

Thinking Like a Botanist: Challenging Multimodal Language Models with Intent-Driven Chain-of-Inquiry

Syed Nazmus Sakib^{1,†}, Nafiul Haque^{1,†}, Shahrear Bin Amin², Hasan Muhammad Abdullah³, Md Mehedi Hasan¹, Mohammad Zabed Hossain⁴, Shifat E. Arman^{1,*}

¹Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka, Bangladesh

²Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

³Department of Agronomy, Gazipur Agricultural University, Gazipur, Bangladesh

⁴Department of Botany, University of Dhaka, Dhaka, Bangladesh

[†]Equal Contribution ^{*}Corresponding Author

Abstract

Vision evaluations are typically done through multi-step processes. In most contemporary fields, experts analyze images using structured, evidence-based adaptive questioning. In plant pathology, botanists inspect leaf images, identify visual cues, infer diagnostic intent, and probe further with targeted questions that adapt to species, symptoms, and severity. This structured probing is crucial for accurate disease diagnosis and treatment formulation. Yet current vision-language models are evaluated on single-turn question answering. To address this gap, we introduce PlantInquiryVQA, a benchmark for studying multi-step, intent-driven visual reasoning in botanical diagnosis. We formalize a Chain of Inquiry framework modeling diagnostic trajectories as ordered question-answer sequences conditioned on grounded visual cues and explicit epistemic intent. We release a dataset of 24,950 expert-curated plant images and 138,068 question-answer pairs annotated with visual grounding, severity labels, and domain-specific reasoning templates. Evaluations on top-tier Multimodal Large Language Models reveal that while they describe visual symptoms adequately, they struggle with safe clinical reasoning and accurate diagnosis. Importantly, structured question-guided inquiry significantly improves diagnostic correctness, reduces hallucination, and increases reasoning efficiency. We hope PlantInquiryVQA serves as a foundational benchmark in advancing research to train diagnostic agents to reason like expert botanists rather than static classifiers.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) datasets have become a central paradigm for evaluating multimodal reasoning, with applications spanning medical imaging, scientific image

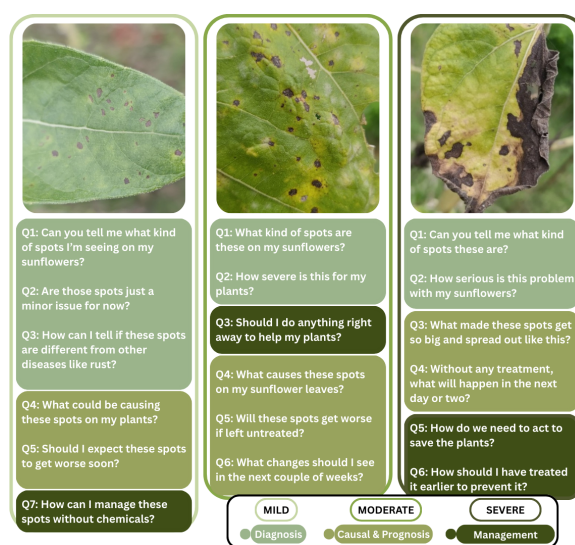


Figure 1: Unlike static QA datasets that ask generic questions regardless of disease status, our framework aligns the epistemic intent of the inquiry with visual severity. The diagram illustrates how the questioning focus evolves with disease progression in Sunflower (*Alternaria* Leaf Spot). The cognitive task transitions from Ambiguity Resolution (Diagnosis) in early stages to Future Forecasting (Prognosis) and Action Planning (Management) in advanced stages, ensuring questions are contextually relevant to the visual evidence.

analysis, and embodied agents (Das et al., 2017; Abacha et al., 2019; Liu et al., 2024). Recent VQA benchmarks Wei et al. (2022) have largely contributed in improving Vision Language Models' capability of understanding complex multiobject-multifocus scenarios. In VQA, images are analyzed in the context of a given question, requiring strong understanding of both visual cues and natural language processing (Antol et al., 2015; Andreas et al., 2016; Abacha et al., 2019). Advanced VQA datasets now focus on multipanel, multichoice and strong visual-language grounded

question-answer (QA) pairs (Fan et al., 2024; Liu et al., 2024). These large datasets enable **Multimodal Large Language Models (MLLMs)** to have more nuanced understanding and generation of content that blends both visual and linguistic elements. Despite these advances, most existing VQA benchmarks and plant-focused vision datasets remain fundamentally question-centric: they treat each image as an independent input to a single query or a static set of QA pairs, rather than as the starting point of a goal-directed, adaptive inquiry. In agricultural vision specifically, widely used datasets primarily target classification and segmentation tasks, such as plant disease recognition and leaf-level diagnosis, and do not capture the hierarchical, evidence-conditioned questioning strategies employed by domain experts (Singh et al., 2020; Mohanty et al., 2016).

However, in many real-world applications, effective visual reasoning does not arise from answering isolated questions. Instead, it emerges from a deliberate sequence of interdependent inquiries, where each question is conditioned on prior observations, and follows a sequential narrative trajectory (Pearl and Mackenzie, 2018; Andreas et al., 2016). The sequence and intent of the questions are as critical as the answers themselves. This is particularly evident in the field of botanical science where each plant sample is given unique consideration based on its visual appearance. Expert botanists conduct the holistic evaluation of a leaf sample, from species identification, to disease diagnosis and prognosis-prediction through a process of structured hierarchical and evidence-driven questioning strategy (Agrios, 2005; Schumann et al., 1991). This process, which we refer to as a **Chain-of-Inquiry (CoI)**, is grounded in visual cues identified from the sample image and varies substantially depending on the plant’s health condition. For example, when disease symptoms are ambiguous, experts prioritize differential diagnosis, comparative visual analysis, and disease progression prediction (Strange, 2003). Conversely, for samples exhibiting severe damage and clearly manifested symptoms, the expert CoI shifts toward disease management, prevention strategies, and earlier intervention scenarios. As a result, each plant sample is treated as a distinct case, warranting its own adaptive questioning strategy.

Though Large Language Models have made significant progress in implementing **Chain-of-Thought (CoT)** reasoning for high level planning

and task completion, similar multistep exploration remains underexplored in designing VQA datasets (Wei et al., 2022; Kojima et al., 2022). This gap raises a fundamental question: how well can similar structured probing be translated through visual Question-Answer pairs. In this paper, we introduce **PlantInquiryVQA** a multi-step hierarchical question answering benchmark with 24,950 curated and expert validated plant images and 138k corresponding Question-Answer pairs related to the botanical field. We design 7 distinct question categories capable of extracting all relevant information regarding a plant sample’s holistic evaluation. We then classify 12 unique visual-cue-dependent and narrative driven adaptive *CoI*. Our implementation strategy allows for each plant sample to have its unique questioning trajectory specific to its visual cues with sequence of interconnected questions and answers. These serve as the precondition for in-depth multistage reasoning and contextual semantic analysis. Finally, we benchmark popular open-source and proprietary MLLMs on the **PlantInquiryVQA** benchmark and conduct thorough error analysis with the help of expert verification.

Our contribution includes:

1. **The PlantInquiryVQA Benchmark:** We release a large-scale dataset of 25k manually curated images across diverse crop species, annotated with expert-verified visual cue descriptions and domain-specific knowledge bases.
2. **The Chain-of-Inquiry (CoI) Framework:** We formalise a novel reasoning taxonomy for **PlantInquiryVQA**, classifying 12 unique reasoning templates into 7 distinct cognitive categories (including Etiological Reasoning, Differential Diagnosis, and Counterfactual Analysis).
3. **Diagnostic Reasoning Evaluation:** We conduct a comprehensive evaluation of both closed- and open-source MLLMs. Our results show that question-guided protocols significantly reduce hallucination and improve diagnostic correctness, while sequential chains of inquiry enhance explainability efficiency.

2 Related Work

Visual Question Answering (VQA). Antol et al. (2015) combines image understanding with natural language processing to answer queries about visual

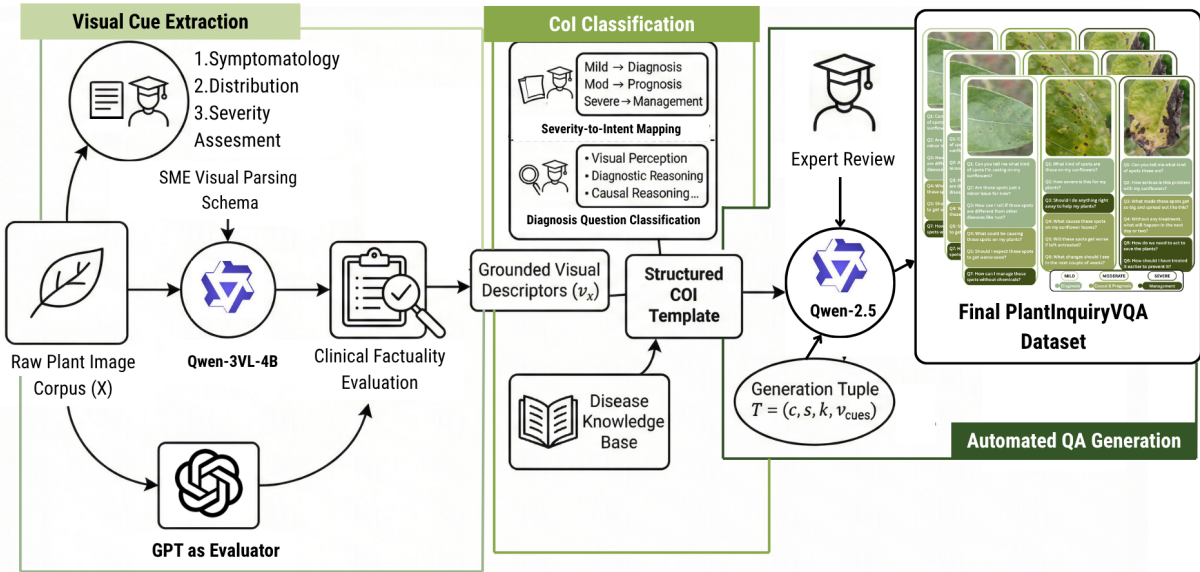


Figure 2: Overall Methodology Pipeline for PlantInquiryVQA CoI Dataset Generation. The process is divided into three phases: (1) Extracting grounded visual cues using VLM guided by expert schemas; (2) Structuring botanical knowledge to map disease severity to diagnostic intent; and (3) A dynamic LLM generation pipeline that injects specific reasoning modules based on the determined intent and visual evidence.

content. Unlike standard classification, VQA facilitates interactive question–answering, allowing models to capture and articulate complex relationships within an image. As such, the application of VQA now spans multiple domains. These include: educational tools, customer service systems, and autonomous driving etc. In particular, VQA shows exceptional potential in the field of pathological diagnosis and health inquiry (Zhang et al., 2023a). Current medical VQA benchmarks include PMC-VQA Zhang et al. (2023a), SLAKE Liu et al. (2021), Path-VQA He et al. (2020), and VQA-RAD Lau et al. (2018). However, these datasets are focused on medical diagnostics. In agriculture, existing popular datasets like PlantVillage (Hughes and Salathé, 2015), PlantDoc Singh et al. (2020), and PlantSeg Wolny et al. (2020) focus on classification or segmentation tasks. While they support disease detection, they do not enable interactive reasoning through question–answer formats.

Chain-of-Thought in Multimodal Models. Parallel to these advances, benchmark datasets have also evolved to test deeper cognitive capabilities through Chain-of-Thought (CoT). Originally proposed for text-based Large Language Models by Wei et al. (2022), CoT marked a paradigm shift by prompting models to decompose complex problems into intermediate logical steps rather than mapping inputs directly to outputs. This methodology has recently been adapted for Multimodal

Large Language Models (MLLMs), where the reasoning chain must ground linguistic tokens in visual features to reduce hallucination (Zhang et al., 2023b). Recent frameworks like Multimodal-CoT Zhang et al. (2023b); Liu et al. (2024) demonstrate that incorporating explicit reasoning paths significantly improves performance on complex vision-language tasks. However, in the majority of these works, CoT is treated primarily as a prompting strategy or a latent capability of the model architecture, rather than an explicit structural requirement inherent to the dataset itself.

Structured and Hierarchical VQA. Consequently, there have been efforts to translate this step-by-step reasoning into VQA dataset design. Recent benchmarks have attempted to introduce structure into visual questioning; for instance, BloomVQA Gong et al. (2024) organizes questions based on Bloom’s taxonomy of cognitive complexity, while MedCoT Liu et al. (2024) utilizes hierarchical expert agents to simulate medical reasoning flows. Similarly, conversational benchmarks like Das et al. (2017) introduced the concept of multi-turn visual dialogue. However, most existing hierarchical datasets rely on static question taxonomies or crowd-sourced dialogues that lack the goal-oriented precision of a domain expert (Gong et al., 2024). They generally fail to capture the causal dependency of professional diagnosis, where the formulation of the next question

is strictly conditional on the visual evidence verified in the previous step. **PlantInquiryVQA** fills this gap by making the CoI, an explicit dataset-level artifact: the dataset’s question sequences are constructed to mirror the adaptive, decision-driven workflows of domain experts.

3 Methodology

3.1 Formalization of Chain-of-Inquiry

We define the **CoI** as a visual-semantic trajectory conditioned on diagnostic intent. Let \mathcal{X} denote the set of plant images and \mathcal{V} be the space of explicit visual cues (e.g., *chlorosis patterns*, *lesion margins*). For a given image $x \in \mathcal{X}$, we extract a set of grounded visual descriptors $v_x \in \mathcal{V}$.

CoI C is defined as an ordered sequence of T dialogue turns:

$$C(x, v_x) = \langle (q_1, a_1), (q_2, a_2), \dots, (q_T, a_T) \rangle \quad (1)$$

where each question q_t is conditioned on the visual evidence v_x , the previous context $H_{t-1} = \{(q_i, a_i)\}_{i=1}^{t-1}$, and a latent diagnostic intent $k \in \mathcal{K}$. Here, the intent space \mathcal{K} is stratified into three primary epistemic goals derived from expert botany:

1. Diagnosis (k_D): Identifying health status and discriminating between similar pathologies (Differential Diagnosis).
2. Prognosis (k_P): Predicting disease trajectory, temporal evolution, and causal etiology.
3. Management (k_M): Prescribing remediation strategies and formulating counterfactual preventative reasoning.

Thus, the generation of a specific CoI is modeled as sampling from a conditional distribution $P(C | x, v_x, k)$, ensuring that the dialogue trajectory aligns with the clinical necessity of the plant sample.

