

# Agri-CM<sup>3</sup>: A Chinese Massive Multi-modal, Multi-level Benchmark for Agricultural Understanding and Reasoning

Haotian Wang<sup>1</sup>, Yi Guan<sup>1</sup>, Fanshu Meng<sup>1</sup>, Chao Zhao  
Lian Yan<sup>1</sup>, Yang Yang<sup>2</sup>, Jingchi Jiang<sup>1\*</sup>

<sup>1</sup>Harbin Institute of Technology <sup>2</sup>Changchun University Of Science And Technology  
{wanght1998, zhaochaocs}@gmail.com, {24s003058, 23b903008}@stu.hit.edu.cn,  
yangyang\_hit@cust.edu.cn, {guanyi, jiangjingchi}@hit.edu.cn

## Abstract

Multi-modal Large Language Models (MLLMs) integrating images, text, and speech can provide farmers with accurate diagnoses and treatment of pests and diseases, enhancing agricultural efficiency and sustainability. However, existing benchmarks lack comprehensive evaluations, particularly in multi-level reasoning, making it challenging to identify model limitations. To address this issue, we introduce Agri-CM<sup>3</sup>, an expert-validated benchmark assessing MLLMs' understanding and reasoning in agricultural management. It includes 3,939 images and 15,901 multi-level multiple-choice questions with detailed explanations. Evaluations of 45 MLLMs reveal significant gaps. Even GPT-4o achieves only 63.64% accuracy, falling short in fine-grained reasoning tasks. Analysis across three reasoning levels and seven compositional abilities highlights key challenges in accuracy and cognitive understanding. Our study provides insights for advancing MLLMs in agricultural management, driving their development and application. Code and data are available at <https://github.com/HIT-Kwoo/Agri-CM3>.

## 1 Introduction

The modernization and intensification of agriculture have led to increasing diversity and adaptability of pests and diseases, making their management increasingly challenging (Singh et al., 2024; Lenné and Wood, 2024). Traditional diagnostic methods primarily rely on manual observation and expert experience, often suffering from delays, low accuracy, and high costs (John et al., 2023). With the rapid advancement of artificial intelligence (AI), particularly multi-modal large language models (MLLMs), agricultural pest and disease management have entered a new era. By integrating multiple data modalities, MLLMs enable intelligent

\*Corresponding Author

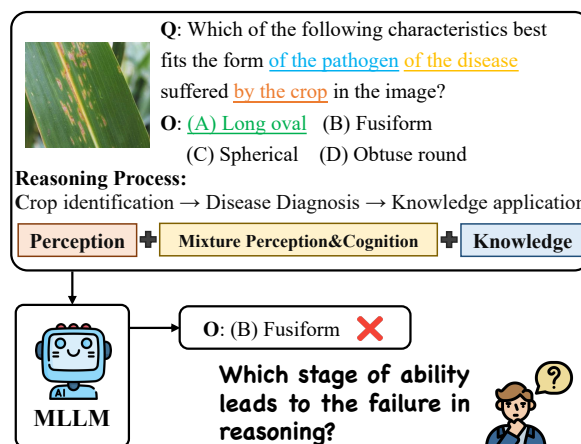


Figure 1: Existing benchmarks fail to identify the specific cause of failure in agricultural complex reasoning, making it difficult to distinguish whether errors stem from perception, mixed perception-cognition reasoning, or knowledge application.

recognition, prediction, and decision support, helping farmers optimize agricultural management, reduce pesticide use, and enhance sustainability (Zhu et al., 2024; Kuska et al., 2024).

