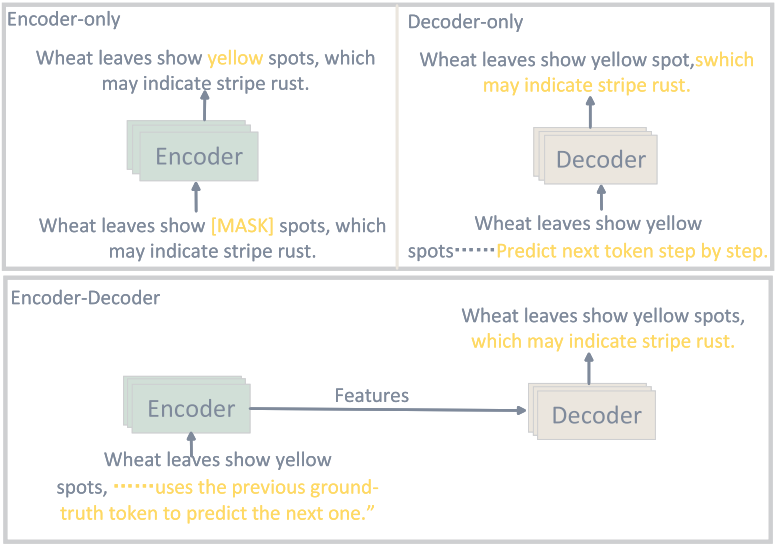

Figure 3: Comparison of model architectures, The figure compares the prediction mechanisms of Encoder-only, Decoder-only, and Encoder-Decoder architectures

Encoder-only architectures (BERT): Employ masked language modeling (MLM), ideal for understanding tasks like text classification and named entity recognition.

Decoder-only architectures (GPT series): Use autoregressive modeling for token-by-token generation, becoming the mainstream for open-domain QA and dialogue systems.

Encoder-decoder architectures (T5, BART): Support tasks with mismatched input and output lengths, widely applied to machine translation, summarization, and multi-turn QA.

In terms of pre-training methods, early language models primarily relied on supervised learning and task-specific annotated data, which limited their scalability. With the accumulation of large-scale corpora, self-supervised learning has become the mainstream approach: masked language modeling (MLM), which predicts masked tokens from surrounding context (BERT); auto-regressive modeling (AR), which generates sequences by predicting the next token (GPT series); and sequence-to-sequence (Seq2Seq) pre-training, which combines masking and generation mechanisms to support more complex conditional generation tasks (T5). In recent years, researchers have also explored contrastive learning and multi-task pre-training to enhance semantic alignment and cross-domain adaptability.

The development of language models can be broadly divided into two stages. In the early stage, pre-trained models such as BERT, built upon large-scale corpora and self-supervised tasks, enabled dynamic contextual representation and became essential tools for knowledge extraction and semantic matching in agriculture. In the generative stage, the GPT series, based on the auto-regressive paradigm, expanded language modeling from understanding to natural language generation, demonstrating remarkable advantages

in dialogue systems and QA tasks. More recently, ultra-large-scale LLMs (GPT-4, LLaMA-3, DeepSeek-V3) have achieved exponential growth in parameter size and training data, enabling cross-task transfer, zero-/few-shot reasoning, application of language models in domain-specific knowledge scenarios.

Meanwhile, the trend toward efficiency and lightweight adaptation has become increasingly prominent: parameter-efficient fine-tuning (PEFT) methods such as LoRA [14] and QLoRA [15] significantly reduce the cost of training and deploying models in agricultural contexts; quantized inference (INT4/INT8) and mixture-of-experts (MoE) [16] architectures substantially improve computational efficiency; and techniques such as retrieval-augmented generation (RAG), knowledge graph injection, and causal chain reasoning extend the capacity of language models in knowledge utilization and interpretability.

2.2 Text-Based Model Adaptation

Foundation models provide strong general understanding and generation abilities, but their direct use in agricultural QA is limited by insufficient domain knowledge, incomplete coverage, and safety concerns.Effective adaptation is therefore essential for practical deployment. As shown in Fig. 4, three complementary strategies are commonly used: Prompt Engineering and Fine-Tuning, Retrieval-Augmented Fine-Tuning and Reliability-Oriented Fine-Tuning.which together enhance the domain expertise, interpretability, and reliability of agricultural QA models.

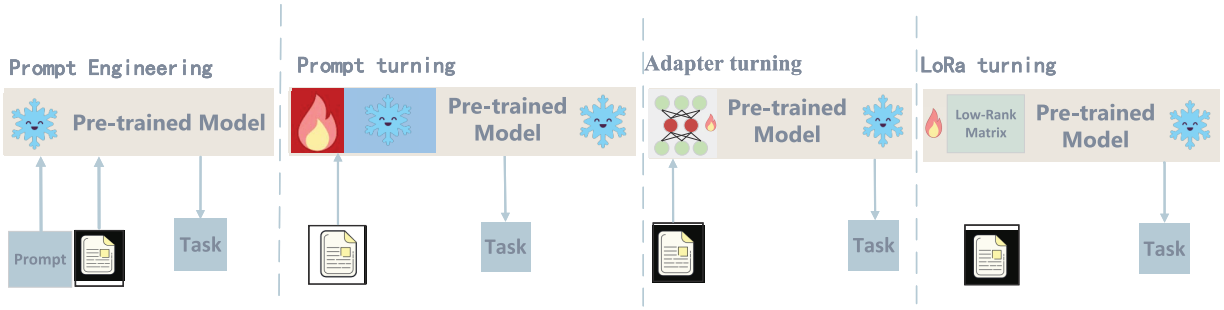

Figure 4: Main approaches for adapting foundation models to agricultural QA, including Parameter-Efficient Fine-Tuning (PEFT) and prompt engineering

2.2.1 Fine-Tuning

Fine-tuning is the most common adaptation approach, including full fine-tuning and parameter-efficient fine-tuning (PEFT). Full fine-tuning updates all model parameters for maximum domain adaptation but requires high computational cost. In contrast, PEFT methods such as LoRA, QLoRA, and Adapter update only a small subset of parameters, enabling efficient transfer of general models to agricultural domains while preserving prior knowledge. Training data for agricultural QA fine-tuning are usually drawn from policy documents, expert QA corpora, and pest–disease case datasets, allowing the model to handle tasks such as disease diagnosis, farming guidance, and policy interpretation more accurately. Beyond these generic adaptation mechanisms, recent work highlights that domain-specific representation learning and multi-granularity signals are particularly important for AQA. Multi-granularity pre-training for text classification, such as the multi-grained pre-trained language model for electric power audit texts [17] and the CL-MIL framework for multi-label classification with multi-granularity information [18], shows that representations can be significantly improved by jointly modeling word-, phrase-, and document-level semantics. Inspired by these ideas, agricultural QA encoders can be adapted to capture domain-specific signals at multiple levels:

- **Term level (word/subword):** explicit tagging of crop varieties, pest and disease names, active ingredients of pesticides, phenological stages (seedling, heading, flowering), and measurement units (kg/ha, ppm). This supports robust entity normalization and lexical matching in the retriever.
- **Phrase level:** modeling recurrent collocations such as "powdery mildew on grape leaves", "late blight under continuous rainfall", or "yellowing at the seedling stage", which often encode fine-grained symptom patterns and environment–disease combinations that are crucial for correct diagnosis.
- **Sentence/question level:** annotating intent types (recognition, how-to, compliance, causal) and risk levels (whether a question involves pesticide dosage or legal constraints), which guides both the retriever (what evidence to retrieve) and the reader or generator (how cautious the answer should be).

In practice, multi-granularity signals can be injected into AQA models via multi-task fine-tuning, where the encoder is jointly optimized for (i) domain-specific classification or tagging tasks (crop/pest type, symptom category, compliance label) and (ii) downstream QA objectives. Such a design is analogous to multi-granularity text classification, but tailored to agricultural corpora (extension bulletins, diagnostic manuals, farmer–expert dialogues). As shown in Table 1, phrase-level and sentence-level representations can be used to construct dense retrievers and re-rankers that are sensitive to agronomic terminology and regional practices, thereby improving robustness under domain shift (new varieties, local dialect terms, or seasonal changes).

Table 1: Domain adaptation strategies for agricultural QA (Refined)

Granularity	Scenario	Examples	Strategy	Impact	Challenge
Word/Term	Pesticide queries; variety search.	Chlorantraniliprole → "Chlor/ant/ran. . ."; Jingke 968.	Vocabulary expansion; KG entity embeddings.	Reduce tokenization errors and improve precision.	Generic tokenizers split chemical and variety names incorrectly.
Phrase	Symptom descriptions.	"yangyu" → potato; "Fire Dragon" → red spider mite; "Burnt tip" → tip blight.	Phrase-masking; synonym contrastive learning.	Robust to dialectal expressions and improves retrieval in technical documents.	Dialect or colloquial terms differ from technical terminology.
Sentence	Complex consultations.	"What disease causes yellowing?"; "What pesticide treats yellowing?"	Prompt-based intent classification; structured symptom queries.	Align response type with user intent.	Similar symptoms but different intents (diagnosis, treatment, compliance).

2.2.2 Prompt Engineering

Prompt engineering aims to steer the behaviour of large language models (LLMs) by designing their input prompts, and has become a key way to improve task performance and controllability without updating

model parameters [19–21]. In agricultural question answering (AQA), queries typically involve domain-specific terminology, contextual variables (crop variety, growth stage, climate conditions), and regulatory constraints. Relying only on generic instructions easily leads to hallucinations, imprecise wording, or non-compliant answers. Consequently, task-specific prompt design is a critical step in making general-purpose LLMs truly "think like" agricultural experts. Existing work broadly groups prompt methods into four categories: zero- and few-shot prompts, chain-of-thought prompts, structured prompts, and constraint/stability prompts, As shown in Table 2.

Table 2: Prompt type comparison

Type	Key feature	Best for	Limitation
Zero-shot	No examples needed	General Q&A	Less precise
Few-shot	Learn from examples	FAQ, templates	Needs examples
CoT	Step-by-step reasoning	Diagnosis, why-questions	May be verbose
Structured	Fixed JSON output	System integration	Less flexible
Safety	Risk control	Pesticide advice	Slower
Meta-prompt	Overall guidance	Multi-turn dialogues	Complex setup

Zero- and few-shotzzz prompting. Zero-shot prompting uses natural language instructions to describe the task and output requirements directly, and is suitable when data are scarce or query formats are highly diverse. In AQA, such prompts can support "on-the-fly" capabilities such as policy explanation or knowledge summarization from extension manuals without constructing a dedicated training set. Few-shot prompting instead provides a small number of exemplar question–answer pairs or labelled examples to teach the model the desired output format and decision criteria, and often yields substantial gains in classification, information extraction, and normalized generation tasks. For instance, in pesticide compliance judgement, only a few "compliant/non-compliant"examples are needed for the model to imitate expert decisions on new cases. Empirical studies further show that example-driven prompts can reduce ambiguity and improve semantic decision consistency in complex professional text tasks [22].

Chain-of-thought prompting. Chain-of-thought (CoT) prompts require the model to output intermediate reasoning steps rather than jumping directly to a final conclusion, which enhances both interpretability and accuracy in complex decision-making tasks [23]. In AQA, CoT prompting is particularly suitable for pest and disease diagnosis, fertilisation planning, and regulation applicability: the model can first organise crop and growth-stage information, then match symptoms and environmental conditions, and finally derive recommendations grounded in technical guidelines or legal documents. For example, Pan et al. [24] propose ChatLeafDisease, a training-free framework that uses GPT-4o with chain-of-thought prompting to classify crop leaf diseases based on textual disease descriptions. By combining a disease knowledge base with step-by-step CoT reasoning and simple self-consistency checks, their system reduces confusion among visually similar diseases and improves the reliability of diagnostic decisions. In a similar spirit, combining CoT with multi-path generation and self-consistency/self-review strategies can make AQA models' behaviour on repeated queries more stable and predictable, which in turn facilitates downstream error analysis and safety

auditing in safety-critical agricultural decision-support settings (e.g., pesticide use, food safety, or regulatory compliance).

Structured prompts and compliance. Structured prompts combine role specification, input normalisation, terminology clarification, and output templates to impose stronger constraints on the model's expression style and reasoning path, making them well suited to tasks that demand professional language and standardised, machine-readable outputs [25]. In agricultural settings, this design can be directly applied to key-point extraction, risk checking, and diagnostic QA. For example, the prompt can require the model to follow a fixed frame such as "problem restatement → evidence source(s) → diagnostic conclusion → risk warnings → disclaimer", or to output a table listing "disease name, supporting clauses, recommended measures, prohibited measures". Such structure not only allows agronomists and regulators to review answers rapidly, but also provides a clear basis for cataloguing common error patterns (misread symptoms, incorrect regulation matching).

Constraint and stability prompts. Constraint prompts explicitly encode rules, limits, or normative requirements into the instructions, encouraging the model to follow domain standards during generation. In agriculture, recent perspectives on LLM-based extension services and food production have already argued that guardrail-like instructions and governance frameworks are essential to prevent harmful or misleading recommendations and to ensure responsible use in real farming systems [26,27]. In AQA, the same ideas naturally extend to pesticide safety and regulatory compliance. Prompts can explicitly enforce constraints such as "do not recommend pesticides outside the given registration list", "dosage must remain within label limits", and "if evidence is insufficient, the model must either abstain or advise consulting local experts", and can further require the model to report confidence or uncertainty. Combined with calibration metrics such as ECE and the Brier score introduced earlier, these constraint-and-stability prompts can effectively suppress dangerous hallucinations and improve the predictability and compliance of AQA systems at deployment time, all without modifying model parameters.

Meta-prompt. Meta-prompting serves as a higher-level framework that orchestrates the model's interaction mode, regulating role consistency and task flow across multi-turn dialogues. Unlike standard prompts that target single queries, meta-prompts define a persistent system persona ("safety-first crop doctor") and a standardized interaction protocol, ensuring the model maintains its expertise and safety boundaries throughout a long session. For instance, a meta-prompt can explicitly instruct the model to sequentially verify geographic location and symptom details before offering any diagnosis, thereby mimicking the rigorous investigative process of human experts. This mechanism is particularly effective for integrating LLMs into larger agricultural decision-support systems, where the model must coordinate with external tools like weather APIs or pest databases. However, while meta-prompting significantly improves context retention and global consistency [28], it introduces a "complex setup" limitation, as crafting robust system instructions that prevent the model from drifting out of character requires extensive testing and optimization.