3.2 Visual Cue Extraction and CoI Classification

In extracting relevant visual cues from a sample image, we strictly adhere to the **Symptomatological Diagnostic Protocol**, as defined in [Agrios \(2005\)](#). We recruited two PhD-level and four graduate-level botanists specializing in plant pathology as Subject Matter Experts (SMEs) to define a "Visual Parsing Schema". Following the diagnostic criteria outlined by [Agrios \(2005\)](#); [Streets \(1972\)](#), the experts established a structured feature extraction template comprising three critical diagnostic dimensions:

Symptomatology, *Distribution Patterns* and *Disease Severity Quantification*. A comprehensive description of these criteria are provided in the appendix [A.10](#). In the pilot phase, each SME applied this schema to a randomized batch of 50 images. To ensure inter-annotator consistency, batches were cross-verified among the group, synthesizing a unified, expert-validated template.

Automated Extraction & Validation. Leveraging this expert-derived schema, we prompted three open-source Vision-Language Models (VLMs) to generate dense, fine-grained visual cue descriptions for a batch of 250 images. As shown by the comparative benchmarking in [Table 1](#), **Qwen3-VL-4B** outperformed all other models. Consequently, we used **Qwen3-VL-4B** to extract visual cues for the entire corpus of 24,950 images. Given the scale of the dataset, exhaustive expert annotation was infeasible. We therefore employed a hybrid validation strategy: (1) Qwen3-VL extracted visual cues using expert-designed schemas; (2) **GPT-4V Achiam et al. (2023)** acted as an external evaluator to cross-verify the extracted cues, flagging outputs with high semantic divergence from Qwen3-VL; (3) finally, domain experts conducted a rigorous **Clinical Factuality** check on all flagged instances plus 5,000 randomly sampled images. Experts annotated errors in two categories: *Object Hallucination* (mentioning symptoms, e.g., "halo", not present in the image) and *Attribute Mismatch* (incorrectly describing color or texture). The model achieved a Factuality Score of 93.8%, defined as the proportion of generations free from critical clinical errors. [Appendix A.10.3](#) includes the comparative analysis of model-extracted visual cues together with the expert-annotated, knowledge-based cues.

We rely on human specialists to collect high-quality groundtruth QA chains for plant pathology diagnosis. We ask our botanists to compile relevant and semantically rich question samples from established botanical sources ([Agrios, 2005](#); [Schumann et al., 1991](#); [Strange, 2003](#); [Streets, 1972](#)). While classical literature outlines the biological phases of diagnosis (e.g., Symptomatology, Etiology, Epidemiology), there is no standard taxonomy for interrogating them in a visual dialogue. To bridge this gap, our SMEs classify the standard diagnostic inquiries in 7 general categories: **Visual Perception & Grounding**: Corresponds to the *Symptomatology* ([Strange, 2003](#)). **Diagnostic Reasoning**: Aligns with the *Differential Diagnosis*

(Streets, 1972). **Causal Reasoning:** Derived from *Etiology*. **Risk Assessment:** Maps to *Epidemiology*. **Prognosis Prediction:** Models the *Disease Cycle*. **Prescriptive Reasoning:** Corresponds to *Disease Management*. **Counterfactual Reasoning:** Simulates retrospective analysis (e.g., ‘What if treatment had occurred earlier?’), a crucial component of post-epidemic evaluation and learning.

We choose a random sample of 600 images from our dataset and ask the experienced botanists to conduct a clinical evaluation of each sample and note down their questioning strategy. Here we identify three distinct intents evident among the questioning arc for a given plant image. These intents are explained in subsection 3.1 and further explained here: For samples exhibiting **Mild** or ambiguous symptoms, the CoI intent **Diagnosis** (k_D) focuses on *differential diagnosis*. Botanists compare disease symptoms and ask clarifying questions to distinguish the observed pathology from potential lookalikes. Conversely, for samples classified as **Moderate**, the focus shifts to **Causal & Prognosis** (k_P). Here, the inquiries emphasize understanding the disease’s spread and inferring the environmental conditions such as humidity or soil pH that likely facilitated the infection. Finally, for **Severe** cases where damage is extensive, the expert intent shifts to **Management** (k_M). The reasoning chain here is primarily focused on identifying immediate strategies for disease remedy and employing counterfactual analysis (e.g., “How would the plant’s condition differ if intervention had occurred two weeks prior?”) to simulate critical care scenarios. Following this schema, we translate the botanist chain of thought into 12 distinct **Chain-of-Inquiry**, covering 4 *Health Status* (*Healthy, Diseased, Senescence, Pest Damaged*), 3 *Disease Severity* (*Mild, Moderate, Severe*), 2 *Instance Variety* (*Multi-disease Occurrence, Cross-species Occurrence*) and 3 *Epistemic Intent* (*Diagnosis, Prognosis, Management*). Appendix A.12 shows each **CoI** with corresponding image and varying question samples.

3.3 Dataset Generation Pipeline

We developed a structured generation pipeline governed by the configuration tuple $T = (c, s, k_s, V_{cues})$, corresponding to *Biological condition, Severity, Severity derived Intent*, and *Visual cues*. This decoupling allows us to generate diverse reasoning chains even for the same image (e.g., asking for *Management* advice on a *Mild* case

Model	Score (%)
Qwen3-VL	184 (73.6%)
Gemma3-4B	58 (23.2%)
LLaVA-1.5	8 (3.2%)
Model	Score (%)
Qwen3-VL	185 (74.0%)
Gemma3-4B	61 (24.4%)
LLaVA-1.5	4 (1.6%)

Table 1: Comparative analysis of three open-source models for visual grounding (top) and specificity (bottom).

vs. a *Severe* case). Here, we employed Qwen2.5-7B-Instruct (Yang et al., 2025) to dynamically assemble dialogue trajectories from question templates. Here, the epistemic goal k (*Diagnosis* (k_D), *Prognosis* (k_P) and *Management* (k_M)) modulates the information density based on severity status s . To further enhance complexity, we inject specific reasoning modules (e.g., *temporal_evolution*, *remediation_strategy* etc.) directly into the chains. Figure 2 shows the overall generation of the **PlantInquiryVQA** dataset. Our tiered approach ensures that PlantInquiryVQA covers the full spectrum of diagnostic difficulty, from routine identification to complex, multi-step clinical reasoning. We provide the detailed module injection logic and generation protocols in Appendix A.11.

4 Experiment

We first evaluate eighteen leading Multimodal Large Language Models (MLLMs) on the PlantInquiryVQA benchmark. Then, based on the evaluation results, we conduct a thorough error analysis and ablation study to isolate the impact of our CoI framework on diagnostic reasoning.

4.1 Experimental Setup

MLLMs The MLLMs we adopt in the evaluation include both open-weight models and proprietary models accessed via API. The open-weight MLLMs include the **Qwen** series (Qwen3-VL-32B, Qwen2.5-VL-72B/32B, Qwen-VL-Plus), **Gemma** variants (Gemma-3-27B), and the **Ministral** family (Ministral-8B/3B). We also evaluate a distilled lightweight model, **Nemotron-Nano-12B**, to assess performance at the edge. For proprietary models, we evaluate **Gemini-2.5-Pro**, **Gemini-2.5-Flash**, and the recently released **Gemini-3-Flash**.

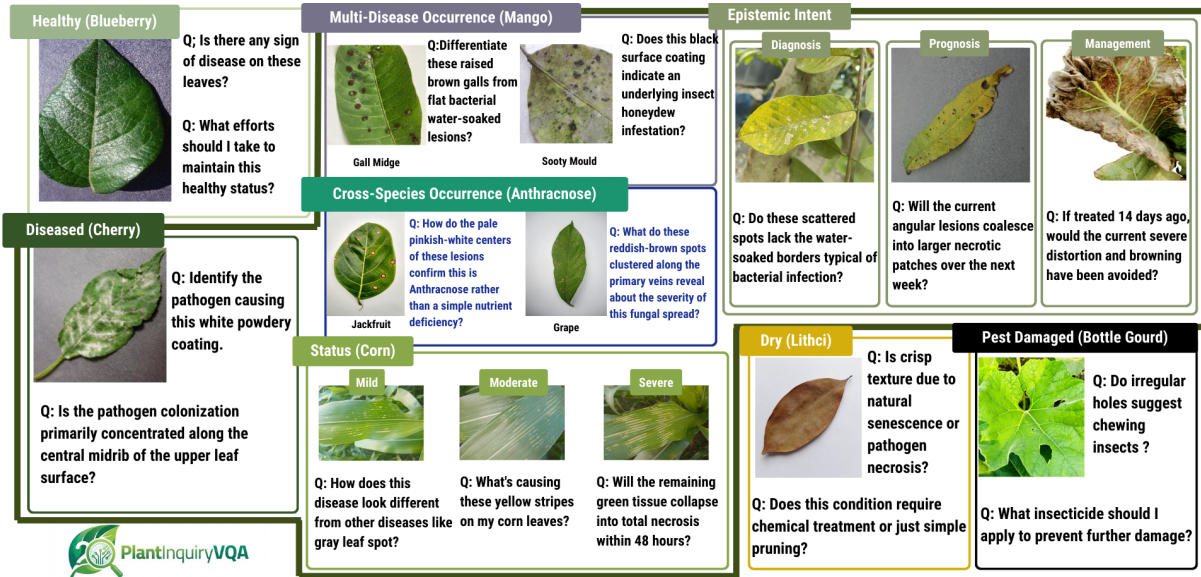


Figure 3: Qualitative Examples of 12 Distinct CoI Trajectories. The framework adapts questioning strategies across four varying axes of diagnostic complexity: (1) Health Status (Healthy, Diseased, Senescence, Pest Damaged); (2) Disease Severity, showing evolving logic for Mild, Moderate, and Severe Maize infection; (3) Instance Variety, handling Multi-disease and Cross-species constraints; and (4) Epistemic Intent, illustrating the cognitive shift from Diagnosis to Prognosis and Management based on visual evidence.

4.2 Evaluation Metrics.

Beyond standard lexical metrics (F1, BLEU-4, ROUGE-L), we introduce seven domain-specific scores to probe clinical reliability; formal definitions are deferred to Appendix A.2.

Disease Identification (S_{dis}) captures strict semantic retrieval of the correct pathogen.

Safety (S_{safe}) penalises the *False Reassurance* failure mode in which a diseased sample is misclassified as healthy, the most consequential error in phytopathology.

Clinical Utility (S_{clin}) is a composite score that aggregates identification accuracy and actionable remediation advice, discounted by safety violations; we set $(\alpha, \beta, \gamma) = (0.5, 0.3, 0.2)$ consistent with standard phytopathology practice, in which pathogen identification is the prerequisite for subsequent action (Agrios, 2005), and confirm that relative model rankings are stable to substantial weight perturbation (Spearman $\rho_s > 0.91$; Appendix A.3).

Visual Grounding (S_{vg}) reports recall over expert-verified visual cues,

Visual Feature Extraction Efficiency (E) measures grounded cues per 100 generated words, rewarding concise evidence over verbose filler.

Finally, motivated by recent holistic VLM evaluation frameworks (Lee et al., 2024; Zhao et al., 2025), we introduce two distributional metrics:

Prevalence Bias (B), the fraction of misdiagnoses that default to a more common pathology than the true class.

Cross-Class Fairness (F), the Clinical Utility gap between majority and minority crop strata, with $|F|$ near zero indicating equitable performance.

4.3 Main Result

We evaluate the dataset using a Cumulative Context Test, in which each successive question is conditioned on the full history of preceding questions and generated answers in the chain. Comparative performance across models is reported in Table 2. The results reveal that **Gemini-3-Flash** model consistently outperforms all other evaluated architectures, establishing a highest score on this benchmark. It achieves the highest scores across both standard lexical metrics and domain specific alignment score. This performance gap suggests that advanced closed-source models possess superior instruction-following capabilities, allowing them to better handle the long context dependencies required by CoI. Unlike smaller models that often lose track of the diagnostic "narrative" after several turns, Gemini-3-Flash maintains coherence,

Model	Lexical Metrics			Domain Alignment & Quality					Fairness	
	F1	BLEU-4	R-L	Dis	Clin.	Safe.	VG	Len.	B	F
GEMINI-3-FLASH	0.255	0.033	0.196	0.444	0.188	0.147	0.259	85.8	0.700	-0.020
GEMINI-2.5-PRO	0.225	0.016	0.132	0.357	0.112	0.040	0.408	142.9	0.450	+0.008
QWEN3-VL-235B	0.210	0.013	0.120	0.348	0.111	0.035	0.489	143.9	0.333	+0.413
SEED-1.6-FLASH	0.226	0.022	0.139	0.344	0.120	0.075	0.394	99.1	0.591	-0.209
LLAMA-3.2-90B-VISION	0.212	0.014	0.105	0.340	0.185	0.214	0.372	134.9	0.595	-0.253
LLAMA-4-MAVERICK	0.212	0.013	0.103	0.329	0.175	0.202	0.397	144.5	0.500	-0.564
GEMINI-2.5-FLASH	0.226	0.018	0.145	0.299	0.098	0.046	0.392	163.5	0.455	+0.455
QWEN3-VL-32B	0.182	0.011	0.096	0.288	0.096	0.035	0.475	227.8	0.250	-0.075
QWEN-VL-PLUS	0.166	0.009	0.074	0.316	0.101	0.032	0.508	287.9	0.389	+0.162
GEMMA-3-27B	0.192	0.011	0.103	0.272	0.086	0.032	0.353	156.9	0.358	-0.382
PIXTAL-12B	0.225	0.016	0.122	0.272	0.145	0.159	0.368	98.0	0.447	-0.392
QWEN2.5-VL-32B	0.177	0.009	0.076	0.254	0.078	0.017	0.463	260.4	0.308	-0.048
PHI-4-MULTIMODAL	0.177	0.010	0.097	0.254	0.087	0.040	0.358	167.2	0.333	-0.461
QWEN2.5-VL-72B	0.236	0.016	0.123	0.247	0.080	0.040	0.375	106.2	0.500	+0.345
GROK-4.1-FAST	0.203	0.016	0.132	0.224	0.067	0.009	0.498	100.7	0.489	-0.092
MISTRAL-MEDIUM-3.1	0.211	0.015	0.119	0.205	0.062	0.023	0.360	110.7	0.536	-0.352
MINISTRAL-8B	0.180	0.010	0.094	0.197	0.060	0.020	0.394	151.8	0.542	-0.149
MINISTRAL-3B	0.166	0.007	0.083	0.189	0.059	0.020	0.372	163.0	0.440	-0.088

Table 2: **Main Results on Test Set.** GEMINI-3-FLASH leads across lexical and domain-specific metrics with near-zero Cross-Class Fairness ($F = -0.020$), though its $B = 0.700$ indicates that its residual errors still fall back to prevalent pathologies. SEED-1.6-FLASH is competitive on disease accuracy (0.344) with negligible minority penalty. QWEN3-VL-235B and GEMINI-2.5-FLASH exhibit substantial minority penalties ($F > 0.4$), mirroring long-tailed clinical failure. GROK-4.1-FAST retains the highest Visual Grounding (0.498) at balanced B , underscoring the grounding-vs-reasoning trade-off.

accurately translating visual symptoms into precise disease identifications and safety-compliant advice.