Despite the impressive performance of state-of-the-art MLLMs (Hurst et al., 2024; Team et al., 2024; Anthropic, 2025) in vision-language tasks within the agricultural domain (Zhu et al., 2024), their practical effectiveness in pest and disease management remains unclear. A systematic benchmark is essential to evaluate their real-world applicability and reasoning capabilities in agricultural management. Existing benchmarks, such as AgriBench (Zhou and Ryo, 2024), focus primarily on broader agricultural vision tasks (e.g., land segmentation and captioning) but lack a dedicated evaluation framework for pest and disease management. Furthermore, current domain-specific benchmarks (Yue et al., 2024; Zhang et al., 2024b) fail to assess models' multi-level reasoning abilities, which are critical for complex decision-making in agricultural contexts. As illustrated in Figure 1, agri-

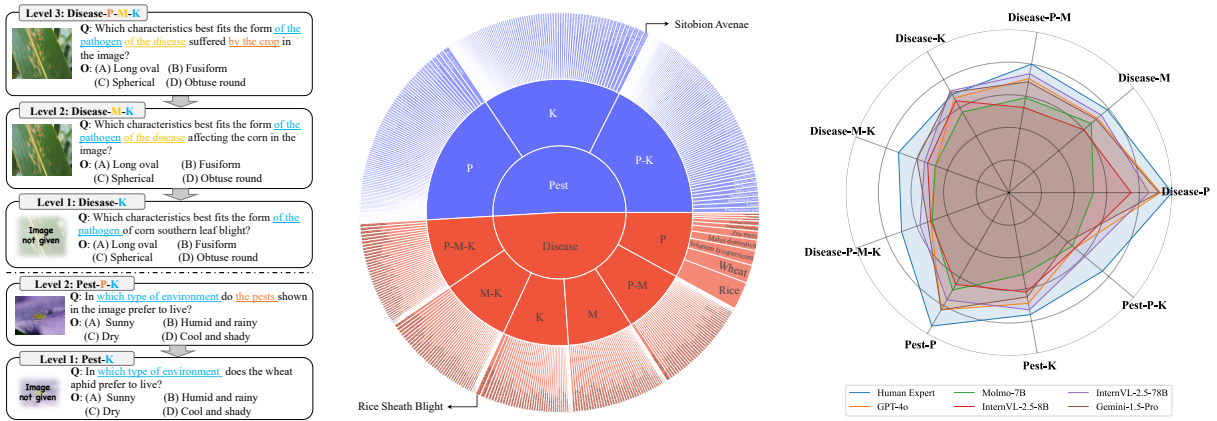


Figure 2: *Left*: The process of deriving high-level data with clearly specified entities into low-level data. *Middle*: Distribution of labels across the entire dataset. *Right*: Performance of 5 MLLMs and experts on the validation set.

cultural multi-modal reasoning tasks involve multiple stages, including crop identification, symptom recognition and disease diagnosis, and knowledge integration and application. Any failure at one stage may lead to an incorrect diagnosis. However, existing benchmarks do not effectively distinguish models’ performance at each reasoning step, making it difficult to diagnose the root causes of errors.

To tackle these challenges, we propose Agri-CM<sup>3</sup>, a novel benchmark designed to provide a comprehensive, hierarchical evaluation of MLLMs in agricultural pest and disease management. Specifically, we decompose complex reasoning tasks into three fine-grained sub-reasoning tasks: (1) Crop identification (Perception Stage). (2) Symptom recognition and disease diagnosis (Mixture of Perception-Cognition Reasoning Stage). (3) Agricultural knowledge integration and application (Knowledge Application Stage). These sub-tasks are then systematically organized as a reasoning hierarchy, as illustrated in Figure 2(Left): Level 1 evaluates individual sub-reasoning tasks. Level 2 assesses combinations of two sub-reasoning tasks. Level 3 examines the whole reasoning pipeline involving all three tasks. Each evaluation is structured as multiple-choice questions (MCQs) with detailed explanations, ensuring a rigorous assessment of MLLMs across different reasoning levels. This hierarchical design allows Agri-CM<sup>3</sup> to systematically analyze models’ performance at various reasoning levels and diagnose their specific weaknesses.