2.2.3 Retrieval-Augmented Fine-Tuning

Agricultural knowledge is highly specialized and rapidly evolving, and general-purpose large models often fail to provide full coverage of domain-specific information, particularly in areas such as input regulation, pest and disease classification, and region-specific policies. To address this limitation, the integration of retrieval-augmented generation (RAG) [28,29] and knowledge graph (KG) [10] have emerged as the main solutions. The mechanism is depicted in Fig. 5. By dynamically invoking domain knowledge bases during the inference process, models are able to generate answers accompanied by references and evidence chains, thereby significantly reducing the risk of hallucinations and enhancing compliance. In the agricultural domain, existing studies have incorporated knowledge graphs such as AgriKG, CropKG, and

PlantKG—embedding knowledge of pests and diseases, crop growth patterns, and agricultural policies into the models—which has effectively improved the specialization, reliability, and traceability of agricultural QA systems.

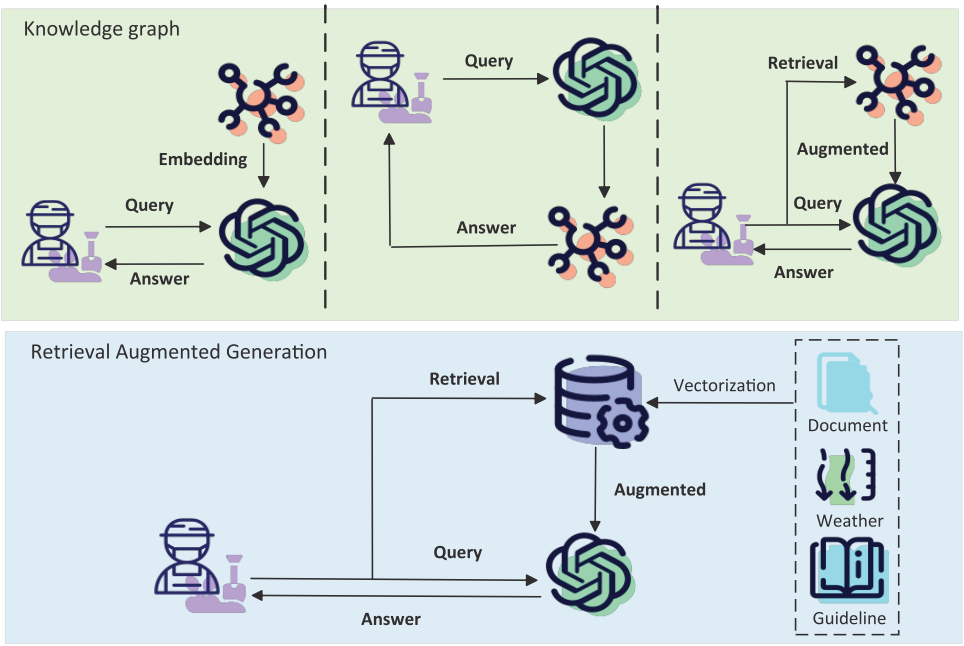

Figure 5: Comparison of knowledge-enhanced QA and retrieval-augmented generation (RAG), highlighting differences in knowledge access, grounding reliability, and error control

2.2.4 Reliability-Oriented Fine-Tuning

In high-risk tasks such as pesticide recommendation and fertilization advice, relying solely on probabilistic model generation is insufficient, as erroneous outputs may lead to severe agricultural losses. To address this, reinforcement learning and causal reasoning have increasingly been introduced to enhance model reliability. Reinforcement learning from human feedback (RLHF) [30] incorporates preference data and compliance annotations to optimize the model's generation strategy, ensuring that its outputs align more closely with expert experience and domain-specific regulations. At the same time, Causal control methods model the relationships 'crop-disease-environment-management practice' as causal chains, a allowing logical and regulatory validation during the reasoning process and ensuring that the generated prescriptions comply with the agronomic principles and regulatory requirements. These approaches not only improve the safety and robustness of the models but also provide theoretical support for the interpretability of agricultural QA systems.

In summary, the adaptation of foundation models in agricultural QA has evolved in multiple layers: from parameter fine-tuning to input prompting, from external knowledge integration to reliability optimization. These methods are gradually converging into a comprehensive framework that integrates fine-tuning, prompting, knowledge augmentation, and causal validation. This trend not only facilitates the practical deployment of models in specialized agricultural tasks but also lays the foundation for building safe, trustworthy, and interpretable intelligent agricultural QA systems.

2.3 Applications of Text-Based QA in Agriculture

2.3.1 Benchmark Datasets and Evaluation Metrics

For agricultural text-based QA, evaluation must jointly consider (i) how answers are scored and (ii) how agricultural knowledge is grounded.

Let $\hat{a}$ denote a system prediction and a a reference answer. Following standard practice in SQuAD-style QA [31], both are normalized before scoring:

1. lowercasing: all characters are converted to lowercase;
2. punctuation removal: punctuation marks (".,!?;:") are stripped;
3. article and stop-word removal (optional): function words such as "a", "an", "the" and language-specific stop-words can be removed when they do not affect factual content;
4. whitespace normalization: multiple spaces are collapsed into a single space.

We write $\widetilde{a} = \mathrm{norm}(a)$ and $\widetilde{\hat{a}} = \mathrm{norm}(\hat{a})$ for the normalized strings. All token-level metrics below are computed on tokenized versions of $\widetilde{a}$ and $\widetilde{\hat{a}}$.

Span-based vs. free-form scoring.

In span-based (extractive) QA, each gold answer is a contiguous span from the context. Evaluation is therefore based on exact span matching and token-level overlap. In free-form QA, answers may be paraphrased sentences or short descriptions; here, token-level F1 and generation metrics (BLEU, ROUGE, BERTScore) are more appropriate, together with human evaluation.

Exact Match (EM).

EM measures whether the normalized prediction is identical to at least one normalized reference answer. Let $\mathcal{D}$ denote the evaluation set with $|\mathcal{D}|$ questions, and $\mathcal{A}(q)$ the set of gold answers for question q. The EM score is

$$\mathrm{EM} = \frac{1}{|\mathcal{D}|} \sum_{q \in \mathcal{D}} \mathbf{1}\Big[\exists a \in \mathcal{A}(q) \text{ s.t. } \mathrm{norm}(\hat{a}_q) = \mathrm{norm}(a)\Big], \tag{1}$$

Token-level Precision, Recall, and F1.

Let $T(\widetilde{a})$ and $T(\widetilde{\hat{a}})$ be the token sets (or multisets) of the normalized gold and predicted answers. The number of overlapping tokens is

$$\mathrm{match} = \left|T(\widetilde{a}) \cap T(\widetilde{\hat{a}})\right|. \tag{2}$$

Token-level precision, recall, and F1 are defined as

$$\mathrm{Precision} = \frac{\mathrm{match}}{|T(\widetilde{\hat{a}})|}, \qquad \mathrm{Recall} = \frac{\mathrm{match}}{|T(\widetilde{a})|}, \tag{3}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall} + \varepsilon}, \tag{4}$$

where ε is a small constant to avoid division by zero. In the presence of multiple gold answers $\{a^{(1)}, \ldots, a^{(m)}\}$, the F1 score for a prediction is typically defined as the maximum token-level F1 over all references:

$$\mathrm{F1}(q) = \max_{j=1,\ldots,m} \mathrm{F1}\left(\hat{a}_q, a^{(j)}\right), \tag{5}$$

and the dataset-level F1 is the average over all questions.

BLEU and ROUGE for short answers (with caution).

BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure n-gram overlap between predictions and references [32,33]. BLEU is given by

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{6}$$

where p_n is the modified n-gram precision, w_n are non-negative weights summing to 1, and BP is the brevity penalty. ROUGE-N is defined as

$$\text{ROUGE-}N = \frac{\sum_{\mathrm{gram}_N \in \mathrm{Ref}} \mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_N)}{\sum_{\mathrm{gram}_N \in \mathrm{Ref}} \mathrm{Count}(\mathrm{gram}_N)}. \tag{7}$$

For short answers (1–3 words, numeric values, or crop names), BLEU and ROUGE can be unstable and misleading, because a single token mismatch (“aphid” vs. “green aphid”) may produce a large drop in n-gram scores even when the prediction is agronomically acceptable. Therefore, in agricultural QA we primarily report EM and token-level F1 for such short answers, and treat BLEU/ROUGE as auxiliary indicators for longer, free-form explanations.

Semantic similarity metrics.

BERTScore measures semantic similarity between prediction and reference in an embedding space [34]. Given token embeddings $e(x)$ from a pretrained language model, the BERTScore for predicted sequence P and reference sequence R is

$$\mathrm{BERTScore}(P,\ R) = \frac{1}{|P|} \sum_{x \in P} \max_{y \in R} \cos\left(e(x),\ e(y)\right), \tag{8}$$

where $\cos(e(x),\ e(y))$ denotes cosine similarity. BERTScore is particularly useful in AQA when agronomic explanations admit paraphrases with similar meaning but different wording.

Classification-style metrics.

For tasks such as intent classification or multiple-choice QA, Accuracy and Mean Average Precision (MAP) remain important. Given TP, FP, TN, and FN as the numbers of true positives, false positives, true negatives, and false negatives, Accuracy is

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{9}$$

In retrieval-augmented QA, the ranking quality of retrieved passages is often evaluated using MAP:

$$\mathrm{MAP} = \frac{1}{Q} \sum_{i=1}^{Q} \mathrm{AP}(q_i), \tag{10}$$

where Q is the number of queries and $\mathrm{AP}(q_i)$ is the average precision for query q_i.

Calibration metrics: ECE and Brier score.

For high-stakes AQA tasks (pesticide or fertilizer recommendations), probability calibration is as important as accuracy. Let $\hat{p}_q \in [0, 1]$ be the model-reported confidence that its answer to question q is correct, and let $y_q \in \{0, 1\}$ indicate whether the answer is actually correct under EM or F1. The Brier score [35] is

$$\text{Brier} = \frac{1}{|\mathcal{D}|} \sum_{q \in \mathcal{D}} \left(\hat{p}_q - y_q\right)^2, \tag{11}$$

where lower values indicate better calibration. The Expected Calibration Error (ECE) [36] partitions predictions into M confidence bins $\{B_m\}_{m=1}^{M}$ and measures the average gap between predicted confidence and empirical accuracy:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{|\mathcal{D}|} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{12}$$

where $\text{acc}(B_m)$ is the average correctness y_q in bin B_m, and $\text{conf}(B_m)$ is the average confidence $\hat{p}_q$ in that bin. In AQA, reporting Brier and ECE helps characterize whether the system is over-confident or under-confident on difficult agricultural queries.

Multiple-answer aggregation.

Many AQA benchmarks allow multiple valid gold answers per question (synonyms, multilingual variants, or acceptable numerical ranges). Aggregation is usually performed in two stages:

1. *Per-answer scoring:* compute EM, F1, and possibly BLEU/ROUGE for each prediction against each reference answer, and keep the maximum score per question, as in the F1 definition above.
2. *Multi-model or multi-sample aggregation:* when multiple responses are generated (via self-consistency or an ensemble of models), final metrics can be computed using:
   - majority voting, where the answer most frequently predicted across samples is scored; or
   - self-consistency, where answers are sampled multiple times and the most semantically consistent cluster (in embedding space) is taken as the final prediction [37].

On top of these metrics, dataset design remains central to text-based AQA (Table 3).

Table 3: Agricultural QA datasets: Question types, languages, and limitations

Name	Q-Types	Lang	Format	Lic	Region	Gaps & Failures
AgriEval	What/Why/How, ID, Proc, Comp	CN	MC, SA	Prop	China	Cross-region, no multi-turn, counterfactual
Agri-342K	What (diag), How (mgmt), Comp	EN	Text, JSON	Open	India	No counterfactual, tool, multi-turn
KisanVaani	What/How (FAQ)	HI, EN	Short, Tmpl	Open	India-R	Weak multi-turn, no tool, seasonal gap
Multi-Ag QA	What/How (X-ling)	EN/HI/PU	Text	Open	S.Asia	Low-res degrade, X-ling halluc
AgriResp	What/How (RT)	Dialects	Struct	Prop	China-N	Complex reasoning fail, no tool

(Continued)

Table 3 (continued)

Name	Q-Types	Lang	Format	Lic	Region	Gaps & Failures
Farmer.Chat	What/Why/How (MT)	20+	Conv	Open	Global	No tool invoke, X-season
Kallaama	What/How (speech)	WO/PU/SE	Sp+Txt	Open	W.Africa	ASR err, dialect var, no multimod

Notes: ID = Identification, Proc = Processing, Comp = Compliance, MC = Multiple Choice, SA = Short Answer, Prop = Proprietary, CN = Chinese, EN = English, HI = Hindi, PU = Punjabi, WO = Wolof, SE = Sereer, RT = Real-time, MT = Multi-turn, X-ling = Cross-lingual, X-season = Cross-season, diag = diagnosis, mgmt = management, Tmpl = Template, Conv = Conversational, Struct = Structured, India-R = Rural India, China-N = Nationwide China, W.Africa = West Africa, Low-res = Low-resource, degrade = degradation, halluc = hallucination, Sp+Txt = Speech+Text, ASR err = ASR errors, multimod = multimodal integration.

As illustrated in Table 3, the landscape of agricultural text-based QA datasets provides the empirical foundation for model development, evolving from static benchmarks to dynamic, interaction-oriented corpora. We categorize these resources into three distinct functional paradigms:

Comprehensive Knowledge Benchmarks. To assess general intelligence in agriculture, AgriEval [38] serves as a holistic benchmark covering diverse domains such as policies, crop protection, and food safety. Unlike simple retrieval tasks, it prioritizes Accuracy and F1 scores to rigorously evaluate a model's comprehensive reasoning and task adaptability across standardized agricultural examinations.

Specialized QA and Practicality. For domain-specific fine-tuning, Agri-342K [39] provides large-scale coverage of pest diagnosis and regulations, utilizing EM and ROUGE metrics to balance retrieval precision with generative fluency. Moving closer to real-world application, KisanVaani [40] derives 22K QA pairs from actual farmer forums, uniquely incorporating expert evaluation to ensure the practical utility and interpretability of answers rather than relying solely on automated metrics.

Interaction and Accessibility (Dialogue & Speech). Recent datasets address the complexity of deployment environments. AgriResponse [41] and Farmer.Chat [42] push the boundary toward conversational AI, with the latter offering over 300,000 multi-turn dialogues to train models on iterative reasoning, evaluated via hybrid metrics (BLEU plus expert judgment). Finally, addressing linguistic barriers, Multilingual Agri QA [43] benchmarks cross-lingual adaptability (English, Hindi, Punjabi), while Kallaama [44] pioneers voice-interactive AQA. By providing 125 h of speech data evaluated via WER and semantic matching, Kallaama supports critical research for serving illiterate farming populations.