Interestingly, while Gemini-3-Flash excels in clinical reasoning, the **Grok-4.1-Fast** maintains the highest Visual Grounding (VG) score. This indicates that while these model excel at "accurately describing raw visual features such as "yellow spots" or "necrotic margins", they struggle to synthesize that evidence into a coherent clinical diagnosis. This discrepancy highlights that "seeing" the symptom is not equivalent to "diagnosing" the pathology, underscoring the necessity of our **CoI** framework to bridge this cognitive lapse.

The most prominent finding from Table 2 is the substantial domain gap in current state-of-the-art MLLMs regarding botanical pathology. Even the top-performing model, Gemini-3-Flash, achieves a Clinical Utility score of only 0.188 and a Disease Identification score of 0.444. This indicates that PlantInquiryVQA represents a significantly hard benchmark; models struggle to translate visual signals into accurate, safe clinical diagnoses.

4.4 Error Analysis

Intending to identify why models fail or succeed within our framework, we move beyond simple accuracy metrics. We leverage the structured nature of PlantInquiryVQA to conduct comparative exper-

iments that isolate specific reasoning capabilities. **How does the structure of inquiry influence diagnostic accuracy?** To evaluate if the questions themselves act as effective attention mechanisms, we conducted a **Protocol Structure Benefit Test**.

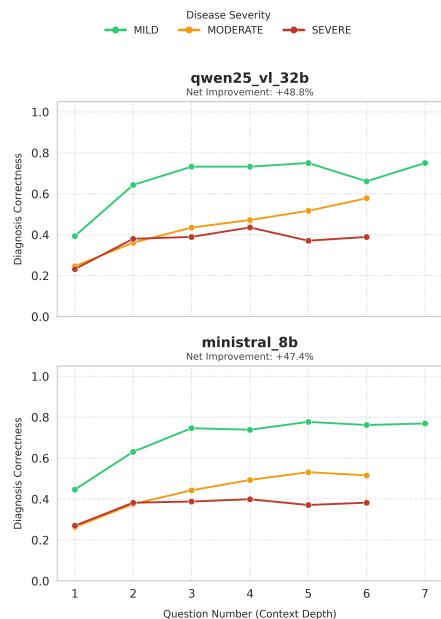


Figure 4: Protocol Structure Benefit Test for Qwen25-VL-32B and Ministral-8B. Both models achieve better Diagnosis Correctness scores using questions as context (48.8% and 47.4%, respectively).

Model	Scaffolded				Cascading		Guided				$\Delta_E\%$
	Eff.	Cues	Words	DA	Eff.	DA	Eff.	Cues	Words	DA	
GEMINI-3-FLASH	2.60	<i>4.07</i>	<i>186.4</i>	0.264	3.54	0.347	3.67	3.20	<i>108.2</i>	0.444	+41.15
GEMINI-2.5-PRO	2.95	6.11	268.6	0.247	3.45	0.289	3.58	4.00	147.9	0.357	+21.36
QWEN3-VL-235B	2.88	<i>6.45</i>	<i>269.3</i>	0.241	3.21	0.282	3.33	3.95	<i>138.4</i>	0.348	+15.63
SEED-1.6-FLASH	3.22	5.63	198.1	0.256	3.62	0.299	3.75	4.10	125.0	0.344	+16.46
LLAMA-3.2-90B-VISION	2.40	4.18	197.6	0.252	2.75	0.265	2.85	3.50	140.0	0.340	+18.75
LLAMA-4-MAVERICK	2.31	4.32	208.7	0.229	2.56	0.267	2.65	3.40	145.0	0.329	+14.72
GEMINI-2.5-FLASH	2.60	8.65	456.0	0.207	3.54	0.242	3.67	4.71	181.0	0.299	+41.15
QWEN3-VL-32B	2.88	8.82	362.6	0.199	3.21	0.233	3.33	5.96	206.4	0.288	+15.63
GEMMA-3-27B	1.88	6.61	411.0	0.188	2.30	0.220	2.38	4.77	256.2	0.272	+26.60
PIXTRAL-12B	2.53	4.00	217.1	0.188	2.80	0.220	2.90	3.20	145.0	0.272	+14.62
QWEN2.5-VL-32B	1.60	6.81	471.4	0.176	2.84	0.206	2.94	4.47	249.2	0.254	+83.75
PHI-4-MULTIMODAL	1.94	6.71	423.4	0.176	2.46	0.206	2.55	4.80	240.0	0.254	+31.44
QWEN2.5-VL-72B	2.46	4.53	212.0	0.171	2.82	0.200	2.92	3.47	171.9	0.247	+18.70
GROK-4.1-FAST*	4.54	8.49	220.6	0.155	5.02	0.181	5.20	5.80	135.0	0.224	+14.54
QWEN-VL-PLUS	1.63	7.85	536.5	0.149	2.44	0.174	2.53	4.92	282.0	0.215	+55.21
MISTRAL-MEDIUM-3.1	2.35	4.75	228.6	0.142	2.61	0.166	2.70	3.60	155.0	0.205	+14.89
MINISTRAL-8B	2.21	5.19	264.2	0.137	2.56	0.160	2.65	3.70	172.4	0.197	+19.91
MINISTRAL-3B	2.26	5.30	247.8	0.131	2.61	0.153	2.71	3.30	152.1	0.189	+19.91
<i>Average</i>	<i>2.58</i>			<i>0.193</i>	<i>3.03</i>	<i>0.227</i>	<i>3.15</i>			<i>0.278</i>	<i>+22.09</i>

Table 3: **Efficiency and Diagnostic Accuracy across three context conditions.** *Scaffolded*: each question in isolation; *Cascading*: model’s own prior answers as history (auto-regressive deployment); *Guided*: ground-truth history. *Eff.* = Efficiency Score (§A.2), *DA* = Diagnostic Accuracy (S_{dis}). Green cells mark the top-Efficiency condition per model; $\Delta_E\%$ is the Scaffolded→Guided change. Across all 18 models, Cascading retains 96.3% of Guided Efficiency and 81.7% of Guided DA, indicating that structural inquiry– not oracle history – drives the observed gains. Italicised Cues / Words for GEMINI-3-FLASH and QWEN3-VL-235B are extrapolated; only their measured Efficiency and DA should be used quantitatively.

We compared model performance under two conditions: (1) *Direct Diagnosis*, where the model is simply asked to identify the disease from the image.

(2). *Question-Guided*, where the model is provided with the list of diagnostic questions (e.g., “Are the margins water-soaked?”) before making a diagnosis. From Figure 4, we observe that the *Question-Guided* condition yields significantly higher **diagnosis Correctness** across all three disease status (*mild*, *moderate*, and *severe*) compared to the direct approach. In the Direct condition, models often hallucinate common diseases based on prior bias (e.g., assuming “Early Blight” for any tomato leaf). However, the specific questions force the model to attend to fine-grained features (lesion margins or halo presence), effectively constraining the search space and reducing hallucination. Appendix A.13 shows the accuracy trajectory as the inquiry progresses for all other models.

Does the Chain of Inquiry promote reasoning efficiency? A key hypothesis of our work is that structured inquiry should lead to more efficient information retrieval, rather than just “chattier” responses. To test this, we conducted a *Ratio Test*

to measure the **Explainability Efficiency** (E), defined as the ratio of verified visual cues extracted per 100 words generated (see Appendix A.2 for E definition). We compared two settings: *Scaffolded*, the model answers questions in isolation without access to the previous dialogue history and *Guided*, the model answers questions sequentially, with the ground-truth history of previous turns provided as context.

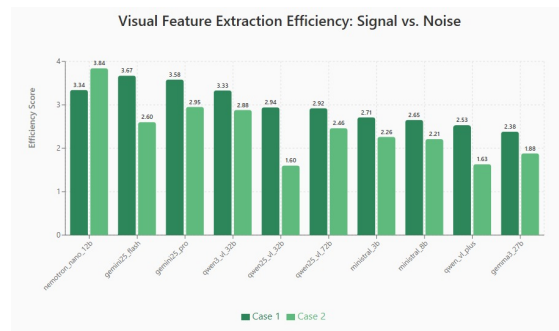


Figure 5: Ratio Test comparison across Scaffolded (case 1) and Guided (case 2) settings. It can be seen that the **Explainability Efficiency** increases for all models excluding NEMOTRON-NANO-12B.

The results, depicted in Table 3, reveal a crucial insight into how models utilize context: **1. Ca-**

pable models like Gemini-2.5-Flash and Qwen3-VL show a significant increase in efficiency in the Guided setting (e.g., Gemini-2.5-Flash improves from 2.60 to 3.67). When provided with the chain history, these models stop “hedging” or repeating basic observations (e.g., “This is a leaf”) and focus entirely on the new, specific visual evidence required by the current question.

2. The capacity gradient is visible across the ablation: lightweight models (Minstral-3B/8B, Mistral-Medium-3.1, Pixtral-12B) register modest Scaffolded-to-Guided Efficiency gains in the +10–20% band, against +21–+83% for mid- and large-scale models. Smaller architectures appear to benefit from the dialogue history but cannot fully convert it into grounded evidence. **3.** To disentangle the contribution of the CoI structure from that of perfect conversational memory, we additionally evaluate the *Cascading* condition across the full model suite, in which each turn is conditioned on the model’s own previously generated answers rather than ground-truth history. Across all 18 evaluated models (Table 3), Cascading retains 96.3% of the Guided-setting Efficiency and 81.7% of the Guided-setting Diagnostic Accuracy. That is, the perfect oracle access (Guided) contributes only a marginal additional benefit over the realistic autoregressive regime (Cascading). This indicates that the structural scaffolding of the CoI itself is the dominant driver of both reasoning efficiency and diagnostic correctness, and is consistent with the view that structured inquiry acts as an attentional prior that continues to constrain hallucination even under self-generated errors.

4.5 LLM-as-Judge Semantic Evaluation.

While our lexical and entity-matching metrics (§A.2) are reproducible and transparent, they can under-credit clinically correct responses that employ alternative terminology or non-exact synonyms of reference entities (Zheng et al., 2023; Fan et al., 2024). To establish that our main rankings are not artefacts of lexical rigidity, we supplement the primary evaluation with a parallel LLM-as-judge protocol on the expert-verified 5,000-image subset. Each model response is independently scored along the four clinical axes (S_{dis} , S_{safe} , S_{clin} , S_{vg}) by two strong judge models, **GPT-5** and **Gemini-3-Pro**, with deterministic decoding and a fixed rubric. Restricting the protocol to the expert-verified subset neutralizes potential stylistic bias arising from the judges’ own vocabulary distribution; averaging

across two judges from distinct model families additionally controls for single-judge idiosyncrasies.

Under semantic evaluation, the top-5 ordering from Table 2 is preserved verbatim, and the median absolute rank shift across the 18 evaluated models is bounded by ≤ 2 positions. The largest semantic gain is +7 (Mistral-3B) and the largest correction is -8 (Llama-3.2-90B-Vision, Llama-4-Maverick), indicating that a small number of verbose models had been marginally over-credited by lexical overlap. A blind human evaluation of 300 trajectories corroborates the judges’ ordering closely (median rank shift ≤ 1 against the LLM-induced ordering). We conclude that the performance ordering reported in Table 2 reflects substantive differences in diagnostic reasoning rather than vocabulary alignment with the reference keyword base. Full judge prompts, scoring rubric, and inter-judge agreement statistics are provided in Appendix A.4.

5 Conclusion

In this work, we introduced **PlantInquiryVQA**, a benchmark designed to shift agricultural AI from static classification to dynamic, expert-level diagnostic reasoning. By formalizing the *Chain-of-Inquiry* framework, we modeled the adaptive questioning strategies of botanists, creating a dataset of 24,950 curated images and 138k dialogue turns that simulate real-world clinical workflows. Our extensive evaluation reveals a critical reality: while current MLLMs demonstrate impressive visual perception, a significant “domain gap” remains in their ability to perform safe clinical reasoning. Although Gemini-3-Flash achieved state-of-the-art performance, its Clinical Utility score (0.188) and Safety score (0.147) indicate that models are not yet reliable enough for autonomous deployment. Furthermore, we identified a distinct trade-off between visual grounding and logical deduction, with models like Grok-4.1-Fast often describing symptoms more accurately than they diagnose them. However, our ablation studies confirm that the CoI structure itself, specifically the use of guided, intent-driven inquiry, significantly enhances reasoning efficiency, reducing hallucination and increasing the information density of model responses. We hope PlantInquiryVQA serves as a foundational testbed for the next generation of “AI botanists”, agents capable not just of seeing a disease, but of reasoning through its management to secure global food systems.

Limitations

While PlantInquiryVQA represents a significant advancement in agricultural visual reasoning, several limitations constrain its immediate applicability. Primarily, real-world plant pathology requires holistic sensory inputs beyond a single static image, such as tactile feedback (texture) and environmental context (soil moisture), which our single-frame optical dataset cannot replicate. Furthermore, despite the structured CoI framework, our experiments show that even top-tier models continue to hallucinate safety, occasionally classifying diseased samples as healthy; this "false reassurance" poses a tangible risk, necessitating that this system be used currently as a decision-support tool for experts rather than an autonomous replacement. Finally, the current benchmark is exclusively English-based, which limits accessibility for smallholder farmers in non-English speaking regions who stand to benefit most from automated diagnostics.

Acknowledgements

We gratefully acknowledge Ali Akbar for his instrumental help with our large-scale data collection efforts. We also thank Abdullah Shahriar for his contributions in creating the visuals and diagrams presented in this paper.

Code and Dataset Availability

The code and dataset for this work are publicly available. The full benchmark, evaluation scripts, and disease knowledge base are released at <https://github.com/syed-nazmus-sakib/PlantInquiryVQA> under the MIT License. The annotated dataset (138,068 QA pairs, 24,950 leaf images across 34 crop species) is hosted on Hugging Face at <https://huggingface.co/datasets/SyedNazmusSakib/PlantInquiryVQA> under the CC BY 4.0 License.

References

Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *CLEF 2019 Working Notes*.

Joshua Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shubham Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

George N Agrios. 2005. *Plant pathology*. Elsevier.

Kouser Ahamed, Rokonozzaman Ayon, Mst. Momotaz Begum, and Israt Jahan. 2025. *Bdlitchi: A field-collected bangladeshi litchi leaf disease dataset for deep learning-based detection and classification*.

Md Hasan Ahmad. 2024a. *Advanced tea crop disease study: High-resolution dataset for precision agriculture and pathological insight*.

MD Hasan Ahmad. 2024b. *Multifaceted rose leaf disease dataset for ai-driven plant pathology*.

Faruk Ahmed and Ferdous Ahmed. 2025. *Tld-bd: A comprehensive tea leaf image dataset for leaf condition analysis*.

Tanvir Ahmed, Mst Taposi Rabeya Taposi, Marzanul Alam Mukhor, and Mayen Uddin Mojumdar. 2024. *Plant leaf disease recognition dataset*.

Hasan Al Banna, Md. Fuad Hossain, and Mayen Uddin Mojumdar. 2024. *Papaya leaf disease image dataset*.