To construct Agri-CM<sup>3</sup>, we curated 3,939 high-quality images covering diverse pest and disease categories, rigorously filtered through model generation and expert validation. We designed 15,901

MCQs based on these images and conducted extensive evaluations on 38 open-source MLLMs and seven proprietary state-of-the-art MLLMs, revealing several **key insights**: (1) *Open-source MLLMs Surpassing Proprietary Models*. (2) *Mixture cognition-perception reasoning tasks are the key factor limiting model performance*. (3) *Multi-modal training boosts performance in single-text tasks*.

In summary, the key contributions of our work are as follows:

- To the best of our knowledge, we are the first multi-modal benchmark for agricultural pest and disease management, covering a wide range of pests and diseases. The benchmark is rigorously filtered and validated by experts, ensuring high quality.
- We decompose complex reasoning tasks and combine them based on practical needs, resulting in a multi-level benchmark. This framework allows for a comprehensive assessment of model capabilities at each reasoning stage, helping identify weaknesses in various steps.
- We evaluated various models and summarized several key insights. Our findings provide valuable guidance for future research and developments in this area.

## 2 Related Work

### 2.1 Multi-modal Large Language Models

MLLMs have introduced a new paradigm for reasoning, offering a natural and intuitive way to interact with users (Zhu et al., 2024). These models

have demonstrated outstanding potential for applications in various vertical domains (Xie et al., 2024; Xiao et al., 2025). Currently, leading proprietary MLLMs have achieved exceptional performance in multi-modal understanding and reasoning tasks. Meanwhile, the research community has developed various open-source MLLMs with multi-modal instruction-tuning datasets (Zhang et al., 2024a). Notable open-source models include the InternVL series (Chen et al., 2024b), LLaVA series (Liu et al., 2024), Qwen-VL series (Wang et al., 2024b), DeepSeek series (Wu et al., 2024), and MiniCPM-V series (Yao et al., 2024), which have been rapidly adopted across diverse domains.

To address domain-specific challenges, researchers have developed specialized MLLMs trained and fine-tuned on domain-specific datasets. For example, medical MLLMs (Li et al., 2023; Moor et al., 2023) and Financial MLLMs (Xie et al., 2024) have been proposed to enhance AI capabilities in their respective fields. In the agricultural domain, Wang et al. (2024a) fine-tuned a multi-modal agricultural model based on LLaVA (Liu et al., 2023b), demonstrating promising potential in pest and disease diagnosis. These advancements highlight the growing impact of MLLMs in agriculture, yet a standardized evaluation framework remains necessary to assess their real-world performance systematically.

## 2.2 MLLM Benchmarks

Despite the rapid progress of MLLMs, systematic benchmarking is equally crucial. It helps evaluate strengths and weaknesses, guiding optimization and accelerating real-world application. Several multi-modal benchmarks, such as MMMU (Yue et al., 2024) and CMMMU (Zhang et al., 2024b), cover multiple vertical domains. However, their domain-specific evaluation capabilities are limited. To address domain-specific needs, researchers have developed specialized benchmarks for evaluating MLLMs in fields such as medicine (Sun et al., 2025; Chen et al., 2024a) and industry (Jiang et al., 2025). In the agricultural domain, Zhou and Ryo (2024) proposed AgriBench, a multi-modal benchmark focusing on landscape understanding tasks. While AgriBench provides a general framework for evaluating MLLMs' multi-modal reasoning in agriculture, it lacks a dedicated assessment for pest and disease management. Furthermore, existing domain-specific benchmarks fail to distinguish which specific reasoning abilities contribute

to model failures in complex reasoning tasks. Identifying these deficiencies at a granular level is essential for improving MLLMs and guiding future model development.

To address this gap, we propose Agri-CM<sup>3</sup>, a benchmark designed to systematically evaluate MLLMs' hierarchical reasoning abilities in agricultural pest and disease management. Unlike existing benchmarks, Agri-CM<sup>3</sup> explicitly evaluates models across multiple reasoning levels, providing fine-grained insights into their strengths and weaknesses.