2.3.2 Applcation

In the development of applying large language models to text-based agricultural QA, three representative paradigms have emerged, as illustrated by Fig. 6: retrieval-based QA, generative QA and hybrid retrieval-generation QA. As summarized in Table 4, retrieval-based QA relies on knowledge base retrieval and answer extraction, emphasizing verifiability and traceability, making it particularly suitable for tasks with high accuracy requirements such as policy and regulation interpretation or market data analysis. Early studies were mostly based on re-pretraining and knowledge injection into BERT models. For example, Rezayi proposed AgriBERT [45], which integrates knowledge graphs into nutritional QA for semantic matching and answer selection; Huang et al. [46] constructed a domain-specific knowledge graph in tea pest and disease scenarios to support QA. With increasing task complexity, retrieval pipelines have continued to evolve. Krishiq-BERT [47] adopted a Retriever–Reader framework to support regional languages; AgAsk [48]

developed by a research group, utilized a BERT variant in combination with a neural ranker and literature retrieval to achieve paragraph-level answer matching and extraction from agricultural documents.

Generative QA, by contrast, leverages large language models to directly produce natural language answers, offering greater flexibility and fluency in open scenarios such as agronomic consultation and causal explanation. Representative examples include Chen et al.'s AgriPrompt [49], which employed dynamic prompt engineering and structured template filling to enhance robustness and interpretability; and Zhang et al.'s IPM-AgriGPT [50], which introduced a generation–evaluation adversarial mechanism and agricultural context reasoning chains for safer prescription generation. The retrieval–generation hybrid paradigm has become the mainstream trend in recent years. This approach uses retrieval to provide evidence chains and generative models to synthesize and articulate answers, thus balancing naturalness and reliability. For example, AgRoBERTa [51] followed an encoder-style extraction/matchinging paradigm for agricultural extensiQA. ShizishanGPT [52] was introduced by a research team, integrating GPT-4 with the agricultural knowledge graph (AgriKG) and external analytical tools to support reasoning and QA in tasks such as cultivation prediction of phenotypes, types, and gene expression. Xiong et al.'s SheepDoctor [53] combined RAG with knowledge graphs for sheep disease diagnosis; Akbar [54] employed FAISS re-ranking within the RAG pipeline to reduce retrieval latency; Wei fine-tuned ChatGLM2-6B-QLoRA-int4 on policy texts and, by integrating local regulatory knowledge bases with retrieval frameworks such as LangChain-ChatGLM, realized lightweight deployment of rural revitalization policy QA. Other works explored ensemble or voting mechanisms, as well as parameter-efficient fine-tuning on large models (InterLM-20B-LoRA) [55] for plant protection QA. The ensemble voting model, proposed by Dofitas [56] employed multi-model integration and a voting mechanism to enhance the stability and cross-task generalization of comprehensive agricultural QA.

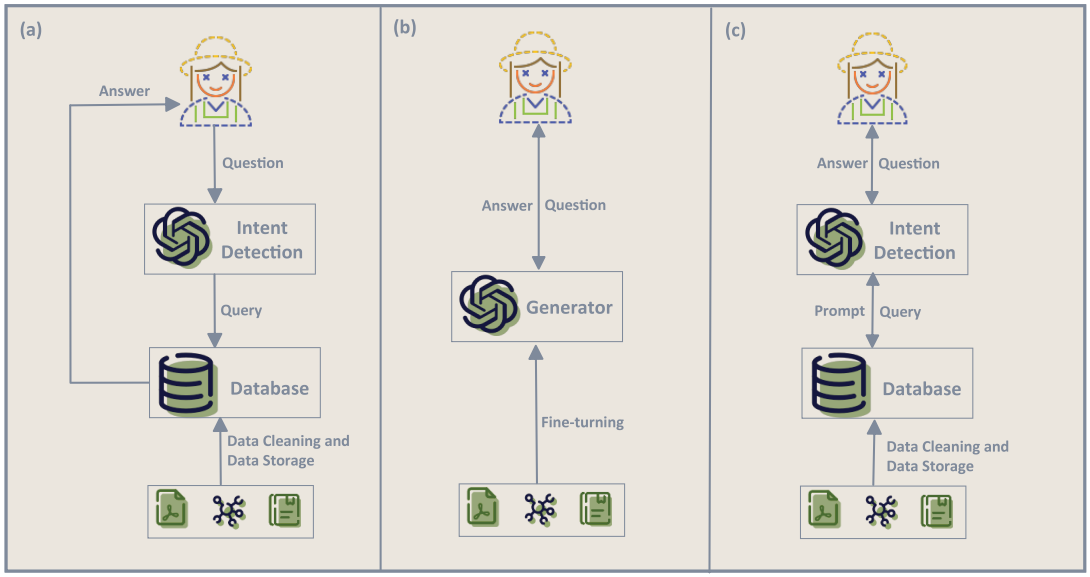

Figure 6: Three categories of knowledge-enhanced AQA frameworks: (**a**) retrieval-based AQA, where an intent detection module routes the query to a domain database and returns a grounded answer; (**b**) hybrid retrieval–generation, where retrieved evidence and external data condition a generator (LLM) that produces the final answer; and (**c**) structured knowledge integration, where intent detection and structured databases jointly condition the LLM so that external knowledge supports both factuality and safety

Table 4: Representative agricultural QA models across generative, retrieval, and hybrid paradigms

Mode	Year	Model name	Base model	Task scenario	Method	Dataset	Performance	Advantages	Limitations
Generative QA	2024	AgriPrompt	BERT, ChatGPT	Agricultural QA (crops/pests)	Dynamic prompts, KB parser, template filling	AgriKB, parsing corpus	KQScore: 0.81	Robustness, interpretability	Template reliance, weak transfer
	2025	IPM-AgriGPT	GPT, LLaMA, BLOOM	Personalized pest management	G–EA, reasoning chain	Pest benchmark	Obj: 0.538; Sub: 4.24; Exp: 4.45	Interpretable reasoning, safer	Complex, slow inference
Retrieval QA	2022	AgriBERT	BERT	Food/nutrition QA	Re-pretraining, KG injection	Agri corpus, USDA, FoodOn	MAP best 44.21%; P@1 best 49.98%	KG boosts QA	Nutrition only, weak generalization
	2023	–	BERT	Tea pest/disease QA	KG construction, few-shot	Germplasm InfoNet	Acc: 90.1; Recall: 90.53%	Few-shot accuracy	Single crop, poor cross-domain
	2024	Krishiq-BERT	DistilBERT, IndicBERT	Multilingual QA (Kannada)	Retriever–Reader	Krishiq	F1: 61.2%	Regional language QA	One language, low perf
	2024	AgAsk	BERT variant	Paragraph-level QA	Neural ranker, retrieval	86k docs, 210 Qs	—	Efficient, precise	Docs only, no execution
Hybrid QA	2024	AgRoBERTa	RoBERTa, GPT-3.5	Extension QA	Re-pretrain, AgEES	AgXQA	EM: 55.2%; F1: 78.9%; BLEU: 52.42%; AgEES: 94.37%	Expert metric, zero-shot	Limited data
	2024	ShizishanGPT	GPT-4	Cultivation QA	GPT-4,AgriKG,tools	100 QA pairs	>General LLMs	Integrates KG/tools	Small, lab-only
	2025	Enhanced RAG	LLaMA-3-8B	Few-shot QA	FAISS RAG	KisanVaani	BLEU ≈; Latency ↓	Faster retrieval, robust	Small dataset
	2025	Ensemble Voting	LLaMA, Agri-BERT, etc.	General QA	Model ensemble, voting	Agri queries corpus	Acc: 93%; BLEU: 53.8%	Stable, generalizable	Complex, not real-time
	2025	InterLM-20B-LoRA	InterLM-20B	Plant protection QA	LoRA fine-tuning	7k papers, 9k QAs	Acc: 97%	Efficient + accurate	Narrow domain
	2025	SheepDoctor	LLaMA2-13B	Sheep disease QA	LoRA, KG, RAG	5987 QAs, 207 diseases	>GPT-4o, Kimi	Animal medicine breakthrough	Vertical, weak transfer
	2024	ChatGLM2-6B-QLoRA	Policy QA	QLoRA, quantization	15k policies	>Base ChatGLM2-6B	Practical policy QA	Narrow scope	

In summary, agricultural text-based QA has been undergoing a paradigm shift from single retrieval or generation approaches toward hybrid models. Retrieval-based QA excels in traceability and precision, generative QA offers greater flexibility and naturalness, while retrieval–generation hybrids achieve a balance between the two, emerging as the key pathway for practical deployment of agricultural QA systems.

2.4 Section Summary

This analysis of LLMs in text-based agricultural QA reveals three emergent paradigms. Retrieval-based QA excels in precision and interpretability for high-stakes tasks. Generative QA offers flexibility for open-ended queries but risks hallucinations. Hybrid QA, integrating both, balances these trade-offs and is becoming the mainstream approach. Advances like pre-training, prompt engineering, and RAG underpin these paradigms, enhancing system adaptability and robustness. However, challenges in knowledge coverage, multi-turn consistency, and domain compliance remain critical for practical deployment.

3 Applications of Vision Models in Visual Diagnosis

With the progress of deep learning and vision foundation models, pest and disease diagnosis has become a core application of computer vision in agriculture. Visual models can recognize lesions and pest features, while their integration with language models enables causal reasoning and prescription generation. This synergy has driven the field from simple recognition toward reasoning and decision making. The representative architecture of the Vision Transformer (ViT) is illustrated in Fig. 7, providing the basis for many subsequent agricultural diagnostic applications. This chapter first reviews the evolution of visual models from CNNs [57] to ViTs [58] and general-purpose architectures, then summarizes their applications in agricultural diagnostic tasks, and finally discusses current challenges and future directions.

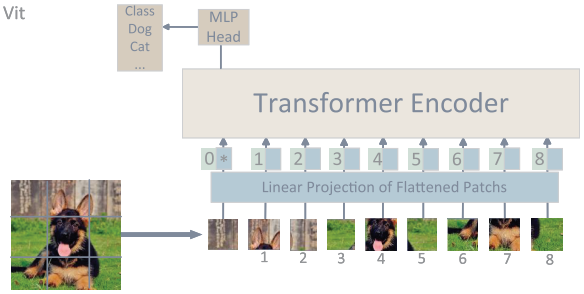

Figure 7: The figure illustrates the typical architecture of Vision Foundation Models (VFMs). In the case of the Vision Transformer (ViT [58]), the input image is divided into patches, linearly projected, and fed into a Transformer encoder. Through the self-attention mechanism, the model performs global feature modeling, and the MLP head outputs the classification results

3.1 Basics of Vision Foundation Models

Vision foundation models (VFMs) refer to general-purpose visual representation models pretrained on large-scale image or multimodal data. Their primary objective is to learn visual feature representations with strong generalization capability, thereby supporting a wide range of tasks such as classification, detection, segmentation, retrieval, and cross-modal reasoning.

In terms of architectural development, VFMs have generally undergone three major stages:

Convolutional Neural Network (CNN) stage: Represented by models such as AlexNet [57], VGG [59], ResNet [60], and Inception [61], CNNs extract hierarchical features from local convolutional operations, progressing from low-level textures to high-level semantics. They have been widely applied in crop disease

classification, remote sensing monitoring, and crop identification, laying the foundation for agricultural visual diagnosis.

Transformer stage: Represented by the Vision Transformer (ViT), which introduces the self-attention mechanism to capture long-range dependencies and enhance fine-grained feature representation. Variants such as Swin Transformer [62] and DeiT [63] further improve computational efficiency and cross-domain generalization, gradually replacing CNNs as the mainstream architecture and demonstrating greater robustness and cross-crop transferability in complex field environments. General-purpose vision and cross-modal stage: With the emergence of models such as the Segment Anything Model (SAM) [64,65] and CLIP-ViT [66], vision models have transcended single-task boundaries. They now offer zero-shot and few-shot adaptability and can be coupled with language models to achieve cross-modal alignment and reasoning.

From the perspective of pretraining paradigms, vision models have evolved from supervised learning to self-supervised learning and multimodal contrastive learning: Supervised pretraining relies on large-scale labeled datasets (ImageNet, COCO) to train CNNs and some ViT models, supporting classification and detection tasks but being highly dependent on manual annotation. Self-supervised pre-training employs image reconstruction, contrastive learning (e.g., SimCLR, MoCo), or masked image modeling (MAE [67], BEiT [68]) on unlabeled data, significantly reducing dependency on annotated resources in agricultural contexts. Multimodal contrastive learning (CLIP, ALIGN [69]) aligns images and text into a shared semantic space, endowing models with cross-modal understanding and reasoning capabilities, which have become vital for agricultural visual diagnosis and QA systems.

In agricultural intelligent QA systems, vision models mainly undertake core tasks such as pest and disease recognition, lesion segmentation, and causal diagnosis [70]: Lesion recognition and segmentation: Automatically identify lesion regions from leaf images and quantify disease severity. Drone-based remote sensing monitoring: Detect anomalies in large-scale field imagery to enable early warning of pest and disease outbreaks. Explainable diagnosis and prescription generation: Combine with knowledge bases or language models to provide diagnostic and treatment recommendations grounded in visual evidence.

Overall, vision foundation models have evolved from CNNs to Transformers and further to general-purpose cross-modal models in terms of architecture, while their pretraining paradigms have shifted from supervised to self-supervised and multimodal learning. This progression has significantly enhanced their transferability and generalization, not only improving diagnostic accuracy but also driving agricultural QA systems from simple "image recognition" toward a closed loop of "cross-modal reasoning" and "prescription generation."

3.2 Vision Model Adaptation

This development has significantly enhanced the transferability and generalization capabilities of vision foundation models, enabling not only improved accuracy in agricultural diagnosis but also promoting the evolution of agricultural QA systems from "image recognition" to "cross-modal reasoning" and "prescription generation."

3.2.1 Prompting and Fine-Tuning

In visual tasks, parameter-efficient fine-tuning has become the mainstream trend. Compared with full-parameter updating, parameter-efficient methods allow training only a small number of additional parameters while keeping the pretrained model frozen, thereby significantly reducing computational and data requirements. A representative approach is Visual Prompt Tuning (VPT) [71], which inserts learnable prompt vectors into the input sequence, enabling the model to quickly adapt to new downstream tasks

without modifying the original network architecture. This mechanism provides a flexible transfer pathway for scenarios such as lesion recognition and crop classification. In addition, prompt-driven generative methods are also being developed. For instance, mechanisms such as progressive masking and template filling can gradually transform visual inputs into textual or structured outputs, achieving controllable generation of task results. Such approaches not only reduce dependence on large-scale annotations but also enhance interpretability and consistency, making them particularly valuable in open-domain agricultural applications.