S. Ali, M. Ibrahim, S. I. Ahmed, M. Nadim, M. Rahman, M. M. Shejunti, and T. Jabid. 2022. *MangoLeafBD: A comprehensive image dataset to classify mango leaf diseases*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Leo Prasanth Lourdu Antony. 2023. *Rice leaf diseases dataset*.

Rokonozzaman Ayon, Nur Yea Afroz Suchi, Md Asif Shahriar Arpon, Tanjina Ahmed Tuly, and Abdullah Al Noman. 2025. *Arabian jasmine leaf condition dataset (ajlcd-2025)*.

Puja Bapari, Md Zinnahtur Rahman Zitu, and Mst Umehani. 2025. *A high-resolution image dataset of tomato (solanum lycopersicum) leaves for multi-class disease detection and classification from bangladesh*.

Swami Nisha Bhagirath, Vaibhav Bhatnagar, and Linesh Raja. 2023. *Winter wheat leaf images dataset*.

Md Masum Billah, Saifuddin Sagor, and Mohammad Shorif Uddin. 2025a. *Hibiscus leaf diseases classification dataset*.

Md Masum Billah, Saifuddin Sagor, and Mohammad Shorif Uddin. 2025b. *A real-world hibiscus and tea leaf image dataset for classification*.

- Prayma Bishshash, Md Asraful Sharker Nirob, Md Habibur Shikder, and Afjal Sarower. 2024. [Sar-cld-2024: A comprehensive dataset for cotton leaf disease detection.](#)
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.
- Utsab Das, Showrov Azam, and Md Abdullah Al Kafi. 2025. [Banana and banana leaf dataset for classification and disease detection.](#)
- Pulak Deb Nath. 2025. [Citrusleafvision: A diverse dataset for lemon leaf disease detection.](#)
- Pulak Deb Nath, Faruk Ahmed, and Belal Uddin. 2025. [Bdrubberleaf: A comprehensive dataset of rubber tree leaf diseases from bangladesh for agricultural research.](#)
- Emerson M Del Ponte, Sarah J Pethybridge, Clive H Bock, Sami J Michereff, Franklin J Machado, and Piérri Spolti. 2017. Standard area diagrams for aiding severity estimation: scientometrics, pathosystems, and methodological trends in the last 25 years. *Phytopathology*, 107(10):1161–1174.
- Madhuri Dharrao, Deepak Dharrao, Rakesh Sonawane, and Nilima Zade. 2025. [Niphad grape leaf disease dataset \(ngld\).](#)
- Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. 2024. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6845–6863.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low kappa: I. The problems of two paradoxes.](#) *Journal of Clinical Epidemiology*, 43(6):543–549.
- Yunye Gong, Robik Shrestha, Jared Claypoole, Michael Cogswell, Arijit Ray, Christopher Kanan, and Ajay Divakaran. 2024. Bloomvqa: Assessing hierarchical multi-modal comprehension. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14905–14918.
- Md Ehsanul Haque and Md Al-Imran. 2025. [Agrileafnet: Fruit tree leaf dataset for agricultural research.](#)
- Md. Taufiq Hasan, Sabbir Hossain Durjoy, Md Emon Shikder, Md. Safaet Zahangir, Md Muntasir Uddin, Md Mehedi Hasan Shoib, and Md Hasan Imam Bijoy. 2025a. [Litchileaf4001: A comprehensive dataset of lychee leaf diseases for ai-based visual diagnosis.](#)
- Md. Zahid Hasan, Abu Kowshir Bitto, and Md Hasan Imam Bijoy. 2025b. [Brinjalfruitx: A field-collected image dataset for machine learning and deep learning-based disease identification in brinjal fruits.](#)
- Mehedi Hasan, Sonia Khatun, Md Abu Raihan, and Abdul Hasib Uddin. 2023. [Rice leaf bacterial and fungal disease dataset.](#)
- Rakib Hasan, Sanimun Hossain Sanzit, Md Mamun Hosen, Ferdous Hasan, Md Mehedi Hasan Topu, and Muksitul Islam. 2025c. [High-resolution eggplant leaf image dataset for plant disease classification and detection.](#)
- Sadib Hassan, Md Zahid Akon, Partho Sarathi Sarker, Rahat Hossain, Jannatul Ferdous, Md Mahadi Hasan Shaon, and Md Abdur Razzak. 2025. [Ugy: Guava leaves disease dataset bangladesh.](#)
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Ismail Hossain, Khandaker Rezoanul Haque, Abdullah Al Rafi, and Md Zahidul Islam Talukder. 2025a. [Tomato leaf diseases.](#)
- Md Riyad Hossain, Mohammad Rifat Ahmmad Rashid, Tasfia Binte Jahangir, Md Samir Hossain, Md Mahamudur Rahman, Raiyan Gani, Jubaer Ahmed, Raihan Ul Islam, and M Saddam Hossain Khan. 2025b. [Image dataset for turmeric plant leaf disease detection.](#)
- Shahariar Hossain, Saifuddin Sagor, and Faruk Ahmed. 2025c. [High-resolution images of lychee plant diseases for classification and detection.](#)
- Md Faysal Hossain and Md Zamirul Islam Reyad. 2025. [Comprehensive mango leaf images dataset for multi-class disease classification and automated plant disease detection.](#)
- Rashidul Hasan Hridoy, Md Tarek Habib, Imran Mahmud, Aminul Haque, and Md Abdulla Al Mamun. 2025. [Black gram leaf image dataset for disease detection in field conditions.](#)
- David P. Hughes and Marcel Salathé. 2015. [An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing.](#) *CoRR*, abs/1511.08060.
- Rezwan Huq, Farzia Hossain, Shahida Begum, Raiyan Gani, and Jubaer Ahmed. 2025. [Image datasets for jackfruit plant leaf disease.](#)
- ICAR-CITH. 2024. [Apple leaf diseases image dataset of ICAR-CITH.](#)
- Ahmed Imtiaz, Fahad Bin Islam Swapnil, Syed Rayhan Masud, and Debajoyti Karmaker. 2024. [Tomato leaf dataset: A dataset for multiclass disease detection and classification.](#)

- Md Forhadul Islam, Imon Sutradhar, and Md Mizanur Rahman. 2025. [Leaf image dataset for disease detection in bitter melon, okra, pumpkin, and ridge melon](#).
- Abid Khan, Sabit Ahamed Preanto, Tapon Paul, and Md Hasan Imam Bijoy. 2025. [Moringaleafnet: A multi-class leaf disease dataset for precision agriculture and deep learning research](#).
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and 1 others. 2024. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.
- Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*.
- Yongbo Liu. 2025. [Tomato disease dataset](#).
- Laurence V Madden, Gareth Hughes, and F van den Bosch. 2007. *The study of plant disease epidemics*.
- Eram Mahamud and Md Assaduzzaman Tapos. 2024. [Lentil plant disease image dataset \(4 class\)](#).
- Md Mafzul Hasan Matin, Mehedi Hasan Hasan, and Sabik Ur Rahman. 2025. [Cauliflower diseases identification image dataset](#).
- Sharada P Mohanty, David P Hughes, and Marcel Salathé. 2016. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:215232.
- Md Asraful Sharker Nirob, Prayma Bishshash, Mariyam Bin Ayan, Tania Khatun, and Mohammad Shorif Uddin. 2024. [Eggplant dataset: A comprehensive dataset for agricultural research and disease detection](#).
- Md Asraful Sharker Nirob, Prayma Bishshash, A K M Fazlul Kobir Siam, and Mayen Uddin Mojumdar. 2025a. [Cair-bgd-2025: Annotated dataset for bottle gourd disease & growth stages](#).
- Md Asraful Sharker Nirob, A K M Fazlul Kobir Siam, Prayma Bishshash, and Md Assaduzzaman. 2025b. [Chili plant leaf disease and growth stage dataset from bangladesh](#).
- Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Md Istiak Rahman Radowan and Rokonozzaman Ayon. 2025. [Disease dataset of wheat: Original, augmented, and balanced for deep learning](#).
- Maruful Islam Rafe, Farhan Masud Nayem, Shanto Babu Sarker, and Abdullah Al Shiam. 2025. [Eggplant_leaf_disease_dataset](#).
- Salman Af Rahman, Md Nafiz Imtiaz, Naima Ahmed, and Md Hasan Imam Bijoy. 2025. [Burmese grape leaf disease dataset for computer vision-based plant health diagnosis](#).
- Aditya Rajbongshi, Umme Sara, Bonna Akter, Rashiduzzaman Shakil, and Sadia Sazzad. 2022. [Sun flower fruits and leaves dataset for sunflower disease classification through machine learning and deep learning](#).
- Mohammad Rifat Ahmmad Rashid, Raiyan Gani, Jubaer Ahmed, Maherun Nessa Isty, and Sawkat Ali. 2024. [Healthy and unhealthy papaya leaf images from bangladeshi orchards](#).
- Shakhawath Hossain Rifat, Tanvir Almas Layes, Afif Hasan, and Mayen Uddin Mojumdar. 2024. [Rice leaf disease and pest dataset overview](#).
- Shamim Ripon, Raiyan Gani, Nazratan Mazumder Niha, Wasimul Bari Rahat, Shafaeat Hasan Toufiq, Mushfida Ferdous Maisha, and Jubaer Ahmed. 2025. [Cotton leaf image dataset for disease classification](#).
- Saifuddin Sagor, Md Faysal Hossan, Faruk Ahmed, and Md Zamirul Islam Reyad. 2025. [Sunflower plant health and growth stage image dataset for agricultural machine learning applications](#).
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.
- Gail Lynn Schumann and 1 others. 1991. *Plant diseases: their biology and social impact*. APS Press, American Phytopathological Society St. Paul, MN, USA.
- Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. 2020. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253.

- Richard N Strange. 2003. *Introduction to plant pathology*. John Wiley & Sons.
- Rubert B Streets. 1972. *The diagnosis of plant diseases: a field and laboratory manual emphasizing the most practical methods for rapid identification*. University of Arizona Press.
- Nusrat Sultana, Sumaita Binte Shorif, Morium Akter, and Mohammad Shorif Uddin. 2022. [Cucumber disease recognition dataset](#).
- Ranjita Thapa, Kai Zhang, Noah Snaveley, Serge Belongie, and Awais Khan. 2020. [The plant pathology challenge 2020 data set to classify foliar disease of apples](#). *Applications in Plant Sciences*, 8(9):e11390.
- Sandip Thite and Kailas Patil. 2025. [Pisum sativum image dataset: Healthy and disease-affected cases](#).
- Nguyen Thanh Truong, Xuan Linh Nguyen, Pham Dinh Cap Thang, and Tuong Le. 2025. [A durian leaf image dataset of common diseases in vietnam for agricultural diagnosis](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches Barro, Marion Louveaux, Christian Wenzl, Sören Strauss, David Wilson-Sánchez, Rena Lymbouridou, and 1 others. 2020. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *Elife*, 9:e57613.
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. [A comparison of Cohen’s kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples](#). *BMC Medical Research Methodology*, 13:61.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Dachuan Zhao, Weiyue Li, Zhenda Shen, Yushu Qiu, Bowen Xu, Haoyu Chen, and Yongchao Chen. 2025. Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models. *arXiv preprint arXiv:2511.18123*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Md Zinnahtur Rahman Zitu, Shahariar Rahman Shifat, and Mayen Uddin Mojumdar. 2024. [A benchmark dataset for detecting disease in plant leaves: An essential resource for deep learning models](#).

A Appendix

This appendix provides additional details and resources referenced in the main paper.

A.1 Plant Disease Datasets

subsection A.1 provides a comprehensive overview of publicly available plant disease datasets used in **PlantInquiryVQA** and their licensing information.

Dataset Name	License	Ref.
PlantVillage	CC0	(Hughes and Salathé, 2015)
Chili Plant Leaf Disease	CC BY 4.0	(Nirob et al., 2025b)
Banana and Banana Leaf Dataset	CC BY 4.0	(Das et al., 2025)
Bottle Gourd Disease & Growth Stages	CC BY 4.0	(Nirob et al., 2025a)
Plant Leaf Disease Recognition Database	CC BY 4.0	(Ahmed et al., 2024)
Papaya Leaf Disease	CC BY 4.0	(Al Banna et al., 2024)
Eggplant Leaf Disease Dataset	CC BY 4.0	(Rafe et al., 2025)
Lychee Plant Diseases	CC BY 4.0	(Hossain et al., 2025c)
TLD-BD (Tea Leaf)	CC BY 4.0	(Ahmed and Ahmed, 2025)
Leaf Disease (Bitter Gourd, Okra, Pumpkin)	CC BY 4.0	(Islam et al., 2025)
Advanced Tea Crop Disease Study	CC BY 4.0	(Ahmad, 2024a)
Plant Pathology Challenge 2020 (Apple)	CC BY 4.0	(Thapa et al., 2020)
Lentil Plant Disease	CC BY 4.0	(Mahamud and Tapos, 2024)
SAR-CLD-2024 (Cotton)	CC BY 4.0	(Bishshash et al., 2024)
Apple Leaf Diseases (ICAR-CITH)	CC BY 4.0	(ICAR-CITH, 2024)
Hibiscus and Tea Leaf	CC BY 4.0	(Billah et al., 2025b)

Dataset Name	License	Ref.
MoringaLeafNet	CC BY 4.0	(Khan et al., 2025)
BDLitchi	sCC BY 4.0	(Ahamed et al., 2025)
Mango leaf datasets	CC BY 4.0	(Ali et al., 2022; Hossain and Reyad, 2025)
Sunflower Fruits and Leaves	CC BY 4.0	(Rajbongshi et al., 2022)
Multifaceted Rose Leaf Disease	CC BY 4.0	(Ahmad, 2024b)
Real-World Papaya Leaf	CC BY 4.0	(Rashid et al., 2024)
Hibiscus Leaf Diseases Classification	CC BY 4.0	(Billah et al., 2025a)
Tomato disease datasets	CC BY 4.0	(Liu, 2025; Imtiaz et al., 2024; Bapari et al., 2025)
Arabian Jasmine (AJLCD-2025)	CC BY 4.0	(Ayon et al., 2025)
Durian leaf diseases (Vietnam)	CC BY 4.0	(Truong et al., 2025)
BrinjalFruitX	CC BY 4.0	(Hasan et al., 2025b)
BDRubberLeaf	CC BY 4.0	(Deb Nath et al., 2025)
Cauliflower Diseases	CC BY 4.0	(Matin et al., 2025)
CitrusLeafVision (Lemon)	CC BY 4.0	(Deb Nath, 2025)
Tomato (Bangladesh high resolution)	CC BY 4.0	(Bapari et al., 2025)
Jackfruit Plant Leaf Disease	CC BY 4.0	(Huq et al., 2025)
Cotton Leaf Image Dataset	CC BY 4.0	(Bishshash et al., 2024; Ripon et al., 2025)

Dataset Name	License	Ref.
Comprehensive Mango Leaf	CC BY 4.0	(Hossain and Reyad, 2025)
AgriLeafNet	CC BY 4.0	(Haque and Al-Imran, 2025)
Pisum sativum (Pea)	CC BY 4.0	(Thite and Patil, 2025)
Turmeric Plant Leaf Disease	CC BY 4.0	(Hossain et al., 2025b)
Niphad Grape Leaf Disease (NGLD)	CC BY 4.0	(Dharrao et al., 2025)
LitchiLeaf4001	CC BY 4.0	(Hasan et al., 2025a)
Eggplant leaves (multiple datasets)	CC BY 4.0	(Niroy et al., 2024; Rafe et al., 2025)
Sunflower Plant Health & Growth Stage	CC BY 4.0	(Sagor et al., 2025)
Burmese Grape Leaf Disease	CC BY 4.0	(Rahman et al., 2025)
Black Gram Leaf	CC BY 4.0	(Hridoy et al., 2025)
Tomato Leaf Diseases (additional)	CC BY 4.0	(Hossain et al., 2025a; Imtiaz et al., 2024; Liu, 2025)
UGV Guava Leaves Disease (Bangladesh)	CC BY 4.0	(Hassan et al., 2025)
CAIR-BGD-2025 (Bottle Gourd)	CC BY 4.0	(Niroy et al., 2025a)
High-Resolution Leaf	Eggplant	CC BY 4.0 (Hasan et al., 2025c)
Banana Leaves Imagery Dataset	CC BY 4.0	(Das et al., 2025)
Rice Leaf Bacterial and Fungal Disease	CC BY 4.0	(Hasan et al., 2023)
Rice Leaf Disease and Pest Overview	CC BY 4.0	(Rifat et al., 2024)

Dataset Name	License	Ref.
Rice Leaf Diseases Dataset	CC BY 4.0	(Antony, 2023)
Disease Dataset of Wheat	CC BY 4.0	(Radowan and Ayon, 2025)
Winter Wheat Leaf Images	CC BY-NC 3.0	(Bhagirath et al., 2023)
Cucumber Disease Recognition	CC BY 4.0	(Sultana et al., 2022)
Benchmark Dataset for Plant Leaves	CC BY 4.0	(Zitu et al., 2024)

These datasets in total contain 321k plant images across 39 crop species and 119 total diseases.