## 3 The Agri-CM<sup>3</sup> Benchmark

### 3.1 Design Principal

Existing benchmarks often evaluate complex reasoning tasks as a whole, especially for tasks involving multiple reasoning steps, such as agricultural knowledge reasoning. This approach fails to accurately identify model strengths and weaknesses at each reasoning stage, leading to a lack of clear guidance for future model improvements.

To address this limitation, we propose Agri-CM<sup>3</sup>, which aims to comprehensively assess the complex reasoning capabilities of models in agricultural pest and disease management through a multi-level evaluation framework. We decompose the complex reasoning task into three key sub-tasks: *i.* Perception (P): The ability to identify crops and pests. *ii.* Mixed Perception-Cognition Reasoning (M): The ability to identify crop symptoms and reason diseases. *iii.* Knowledge Application (K): The ability to integration and apply agricultural knowledge. These key abilities are then combined into three levels based on task-specific needs:

**Level 1:** Evaluates model performance on a single sub-reasoning task.

**Level 2:** Assesses the model's ability to perform a combination of two sub-reasoning tasks.

**Level 3:** Evaluates the model's performance across the full reasoning chain.

Through this layered design, Agri-CM<sup>3</sup> provides a comprehensive evaluation of a model's capabilities at different reasoning stages, offering detailed feedback on performance at each level. The overall data collection and generation process is illustrated in Figure 3.

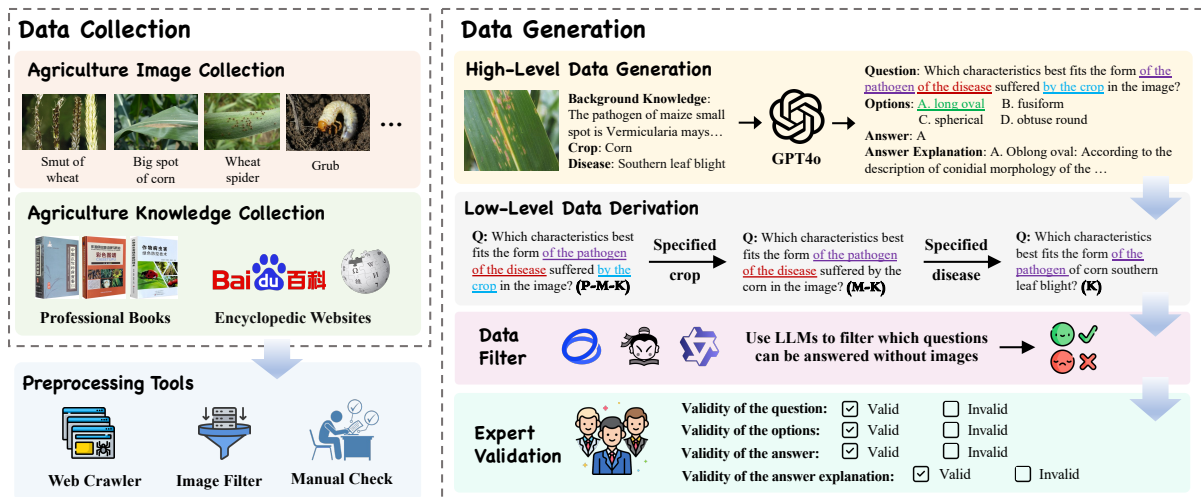


Figure 3: Overview of the Agri-CM<sup>3</sup> generation process. *Left*: The collection and processing of images and background knowledge. *Right*: Multi-level data generation, followed by filtering using LLMs, and final validation by agricultural experts.

### 3.2 Data Collection and Preprocessing

Traditional agricultural visual datasets typically focus on a single crop or pest type. However, an expert in agricultural pest and disease management should be capable of identifying and diagnosing various pests and diseases. To address this, we extensively collected existing agricultural pest and disease image datasets, resulting in 21 datasets, including 11 crops, 45 diseases, and 107 pests.