3.2.2 Retrieval-Based Fine-Tuning

In complex tasks, relying solely on the internal representations of vision models is often insufficient to cover the necessary domain knowledge, resulting in limitations in diagnosis and reasoning. To address this, the idea of Retrieval-Augmented Generation (RAG) has been extended to visual modeling. The core of RAG is to incorporate external knowledge bases or documents as additional conditional inputs, which are then combined with the model's internal representations to enhance the accuracy and reliability of generation. The recent VisRAG [72] introduced this framework into vision–language models, directly using document images as retrieval and generation units, thereby avoiding potential information loss from traditional OCR processes. In cross-modal QA and image–text diagnostic tasks, VisRAG achieved notable improvements in performance, demonstrating the effectiveness of vision–retrieval integration. More broadly, approaches such as FiD [73] and Atlas [74] provide valuable references for visual tasks, showing that retrieval and fusion mechanisms can extend the knowledge boundaries of models, thereby enhancing their cross-domain transferability and reasoning capabilities.

3.2.3 Alignment-Based Fine-Tuning

The deep integration of vision models and language models relies on alignment mechanisms, which constitute one of the core challenges in visual question answering systems. A representative approach is Q-Former, which introduces learnable query vectors to interact with a frozen visual encoder, thereby generating compact and semantically rich representations that can be directly connected to large-scale language models. In addition, cross-modal attention mechanisms (BLIP-2 [75], InstructBLIP [76]) further achieve efficient fusion of image and text features through interactive attention computation, while shared semantic space alignment leverages large-scale image–text contrastive learning to establish unified multimodal representations during pretraining. These methods collectively enhance cross-modal reasoning capabilities, enabling vision models to participate more naturally in complex QA and reasoning processes, and providing solid technical support for downstream agricultural tasks such as diagnosis, description generation, and prescription recommendation.

3.3 Applications of Vision-Based QA in Agriculture

3.3.1 Evaluation Metrics and Datasets for Agricultural Visual QA

Similar to text-based AQA, evaluation of agricultural visual QA must distinguish between two levels: (i) *perception quality*, i.e., how accurately the model detects and segments diseased regions or anomalies from images, and (ii) *answer quality*, i.e., how reliably the model produces correct and stable textual responses conditioned on visual evidence. Accordingly, we group commonly used metrics into three categories: localization and segmentation metrics, detection metrics, and visual QA generation metrics.

(1) Localization and segmentation metrics.

For pixel-level lesion segmentation or region-level anomaly mapping, Intersection over Union (IoU) and Dice coefficient are widely used.

Intersection over Union (IoU). Given a predicted region B_p and ground-truth region B_{gt}, IoU is defined as

$$\text{IoU} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}, \tag{13}$$

where $|\cdot|$ denotes the area (number of pixels). IoU can be computed per image and then averaged over the dataset, or computed per class and then macro-averaged across classes. High IoU indicates accurate spatial alignment between predicted and true lesion areas, which is critical for downstream diagnosis and treatment recommendation.

Dice coefficient. Dice is another overlap-based metric, defined as

$$\text{Dice} = \frac{2|B_p \cap B_{gt}|}{|B_p| + |B_{gt}|}. \tag{14}$$

Compared to IoU, Dice is more sensitive to small objects and is therefore often preferred in fine-grained lesion segmentation, especially when diseased regions occupy only a small portion of the image.

(2) Detection metrics: AP and mAP.

For bounding-box based detection of diseased leaves, weeds, or remote-sensing anomalies, average precision (AP) and mean average precision (mAP) are standard metrics. Let $P(R)$ denote the precision–recall curve for a given class, then

$$\text{AP} = \int_0^1 P(R)\, dR, \tag{15}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i, \tag{16}$$

where AP_i is the AP of the i-th class and N is the total number of classes. In agricultural settings, mAP is the primary indicator of whether the model can reliably localize diseased plants, nutrient deficiency symptoms, or field anomalies across diverse conditions.

(3) Visual QA generation and reasoning metrics.

On top of perception metrics, agricultural visual QA requires evaluation of answer correctness and reasoning quality.

VQA Accuracy. Given an evaluation set D of visual QA pairs, VQA Accuracy is defined as

$$\text{VQA Accuracy} = \frac{1}{|D|} \sum_{q \in D} \mathbf{1}[\hat{a}_q \in A(q)]. \tag{17}$$

where $\hat{a}_q$ is the model prediction for question q, $A(q)$ is the set of acceptable gold answers. In consensus-based protocols (multiple human annotations), the indicator can be replaced by a soft scoring rule that gives partial credit when the prediction matches a subset of annotators.

Token-level F1 and generation metrics. For free-form answers (explanations of disease causes, treatment plans, or risk descriptions), token-level Precision, Recall, and F1 can be computed in the same way, by treating the textual answer as a short sequence and measuring word overlap with one or multiple references. BLEU, ROUGE, and BERTScore are also used as auxiliary indicators for longer explanatory answers, while human experts judge factuality and agronomic soundness.

Consistency. Consistency evaluates the logical coherence and stability of answers under paraphrased or repeated queries. Let $\mathcal{Q}(x)$ denote a set of questions that refer to the same underlying image x (rephrasings or follow-up questions). A simple consistency metric is

$$\text{Consistency} = \frac{\text{Num}_{\text{consistent}}}{\text{Num}_{\text{total}}}, \tag{18}$$

where $\text{Num}_{\text{consistent}}$ counts how many question sets in $\{\mathcal{Q}(x)\}$ receive mutually compatible answers, and $\text{Num}_{\text{total}}$ is the total number of such sets. In high-stakes AQA scenarios, high consistency is as important as high accuracy, since erratic behaviour undermines farmer trust and complicates safety auditing.

As summarized in Table 5, research on agricultural visual QA and multimodal diagnosis relies on a heterogeneous set of benchmarks. PlantVillage VQA is a canonical image–text benchmark covering 14 crops and 38 diseases, and is typically evaluated using Accuracy, IoU, and F1 to jointly assess lesion recognition and QA generation [77]. Multimodal diagnostic datasets such as CDDM provide hundreds of thousands of images and over one million QA pairs, placing greater emphasis on cross-modal consistency and natural language generation quality [78]. Field-based datasets such as Agriculture-Vision focus on large-scale farmland images with anomaly annotations, and are primarily used for segmentation and detection tasks evaluated by mAP and IoU [79]. Crop-specific datasets, exemplified by the Coffee Leaf Dataset, concentrate on single-crop disease scenarios and are often combined with knowledge-enhanced models to support prescription generation and fine-grained management decisions [80]. Recently, extended resources such as PlantPAD-PlanText, WDSD, Potato Disease Dataset, and Cane-MultiModal Dataset provide richer question types and longer textual descriptions, serving as testbeds for more advanced visual–language reasoning [81]. Together, these metrics and datasets provide the empirical foundation for evaluating agricultural visual QA systems from both perception and answer-quality perspectives.

Table 5: Representative datasets for agricultural VQA and multimodal diagnosis

Name	Type	Content scope	Common metrics	Source
PlantVillage VQA	Image–Text Pairs	55K images and 193K QA pairs, 14 crops, 38 diseases	VQA Accuracy, IoU, F1	[77]
CDDM	Multimodal (Image + Text)	137K images and 1 M QA pairs for disease diagnosis	VQA Accuracy, Cross-modal Consistency, BLEU	[78]
Agriculture-Vision	Remote sensing images	95K farmland RGB+NIR images with anomaly annotations	IoU, mAP, Segmentation Accuracy	[79]
Coffee Leaf Dataset	Segmentation images	Coffee leaf disease images (Karnataka dataset)	mAP, Prescription Accuracy	[80]
PlantPAD-PlanText	Image–Text Pairs	8200 disease images with auto-diagnostic QA pairs	Accuracy, Consistency, BLEU	[81]

3.3.2 Applications

With the deep integration of vision foundation models and large language models, agricultural visual diagnostic tasks have gradually evolved into multiple technical pathways, with core differences mainly reflected in cross-modal alignment mechanisms and prompt-driven generation methods. These pathways are exemplified in Fig. 8, which outlines both cross-modal fusion and prompt-based paradigms, as summarized in Table 6. First, in terms of cross-modal alignment and fusion, related approaches focus on establishing close connections between image and text feature spaces to enhance cross-modal reasoning capabilities. For example, Lan et al. developed a VQA model for fruit tree disease decision-making, integrating ResNet-152 and BERT with a co-attention fusion module, achieving effective multimodal reasoning on a self-built dataset [82]. AgriVLM [83] leverages Vision Transformer (ViT) to extract pest and disease image features and achieves cross-modal alignment with ChatGLM through Q-Former, achieving promising results in downstream diagnostic tasks. CroMIC-QA [84] integrates MobileNetV2 with GRU to build a shared semantic space, and combines cross-modal attention and answer re-ranking strategies, demonstrating effectiveness over traditional VQA models on agricultural community QA datasets. ILCD [85] further introduces multi-head co-attention (MCA) and bias correction mechanisms, enabling textual semantics to guide visual recognition while enhancing attribute diagnosis through knowledge-based guidance, and showed competitive performance on the CDwPK-VQA dataset. In addition, PotatoGPT [86] employs a multi-branch residual ViT and TextCNN fusion architecture in potato disease recognition, highlighting strong potential for cross-crop generalization. MA3 [87] proposes a multimodal agent framework, in which CLIP-ViT and a BERT-based router coordinate classification, detection, and expert models for multi-task decision-making, underscoring the potential of cross-modal fusion in complex agricultural decision-making.

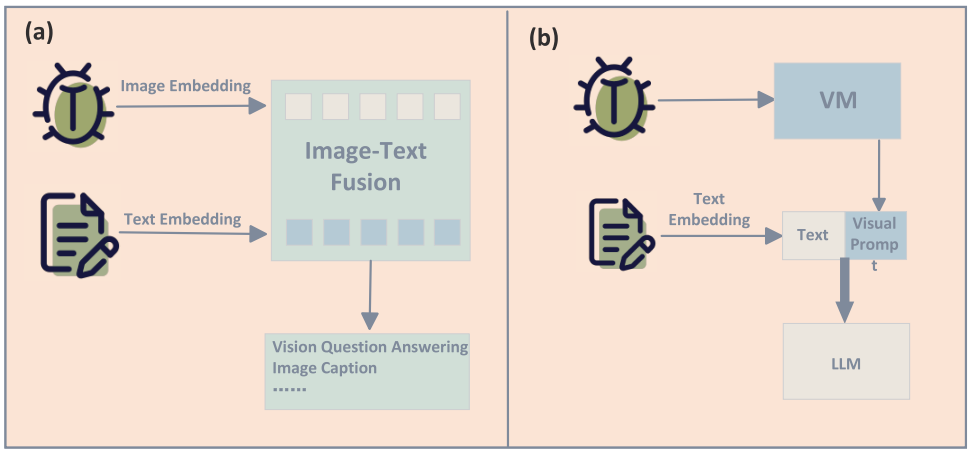

Figure 8: Two technological pathways in agricultural visual diagnosis: (**a**) cross-modal alignment between visual features and domain knowledge; and (**b**) collaborative pipelines where specialised vision models and large language models interact to produce diagnostic explanations and recommendations

Table 6: Representative multimodal agricultural QA models

Method	Year	Model name	Base model	Method	Task	Dataset	Performance	Advantages	Limitations
Cross-modal alignment fusion	2023	Fruit Tree VQA	ResNet-152 + BERT	MCA, Tucker fusion	Fruit tree disease QA	Self-built dataset	Acc 86.36%	Deep image–text fusion; effective reasoning	Small dataset; weak cross-crop generalization
	2024	AgriVLM	ViT + ChatGLM	Q-Former alignment, LoRA	Pest/disease QA	Self-built dataset	Acc >90%	Good alignment; efficient fine-tuning	Limited generalization; poor field transfer
	2024	CroMIC-QA	MobileNetV2 + GRU	Shared space, cross-modal attention, re-ranking	Community QA	CroMIC-QA-Agri	MAP 0.89, MRR 0.90	Strong small-sample matching; re-ranking boosts	Small dataset; poor high-res adaptation
	2025	PotatoGPT	MSC-ResViT + TextCNN	Three-branch residual fusion	Potato disease recognition	Potato dataset	Acc 98.43%	Highest accuracy; strong cross-crop	Single crop; low real-time ability
	2025	MA3	CLIP-ViT + BERT Router + Qwen	Multi-Agent (classifier, Router, experts)	Sugarcane disease QA/decision	Sugarcane dataset	Cls 96.2%; Router >99%	Lightweight Router; flexible framework	Focused on sugarcane; weak generalization
	2025	ILCD	Inception-v4 + LSTM	MCA, BiBa, knowledge-guided	Multi-attribute diagnosis	CDwPK-VQA	Acc 86.06%	Bias correction; knowledge-guided	Complex; high training cost
Template prompt generation	2024	PlanText	ViT + GPT-2	Mask prompting, template filling	Disease description	PlantPAD-PlanText	Achieves the best results on Morph, Sit, Area, Addr, B@4, and M.	Consistent, controllable text; >GPT-4/BLIP	Template dependent; limited transfer
	2025	WDLM	Grounded-SAM + VLM	Segmentation, LoRA, template, reasoning chain	Wheat diagnosis and prescription	WDSD	Better than CNN or Transformer; mobile-ready	Interpretable, robust	Latency; KB dependent
	2025	PestGPT	YOLOv8 + LLM	Detection + CoT + IoT	Pest/disease diagnosis	IoT + image dataset	Avg acc77.33%Avg Prec 44.02% Avg Rec44.00%F1 Score 44.01%	CoT + IoT closed-loop	High compute cost; unstable rare disease
	2024	Coffee-RAG	YOLOv8 + GPT-3.5 + RAG	Segmentation, RAG, text generation	Coffee leaf prescription	Karnataka dataset	—–	Small; coffee-only	

Second, regarding template- and prompt-driven generation methods, these approaches emphasize the role of textual prompts or masking guidance in visual processing and diagnostic generation. A typical example is PlanText [81], which employs progressive masking prompts to map lesion colors and textures into GPT-2 template slots, generating structured diagnostic descriptions with performance surpassing GPT-4 and BLIP. WDLM [88] integrates segmentation results from Grounded-SAM with a LoRA-fine-tuned language model, generating reasoning chains through prompt engineering to enable interpretable wheat disease diagnosis. Coffee-RAG [89] combines YOLOv8 segmentation with retrieval-augmented generation (RAG), allowing the language model to produce prescription recommendations with improved interpretability and practicality in coffee leaf disease tasks. Similarly, PestGPT [90] integrates YOLOv8 detection, language generation, and chain-of-thought (CoT) reasoning with IoT data, forming a closed-loop field diagnostic system covering recognition, reasoning, and prescription generation, further demonstrating the advantages of prompt-driven approaches in interpretable diagnosis.