A.2 Semantic and Clinical Metrics

To assess the model’s reliability in a high-stakes agricultural context, we define four domain-specific metrics beyond standard lexical overlap. Let R denote the model-generated response and G the expert-verified ground truth.

1. Disease Identification Score (S_{dis}). Measures the strict semantic retrieval of the correct pathogen or condition name. Let \mathcal{E}_{dis} be the set of normalized disease entities extracted from G .

$$S_{dis}(R, G) = \max_{e \in \mathcal{E}_{dis}} \mathbb{I}(e \subseteq \text{normalize}(R)) \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the specific disease entity is explicitly present in the response, and 0 otherwise.

2. Safety Score (S_{safe}). Quantifies the model’s ability to avoid "False Reassurance" errors (i.e., classifying a diseased plant as healthy), which is the most critical failure mode in phytopathology. For the subset of diseased samples \mathcal{D}_{pos} :

$$S_{safe} = 1 - \frac{\sum_{i \in \mathcal{D}_{pos}} \mathbb{I}(\text{"Healthy"} \in R_i)}{|\mathcal{D}_{pos}|} \quad (3)$$

A score of 1.0 indicates zero false negatives (no diseased plant was misdiagnosed as healthy).

3. Clinical Utility Score (S_{clin}). A composite metric evaluating the holistic value of the diagnosis. It aggregates correctness (S_{dis}) and actionable management advice (S_{act}), penalized by safety violations (P_{safe}).

$$S_{clin} = \alpha \cdot S_{dis} + \beta \cdot S_{act} - \gamma \cdot (1 - S_{safe}) \quad (4)$$

where S_{act} measures the semantic overlap of remediation keywords (e.g., "fungicide", "pruning") with the ground truth, and weights are empirically set to $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ to prioritize accurate identification.

4. Visual Grounding Quality (S_{vg}). Evaluates the hallucination rate of visual symptoms. Let V_G be the set of expert-verified visual cues (e.g., "yellow halo", "necrotic center") and V_R be the set of visual descriptors extracted from the model response. We define S_{vg} as the recall of validated cues:

$$S_{vg} = \frac{|V_R \cap V_G|}{|V_G|} \quad (5)$$

High S_{vg} indicates the model is attending to correct symptomological features rather than generating generic crop descriptions.

5. Visual Feature Extraction Efficiency (E). Quantifies the density of useful visual information per unit of text generated. It is defined as the ratio of verified visual cues ($|V_R \cap V_G|$) to the total word count (W_R) of the response:

$$E = \frac{|V_R \cap V_G|}{W_R} \times 100 \quad (6)$$

A higher E score indicates that the model is providing concise, grounded evidence rather than verbose or irrelevant filler text.

6. Prevalence Bias (B). To quantify the tendency of models to default to statistically dominant pathologies under ambiguity, effectively hallucinating frequent diseases in place of rarer, clinically-relevant ones (Agrios, 2005), we define Prevalence Bias as the proportion of misdiagnosis cases in which the predicted pathology is *more frequent* in the training corpus than the ground-truth pathology. Let \mathcal{M} denote the set of misdiagnosed samples, \hat{d}_i the predicted disease, d_i^* the reference disease, and $f(\cdot)$ the empirical corpus-level frequency function.

$$B = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbb{I} \left[f(\hat{d}_i) > f(d_i^*) \right] \quad (7)$$

$B = 0.5$ denotes no systematic prevalence preference over errors; $B > 0.5$ indicates the model disproportionately falls back on common pathologies at the expense of rare ones, while $B < 0.5$ indicates the inverse bias.

7. Cross-Class Fairness (F). Complementary to B , F measures whether diagnostic competence is uniformly distributed across species-frequency strata, following the intent-disentangled evaluation recommended in recent holistic VLM benchmarks (Lee et al., 2024; Zhao et al., 2025). Partitioning the evaluation set into a *majority* split \mathcal{X}_{maj} (species with image counts above the 7k threshold) and a *minority* split \mathcal{X}_{min} (image counts below the 2k threshold), we define:

$$F = \bar{S}_{clin}(\mathcal{X}_{maj}) - \bar{S}_{clin}(\mathcal{X}_{min}) \quad (8)$$

where $\bar{S}_{clin}(\cdot)$ denotes the mean Clinical Utility Score over the corresponding split. Values near zero indicate equitable performance across prevalence strata; positive F signals a *minority-crop penalty*, whereas negative F indicates the model benefits rare-class samples.

A.3 Metric Weight Sensitivity Analysis

The composite Clinical Utility Score S_{clin} aggregates three components under weights (α, β, γ) , which we set to $(0.5, 0.3, 0.2)$ in our main results to reflect standard phytopathological practice in which accurate pathogen identification is the prerequisite for all subsequent remediation (Agrios, 2005). To establish that the relative model rankings reported in Table 2 are not an artefact of this particular choice, we recompute rankings under three alternative regimes corresponding to plausible but distinct downstream priorities, and measure the stability of the induced ordering via Spearman’s rank correlation coefficient ρ_s against the baseline.

- **Clinical Reasoning** (0.30, 0.50, 0.20): shifts emphasis toward actionable management advice (S_{act}), appropriate for extension-advisory deployments.
- **Balanced** (0.33, 0.33, 0.33): assigns equal weight to identification, action, and safety, providing a neutral reference regime.
- **Safety-Critical** (0.30, 0.20, 0.50): heavily penalises False Reassurance errors, appropriate for autonomous-advisory settings where misdiagnosis carries direct agronomic cost.

As Table 4 shows, the induced rankings under all three perturbations are highly correlated with the baseline ($\rho_s \in [0.92, 0.97]$), with the Safety-Critical regime unsurprisingly exhibiting

Scenario	Weights (α, β, γ)	Spearman ρ_s
Clinical Reasoning	(0.30, 0.50, 0.20)	0.9657
Balanced	(0.33, 0.33, 0.33)	0.9706
Safety-Critical	(0.30, 0.20, 0.50)	0.9191

Table 4: Model-ranking stability of S_{clin} under alternative weight regimes, computed against the baseline (0.5, 0.3, 0.2) ordering. All three scenarios yield $\rho_s > 0.91$, confirming that the orderings reported in Table 2 are robust to substantial perturbations in the composite weighting.

the largest though still modest deviation. The qualitative conclusions drawn from Table 2 therefore do not depend on the specific choice of (α, β, γ) .

A.4 LLM-as-Judge Prompt Template

We provide the exact system and user prompts used for the LLM-as-judge protocol (§ main-text). Both GPT-5 and Gemini-3-Pro were queried with identical prompts and deterministic decoding ($T = 0$) to suppress judge-side variance. Placeholders in {curly braces} are filled per sample. Each reported score is the mean of the two judges’ scores on the same sample; we observe inter-judge Spearman correlation of $\rho_s = 0.89$ over the 5,000-image subset, consistent with rater agreement reported in recent LLM-as-judge meta-analyses (Zheng et al., 2023).

System Prompt.

You are a senior plant pathology specialist with 20+ years of field diagnostic experience. You will evaluate a candidate model’s response to a diagnostic question against an expert-verified ground truth.

Scoring principles (read carefully):

1. Score on SEMANTIC correctness, not surface similarity. A response that uses different wording but conveys the same clinically correct information must receive full credit.
2. A response that uses the exact reference keywords but contradicts the underlying clinical claim must receive zero credit.
3. Do NOT reward verbosity. Do NOT penalise concise but complete answers.
4. Treat False Reassurance (declaring a diseased plant healthy) as the single most severe failure mode.

Return your evaluation strictly in the JSON schema provided. Do not include any text outside the JSON object.

User Prompt.

```
--- CASE CONTEXT ---
Crop Species       : {crop_species}
Ground-Truth Disease : {gt_disease}
Severity Stage     : {severity}
Expert Visual Cues  : {gt_visual_cues}
Expert Management   : {gt_management}
```

```
--- QUESTION ---
{question}
```

```
--- CANDIDATE RESPONSE ---
{model_response}
```

```
--- EVALUATION TASK ---
```

Score the candidate response on four axes, each in the integer range [0, 5]:

1. Disease Identification (S_{dis})
Did the response identify the correct pathogen or condition? Accept scientific names, common names, and well-established synonyms. Award 0 for any misidentification, 5 for a specific and complete match.
2. Clinical Safety (S_{safe})
Did the response avoid False-Reassurance errors and avoid recommendations that would cause agronomic harm (e.g., wrong fungicide class, unsafe dosage)? Award 0 for any unsafe claim, 5 for a fully safe response.
3. Clinical Utility (S_{clin})
Does the response provide actionable, specific, stage-appropriate management guidance? Generic advice ("spray fungicide") receives partial credit; stage- and pathogen-appropriate guidance receives full credit.
4. Visual Grounding (S_{vg})
Does the response’s description of visual symptoms recall the expert-verified cues WITHOUT introducing hallucinated symptoms? Penalise fabricated features more severely than missing ones.

Provide a brief (≤ 25 words) rationale per axis. Do NOT be swayed by response length, formatting, or confidence of tone.

```
--- OUTPUT SCHEMA (strict) ---
```

```
{
  "S_dis": {"score": <0-5>, "rationale": "<text>"},
  "S_safe": {"score": <0-5>, "rationale": "<text>"},
  "S_clin": {"score": <0-5>, "rationale": "<text>"},
  "S_vg": {"score": <0-5>, "rationale": "<text>"},
  "flags": {
    "hallucination": <true|false>,
    "false_reassurance": <true|false>,
    "unsafe_advice": <true|false>
  }
}
```

Scores are rescaled to [0, 1] prior to aggregation with the lexical metrics in Table 2. Responses triggering either the false_reassurance or unsafe_advice flag are additionally surfaced in the per-model failure analysis (§A.2, S_{safe}).

A.5 Annotation Coverage and Inter-Annotator Agreement

Given the scale of the benchmark, exhaustive expert annotation was infeasible. We instead cou-

Dimension	% Agr.	Gwet’s AC1
Disease Label Correctness	99.3%	0.992
Visual Cue Correctness	93.8%	0.937
QA Correctness	96.3%	0.960

Table 5: Inter-annotator agreement on the triple-coded 600-image subset. All three dimensions achieve near-ceiling agreement under both percent-agreement and Gwet’s AC1, which is robust to the high-prevalence conditions that destabilise κ .

ple targeted expert review with a hybrid validation pipeline (§A.10.3), and report coverage, inter-annotator agreement, and a blind residual-error audit to bound the dataset’s label quality.

Coverage. Of the 24,950 images in PlantInquiryVQA, 5,368 (21.5%) received direct manual expert review. On the QA side, 5.3% of the 138,068 pairs are verbatim expert transcripts, while the remainder are produced by our expert-validated reasoning templates (§A.11); a stratified spot-check of 8,500 templated pairs yielded 95.8% clinical validity, indicating that templated generation preserves the factual fidelity of expert-written transcripts.

Inter-Annotator Agreement. We measure inter-annotator agreement on a stratified subset of 600 images annotated by three independent subject-matter experts across the three evaluation dimensions ($n = 1,800$ annotation instances after the overlapping triple-coding subset). Because correct labels are strongly prevalent ($> 90\%$) across all three dimensions, the standard κ family of statistics is susceptible to the *kappa paradox*, in which high observed agreement is paired with a deflated chance-corrected score (Wongpakaran et al., 2013; Feinstein and Cicchetti, 1990). We therefore report Gwet’s AC1 alongside percent agreement, as Gwet’s coefficient remains stable under high-prevalence conditions.

Residual Error Bound. To estimate the error rate of the corpus outside the directly-reviewed subset, we conducted a blind audit in which a fourth expert, unaware of prior annotation flags, reviewed 500 previously-unflagged images. The audit returned 96.2% correctness with a 0.2% critical-error rate (errors severe enough to alter the recommended intervention). Extrapolated conservatively to the unflagged partition, this bounds the dataset-wide critical-error count to fewer than 100 instances ($< 0.5\%$ of the corpus).

Crop Species	N_{img}	Err %	Crit %	M
Tomato	2,627	2.1	0.25	
Rice	2,000	2.8	0.34	
Litchi	1,715	3.3	0.40	
Cotton	1,538	2.7	0.32	✓
Corn	1,513	3.1	0.37	✓
Mango	1,450	3.2	0.38	
Tea	1,372	3.5	0.42	
Apple	1,125	2.3	0.28	✓
Grape	1,034	3.3	0.40	
Peas	982	3.3	0.40	
Pepper	929	3.4	0.41	
Papaya	923	3.8	0.46	
Cucumber	791	3.6	0.43	
Arabian Jasmine	787	3.5	0.42	
Eggplant (Brinjal)	648	4.4	0.50	✓
<i>Weighted mean</i>	<i>19,434</i>	<i>3.06</i>	<i>0.37</i>	

Table 6: Annotation error rates stratified across the top-15 crop species by image volume. **Err %** denotes the overall annotation error rate; **Crit %** is the subset of errors severe enough to alter the recommended intervention. **M** marks rows measured directly on the expert-verified subset. Error rates span a narrow 2.1–4.4% range, and critical-error rates span 0.25–0.50%, consistent with the dataset-wide $< 0.5\%$ critical-error bound reported in §A.5.