To ensure the diversity, we applied K-means (Arthur and Vassilvitskii, 2007) clustering to group images and filter out highly similar images. We filtered out low-resolution images to ensure that details are clearly visible, resulting in a final average resolution of  $1,800 \times 1,500$  for the retained images. Additionally, we enlisted professional agricultural experts to filter out unusable images manually. To prevent data leakage, we did not directly use image-question pairs from existing datasets. Instead, we collected images and their corresponding captions and employed GPT-4o to generate image-text question-answer pairs and other options. Besides, we sourced domain-specific agricultural knowledge from professional agricultural books and reputable encyclopedic websites, ensuring the model generated questions based on accurate, relevant knowledge. See the Appendix A.1 for more details.

### 3.3 Data Generation and Expert Validation

To ensure the accuracy and diversity of the generated data, we not only provide crop, disease, and pest labels but also select relevant background

knowledge to generate high-quality knowledge reasoning data. To facilitate manual validation of the generated data, we include answer explanation in the generated data, ensuring that each question is supported by a clear and justified rationale.

**High-level Data Generation.** We begin by generating high-level knowledge reasoning data based on agricultural pest and disease images and relevant background knowledge. The tasks include Disease-P-M-K, Disease-P-M, and Pest-P-K. These tasks require the model to perform multiple reasoning steps, including crop and pest identification, symptom recognition and disease diagnosis, followed by final knowledge reasoning. Pest identification involves only perceptual stage, with high-level data considering only P-K.

**Low-level Data Derivation.** After generating high-level data, we derive low-level data by explicitly replacing entities in the high-level data questions. This method allows us to generate data at different reasoning stages, thus providing more granular tasks that align with the evaluation needs of each reasoning level.

**Data Filtering.** To ensure that the generated data relies heavily on the image information, we use three LLMs to answer questions based on the question and options. We then filter out the data where two models provide correct answers, as these are deemed to have low dependency on the images. This ensures that the final dataset accurately reflects the image content.