3.4 Section Summary

This section reviewed the role of vision models in agricultural visual diagnosis. The progression from CNNs to versatile multimodal foundation models has significantly improved their transferability and generalization for agricultural tasks. Methodologically, approaches like parameter-efficient fine-tuning (PEFT) and retrieval-augmented generation (RAG), supported by diverse datasets, have enabled advanced applications from lesion recognition to prescription generation. In conclusion, while vision models are now a key driver for intelligent agricultural diagnosis, further progress is essential to overcome challenges in robustness against low-quality imagery, cross-crop transferability, and adherence to domain-specific standards.

4 Applications of Multimodal Foundation Models in Multimodal Understanding Tasks

With the rise of multimodal foundation models, agricultural QA has shifted from single-modal methods to cross-modal integration. By combining images, text, speech, and sensor data, multimodal models enable joint representation and reasoning, supporting tasks such as pest and disease diagnosis, agronomic consultation, and yield prediction. This transition drives agricultural systems toward multi-source fusion and integrated decision-making.

This chapter reviews the foundations and representative architectures of multimodal models, summarizes their application pathways (standard fine-tuning, knowledge-augmented fine-tuning, and reliability-oriented fine-tuning).

4.1 Foundations of Multimodal Models

Multimodal Models (MMs) are artificial intelligence systems capable of simultaneously processing and integrating heterogeneous modalities such as text, images, speech, and sensor signals. Their core objective is to achieve semantic alignment and joint reasoning across modalities, thereby overcoming the limitations of single-modality models and supporting more complex cognitive and decision-making tasks.

The modality encoders (ViT, CLIP) are responsible for extracting high-dimensional features from images, speech, or other inputs; the connector (such as an MLP, Q-Former, or cross-modal attention mechanism) maps these heterogeneous features into a shared representation space; and the decoder, usually a large language model, performs unified reasoning and natural language generation. The overall architecture is illustrated in Fig. 9.

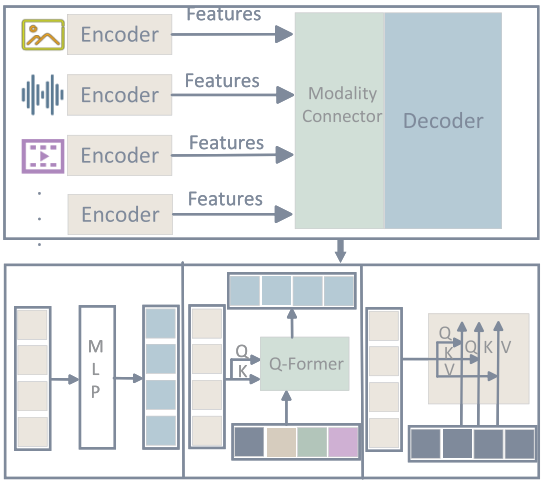

Figure 9: General architecture of multimodal foundation models, showing how visual, textual, and auxiliary sensor features are jointly encoded, aligned, and fused for reasoning over complex agricultural problems

Depending on the depth of information interaction, fusion strategies can be classified into early fusion, middle fusion, and late fusion. Early fusion relies on direct concatenation or shared embedding spaces for coarse-grained alignment. Middle fusion leverages cross-modal attention mechanisms to enable fine-grained interactions, while late fusion integrates the outputs of individual modalities at the decision stage, prioritizing computational efficiency and scalability.

Over the course of development, multimodal modeling has evolved from shallow feature concatenation to deep interactive reasoning. Early research largely depended on simple feature fusion, enabling basic image–text matching and QA. With the emergence of cross-modal attention and Q-Former, middle-fusion paradigms became mainstream, offering stronger reasoning capabilities while maintaining efficiency. More recently, large-scale general-purpose multimodal models such as GPT-4V [91], LLaVA [92], and DeepSeek-VL2 [93] have achieved unified modeling across text, images, and speech, extending to open-domain, multi-source information processing and marking the transition to large-scale, generalizable multimodal intelligence.

4.2 Multimodal Model Adaptation

4.2.1 Prompt-Based and Parameter-Efficient Fine-Tuning

In multimodal tasks, parameter-efficient fine-tuning and prompt-driven methods also demonstrate broad application potential. By introducing techniques such as LoRA, Adapters, or Prompt Tuning at the input level of different modalities (vision, text, or even speech), task adaptation can be rapidly achieved while keeping the majority of the pretrained model frozen. Representative work includes InstructBLIP, which incorporates instruction tuning on top of BLIP-2, enabling the model to better understand cross-modal instructions and perform diverse tasks; and LLaVA, which connects visual projection layers with language models and employs instruction fine-tuning to support high-quality image–text dialogue. These approaches reduce computational costs while significantly improving controllability and interpretability in complex tasks such as cross-modal question answering and agricultural diagnostic consultation.

4.2.2 Retrieval-Augmented Fine-Tuning

Retrieval-augmented methods in multimodal settings further enhance the reliability and knowledge coverage of models in real-world applications. Unlike unimodal RAG, multimodal RAG can simultaneously leverage images, text, and structured data, providing additional conditional inputs from external knowledge bases or documents. Representative works include Atlas and RETRO, which integrate retrieval mechanisms during both training and inference, substantially improving performance in knowledge-intensive tasks. In fields such as medicine and agriculture, similar frameworks have been applied to combine image-based diagnosis with knowledge-base matching, ensuring that generated answers are not only accurate but also traceable and compliant. Looking ahead, multimodal RAG approaches that combine image, text, and sensor data are expected to become a key solution to challenges of data scarcity and outdated knowledge in agricultural QA systems.

4.2.3 Reliability-Oriented Fine-Tuning

As multimodal models are increasingly applied to high-stakes domains such as healthcare and agriculture, ensuring their reliability and safety has become a critical concern. Recent studies have proposed reliability-oriented fine-tuning strategies such as reinforcement learning from human feedback (RLHF), consistency regularization, and chain-of-thought (CoT) consistency constraints. For example, incorporating expert scoring mechanisms or consistency metrics into multimodal QA can effectively suppress model hallucinations, improving the stability and trustworthiness of answers. In multi-source data fusion tasks, enforcing logical consistency across modalities can further enhance the rigor of cross-modal reasoning. These methods not only improve the scientific validity and regulatory compliance of generated results but also provide a solid foundation for deploying multimodal models in practical agricultural scenarios.

4.3 Applications of Multimodal Reasoning Tasks

4.3.1 Evaluation Metrics and Multimodal Datasets

In multimodal question answering (QA), the evaluation framework largely inherits metrics from text-based and visual AQA (Accuracy, F1, BLEU, ROUGE, IoU, mAP), but additionally needs to capture *cross-modal consistency* and *application-oriented performance*, such as task completion and system efficiency. Below we highlight several complementary metrics that are particularly relevant to agricultural multimodal QA.

(1) Task success and execution metrics.

For agent-style or tool-using multimodal systems, the ultimate goal is often to *complete* a task (generate a valid prescription, trigger the correct actuator, or finish a workflow) rather than merely output a correct answer string. Task Success Rate (TSR) measures the proportion of tasks that are completed according to a predefined success criterion:

$$\text{Task Success Rate} = \frac{\text{Num}_{\text{success tasks}}}{\text{Num}_{\text{total tasks}}}, \tag{19}$$

where $\text{Num}_{\text{success tasks}}$ is the number of tasks that meet domain-specific constraints (valid recommendation, no safety violations, correct tool invocation), and $\text{Num}_{\text{total tasks}}$ is the total number of tasks. TSR is widely used in instruction-following and decision-support scenarios, and complements token-level or span-level metrics by directly reflecting practical executability.

(2) Regression metrics for continuous targets.

Many multimodal agricultural applications involve predicting continuous variables, such as yield, biomass, or physiological indices, from fused image, text, and sensor data. In such settings, coefficient of determination (R^2) and Root Mean Squared Error (RMSE) are commonly reported.

Given N samples with ground-truth values y_i and predictions $\hat{y}_i$, and $\bar{y}$ denoting the mean of $\{y_i\}_{i=1}^{N}$, R^2 and RMSE are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{20}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{21}$$

Higher R^2 and lower RMSE indicate better agreement between predictions and ground truth. In breeding or yield-forecasting scenarios, these metrics quantify how well multimodal models capture complex genotype–environment–management interactions.

(3) System-level and deployment metrics.

For edge or IoT-oriented multimodal AQA systems, *latency*, *energy consumption*, and *edge reliability* are also critical. Latency measures end-to-end response time, energy consumption tracks power usage on constrained devices, and edge reliability reflects the fraction of queries that can be answered within resource and timing constraints. These metrics are increasingly reported in studies that evaluate lightweight multimodal models or edge deployment frameworks, and complement accuracy-oriented metrics by reflecting real-world feasibility.

In agricultural multimodal QA research, the development of dedicated datasets provides the empirical basis for training and evaluating models under diverse modality combinations and application scenarios, as summarized in Table 7. Agri-LLaVA offers more than 400K image–text instruction pairs covering 221 pests and diseases, enabling large-scale multimodal dialogue and QA over crop protection scenarios [94]. AgroInstruct contains 70K expert-aligned instruction dialogues, supporting instruction fine-tuning and explicit knowledge injection into foundation models [95]. AgMMU is designed as a comprehensive multimodal benchmark with thousands of questions spanning images, tables, and text, and focuses on cross-task reasoning and cross-modal consistency evaluation [96].

Table 7: Representative datasets for multimodal agricultural QA and decision support

Name	Type	Content scope	Common metrics	Source
Agri-LLaVA Dataset	Image–Text Instructions/Multimodal QA	400K image–text instruction pairs, covering 221 types of pests and diseases	VQA Accuracy, BLEU, expert evaluation	[94]
AgroInstruct	Instruction data	70K expert-aligned agricultural instruction dialogues	BLEU, ROUGE, task completion rate	[95]

(Continued)

Table 7 (continued)

Name	Type	Content scope	Common metrics	Source
AgMMU	Multimodal benchmark	5460 MCQs and 5460 OEQs across images, tables, and text for agricultural reasoning	VQA Accuracy, cross-modal consistency, reasoning accuracy	[96]
Tri-Modal Transformer Dataset	Image, Text, and Sensor	7800 samples for tri-modal disease diagnosis	Precision, Recall, Accuracy	[97]
ROOT	Multimodal comparison	100 images, 100 matrices, and 100 QA samples	Image/table/QA scores, latency, energy consumption	[98]
Agri-QA Net Dataset	Image, Text, and Speech	10,000 text, 2000 speech, and 3000 images in a cabbage scenario	Accuracy, F1, cross-modal attention analysis	[99]
Farm-LightSeek Dataset	IoT with Image and Edge computing	Two real IoT farm datasets for lightweight edge multimodal validation	Edge reliability, energy consumption, task completion rate	[100]

The Tri-Modal Transformer dataset integrates sensor measurements with images and textual descriptions, allowing researchers to quantify the benefit of multi-source fusion for disease diagnosis [97]. ROOT provides balanced sets of images, matrices, and QA samples, and is specifically used to compare models along accuracy, latency, and energy consumption dimensions [98]. Agri-QA Net incorporates text, images, and speech in a cabbage disease management scenario, advancing voice-interactive QA and robustness under field-acoustic conditions [99]. Finally, Farm-LightSeek includes real-world IoT farm datasets that couple multimodal sensing with on-device and edge validation, enabling evaluation of edge reliability, energy usage, and task completion rate for lightweight deployment scenarios [100]. Taken together, these datasets and metrics constitute the experimental backbone of agricultural multimodal QA, linking model-level performance to practical decision-support and deployment considerations.

4.3.2 Applications

In agricultural multimodal QA tasks, different modality combinations have given rise to diverse application paradigms, as summarized in Table 8. The most common is the vision-language paradigm, which relies primarily on image-text instructions to support pest and disease diagnosis, as well as field quality assurance. A representative example is Agri-LLaVA [94], built on the LLaVA-1.5 framework with approximately 400,000 agricultural vision–language instruction pairs. Through a two-stage fine-tuning process, it successfully adapts from general-purpose dialogue to agriculture-specific multimodal dialogue, achieving significant improvements on both VQA and chatbot benchmarks. Similarly, AgroGPT [95] synthesizes instructions from visual data and applies LoRA fine-tuning, achieving superior fine-grained recognition performance compared to general LLM in datasets, such as PlantVillage and PlantDoc. WBLM [101]

further extends this paradigm by combining InternVL2-8B with multi-stage training and cross-domain knowledge integration, achieving strong adaptability in wheat breeding QA tasks, with an R^2 of 0.821 for yield prediction across multi-source knowledge bases. Another paradigm is the vision–language–speech framework, which enhances human–machine interaction through voice input, making it suitable for real-time field QA. For example, Agri-QA Net, built on Qwen-VL-Chat-Int4, incorporates quantization-based inference and dynamic cross-modal attention to support trimodal inputs, achieving strong precision in cabbage-related QA tasks. With the increasing importance of environmental factors in the diagnosis of pests and diseases, the vision-language-sensor paradigm has gained increasing attention. Tri-Modal Transformer [97] is the first to align visual, textual, and sensor data, achieving strong performance on a plant disease diagnosis dataset with high precision and recall. Farm-LightSeek [100] further advances this paradigm by leveraging knowledge distillation to compress LLMs/MLLMs into lightweight edge models that interact with IoT data streams, demonstrating reliable performance on two real-world smart farm datasets under low-power conditions. Beyond these, researchers have also explored extended multimodal paradigms incorporating tabular data, remote sensing imagery, and cross-lingual information. For example, AI Plant Doctor [102] leverages GPT-4V in combination with a pesticide knowledge base to perform disease diagnosis and medication consultation under zero-shot conditions. ROOT [98] employs prompt engineering to orchestrate general-purpose multimodal models such as Claude-3.5, GPT-4o, and Gemini-1.5. On the GARDeN dataset, it simultaneously handles image recognition, table parsing, and QA, while also providing quantitative comparisons of latency and cost. Overall, the vision–language paradigm emphasizes large-scale instruction tuning and lightweight adaptation, focusing on knowledge injection and generalization. The vision–language–speech paradigm strengthens interactivity and practical usability, while the vision–language–sensor paradigm highlights causal reasoning and environmental adaptability. Other extended multimodal paradigms further expand task boundaries but remain constrained by limited dataset scale, weak cross-domain transferability, and instability in model performance.