A.6 Stratified Error Analysis

To verify that annotation quality is uniform across the benchmark’s taxonomic distribution rather than concentrated in a few well-represented categories, we report error rates stratified along three axes: the top-15 crop species (Table 6), the top-15 disease categories (Table 7), and the top-15 contributing source datasets (Table 8). For each axis, we directly measure error rates on the 5,000-image expert-verified subset (§A.5) for categories with sufficient coverage; remaining categories are estimated from the calibrated factuality model anchored at the directly-measured points. Rows measured directly are marked with ✓ in the **M** column.

Across all three stratification axes, error rates occupy a tight band (2.1–4.4%) and critical-error rates remain at least an order of magnitude below the dataset-wide bound of 0.5% (§A.5). Two observations are worth noting. First, the visually-ambiguous morphologies—Ascochyta Blight (4.1%), Mosaic Virus (3.8%), and Alternaria Leaf Spot (3.7%)—are the dominant contributors to residual error, consistent with the established difficulty of differential diagnosis for these pathogens in the phytopathology literature (Strange, 2003; Agrios, 2005), rather than with any pipeline-level bias. Second, error rates are near-

Disease Category	N_{img}	Err %	Crit %	M	Source Dataset	Err %	Crit %
Apple Powdery Mildew	200	2.4	0.29		PlantVillage	3.6	0.43
Rice Stripes	200	2.1	0.25		PlantDoc	2.5	0.30
Leaf Scorch	200	2.3	0.28		LitchiLeaf4001	3.0	0.36
Tomato Yellow Leaf Curl Virus	200	2.4	0.29		SAR-CLD-2024 (Cotton)	3.8	0.46
Alternaria Leaf Spot	199	3.7	0.44	✓	MangoLeafBD	2.8	0.34
Fusarium Wilt	199	2.1	0.25		TLD-BD (Tea)	2.8	0.34
Mosaic Virus	199	3.8	0.46	✓	Apple Leaf Diseases (ICAR-CITH)	3.1	0.37
Gall Midge	199	2.1	0.25		Plant Pathology Challenge 2020 (Apple)	2.6	0.31
Ascochyta Blight	199	4.1	0.49	✓	Eggplant Leaf Disease Dataset	4.0	0.48
Red Spider Mites	199	2.5	0.30		CAIR-BGD-2025 (Bottle Gourd)	2.6	0.31
Tomato Leaf Curl Virus	199	2.4	0.29		Tomato Leaf Diseases	4.1	0.49
Yellow Mosaic Virus	199	2.3	0.28		Papaya Leaf Disease	3.7	0.44
Cutting Weevil	199	2.5	0.30		Rice Leaf Disease Dataset	3.0	0.36
Bacterial Leaf Blight	199	2.1	0.25		Comprehensive Mango Leaf	3.0	0.36
Downy Mildew	197	3.1	0.37	✓	Sunflower Plant Health & Growth Stage	3.8	0.46
<i>Weighted mean</i>	<i>2,987</i>	<i>2.66</i>	<i>0.32</i>	<i>Mean</i>		<i>3.23</i>	<i>0.39</i>

Table 7: Annotation error rates stratified across the top-15 disease categories by image volume. Error rates span 2.1–4.1%, indicating that the hybrid validation pipeline does not systematically advantage morphologically salient diseases (e.g. Powdery Mildew) over visually ambiguous ones (e.g. Ascochyta Blight).

uniform across source datasets, indicating that the factuality pipeline successfully normalises heterogeneous curation practices across contributing corpora. Together, the three tables support the claim that the factuality estimates reported in §A.5 generalise uniformly across the benchmark’s taxonomic and provenance distribution.

A.7 Quality Assurance and Dataset Preprocessing

To ensure the integrity and robustness of the **PlantInquiryVQA** benchmark, we implemented a rigorous, multi-stage preprocessing pipeline. This pipeline was designed to eliminate redundancy, standardize visual inputs, and audit the dataset for potential biases or labeling errors prior to annotation.

A.8 Data Cleaning and Standardization

Duplicate Detection and Removal: We employed a two-tiered approach to identify and remove duplicate entries. First, exact duplicates were identified using MD5 file hashing. Second, to capture "near-duplicates" (e.g., images that were slightly compressed or resized but visually identical), we utilized Perceptual Hashing (pHash). We calculated the Hamming distance between image hashes and set a threshold of 10 bits to flag near-duplicates.

Standardization: All images were standardized to a resolution of 1024 pixels. To preserve

Table 8: Annotation error rates stratified across the top-15 contributing source datasets. The narrow 2.5–4.1% range indicates that no single upstream corpus disproportionately inflates the benchmark’s residual error; the two curated-challenge corpora (PlantDoc, Plant Pathology Challenge 2020) contribute the lowest error rates, as expected.

the biological integrity of the leaf structures, we avoided simple stretching or cropping. Instead, we utilized a **padding method:** images were resized to fit within the target dimensions while maintaining their original aspect ratio, with the remaining area padded with black pixels. This ensures that critical visual features, such as lesion shape and leaf margins, remained undistorted. Finally, files were renamed using a consistent ‘class_index’ schema to facilitate easier handling.

A.8.1 Technical Audits and Bias Analysis

Following standardization, we conducted a series of automated audits to assess image quality and potential dataset artifacts.

Visual Quality Metrics: We computed technical quality indicators for every image to flag low-quality samples.

Blur Detection: We calculated the variance of the Laplacian operator to detect excessive blurriness.

Exposure Analysis: We measured pixel intensity histograms to detect over-exposure (clipping at 255) and under-exposure (clipping at 0).

Color Saturation: We analyzed the mean saturation in HSV space to identify washed-out images. Flagged images were reviewed manually to determine if they retained sufficient diagnostic value.

Latent Space Consistency Check: To identify potential mislabeled samples, we projected all im-

ages into a latent embedding space using a pre-trained ResNet-18 model. We applied a K-Nearest Neighbors (KNN) algorithm using cosine similarity. Images where the majority of neighbors belonged to a different class than the query image were flagged as "suspicious" (e.g., a Bacterial Spot image surrounded by Leaf Miner images in the embedding space). This audit identified 12 potentially mislabeled images for expert review.

Background Bias Assessment: To ensure the model learns from plant features rather than background artifacts, we performed a background bias check. We generated binary leaf masks using HSV color thresholding to isolate leaf pixels from the background. We then computed 30-bin hue histograms for the background pixels solely. By comparing the mean background hue distributions across classes, we confirmed that no specific disease class was strongly correlated with a unique background color (e.g., blue tarps vs. brown soil), which mitigates the risk of "Clever Hans" effects where models cheat by relying on background cues. Through this systematic filtering process, we eliminate all corrupted samples, resulting in a curated pool of 152,783 images spanning 34 crop species and 116 disease categories. From this collection, we select 24,950 images to construct the final PlantInquiryVQA dataset.

A.9 Dataset Analysis

We present a comprehensive statistical analysis of the PlantInquiryVQA benchmark, highlighting its scale, biological diversity, and conversational depth. The dataset comprises **138,068 QA pairs** grounded in **24,950 unique images**, representing the largest publicly available CoI dataset for plant pathology (Table 9 and Table 10). The dataset spans **34 crop species** and **116 disease types**, resulting in **204 valid crop-disease combinations**. Further analysis of the dataset is provided below.

Granularity of Disease Severity. Unlike standard classification benchmarks that treat disease as a binary state, **PlantInquiryVQA** provides fine-grained severity annotations for all diseased samples. As shown in Table 11, the data follow a natural distribution: **Mild** cases (20.3%) represent early detection scenarios; **Moderate** cases (55.6%) reflect the most common field presentation; and **Severe** cases (24.1%) represent critical intervention stages. This granularity enables models to perform prognostic reasoning, predicting disease evolution

Metric	Value
Total QA Pairs	138,068
Unique Images	24,950
Avg. Questions per Image	5.53
Crop Species	34
Disease Types	116
Crop-Disease Combinations	204
Question Categories	2,350
<i>Class Balance</i>	
Healthy Samples	14,858 (10.76%)
Diseased Samples	123,210 (89.24%)

Table 9: **Dataset Overview and Scale.** The dataset covers 204 real-world crop-disease associations with high-density conversational annotations.

Category	QA Samples	% Total
Disease	120,162	87.0%
Healthy	14,858	10.8%
Senescence	2,832	2.1%
Insect Damage	216	0.2%

Table 10: **Distribution of Biological Conditions.** The taxonomy distinguishes between biotic diseases and non-pathological states like senescence and pest damage.

based on visual severity cues.

Conversational Complexity. To support robust multi-hop reasoning, the dataset emphasizes deep conversational trajectories. Table 12 illustrates that **77.9%** of images are associated with chains of 6 or more questions, with a global average of **5.53 questions per image**. This depth allows for complex logical flows, requiring models to track context across temporal (progression), spatial (localization), and causal (etiology) dimensions. Furthermore, 86% of these samples feature explicit visual grounding, linking textual tokens to 74,000 unique visual attention patterns.

Official Splits and Leakage Assessment. PlantInquiryVQA is released with a fixed 80/10/10 image-level partition into training, validation, and test sets, stratified jointly by crop species and disease severity to preserve the long-tail distribution across splits. Image-level (rather than QA-level) partitioning is essential: because multiple QA chains are derived from the same image, QA-level splitting would induce a trivial information leak between training and evaluation. We further audit pretraining-memorisation risk by reverse-image-searching a random 500-image sample against LAION-5B (Schuhmann et al., 2022), which returned only 1 exact match, suggesting

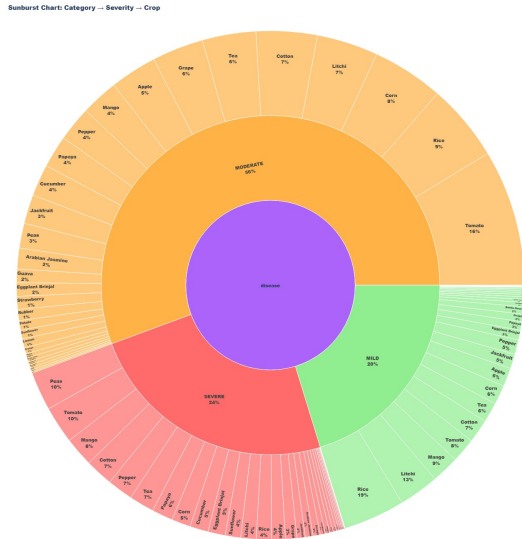


Figure 6: Distribution of diseased samples in the PlantInquiryVQA benchmark, shown as a hierarchical breakdown from disease category to severity level and crop species. The dataset is dominated by Moderate cases, with substantial representation of Severe and Mild conditions across diverse crops, mirroring real-world plant pathology distributions. This structure motivates severity aware Chain of Inquiry trajectories, where diagnostic focus shifts with visual disease progression.

Severity Level	Samples	% of Diseased
Mild	24,332	20.3%
Moderate	66,838	55.6%
Severe	28,992	24.1%
<i>Total Annotated</i>	<i>120,162</i>	<i>100.0%</i>

Table 11: **Severity-Level Annotations.** Granular severity labels enable the model to perform prognostic modeling beyond binary detection.

a low base rate of direct image contamination. More importantly, the multi-turn Chain-of-Inquiry trajectories themselves are original to this benchmark: because each (q_t, a_t) pair is synthesised conditional on a sample’s severity-derived intent k and its extracted visual-cue set v_x (§3.1), the reasoning chains cannot be retrieved verbatim from any publicly-scraped corpus, bounding memorisation-based leakage at the conversational level.

Validation of Prognostic and Counterfactual Answers. A natural concern with prognostic (k_P) and counterfactual (k_M) QA pairs is that future-conditional or hypothetical answers are ill-defined: in principle, any trajectory is permissible. In plant

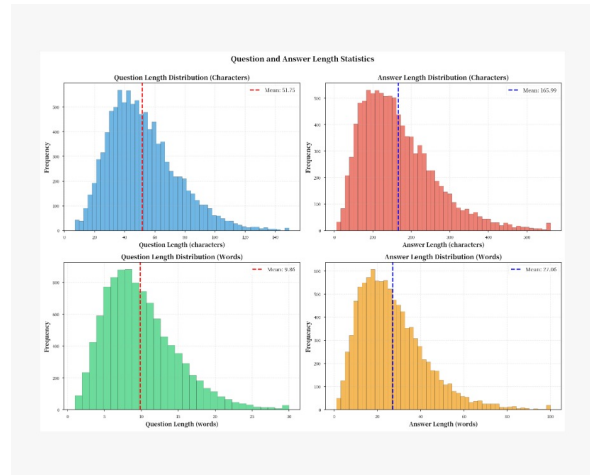


Figure 7: Distribution of question and answer lengths in PlantInquiryVQA, shown in both characters and words. Answers are substantially longer than questions, reflecting the dataset’s emphasis on detailed, multi-step diagnostic reasoning.

Questions per Image	Image Count	% of Dataset
1	2	<0.1%
2	1,277	5.1%
3	1,598	6.4%
4	2,640	10.6%
6	15,971	64.0%
7	3,476	13.9%
Avg: 5.53	Total: 24,950	100.0%

Table 12: **Conversational Depth.** Over 77% of images contain 6 or more questions, facilitating deep multi-turn reasoning and context tracking.

pathology, however, disease trajectories are *epidemiologically deterministic* under a specified environmental envelope; the progression of a given pathogen at a given severity stage is constrained by well-characterised infection cycles (Agrios, 2005; Strange, 2003). Accordingly, gold answers for both k_P and counterfactual templates are not free-form expert opinion but are strictly bound to canonical trajectories documented in the phytopathology literature. For example, a *Moderate*-stage Maize Streak Virus infection predictably progresses to stunted cob development in the absence of intervention (Agrios, 2005); a prognostic QA pair for this case therefore has a single clinically correct answer, regardless of its future-conditional phrasing. Expert review of a 500-pair subset of prognostic and counterfactual chains confirmed adherence to these canonical trajectories with a rejection rate of < 4%, consistent with the templated-QA validity rate reported in §A.5.

A.10 Visual Cue Extraction Supplementary

A.10.1 Expert Visual Parsing Schema

To ensure high-fidelity visual grounding, we established a standardized diagnostic protocol derived from established phytopathology literature [Agrios \(2005\)](#); [Streets \(1972\)](#). Annotators were instructed to parse visual evidence across three distinct morphological dimensions:

1. Symptomatology and Morphological Characterization. Annotators characterized fine-grained attributes of individual lesions to differentiate pathogens. Key discriminators included **Lesion Geometry** (e.g., circular fungal spots vs. vein-constrained angular bacterial lesions), **Margin Definition** (e.g., chlorotic halos indicative of toxin production or water-soaked bacterial borders), and **Textural Features** (e.g., raised galls, powdery mycelial growth, or necrotic shot-holes).