Model	Validation Total	Test Total	Disease						Pest		
			P	M	P-M	K	M-K	P-M-K	P	K	P-K
Random Choice	23.22	25.14	25.44	26.04	25.44	24.15	24.31	26.61	25.04	24.94	24.89
Expert Performance	79.56	-	-	-	-	-	-	-	-	-	-
<b>Multi-Modal Large Language Model(MLLMs): Text + Image as Input</b>											
LLaVA-1.5-13B (Liu et al., 2023a)	41.33	38.19	43.87	53.00	52.41	40.52	42.54	42.38	23.40	33.31	36.47
Phi-3.5-Vision (Abdin et al., 2024)	46.11	41.69	39.14	57.99	53.85	44.10	44.66	45.18	30.01	38.50	40.29
Deepseek-VL2 (Wu et al., 2024)	46.33	42.91	36.18	47.93	41.5	48.12	48.47	45.23	37.28	44.17	42.56
Yi-VL-34B (AI et al., 2024)	48.78	45.98	49.79	52.83	38.46	51.25	48.43	44.25	42.68	46.50	44.76
CogVLM2 (Hong et al., 2024)	48.11	46.37	76.84	45.65	43.11	39.45	39.53	37.24	58.92	38.21	40.91
Cambrian-34B (Tong et al., 2024)	56.00	49.16	74.13	54.35	45.98	53.13	45.41	38.71	53.99	47.91	39.11
MiniCPM-V2.6 (Yao et al., 2024)	54.67	52.24	65.34	58.58	54.35	49.02	46.70	45.96	64.05	44.85	45.42
GLM4V-9B (GLM et al., 2024)	56.00	52.65	<b>86.73</b>	55.62	56.04	50.72	47.84	47.71	57.79	43.32	44.17
LLaVA-1.6-34B (Liu et al., 2024)	57.78	52.71	59.09	59.51	54.35	57.60	51.96	50.47	54.58	48.51	47.76
Qwen2.5-VL-7B (Bai et al., 2025)	56.44	53.67	76.16	53.85	50.38	49.46	45.33	43.75	73.51	45.70	44.87
LLaVA-OneVision-7B (Li et al., 2024)	57.00	53.92	70.33	58.66	53.00	51.25	50.78	50.54	64.87	45.86	46.41
Molmo-7B (Deitke et al., 2024)	55.89	54.76	54.69	66.19	59.93	52.50	49.92	49.77	67.41	47.91	47.54
InternVL-2.5-8B (Chen et al., 2024b)	59.67	55.73	70.58	62.13	57.06	55.37	52.43	50.54	64.16	49.20	48.13
Qwen2-VL-7B (Wang et al., 2024b)	60.00	56.32	<u>85.88</u>	58.92	57.82	54.20	47.88	46.27	70.34	47.43	46.26
InternVL-2.5-26B (Chen et al., 2024b)	64.78	61.97	74.30	71.09	66.61	63.69	57.06	56.07	72.57	54.91	51.50
Molmo-72B (Deitke et al., 2024)	66.33	63.13	63.14	73.96	71.01	<b>66.01</b>	60.75	59.56	68.86	<b>58.37</b>	55.61
LLaVA-OneVision-72B (Li et al., 2024)	69.00	64.37	78.61	72.19	68.98	65.38	60.99	58.79	70.07	57.32	<b>57.66</b>
Qwen2.5-VL-72B (Bai et al., 2025)	67.78	64.70	84.02	68.64	67.03	63.33	56.86	54.46	<b>79.11</b>	55.47	53.41
Qwen2-VL-72B (Wang et al., 2024b)	<b>72.00</b>	64.81	80.98	73.29	<u>72.36</u>	<u>65.74</u>	<b>61.77</b>	<u>59.95</u>	72.07	55.07	56.27
InternVL-2.5-38B (Chen et al., 2024b)	68.11	<u>65.31</u>	74.81	<b>76.67</b>	70.16	65.65	<u>61.30</u>	<b>60.26</b>	<u>75.04</u>	56.84	<u>56.89</u>
InternVL-2.5-78B (Chen et al., 2024b)	<u>69.89</u>	<b>65.49</b>	79.54	<b>76.67</b>	<b>72.78</b>	65.56	60.44	59.33	73.04	<u>58.00</u>	56.38
Qwen-VL-PLUS (Wang et al., 2024b)	51.67	51.54	80.56	47.76	48.27	47.14	45.29	42.69	68.39	40.23	45.42
Yi-Vision (AI et al., 2024)	61.56	56.59	60.27	58.66	55.45	60.82	53.61	51.01	70.31	51.85	48.35
Qwen-VL-MAX (Wang et al., 2024b)	59.44	57.97	76.84	60.19	60.19	62.25	52.04	48.83	66.12	53.46	49.67
GPT-4o-mini (Hurst et al., 2024)	63.00	58.21	79.88	<b>69.99</b>	67.12	53.49	52.28	48.37	68.78	48.87	47.80
Gemini-1.5-flash (Team et al., 2024)	64.67	61.73	<u>87.57</u>	68.55	<u>69.74</u>	57.33	52.51	<u>52.72</u>	<b>80.40</b>	49.92	47.73
GPT-4o (Hurst et al., 2024)	<u>67.33</u>	<u>63.64</u>	83.85	69.48	68.30	<b>65.65</b>	<u>54.71</u>	49.53	<u>80.09</u>	<u>56.84</u>	<u>51.10</u>
Gemini-1.5-pro (Team et al., 2024)	<b>67.78</b>	<b>64.73</b>	<b>89.18</b>	<u>69.82</u>	<b>70.33</b>	<u>62.43</u>	<b>57.30</b>	<b>54.20</b>	78.72	<b>57.36</b>	<b>52.49</b>
<b>Large Language Models (LLMs): Only Text as Input</b>											
GLM4-9B (GLM et al., 2024)	<u>41.89</u>	35.37	34.66	38.55	27.98	45.26	40.03	38.72	21.56	41.67	36.91
Qwen2.5-7B (Hui et al., 2024)	40.67	35.37	<u>36.43</u>	<u>44.04</u>	19.78	<u>48.21</u>	40.89	37.48	20.74	42.04	36.69
InternLM2.5-7B (Cai et al., 2024)	40.44	36.55	<b>44.55</b>	39.73	32.12	46.33	39.72	38.34	22.38	<u>42.12</u>	35.48
GPT-3.5-Turbo (OpenAI, 2023)	39.11	<u>37.01</u>	32.04	<b>44.72</b>	<u>32.29</u>	45.80	<u>43.33</u>	<u>39.04</u>	<b>23.28</b>	41.55	<u>39.11</u>
GPT-4-Turbo (Achiam et al., 2023)	<b>47.11</b>	<b>41.84</b>	33.39	43.20	<b>35.08</b>	<b>57.07</b>	<b>52.83</b>	<b>46.27</b>	<u>23.08</u>	<b>50.36</b>	<b>44.21</b>