Table 8: Representative multimodal models for agricultural QA (2023–2025)

Year	Model	Modality	Method	Task	Results	Advantages	Limitations
2023	Tri Modal Transformer	Image, Text, Sensor	Multimodal Transformer, adaptive loss, tri-modal alignment	Disease diagnosis (fusion)	Precision = 96; Recall = 93; Accuracy = 94	Validated sensor semantics; high fusion accuracy	Requires high-quality sensor data; high computational cost
2024	AI Plant Doctor	Image, Text	GPT-4V zero-shot reasoning, prompt engineering, knowledge-base retrieval	Disease diagnosis and pesticide consultation	–	Combination of prompting and structured KB; easy deployment	Lacks domain-specific fine-tuning; dependent on KB coverage
2024	Agri-LLaVA	Image, Text	LLaVA-1.5 framework, large-scale image–text instruction data, two-stage supervised fine-tuning	Multimodal dialogue and visual QA	Chatbot = 55.4; VQA = 60.05	Two-stage knowledge injection; large-scale dialogue training	High inference latency; requires large computational resources

(Continued)

Table 8 (continued)

Year	Model	Modality	Method	Task	Results	Advantages	Limitations
2024	ROOT	Image, Text	Prompt engineering, direct use of general-purpose multimodal models	Image recognition, matrix parsing, farmer QA	Image = 8.80; Matrix = 9.00; QA = 8.57	Zero-shot adaptation across tasks; cross-task feasibility	Small-scale dataset; synthetic evaluation tasks
2024	WBLM	Image, Text	Multi-stage training (SFT→RAG →RLHF), knowledge-base integration	Wheat breeding cross-domain multimodal QA	Yield prediction R^2 = 0.821; RMSE = 489.25 kg/ha	Three-stage closed-loop training; strong cross-domain adaptability	Very high training cost; relies on multi-source KBs
2024	Farm-LightSeek	Image, Text, Sensor	Knowledge distillation, lightweight models, IoT integration	IoT-based perception–decision–action	Verified on real IoT farm datasets	Lightweight deployment; feasible on edge devices	Limited task scope; weak cross-crop generalization
2025	AgroGPT-7B/3B	Image, Text	LoRA fine-tuning, visual instruction synthesis, CLIP/SigLIP alignment	Fine-grained recognition and dialogue	Disease recognition = 51.37/30.82; Pest recognition = 45.58	Low-cost domain model construction; efficient adaptation	Limited to visual tasks; lacks sensor/text fusion
2025	Agri-QA Net	Image, Text, Speech	Quantized inference, dynamic cross-modal attention	Multimodal QA for cabbage crops	Accuracy = 89.5; F1 = 89.6	Lightweight inference; dynamic cross-modal attention	Small dataset; specific to one crop

4.4 Section Summary

Multimodal foundation models are advancing agricultural QA by integrating vision, language, and sensor data for complex tasks. Their evolution toward general-purpose architectures, combined with adaptation strategies like fine-tuning and RAG, has enhanced domain-specific performance and robustness. However, practical deployment is still hindered by critical challenges such as data scarcity, poor transferability, and model instability. Overcoming these issues through better datasets, alignment, and lightweight deployment is required to unlock their full potential.

5 Evaluation and Diagnostic Framework for Agricultural QA

5.1 Standardized Evaluation Protocols

Currently, the lack of unified evaluation standards in AQA leads to unfair comparisons. To address this, we propose a standardized evaluation protocol (summarized in Table 9) that future research should adhere to, focusing on clear experimental settings, rigorous parameter reporting, and bias control in automated evaluation.

Table 9: Standardized AQA evaluation protocols (Four-stage framework)

Stage	Core metrics	Reporting mandates	Bias mitigation
Knowledge injection	Accuracy, F1-Score.	**Distinguish Paradigm:** (1) Closed-book (Parametric). (2) Open-book (RAG/Agent).	**Context Fairness:** Do not compare closed-book and RAG without identifying data advantage.
Retrieval	Recall@k, NDCG@k, MRR.	**Retriever Specs:** (1) Model (BM25/dense). (2) Chunk size. (3) Top-k setting.	**Noise Control:** Report if re-ranker or filtering is used.
Generation	Win-rate, GPT-Score, Correctness.	**Judge Prompts:** Must disclose the exact evaluation prompts.	**Position Bias:** Use pairwise evaluation and swap A/B order.
Human evaluation	Factuality, Safety, Actionability.	**Annotator Profile:** Agronomists and laypersons.	**Double-Blind:** Hide model identity to prevent branding bias.

Distinguishing Experimental Paradigms. A common pitfall in current literature is the ambiguous comparison between models with different knowledge access. As defined in the first tier of Table 9, researchers must explicitly categorize their experiments into Closed-book (relying solely on internal parametric memory) or Open-book (accessing external vector databases or KGs). Comparing a RAG-enhanced model against a closed-source baseline without explicit disclaimer creates a disparity in information accessibility.

Transparency in Retrieval Configurations. For RAG-based systems, the performance is heavily dependent on retrieval hyperparameters. We mandate that evaluation reports must disclose the Retrieval Configuration (in Table 9), specifically: the retrieval algorithm (Sparse BM25 vs. Dense Retrievers like BGE), the Chunk Size (256 vs. 512 tokens), and the number of retrieved contexts (Top-k). Omitting these details renders the implementation unreproducible and prevents the community from analyzing whether improvements stem from the LLM or the retriever.

Mitigating Bias in LLM-as-a-Judge. Given the high cost of expert evaluation, using advanced LLMs (GPT-4) as judges is becoming standard. However, these judges exhibit Positional Bias favoring the first answer and Verbosity Bias favoring longer answers. To ensure fairness, the evaluation protocol must implement Position Swapping (evaluating both Order A-B and Order B-A) and report the win-rate average. furthermore, for critical safety claims, human evaluation by qualified agronomists remains indispensable and cannot be fully replaced by automated metrics.

Differentiating Human Evaluation Levels. While automated metrics provide scalability, they cannot fully capture the actionability and safety of agricultural advice. As outlined in the final tier of Table 9, standardized protocols must strictly distinguish between "Layperson Evaluation"(judging fluency and coherence) and "Expert Evaluation" (judging agronomic correctness). We advocate for a Double-Blind Review process where annotators—preferably certified agronomists or extension agents—evaluate anonymized responses. This ensures that the assessment focuses on whether the advice is practically implementable by farmers, free from model branding bias or hype.

5.2 Diagnostic Framework and Error Analysis

While Section 5.1 outlines the procedural protocols for evaluation, a robust diagnostic framework requires a theoretical taxonomy to categorize the nature and severity of model failures. We propose a three-dimensional framework for analyzing errors in agricultural LLMs like Table 10.

Table 10: Proposed diagnostic taxonomy: Error types, examples, and severity grading

Error dimension	Failure mechanism	Agricultural scenario example	Severity
Safety & Factuality (Critical Failures)	Hallucinated recommendation	Suggesting a pesticide that is banned or non-existent.	High
	Numerical calculation	Correctly citing a dosage rate (5 g/L) but failing to calculate the total amount for a 200 L tank.	High
Spatio-Temporal (Context Mismatch)	Temporal decay	Providing pest control guidelines from 2015 that contradict 2024 regulations.	Med-High
	Spatial/Regional mismatch	Recommending a crop variety suited for humid tropics to a farmer in an arid zone.	Medium
Surface Level (Benign Errors)	Stylistic inconsistency	Using overly academic jargon instead of farmer-friendly colloquialisms.	Low
	Imprecise description	Describing a symptom color as “light green” instead of “chlorotic yellow” (conceptually close).	Low

Severity Grading: Benign and Critical Errors. Unlike general domain chat-bots, agricultural errors carry varying levels of risk.

- *Benign Errors*: Minor hallucinations in linguistic style or non-critical descriptors (describing a leaf as “light green” instead of “pale green”).
- *Critical Errors*: Safety violations, such as recommending banned pesticides, incorrect dilution ratios, or misidentifying pathogens with similar visual symptoms.

Evaluation metrics must strictly penalize Critical Errors with a higher weight than Benign Errors, rather than averaging them into a single accuracy score.

Spatio-Temporal Misalignment. A unique challenge in agriculture is that “correctness” is time- and location-dependent. A recommendation valid in 2020 (a specific pesticide) might be banned in 2024; a treatment effective in arid climates might fail in humid regions. We define *Temporal Decay* as an error type where the model retrieves outdated knowledge, and Spatial Mismatch where advice is applied to the wrong geographic context.

Reasoning-Retrieval Dissociation. Our analysis reveals that many SOTA models exhibit a “Reasoning-Retrieval Dissociation.” The model may successfully retrieve the correct technical manual (high Recall) but fail to extract the specific numerical value required for calculation (low Reasoning). This suggests that future optimization should focus on Chain-of-Thought (CoT) prompting specifically designed to bridge the gap between retrieved text and numerical application.

The evaluation protocol and diagnostic taxonomy proposed in this section together provide a foundation for building more rigorous and actionable AQA benchmarks. On the one hand, the four-stage evaluation framework (Knowledge Injection, Retrieval, Generation, and Human Evaluation) clarifies what should be measured at each stage of the pipeline and how experimental settings must be reported to ensure reproducibility and fair comparison. On the other hand, the error-oriented diagnostic framework highlights safety-critical, spatio-temporal, and surface-level failure modes that are specific to agricultural decision-making, and thus cannot be adequately captured by generic accuracy metrics alone.

These insights naturally extend to an agricultural QA roadmap along four dimensions: benchmark construction, evaluation protocols, safety norms, and deployment considerations. For benchmark construction, future datasets should be designed to expose the error types in Table 10, including long-tail pests, region- and time-sensitive regulations, and numerically intensive recommendation scenarios. For evaluation protocols, standardized reporting of retrieval settings, judge prompts, and human annotator profiles (Table 9) should become mandatory to make cross-paper comparison meaningful. For safety norms, dynamic compliance checkers and double-blind expert evaluation must be integrated into the assessment of high-risk outputs such as pesticide and fertilizer recommendations. Finally, for deployment, metrics related to latency, energy consumption, and robustness under field conditions should be incorporated into future benchmarks, so that AQA systems are evaluated not only as models in isolation but as end-to-end tools that can be reliably deployed in real agricultural environments.

6 Critical Challenges, Safety Alignment, and Future Directions

6.1 Summarizing Service Paradigms Based on Model Classification Methods

This chapter reviews the application of multimodal foundation models (FMs) in agricultural question answering (QA). By integrating vision, language, speech, and sensor data, multimodal models now support complex agricultural scenarios such as pest and disease diagnosis, agronomic consultation, and yield prediction. Model architectures have evolved from contrastive-learning frameworks (CLIP, ALIGN), to alignment–generation models (BLIP-2, Flamingo), and further to large-scale multimodal systems such as GPT-4V, LLaVA, and DeepSeek-VL2. Adaptation strategies including efficient fine-tuning, retrieval-augmented generation (RAG), and reliability optimization have improved domain specialization and robustness.

As summarized in Table 11, three major paradigms can be identified:

Text-based QA—LLMs adapted via pretraining, LoRA/QLoRA, RAG, and prompting are suited for policy QA and agronomic texts, but remain limited by long-tail coverage, hallucinations, and weak context retention.

Visual diagnostic QA—vision models such as ViT and YOLOv8 with fine-tuning and prompt-driven alignment perform strongly in pest and disease recognition and symptom inference, but are sensitive to image quality and perform poorly on rare diseases.

Multimodal QA—fusion of images, text, and sensors through multi-stage training (SFT, RAG, RLHF) enables dialogue-based diagnosis and yield prediction, yet remains constrained by limited data, high computational cost, and inference latency.

Overall, these paradigms are complementary: LLMs excel at language understanding, visual models at perception and recognition, and multimodal models at cross-modal reasoning. Together, they are driving agricultural QA toward more comprehensive, interpretable, and deployable systems, although progress in dataset construction, cross-modal alignment, and lightweight deployment remains needed.

Table 11: Paradigms of agricultural QA tasks

Attribute	Text-based QA	Visual diagnostic QA	Multimodal QA
Main models	LLMs	Vision + LLMs / MLLMs	Multimodal Models
Service characteristics	Domain pretraining, LoRA/QLoRA, RAG, prompt engineering	Image feature extraction, LoRA/VPT fine-tuning, cross-attention alignment, prompt-driven fusion	Image–text finetuning; SFT + RAG + RLHF
Typical applications	Policy QA, nutrition matching, text-based pest/disease diagnosis	Pest/disease recognition, lesion segmentation, symptom inference	Multimodal dialogue, field monitoring, yield prediction
Advantages	Strong language understanding and generation; fast task adaptation	Fine-grained image recognition; zero-shot or cross-crop transfer	Strong cross-modal reasoning; text+image+sensor integration
Limitations/Challenges	Limited long-tail coverage; hallucinations; weak context retention	Sensitive to image quality; poor rare-disease accuracy; low interpretability	High training cost; large data/compute dependence; high latency

6.2 Safety, Reliability, and Hallucination Control

While foundation models have demonstrated impressive capabilities in agricultural reasoning, their deployment in real-world farming scenarios is hampered by stochastic generation errors. Unlike general-domain tasks, agricultural QA operates in a high-stakes environment where incorrect advice can lead to economic loss or safety hazards. As systematically outlined in Table 12, future AQA research must limit model behavior across four critical dimensions: hallucination control, risk mitigation, uncertainty quantification, and ethical alignment.

1. Hallucination Control (Sourcing and Refusal). Generative models often fabricate non-existent pesticide names or cite obsolete technical manuals. To curb this, strictly grounding answers in retrieved evidence (RAG) is a prerequisite. More importantly, models must be equipped with a refusal mechanism. When the retrieval system yields low-similarity documents (score < 0.7) or evidence is conflicting, specific prompting strategies (as shown in Table 12) should trigger the model to explicitly state "Insufficient Information" rather than hallucinating a plausible but incorrect response.

2. Risk Mitigation (Safety Guardrails). Agricultural safety involves strict adherence to regulations regarding banned substances, dosage limits, and compatibility. Prompt engineering alone often fails to prevent "jailbreaks." Therefore, effective risk mitigation requires neuro-symbolic guardrails integrating external rule engines that filter LLM outputs against legal databases (MRL standards). Furthermore, automated disclaimer injection and human-in-the-loop routing mechanisms are essential to ensure that high-risk prescriptions are verified by experts before reaching farmers.