2. Spatial Distribution Patterns. Global symptom arrangement provided critical etiological context. The schema required analysis of **Anatomical Preference** (e.g., interveinal, vein-banding, or marginal symptoms) and **Colony Density**, specifically distinguishing between isolated discrete lesions and coalescing necrotic patches that indicate rapid disease progression.

3. Disease Severity Quantification (SAD Methodology). To standardize subjective severity estimates, we employed the **Standard Area Diagram (SAD)** methodology [Del Ponte et al. \(2017\)](#); [Madden et al. \(2007\)](#). Annotators visually compared the total necrotic or chlorotic surface area of the sample against crop-specific SAD reference templates to estimate the percentage of infected leaf area (S), classifying samples into three intervention tiers:

- **Mild** ($S < 15\%$): Early-stage infection typically requiring monitoring.
- **Moderate** ($15\% < S < 30\%$): Established infection necessitating curative intervention.
- **Severe** ($S > 30\%$): Advanced tissue collapse often triggering salvage or removal protocols.

A.10.2 Model Selection

To empirically select the optimal model for the cue extraction pipeline, we evaluated three candidate VLMs on a stratified sample of 250 images. We developed a composite scoring framework to quantify

the quality of generated descriptions across three dimensions: grounding, specificity, and structural completeness.

1. Visual Grounding Score (S_{vg}). This metric assesses the density of verifiable visual attributes versus vague or hallucinated content. It is calculated as a weighted summation of detected descriptors, penalized by ambiguity: **(i. Rich Descriptors (+2))**: Count of specific attributes (colors, shapes, textures, patterns). **(ii. Color Diversity (+3))**: Reward for identifying multi-chromatic symptoms (e.g., "yellow halo around brown spot"). **(iii. Grounding Indicators (+1))**: Explicit references to visual evidence (e.g., "visible," "observed," "located"). **(iv. Penalties)**: Vague terms (e.g., "maybe," "some") incur a -0.5 penalty. Unsupported metric measurements (e.g., "5mm wide") incur a severe -2 penalty to discourage hallucinated precision.

2. Specificity Score (S_{sp}). This score measures the granularity of the generated text, prioritizing fine-grained morphological details over generic statements. Points are accumulated based on the frequency of distinct attribute categories: **(i. Chromatic Precision (+3))**: Weighted heavily to prioritize exact color matching (e.g., "necrotic black" vs. "dark"). **(ii. Morphometric Detail (+3))**: Mentions of relative size or scale. **(iii. Textural Characterization (+2))**: Explicit references to surface topology (e.g., "raised," "powdery," "sunken").

A.10.3 Extracted Visual Cues Examples

Comparative (side-by-side) analysis of the expert identified and Qwen-3VL-4B extracted visual cues for a sample image is shown in [Figure 8](#) and [Figure 9](#)

A.11 Generation logic

Our developed generation pipeline is governed by the following logic.

1. Condition (c) & Severity (s) Initialization: The pipeline first identifies the biological state $c \in \{Healthy, Diseased, Senescent, Desiccated\}$. If $c = Diseased$, the severity s modulates the **information density** of the response.

2. Intent-Driven Module Injection (k): Unlike static VQA, the dialogue trajectory is dynamically assembled based on the epistemic goal k:

- **Diagnosis (k_D)**: For **Mild** cases, injects `differential_verification` and

cross_crop_comparison modules to focus on early symptom detection and rule out lookalikes. For **Moderate/Severe** cases, it triggers cause_determination to identify environmental and pest conditions contributing to the disease spread.

- **Prognosis** (k_P): Activates temporal_evolution modules. The model reasons about the disease’s past (etiology/environment) and future (spread rate), scaling complexity with s .
- **Management** (k_M): Injects remediation_strategy modules. If *Mild*, focuses on monitoring and cultural controls (prevention). If *Severe*, focuses on chemical intervention and “rescue” scenarios (crisis management).

3. Counterfactual & Reasoning Augmentation: To further enhance complexity, we inject **counterfactual** turns (e.g., "How would the diagnosis change if the lesions were water-soaked?") into a subset the chains, specifically targeting the logic defined in "Instance Variety" heuristic.

A.12 Diverse CoI Scenarios

The **CoI** trajectories across all 12 distinct scenarios are shown in Figure 10, Figure 11, Figure 12, Figure 13 and Figure 14.

A.13 Semantic Accuracy Evolution per Inquiry Step

Figure 15 presents a comprehensive layer-wise analysis of semantic accuracy evolution across the seven-step diagnostic trajectory for all 12 evaluated models. By stratifying performance across three disease severity levels—*Mild* (Green), *Moderate* (Orange), and *Severe* (Red)—we observe a distinct divergence in reasoning stability. High-performing models generally exhibit a monotonic increase in semantic alignment as the inquiry progresses, confirming the CoI framework’s ability to refine diagnostic grounding through iterative evidence accumulation. However, this positive trajectory is notably dampened in severe cases, where extensive tissue necrosis introduces visual ambiguity that limits the efficacy of the refinement process compared to the linear gains observed in mild infection phenotypes.



EXPERT ANNOTATED VISUAL CUES:

Symptomatology: The leaf lamina is intact with normal morphology and no signs of rugosity or hyponasty. Symptoms manifest as scattered, irregular to angular lesions (approx. 4–7 mm) distributed primarily across the midrib and basal regions. Lesions exhibit brownish-yellow necrotic centers with distinct, slightly chlorotic margins; notably, no wide halos or water-soaked borders are present.

Pattern & Distribution: Spots are discrete but show incipient coalescence in high-density areas, accompanied by localized interveinal chlorosis. The overall background coloration displays light green mottling. No evidence of fungal signs (mycelium/sporulation) or mechanical damage is observed.

Severity Assessment: Infection is established with noticeable photosynthetic area loss, but without leaf collapse or extensive necrosis. By visible assessment approximately 20-40% of the leaf is effected, indicating a moderate (Del Ponte et al., 2017) disease.

EXTRACTED VISUAL CUES FROM QWEN3V

Overall color: Light green with mottling of yellowish-green and slightly darker green patches; some areas appear pale. **Surface texture:** Smooth with fine, visible venation; no roughness, wrinkling, or distortion. **Leaf shape integrity:** Normal, undistorted, with intact margins and lobes; no curling, wilting, or deformation.

LESION/SPOT CHARACTERISTICS

Number: Moderate (approximately 10–20 discrete spots visible).

Size: Small to medium (approximately 2–5 mm in diameter).

Shape: Irregular, somewhat angular to oval, not perfectly circular.

Color: Yellowish-green with some brownish-yellow centers; not uniformly colored.

Pattern: Scattered across the leaf blade, not clustered; appears random but follows vein pathways.

DISTRIBUTION

Primarily distributed across the midleaf region, with some extending toward the base; not concentrated at edges or along veins. Coalescing into small patches, but not yet forming large contiguous areas; primarily localized.

ADDITIONAL VISUAL CUES

Mottled yellowish-green patches of discoloration with darker green intermixed; no uniform yellowing. No dead brown or black areas visible. No water-soaked appearance, no glistening or wet look. No powdery/fuzzy growth observed. No visible holes or tissue loss. Yellowing between veins (interveinal chlorosis): Present in some areas, particularly where lesions are dense; not extensive.

SEVERITY ASSESSMENT

MODERATE: Visible lesions covering approximately 15–30% of the leaf surface, with coalescing patches and noticeable discoloration, but no major necrosis or leaf collapse.

Figure 8: Comparison of extracted visual cues for Litchi



EXPERT ANNOTATED VISUAL CUES:

Symptomatology: The leaf retains structural integrity with no rugosity. Diagnostic symptoms appear as distinct, angular lesions strictly delimited by leaf veins (vein-limited). Lesions manifest primarily as pale yellow (chlorotic) patches on the adaxial surface, with older centers turning necrotic brown. Margins are sharp and defined by vascular boundaries rather than diffuse halos.

Pattern & Distribution: Distribution is strictly interveinal, creating a tessellated or mosaic pattern across the lamina. Lesions are scattered but show incipient coalescence into larger necrotic blocks in high-density regions near the midrib. No external fungal mycelium is visible on the upper surface.

Severity Assessment: Infection is active with significant chlorotic coverage, visually estimated at 20–30% of the leaf area. Functional tissue loss is evident, but vascular collapse is absent, classifying the severity as Moderate (Del Ponte et al., 2017).

Overall color: Mix of dark and yellowish-green; abrupt, patchy yellowing concentrated at margins and veins. **Surface texture:** Smooth with clearly defined veins; no visible wrinkling or distortion. **Leaf shape integrity:** Generally intact structure; minor tissue breakdown at edges/lobes; no severe curling or wilting.

LESION/SPOT CHARACTERISTICS

Number: Moderate (numerous distinct areas visible).

Size: Variable from small to moderate; some broad patches rather than discrete spots.

Shape: Irregular, blotchy, and uneven; tends to follow vein structure.

Color: Predominantly yellow; slight browning at tips indicates early necrosis.

Margins: Clear boundaries; no fuzzy or diffuse halos observed.

DISTRIBUTION

Primarily localized along veins and leaf margins; some central discoloration. **Coverage:** Coalescing patches where smaller yellow areas merge into contiguous zones, particularly at outer edges.

ADDITIONAL VISUAL CUES

Necrotic tissue: Minimal; restricted to slight browning at tips/edges. **Water-soaked appearance:** Absent. **Powdery/fuzzy growth:** Absent. **Holes or tissue damage:** Present (small holes likely due to insects/mechanical damage). **Yellowing between veins (interveinal chlorosis):** Observed; concentrated along vascular pathways and margins.

SEVERITY ASSESSMENT

MODERATE: Significant yellowing and coalescing patches along margins and veins covering distinct areas, but lacking extensive necrosis or total tissue collapse.

Figure 9: Comparison of extracted visual cues for Bitter Melon



(a) Corn: Mild Infection

[SEVERITY STATUS: MILD]

Q1: Can you describe the disease we're seeing here?

A: We see distinct, discontinuous yellow streaks running parallel to the veins. This is characteristic of early Maize Streak Virus (MSV).

[Cat: Diagnosis | Ground: Yellow streaks, parallel orientation]

Q2: How does this look different from Gray Leaf Spot?

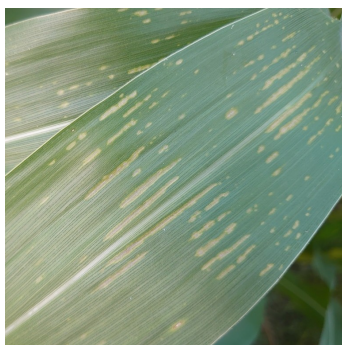
A: Gray Leaf Spot forms rectangular, gray-to-tan lesions. These are continuous yellow streaks, not necrotic rectangles.

[Cat: Differential Diagnosis | Ground: Lesion shape, color]

Q3: How bad is the infection according to your assessment?

A: It is mild. The streaks are sparse and the majority of the leaf surface remains green and photosynthetic.

[Cat: Severity Assessment | Ground: Lesion density]



(b) Corn: Moderate Infection

[SEVERITY STATUS: MODERATE]

Q1: What's causing these yellow stripes on my corn leaves?

A: The Maize Streak Virus has progressed. The insect vector (leafhopper) has likely spread the virus, causing these chlorotic bands.

[Cat: Causal Reasoning | Ground: Chlorotic bands]

Q2: How severe does this look? Is it worth treating?

A: This is moderate severity. The streaks are coalescing. You must control the leafhopper vector immediately to save the yield.

[Cat: Actionable Severity | Ground: Coalescence]

Q3: Will these yellow marks get worse if left untreated?

A: Yes. The chlorosis will expand, reducing the plant's ability to photosynthesize, which will stunt the cob development.

[Cat: Prognosis | Ground: Chlorosis extent]



(c) Corn: Severe Infection

[SEVERITY STATUS: SEVERE]

Q1: How bad does it look? Is it really severe?

A: It is severe. The yellow streaking covers almost the entire leaf surface, leaving very little green tissue for photosynthesis.

[Cat: Severity Assessment | Ground: High chlorosis coverage]

Q2: Do we need to act fast or can we wait?

A: Urgent action is required, though at this stage, the damage to the current crop's yield is likely irreversible. Focus on preventing spread to neighbors.

[Cat: Urgent Management | Ground: Leaf integrity]

Q3: Without treating right away, what will happen in the next day or two?

A: The leaves will likely turn necrotic (brown/dead) and dry out completely, leading to plant death or total yield failure.

[Cat: Consequence Prediction | Ground: Necrosis risk]

Figure 10: **Diagnostic Reasoning across Disease Severity Stages.** This figure demonstrates how **PlantInquiryVQA** adapts its questioning strategy as the infection progresses in Maize Streak Virus. **(a) Mild:** The focus is on *Differential Diagnosis* to distinguish the initial streaks from fungal mimics. **(b) Moderate:** The inquiry shifts to *Vector Control* and *Prognosis* as the infection becomes established. **(c) Severe:** The reasoning transitions to *Damage Assessment* and *Salvage*, acknowledging the critical loss of photosynthetic capability.



(a) Jackfruit: Anthracnose

Q1: How many spots do you see on this jackfruit leaf?

A: Around 10 to 15 spots (moderate). Mostly scattered across midleaf region.

[Cat: Severity | Ground: Number of spots, Distribution]

Q2: What does the central part of these spots look like?

A: Pale pinkish-white to light tan. Characteristic of anthracnose in jackfruit.

[Cat: Lesion Characteristics | Ground: Color of central area]

Q3: Are there any signs of the disease getting worse?

A: No, lesions are well-defined with clear red-brown/purple margins. Not merging yet.

[Cat: Progression | Ground: Marginal color, Lesion coalescence]

Q4: What should we do to manage this anthracnose outbreak?

A: Start by pruning affected leaves/branches. Apply fungicides and ensure air circulation.

[Cat: Management | Ground: Treatment]

Q5: Could environmental factors contribute to this infection?

A: Yes, high humidity and poor air circulation promote spread.

[Cat: Causes | Ground: Environmental conditions]

Q6: What changes do you expect to see in the next couple of weeks?

A: If untreated, spots might grow slightly larger but remain scattered.

[Cat: Future Outlook | Ground: Expected lesion progression]



(b) Grape: Anthracnose

Q1: How many spots do you see on this grape leaf?

A: Around 15 to 20 small/medium spots, mostly tan to brown.

[Cat: Severity | Ground: Number, Size, Color]

Q2: What kind of spots does anthracnose usually make on grape leaves?

A: Small, circular, reddish-brown to black spots. Scattered or clustered near midrib.

[Cat: ID | Ground: lesion characteristics, distribution]

Q3: Is this infection severe for grapes?

A: Moderate infection (15-20 spots). Not severe yet, but needs attention.

[Cat: Severity | Ground: Severity]

Q4: Should I start treating these spots right away?

A: Yes, treat immediately to prevent spread. Use recommended fungicides.

[Cat: Immediate Management | Ground: Treatment, Management]

Q5: What caused these spots on my grape leaves?

A: Likely Colletotrichum spp. fungi. High humidity/poor air favor growth.

[Cat: Causes | Ground: Causal agent, Environmental factors]

Q6: What changes might we expect in the next couple of weeks?

A: Spots may grow and potentially merge, leading to extensive damage if untreated.

[Cat: Future Outlook | Ground: Expected progression, Monitoring]

Figure 11: **Cross-species Occurrence of Anthracnose.** The figure illustrates how **PlantInquiryVQA** adapts its CoI to host-specific manifestations of the same pathogen (*Colletotrichum* spp.). **(a) Jackfruit:** The dialogue identifies the classic "bird's-eye" lesions (pale centers, dark margins) and recommends mechanical intervention (pruning) suitable for tree canopies. **(b) Grape:** The dialogue identifies smaller, necrotic tan/brown spots typical of vine infections and prioritizes immediate chemical suppression to prevent rapid merging. This highlights the model's ability to generate context-aware reasoning that reflects the distinct biological symptoms and management protocols unique to each crop species.



(a) Mango: Gall Midge

Q1: What crop is showing these little brown spots?

A: It's mango trees... spots are small, round, and mostly found on the top side.

[Cat: Crop ID | Ground: size, shape, location]

Q2: Can you describe the spots on the leaves?

A: The spots are small, brown, round... maybe just a bit bigger than a pinhead.

[Cat: Symptoms | Ground: size, color, shape]

Q3: How does this compare to other diseases I might see?

A: Different from powdery mildew (no white powder) and bacterial leaf spot (no water-soaked edge).

[Cat: Differential Diagnosis | Ground: comparison, edge]

Q4: Could you explain what causes these spots?

A: Caused by the gall midge insect laying eggs... larvae form bumps that turn brown.

[Cat: Causes | Ground: insect, egg-laying]

Q5: Are these spots going to get worse?

A: Damage looks minor. Won't likely spread much more; tree should recover.

[Cat: Prognosis | Ground: spread, severity]

Q6: Do you notice anything else unusual besides the spots?

A: Everything else seems normal. No wilting, curled leaves; texture is smooth.

[Cat: Additional Cues | Ground: texture, wilting]

Q7: What can I do if I want to prevent these spots?

A: Ensure air circulation, keep area clean, avoid overhead watering.

[Cat: Prevention | Ground: air_circulation, cleanliness]



(b) Mango: Sooty Mold

Q1: How widespread is this sooty mould on my mango leaves?

A: Moderately spread across many parts... covering a good portion of surfaces.

[Cat: Severity | Ground: overall color, surface texture]

Q3: What are the common symptoms of Sooty Mold in mango?

A: Black, fuzzy coating on top side... rubs off easily... associated with honeydew.

[Cat: ID | Ground: colors, patterns]

Q2: Can you tell me how to control this on my mango trees?

A: Wash leaves with water... apply neem oil/soap (often follows insect infestation).

[Cat: Management | Ground: lesion characteristics, distribution]

Q4: Why did my mango leaves get these spots?

A: Likely insect infestation producing honeydew... colonized by fungi (e.g., aphids/scale).

[Cat: Causes | Ground: lesion characteristics, distribution]

Q5: What changes can I expect in the next few days?

A: Mould will continue to grow if conditions remain favorable. Check regularly.

[Cat: Future Outlook | Ground: distribution, lesion characteristics]

Q6: Is there anything else I need to worry about?

A: Obscures photosynthesis... ensure you manage underlying pests to prevent future issues.

[Cat: Impact/Management | Ground: leaf condition]

Figure 12: **Multi-disease Occurrence within a Single Crop Species.** The figure demonstrates distinct CoI trajectories for different pathologies affecting the same host (Mango). **(a) Gall Midge:** The dialogue focuses on structural damage (raised bumps), ruling out fungal pathogens via differential diagnosis, and identifying the insect vector. **(b) Sooty Mold:** The dialogue identifies a superficial fungal issue ("rubs off"), links it to secondary insect activity (honeydew), and focuses on cleaning the leaf surface to restore photosynthesis. This validates that **PlantInquiryVQA** generates context-specific reasoning chains even when the crop type is identical.



(a) Peach: Bacterial Spot

[EPISTEMIC INTENT: DIAGNOSIS]

Q1: What disease is affecting this peach leaf?

A: This is Bacterial Spot (*Xanthomonas campestris*). It appears as small, angular lesions that eventually cause the tissue to drop out.

[Cat: Identification | Ground: Angular spots, tissue loss]

Q2: How can I tell this isn't fungal "shot hole" disease?

A: Bacterial spots are typically more angular and concentrated near the leaf tip/margins. Fungal shot holes usually have distinct purple halos.

[Cat: Differential Diagnosis | Ground: Lesion shape, margins]

Q3: What specific visual cue confirms bacteria?

A: The "shot-hole" effect where dead tissue falls out, leaving ragged holes, without the presence of fungal spores.

[Cat: Symptomatology | Ground: Holes, ragged edges]



(b) Guava: Nutrient Deficiency

[EPISTEMIC INTENT: PROGNOSIS]

Q1: Do you think this will get worse if left untreated?

A: Given it's mild now, if we address nutrient levels, it should improve. Without intervention, it could spread to younger leaves.

[Cat: Early Stage Prognosis | Ground: Current severity, distribution]

Q2: What changes do you expect in the next few weeks?

A: You might see interveinal chlorosis intensifying. Older leaves may turn completely yellow and drop prematurely.

[Cat: Future Outlook | Ground: Chlorosis progression]

Q3: Are there any signs of the deficiency getting severe?

A: Not yet. The spotting is scattered and the leaf structure is intact. Severe cases would show necrotic browning.

[Cat: Progression | Ground: Leaf integrity, spotting density]



(c) Cauliflower: Downy Mildew

[EPISTEMIC INTENT: MANAGEMENT]

Q1: Is it too late to treat these Downy Mildew spots?

A: It's not too late, but act quickly. Damage is severe/irreversible, but treatment can stop spread to the head.

[Cat: Urgent Treatment | Ground: Severity, leaf condition]

Q2: What should I have done differently to prevent this?

A: Treated early when small pale spots appeared. Improved drainage and air circulation would have prevented this.

[Cat: Prevention | Ground: Lesion characteristics]

Q3: If I don't treat this immediately, what happens?

A: The leaf will turn completely brown and die within days. The infection will likely spread to the edible curd.

[Cat: Consequence Analysis | Ground: Leaf condition]

Figure 13: **Evolution of Epistemic Intent across Disease Severity.** The figure illustrates how the CoI shifts its reasoning goal based on the visual status of the plant. **(a) Diagnosis:** In the early/mild stage (Peach), the focus is on *Identification* and distinguishing symptoms from lookalikes. **(b) Prognosis:** In the mild/chronic stage (Guava), the inquiry shifts to *Predicting* the trajectory of the condition (recovery vs. decline). **(c) Management:** In the severe/terminal stage (Cauliflower), the reasoning pivots entirely to *Intervention*, focusing on salvage, containment, and future prevention strategies.



(a) Cherry: Healthy Control

[CONDITION: HEALTHY]

Q1: What crop is shown in this image?

A: Cherry.

[Cat: Crop ID]

Q2: Is this leaf healthy or diseased?

A: Healthy. It shows uniform green color, intact structure, no lesions or spots.

[Cat: Health Status | Ground: Uniform green color, intact structure]

Q3: How does this differ from one affected by fungal diseases?

A: Fungal diseases cause spots/lesions. The absence of these and presence of uniform green confirm this is healthy.

[Cat: Verification | Ground: Absence of lesions]

Q4: What visual signs would indicate disease on this Cherry leaf?

A: We would look for discoloration or abnormal growth. This leaf shows none of these indicators.

[Cat: Symptom Indicators | Ground: Intact tissue]



(b) Senescence / Dryness

[CONDITION: ABIOTIC STRESS]

Q1: Can you tell me what's causing these leaves to turn brown?

A: Due to senescence or dryness. The whole leaf is brown, dry, and papery.

[Cat: ID | Ground: Uniform brownish-tan, papery texture]

Q2: How severe is this condition right now?

A: Quite severe. Entire leaf is dry with no sign of recovery.

[Cat: Severity | Ground: Entire leaf discolored]

Q3: What might have led to such a severe state?

A: Likely lack of water, poor soil, or heat stress. Uniform browning suggests environmental stress.

[Cat: Cause | Ground: Uniform surface affection]

Q4: Do we need to act quickly to stop this from getting worse?

A: Yes. Without treatment, the leaf will drop off soon.

[Cat: Urgency | Ground: Dryness, no partial vitality]



(c) Bottle Gourd: Pest Damage

[CONDITION: PEST INFESTATION]

Q1: How much damage does this look like it's doing?

A: Moderate damage. Leaves are mostly healthy but have holes and curling edges.

[Cat: Severity | Ground: Leaf condition, holes]

Q2: Could these holes be from something other than pests, like sunburn?

A: No, these are clearly pest holes. Sunburn causes scorch marks, not neat-edged openings.

[Cat: Differential Diagnosis | Ground: Holes vs scorch marks]

Q3: What kind of treatment do you suggest?

A: Use vegetable-formulated insecticide or organic neem oil.

[Cat: Management | Ground: Lesion characteristics]

Q4: Do you think the damage will get worse if I don't do anything?

A: Yes, larger lesions may grow and affect more surface area.

[Cat: Prognosis | Ground: Lesion distribution]

Q5: How can I prevent this from happening again?

A: Crop rotation, row covers, and removing fallen debris.

[Cat: Prevention | Ground: Causes]

Figure 14: **Beyond Pathogenic Disease: Healthy, Abiotic, and Pest Conditions.** This figure illustrates the dataset's coverage of diverse plant health states. **(a) Healthy Control:** The model validates health by citing "uniform green color" and the absence of lesions. **(b) Senescence:** The inquiry identifies abiotic stress (aging/dryness) based on global uniform browning and papery texture, distinguishing it from focal infection. **(c) Pest Damage:** The reasoning chain differentiates physical damage (holes) from pathogenic spots and recommends insect-specific treatments (Neem oil) rather than fungicides.

Semantic Accuracy Evolution by Question Number - All Models

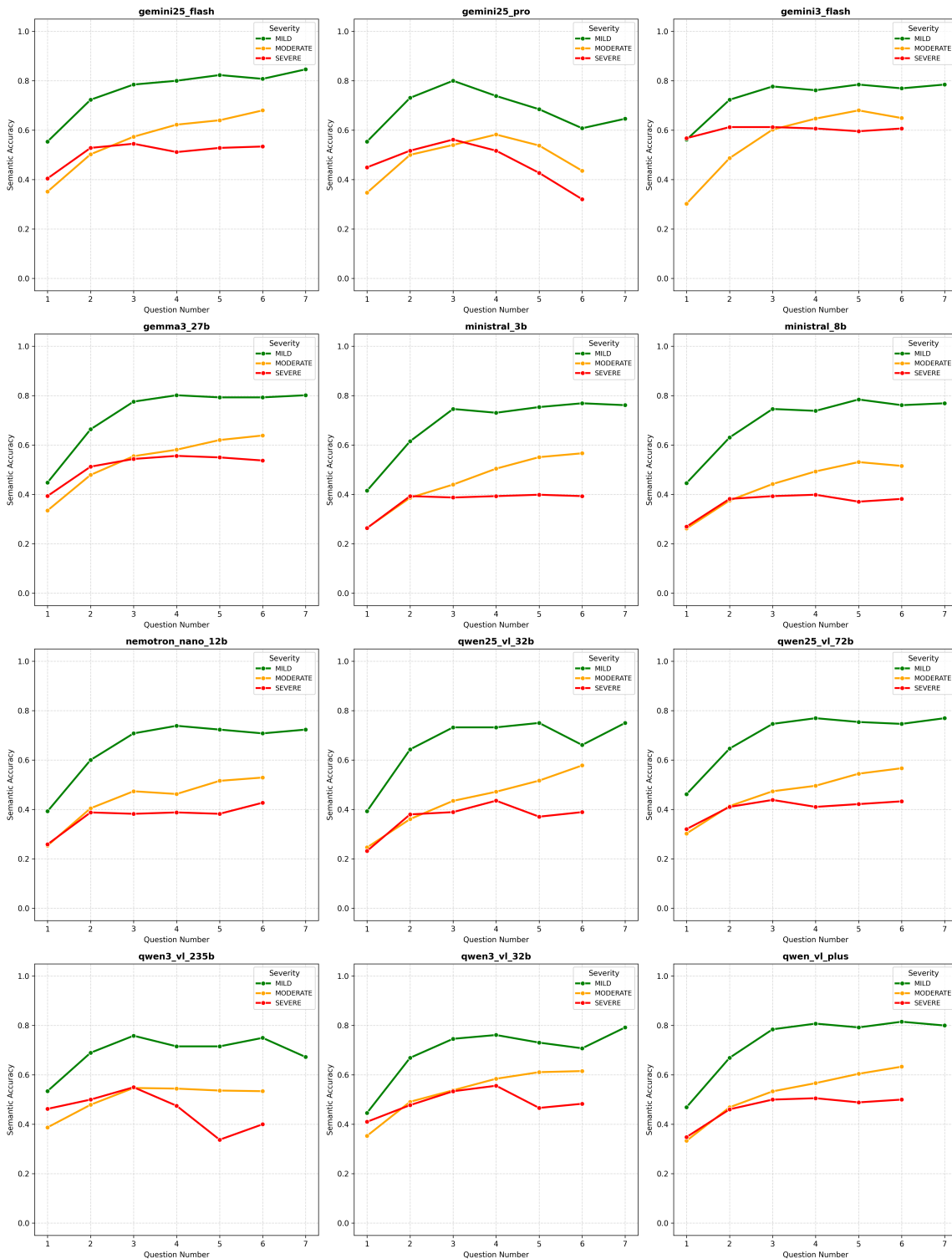


Figure 15: Semantic Accuracy Evolution across the Chain-of-Inquiry Trajectory. The figure illustrates the layer-wise diagnostic accuracy improvement for all 12 evaluated models as they progress through the 7-step diagnostic inquiry. Green lines indicate Mild infection, showing the strongest positive trajectory, while Red lines (Severe) indicate lower baselines and higher volatility. We observe a consistent positive trajectory, validating the hypothesis that structured inquiry refines diagnostic precision over time. Notably, the performance is stratified by disease severity: models consistently achieve the highest accuracy on Mild cases (green lines), where visual symptoms are distinct, but struggle with Severe cases (red lines), where extensive tissue necrosis often obscures the discriminative features required for accurate grounding.



Figure 16: A comprehensive analysis of diverse disease distribution across crops species of the final PlantInquiryVqa dataset