3. Uncertainty and Reliability (Confidence Reporting). A deterministic answer is dangerous when the input evidence is ambiguous. Reliable AQA systems must move beyond single-choice outputs to provide. This includes reporting calibrated probabilities and verbalized confidence scores. Future work should explore conformal prediction to provide mathematical guarantees for prediction intervals, helping users judge the credibility of the advice.

4. Ethics and Values (Ecological Alignment). Beyond correctness, AQA models serve a directive role in farming practices. Consequently, they must be aligned with broader ecological values and ethics. Instead of defaulting to chemical solutions, models should be instruction-tuned to prioritize sustainable practices, such as Integrated Pest Management (IPM) and biological control. This ensures that the widespread use of AI assistants contributes to long-term agricultural sustainability rather than exacerbating chemical dependency.

Table 12: Strategies for safety, hallucination control, and ethical alignment

Dimension	Control mechanism	Prompt engineering	Future research
Hallucination control	**Sourcing & Refusal:** Ensuring answers are grounded in retrieved docs; refusing when evidence is low.	**RAG Prompt:** "Answer strictly based on provided context. State "Insufficient Information" if unsure." Score Threshold: Refusal if score < 0.7.	**Verifiable Gen:** Loss functions penalizing ungrounded claims. Granular Refusal: Partial answering.
Risk mitigation	**Safety Guardrails:** Preventing illegal pesticide advice or toxic combinations.	**Constraint Prompts:** "Prohibition: Do not recommend restricted pesticides." Disclaimer: Auto-adding legal texts.	**Neuro-Symbolic:** Integrating rule engines (MRL limits). **Human-in-the-loop**: Expert routing.
Uncertainty/ Reliability	**Confidence Reporting:** Quantifying the model's certainty.	**Verbalized:** "Rate confidence (1–5) and explain." **Calibrated Prob**: Top-3 disease probabilities.	**Conformal Prediction:** Mathematical guarantees for prediction intervals.
Ethics & Values	**Value Alignment:** Prioritizing sustainability.	**Instruction:** "Prioritize IPM and biological control over chemicals."	**Ethical Fine-tuning:** Training on sustainable farming principles.

6.3 Future Research Directions for Agricultural QA

Based on the review of text-based, visual, and multimodal agricultural QA tasks, it is evident that although foundation models have achieved remarkable progress, their further development is still constrained by several common challenges. Future research should focus on the following aspects to advance agricultural QA systems from being merely "usable" to becoming "trustworthy, reliable, and efficient," while at the same time guiding the construction of benchmarks, evaluation protocols, and safety norms.

First, in **text-based QA**, research must go beyond knowledge retrieval and surface-level generation, and move toward causal reasoning and compliance assurance. On the one hand, integrating causal chains from agricultural knowledge graphs into the reasoning process can allow models to produce more logical and interpretable answers that can be systematically evaluated in reasoning-oriented benchmarks. On the other hand, dynamic compliance checkers should be established to examine outputs related to prescriptions such as pesticide or fertilizer recommendations, ensuring strict alignment with the latest regulations and safety standards; these components should be explicitly reflected in future benchmark design and evaluation protocols for high-risk agronomic decisions.

Second, in **visual diagnosis**, the key challenge lies in improving cross-domain generalization and adaptive perception in real-world farmlands. Future efforts should improve robustness against low-quality or long-tailed images, ensuring reliable performance under poor lighting, complex backgrounds, or rare pest and disease scenarios. This calls for curated visual benchmarks that explicitly cover cross-region, cross-variety, and long-tailed conditions, along with protocol guidelines on train–test splits and domain shift settings. In addition, combining vision models with embodied platforms such as drones and agricultural robots can support active perception and continuous learning, raising new requirements for online evaluation, safety constraints, and deployment-oriented metrics such as latency and energy consumption.

Finally, in **multimodal reasoning**, the priority is achieving deep semantic alignment between modalities and enabling decision-driven intelligence. Dynamic cross-modal attention mechanisms and lightweight mixture-of-experts (MoE) architectures should be explored to flexibly prioritize visual, sensor, or textual information depending on the task, and future benchmarks should include multi-source, temporally evolving scenarios that stress long-horizon reasoning. Furthermore, the development of multimodal agents should be encouraged, allowing systems not only to answer questions, but also to call external tools and databases, simulate the outcomes of agricultural management strategies, and provide calibrated recommendations under explicit safety and compliance constraints. Taken together, these directions outline an **agricultural QA roadmap** centered on rigorous benchmark construction, standardized evaluation protocols, and safety- and deployment-aware system design for foundation-model–based AQA.

Beyond these conceptual directions, we further propose an actionable Agricultural QA Roadmap that outlines how future research should advance from benchmark design to real-world deployment. At the data and task-definition level, the community should prioritize the construction of open, reproducible benchmarks that cover text-based, visual, and multimodal scenarios, together with consistent annotation standards and well-defined task boundaries. At the methodology and evaluation level, unified evaluation protocols are needed to ensure comparability across models, requiring explicit reporting of retrieval settings, safety constraints, and human evaluation procedures. At the system and safety level, agricultural QA models must incorporate regulatory alignment, uncertainty estimation, audit logs, and compliance checkers, especially for high-risk recommendations involving pesticide or fertilizer use. Finally, at the engineering and deployment level, future work should address practical constraints—including latency, energy consumption, and varying compute environments—by exploring model compression, inference acceleration, and edge–cloud collaborative deployment. Progress along these four layers will allow agricultural QA systems to evolve from isolated model capabilities toward robust, field-ready intelligent solutions.

7 Conclusion

The main contribution of this survey is to adopt a task-paradigm–driven view that jointly analyzes how foundation models are applied across textual, visual, and multimodal AQA tasks, thereby constructing a unified framework that links datasets, model paradigms, evaluation protocols, and application scenarios. This integrated perspective makes it easier to compare heterogeneous systems, to understand trade-offs between retrieval-based and generation-based pipelines, and to transfer design insights across tasks and modalities. Looking ahead, we argue that progress will increasingly depend on a clearer agricultural QA roadmap that connects methodological advances with concrete infrastructure. In particular, future work should (i) develop open and reproducible benchmarks that cover text-based, visual, and multimodal scenarios under realistic domain shifts; (ii) establish standardized evaluation protocols that require transparent reporting of retrieval configurations, safety strategies, and human evaluation procedures; (iii) build safety and compliance norms, including dynamic checkers for prescription-related outputs and audit mechanisms for high-risk recommendations; and (iv) address deployment considerations, such as latency, energy consumption, and

heterogeneous compute environments through model compression, inference acceleration, and edge–cloud collaboration. These directions, summarized in Fig. 10, outline a path for moving AQA systems from isolated model improvements toward trustworthy, reliable, and efficient decision-support tools. We hope that this work can serve as a reference for future research on model architecture, benchmark and dataset construction, evaluation and safety protocols, and end-to-end system design in multi-task AQA, and that it will accelerate the practical impact of foundation-model-based AQA technologies in smart agriculture and sustainable development.

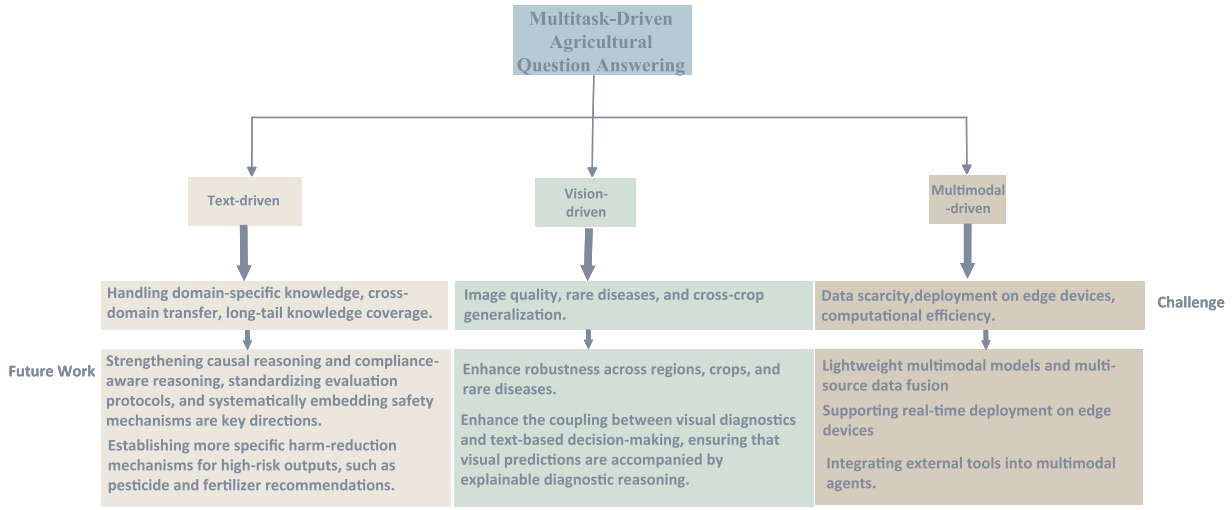

Figure 10: Summary of key challenges and future research directions for multitask agricultural QA, including safety, domain generalization, multimodal alignment, and deployment constraints

The main contribution of this survey is to adopt a task-paradigm–driven view that jointly analyzes how foundation models are applied across textual, visual, and multimodal AQA tasks, thereby constructing a unified framework that links datasets, model paradigms, evaluation protocols, and application scenarios. This integrated perspective makes it easier to compare heterogeneous systems, to understand trade-offs between retrieval-based and generation-based pipelines, and to transfer design insights across tasks and modalities. Looking ahead, we argue that progress will increasingly depend on a clearer agricultural QA roadmap that connects methodological advances with concrete infrastructure. In particular, future work should (i) develop open and reproducible benchmarks that cover text-based, visual, and multimodal scenarios under realistic domain shifts; (ii) establish standardized evaluation protocols that require transparent reporting of retrieval configurations, safety strategies, and human evaluation procedures; (iii) build safety and compliance norms, including dynamic checkers for prescription-related outputs and audit mechanisms for high-risk recommendations; and (iv) address deployment considerations such as latency, energy consumption, and heterogeneous compute environments through model compression, inference acceleration, and edge–cloud collaboration.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Ningxia Natural Science Foundation (2025AAC050001); the Scientific Research Startup Project for Full-Time Introduced High-Level Talents in Ningxia (2024BEH04130); the National Natural Science Foundation of China (32460444); the Ningxia Hui Autonomous Region Key Research and Development Program (2024BBF0101302, 2023BDE02001) and Supported by the Special Fund for Basic Research Business of Central Universities of North Minzu University (2025BG234, 2023ZRLG12).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology: Jianping Liu, Changxu Zhao; Formal analysis and investigation: Changxu Zhao, Xiaofeng Wang, Haiyu Ren, Pan Liu; Data collection and curation: Qiantong Wang, Wei Sun, Libo Liu; Writing—original draft preparation: Changxu Zhao; Writing—review and editing: Jianping Liu, Xiaofeng Wang; Supervision and funding acquisition: Jianping Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: No new data were generated or analysed in support of this research. All data supporting the findings of this study are available from the corresponding references cited in the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Nedelciu CE, Ragnarsdottir KV, Schlyter P, Stjernquist I. Global phosphorus supply chain dynamics: Assessing regional impact to 2050. Glob Food Secur. 2020;26:100426. doi:10.1016/j.gfs.2020.100426.
2. Li J, Xu M, Xiang L, Chen D, Zhuang W, Yin X, et al. Foundation models in smart agriculture: basics, opportunities, and challenges. Comput Electron Agric. 2024;222:109032. doi:10.1016/j.compag.2024.109032.
3. Li H, Wu H, Li Q, Zhao C. A review on enhancing agricultural intelligence with large language models. Artif Intell Agric. 2025;15(4):671–85. doi:10.1016/j.aiia.2025.05.006.
4. Chen D, Fisch A, Weston J, Bordes A. Reading Wikipedia to answer open-domain questions. In: Barzilay R, Kan MY, editors. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017. p. 1870–9.
5. Aviles Toledo C, Crawford MM, Tuinstra MR. Integrating multi-modal remote sensing, deep learning, and attention mechanisms for yield prediction in plant breeding experiments. Front Plant Sci. 2024;15:1408047. doi:10.3389/fpls.2024.1408047.
6. Delfani P, Thuraga V, Banerjee B, Chawade A. Integrative approaches in modern agriculture: IoT, ML and AI for disease forecasting amidst climate change. Precis Agric. 2024;25(5):2589–613. doi:10.1007/s11119-024-10164-7.
7. Sen LTH, Phuong LTH, Chou P, Dacuyan FB, Nyberg Y, Wetterlind J. The opportunities and barriers in developing interactive digital extension services for smallholder farmers as a pathway to sustainable agriculture: a systematic review. Sustainability. 2025;17(7):3007. doi:10.3390/su17073007.
8. Osinga SA, Paudel D, Mouzakitis SA, Athanasiadis IN. Big data in agriculture: between opportunity and solution. Agric Syst. 2022;195:103298. doi:10.1016/j.agsy.2021.103298.
9. Shafik W, Tufail A, Namoun A, De Silva LC, Apong RAAHM. A systematic literature review on plant disease detection: motivations, classification techniques, datasets, challenges, and future trends. IEEE Access. 2023;11:59174–203. doi:10.1109/ACCESS.2023.3284760.
10. Wang H, Zhao R. Knowledge graph of agricultural engineering technology based on large language model. Displays. 2024;85:102820. doi:10.1016/j.displa.2024.102820.
11. Yang T, Mei Y, Xu L, Yu H, Chen Y. Application of question answering systems for intelligent agriculture production and sustainable management: a review. Resourc Conservat Recycl. 2024;204:107497. doi:10.1016/j.resconrec.2024.107497.
12. Vizniuk A, Diachenko G, Laktionov I, Siwocha A, Xiao M, Smoląg J. A comprehensive survey of retrieval-augmented large language models for decision making in agriculture: unsolved problems and research opportunities. J Artif Intell Soft Comput Res. 2025;15(2):115–46. doi:10.2478/jaiscr-2025-0007.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inform Process Syst. 2017;30:5998–6008.
14. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021.

15. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS '23). Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 10088–115.
16. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv:1701.06538. 2017.
17. Meng Q, Song Y, Mu J, Lv Y, Yang J, Xu L, et al. Electric power audit text classification with multi-grained pre-trained language model. IEEE Access. 2023;11:13510–8. doi:10.1109/ACCESS.2023.3240162.
18. Li F, Su P, Duan J, Xiao W. Towards better representations for multi-label text classification with multi-granularity information. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA, USA: ACL; 2023. p. 9470–80.
19. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns. 2025;6(6):101260. doi:10.1016/j.patter.2025.101260.
20. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: techniques and applications. arXiv:2402.07927. 2024.
21. Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Prompt engineering in large language models. In: International Conference on Data Intelligence and Cognitive Informatics. Cham, Switzerland: Springer; 2023. p. 387–402.
22. Arvidsson S, Axell J. Prompt engineering guidelines for LLMs in Requirements Engineering. arXiv:2311.17647. 2023.
23. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. 2024;7(1):41. doi:10.1038/s41746-024-01029-4.
24. Pan J, Zhong R, Xia F, Huang J, Zhu L, Yang Y, et al. ChatLeafDisease: a chain-of-thought prompting approach for crop disease classification using large language models. Plant Phenomics. 2025;7(3):100094. doi:10.1016/j.plaphe.2025.100094.
25. Velásquez-Henao JD, Franco-Cardona CJ, Cadavid-Higuita L. Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. Dyna. 2023;90(230):9–17. doi:10.15446/dyna.v90n230.111700.
26. Tzachor A, Devare M, Richards C, Pypers P, Ghosh A, Koo J, et al. Large language models and agricultural extension services. Nature Food. 2023;4(11):941–8. doi:10.1038/s43016-023-00867-x.
27. De Clercq D, Nehring E, Mayne H, Mahdi A. Large language models can help boost food production, but be mindful of their risks. Front Artif Intell. 2024;7:1–17. doi:10.3389/frai.2024.1326153.
28. Hou Y, Bishop JR, Liu H, Zhang R. Improving dietary supplement information retrieval: development of a retrieval-augmented generation system with large language models. J Med Internet Res. 2025;27:e67677.
29. Saha Roy R, Schlotthauer J, Hinze C, Foltyn A, Hahn L, Kuech F. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. New York, NY, USA: ACM; 2025. p. 1040–3. doi:10.1145/3701551.3704126.
30. Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. arXiv:250112948. 2025.
31. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2018. p. 784–9.
32. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2002. p. 311–8.
33. Lin CY. Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out. Stroudsburg, PA, USA: ACL; 2004. p. 74–81.
34. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv:1904.09675. 2020.

35. Glenn WB. Verification of forecasts expressed in terms of probability. Monthly Weather Rev. 1950;78(1):1–3.
36. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International Conference on Machine Learning. London, UK: PMLR; 2017. p. 1321–30.
37. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171. 2023.
38. Yan L, Wang H, Tang C, Liu H, Sun T, Liu L, et al. AgriEval: a comprehensive Chinese agricultural benchmark for large language models. arXiv:2507.21773. 2025.
39. Sarma D, Dasgupta M, Deka V, Das D, Rashed MG. Agricultural question answering dataset creation and its evaluation through BERT based experimental analysis. In: 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC). Piscataway, NJ, USA: IEEE; 2025. p. 1209–13. doi:10.1109/ISACC65211.2025.10969374.
40. KisanVaani AI. Griculture QA English Only; 2025. Contains around 22.6K QA pairs on crop and soil management. Hugging Face dataset. [cited 2025 Dec 10]. Available from: https://huggingface.co/datasets/KisanVaani/agriculture-qa-english-only.
41. Godara S, Bedi J, Parsad R, Singh D, Bana RS, Marwaha S. AgriResponse: a real-time agricultural query-response generation system for assisting nationwide farmers. IEEE Access. 2023;12:294–311. doi:10.1109/access.2023.3339253.
42. Singh N, Wang'ombe J, Okanga N, Zelenska T, Repishti J, Mishra S, et al. Chat: scaling AI-powered agricultural services for smallholder farmers. arXiv:2409.08916. 2024.
43. Zhao B, Jin W, Del Ser J, Yang G. ChatAgri: exploring potentials of ChatGPT on cross-linguistic agricultural text classification. Neurocomputing. 2023;557:126708. doi:10.1016/j.neucom.2023.126708.
44. Gauthier E, Ndiaye A, Guissé A. Kallaama: a transcribed speech dataset about agriculture in the three most widely spoken languages in senegal. In: Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024. Stroudsburg, PA, USA: ACL; 2024. p. 10–9.
45. Rezayi S, Liu Z, Wu Z, Dhakal C, Ge B, Zhen C, et al. AgriBERT: knowledge-infused agricultural language models for matching food and nutrition. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI-22). Peterborough, UK: International Joint Conferences on Artificial Intelligence Organization; 2022. p. 5150–6.
46. Huang Q, Tao Y, Ding S, Liu Y, Marinello F. An intelligent q&a module for tea diseases and pests based on automatic knowledge graph construction. In: 2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor). Piscataway, NJ, USA: IEEE; 2023. p. 66–70. doi:10.1109/MetroAgriFor58484.2023.10424245.
47. Ajawan P, Desai V, Kale S, Patil S. Krishiq-BERT: a few-shot setting BERT model to answer agricultural-related questions in the Kannada Language. J Instit Eng (India): Series B. 2024;105(2):285–96. doi:10.1007/s40031-023-00952-6.
48. Koopman B, Mourad A, Li H, Vegt A, Zhuang S, Gibson S, et al. AgAsk: an agent to help answer farmer's questions from scientific documents. Int J Digit Libraries. 2024;25(4):569–84. doi:10.1007/s00799-023-00369-y.
49. Chen T, Chen X, Qian Y, Zheng L, Li H, Zhao J, et al. AgriPrompt: a method to enhance ChatGPT for agricultural question answering. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Piscataway, NJ, USA: IEEE; 2024. p. 2436–41.
50. Zhang Y, Fan Q, Chen X, Li M, Zhao Z, Li F, et al. IPM-AgriGPT: a large language model for pest and disease management with a G-EA framework and agricultural contextual reasoning. Mathematics. 2025;13(4):566.
51. Kpodo J, Kordjamshidi P, Nejadhashemi AP. AgXQA: a benchmark for advanced Agricultural Extension question answering. Comput Electron Agric. 2024;225:109349. doi:10.1016/j.compag.2024.109349.
52. Yang S, Liu Z, Mayer W, Ding N, Wang Y, Huang Y, et al. ShizishanGPT: an agricultural large language model integrating tools and resources. In: International Conference on Web Information Systems Engineering (WISE 2024). Vol. 15439 of Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2024. p. 284–98.
53. Xiong J, Zhou Y, Tian F, Ni F, Zhao L. SheepDoctor: a knowledge graph enhanced large language model for sheep disease diagnosis. Smart Agric Technol. 2025;11:101001. doi:10.1016/j.atech.2025.101001.

54. Akbar NA. Are re-ranking in retrieval-augmented generation methods impactful for small agriculture qa datasets? a small experiment. In: BIO Web of Conferences. Vol. 167. Les Ulis, France: EDP Sciences; 2025. 01001 p. doi:10.1051/bioconf/202516701001.
55. Xiong J, Pan L, Liu Y, Zhu L, Zhang L, Tan S. Enhancing plant protection knowledge with large language models: a fine-tuned question-answering system using LoRA. Appl Sci. 2025;15(7):3850. doi:10.3390/app15073850.
56. Dofitas C, Kim YW, Byun YC. Advanced agricultural query resolution using ensemble-based large language models. IEEE Access. 2025;13:34732–46. doi:10.1109/ACCESS.2025.3541602.
57. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097–105.
58. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16 x 16 Words: transformers for Image Recognition at Scale. arXiv:2010.11929. 2021.
59. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2015.
60. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2016. p. 770–8. doi:10.1109/CVPR.2016.90.
61. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2015. p. 1–9. doi:10.1109/CVPR.2015.7298594.
62. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 9992–10002.
63. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 10347–57.
64. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 4015–26.
65. Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, et al. SAM 2: segment anything in images and videos. arXiv:2408.00714. 2024.
66. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 8748–63.
67. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 16000–9.
68. Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-training of image transformers. arXiv:2106.08254. 2022.
69. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. London, UK: PMLR; 2021. p. 4904–16.
70. Htun NN, Rojo D, Ooge J, De Croon R, Kasimati A, Verbert K. Developing visual-assisted decision support systems across diverse agricultural use cases. Agriculture. 2022;12(7):1027. doi:10.3390/agriculture12071027.
71. Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2022. p. 709–27.
72. Yu S, Tang C, Xu B, Cui J, Ran J, Yan Y, et al. VisRAG: vision-based retrieval-augmented generation on multi-modality documents. arXiv:2410.10594. 2025.
73. 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. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 874–80.

74. Izacard G, Lewis P, Lomeli M, Hosseini L, Petroni F, Schick T, et al. Atlas: few-shot learning with retrieval augmented language models. J Mach Learn Res. 2023;24(251):1–43.
75. Li J, Li D, Savarese S, Hoi S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. London, UK: PMLR; 2023. p. 19730–42.
76. Dai W, Li J, Li D, Tiong A, Zhao J, Wang W, et al. Instructblip: towards general-purpose vision-language models with instruction tuning. Adv Neural Inform Process Syst. 2023;36:49250–67.
77. Nazmus Sakib S, Haque N, Zabed Hossain M, Arman SE. PlantVillageVQA: a visual question answering dataset for benchmarking vision-language models in plant science. arXiv.2508.17117. 2025.
78. Liu X, Liu Z, Hu H, Chen Z, Wang K, Wang K, et al. A multimodal benchmark dataset and model for crop disease diagnosis. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2024. p. 157–70.
79. Chiu MT, Xu X, Wei Y, Huang Z, Schwing AG, Brunner R, et al. Agriculture-vision: a large aerial image database for agricultural pattern analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2020. p. 2828–38.
80. Stunning Vision AI. Coffee leaf diseases YOLO dataset [Internet]. 2023 [cited 2024 Aug 13]. Available from: https://www.kaggle.com/datasets/stunningvisionai/coffee-leaf-diseases-yolo.
81. Zhao K, Wu X, Xiao Y, Jiang S, Yu P, Wang Y, et al. PlanText: gradually masked guidance to align image phenotypes with trait descriptions for plant disease texts. Plant Phenomics. 2024;6:0272. doi:10.34133/plantphenomics.0272.
82. Lan Y, Guo Y, Chen Q, Lin S, Chen Y, Deng X. Visual question answering model for fruit tree disease decision-making based on multimodal deep learning. Front Plant Sci. 2023;13:1064399. doi:10.3389/fpls.2022.1064399.
83. Yu P, Lin B. A framework for agricultural intelligent analysis based on a visual language large model. Appl Sci. 2024;14(18):8350. doi:10.3390/app14188350.
84. Qian S, Liu B, Sun C, Xu Z, Ma L, Wang B. CroMIC-QA: the cross-modal information complementation based question answering. IEEE Trans Multimed. 2023;26:8348–59. doi:10.1109/TMM.2023.3326616.
85. Zhao Y, Wang S, Zeng Q, Ni W, Duan H, Xie N, et al. Informed-learning-guided visual question answering model of crop disease. Plant Phenomics. 2024;6:0277. doi:10.34133/plantphenomics.0277.
86. Zhu H, Shi W, Guo X, Lyu S, Yang R, Han Z. Potato disease detection and prevention using multimodal AI and large language model. Comput Electron Agric. 2025;229:109824. doi:10.1016/j.compag.2024.109824.
87. Xu Z, Xu J, Zhang M, Wang P, Deng C, Liu CL. Multimodal agricultural agent architecture (MA3): a new paradigm for intelligent agricultural decision-making. arXiv:2504.04789. 2025.
88. Zhang K, Ma L, Cui B, Li X, Zhang B, Xie N. Visual large language model for wheat disease diagnosis in the wild. Comput Electron Agric. 2024;227:109587. doi:10.1016/j.compag.2024.109587.
89. Kumar SS, Khan AKMA, Banday IA, Gada M, Shanbhag VV. Overcoming llm challenges using rag-driven precision in coffee leaf disease remediation. In: 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS). Piscataway, NJ, USA: IEEE; 2024. p. 1–6. doi:10.1109/ICETCS61022.2024.10543859.
90. Yuan Z, Liu K, Peng R, Li S, Leybourne D, Musa N, et al. PestGPT: leveraging large language models and IoT for timely and customized recommendation generation in sustainable pest management. IEEE Internet Things Mag. 2024;8(1):26–33. doi:10.1109/IOTM.001.2400036.
91. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv:2303.08774. 2023.
92. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inform Process Syst. 2023;36:34892–916.
93. Wu Z, Chen X, Pan Z, Liu X, Liu W, Dai D, et al. Deepseek-vl2: mixture-of-experts vision-language models for advanced multimodal understanding. arXiv:2412.10302. 2024.
94. Wang L, Jin T, Yang J, Leonardis A, Wang F, Zheng F. Agri-llava: knowledge-infused large multimodal assistant on agricultural pests and diseases. arXiv:2412.02158. 2024.
95. Awais M, Alharthi AHSA, Kumar A, Cholakkal H, Anwer RM. Agrogpt: efficient agricultural vision-language model with expert tuning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE; 2025. p. 5687–96.

96. Gauba A, Pi I, Man Y, Pang Z, Adve VS, Wang YX. AgMMU: a comprehensive agricultural multimodal understanding benchmark. arXiv:2504.10568. 2025.
97. Lu Y, Lu X, Zheng L, Sun M, Chen S, Chen B, et al. Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems. Plants. 2024;13(7):972. doi:10.3390/plants13070972.
98. Sengupta A. ROOT: reasoning over multimodal agricultural observation data with transformers for effective plant anomaly management. In: 2024 IEEE MIT Undergraduate Research Technology Conference (URTC). Piscataway, NJ, USA: IEEE; 2024. p. 1–5.
99. Wu H, Zhao C, Li J. Crop knowledge question answering system based on the multimodal fusion large model architecture Agri-QA Net. Smart Agric. 2025;7(1):1–10.
100. Jiang D, Shen Z, Zheng Q, Zhang T, Xiang W, Jin J. Farm-LightSeek: an edge-centric multimodal agricultural IoT data analytics framework with lightweight LLMs. IEEE Internet Things Mag. 2025;8(5):72–9. doi:10.1109/MIOT.2025.3575887.
101. Yang G, Li Y, He Y, Zhou Z, Ye L, Fang H, et al. Multimodal large language model for wheat breeding: a new exploration of smart breeding. ISPRS J Photogram Remote Sens. 2025;225:492–513. doi:10.1016/j.isprsjprs.2025.03.027.
102. Hue Y, Kim JH, Lee G, Choi B, Sim H, Jeon J, et al. Artificial intelligence plant doctor: plant disease diagnosis using Gpt4-vision. Res Plant Dis. 2024;30(1):99–102. doi:10.5423/RPD.2024.30.1.99.