Table 1: The zero-shot result of different models in Agri-CM<sup>3</sup>. The best results are highlighted in **bold**, while the second-best is underlined. The abbreviation in Agri-CM<sup>3</sup> means P: Perception, M: Mixture of Perception and Cognition Reasoning, and K: Knowledge Application. Test Total represents the average of all test instances.

**Expert Validation.** To guarantee the quality of the generated data, we employed two agricultural experts for manual validation. Initially, we randomly selected 200 data samples for annotation, including questions, options, answers and answers explanations. By evaluating the consistency of their annotations, if the consistency is below 90%, we retrain the experts on the inconsistent annotated data and reassess their consistency until it exceeds 90%. Only after reaching this threshold can we proceed with the formal annotation process. For detailed annotation procedures, please refer to the Appendix A.3.

### 3.4 Data Analysis

**Base Statistics.** The Agri-CM<sup>3</sup> contains 3,939 high-quality images and 15,901 multiple-choice

questions across three reasoning levels, six disease tasks, and three pest tasks. 900 samples, 100 from each subclass, form the validation set, with the remaining 15,001 used for testing. More details in Appendix A.4.

**Multi-modal diversity.** Figure 11 illustrates the semantic space representation of images encoded by CLIP (Radford et al., 2021). The results demonstrate that Agri-CM<sup>3</sup> spans a broad semantic space, highlighting its rich semantic diversity and comprehensive coverage.

## 4 Experiments

### 4.1 Experiments Setting

In this study, we conduct a systematic evaluation of various MLLMs, including 38 open-source and

seven proprietary models, to assess their performance on the Agri-CM<sup>3</sup> benchmark. All evaluations are conducted in a zero-shot setting, meaning that the models do not receive any example prompts during reasoning and have not been trained explicitly on tasks related to the benchmark. This approach allows us to measure the generalization ability of MLLMs in agricultural pest and disease management and evaluate their reasoning capabilities on unseen tasks. All experiments are executed on 4 NVIDIA H800 GPUs (80GB memory). Detailed experimental settings and hyperparameter configurations are provided in the Appendix B.

## 4.2 Main Results

In this section, we comprehensively compare various LLMs and MLLMs using the Agri-CM<sup>3</sup> benchmark. For clarity, we present results from selected models in Table 1, while the complete experimental results and data can be found in the Appendix A.4 and C. Based on these results, we summarize the following key findings:

**Agri-CM<sup>3</sup> remains a challenging benchmark.** The results from Table 1 show that 26 out of the 35 MLLMs achieved a performance below 60%. The top-performing model on the test set reached 65.49%, demonstrating that real-world applications in agricultural pest and disease management still face significant challenges. Specifically, the models showed limited success in high-level reasoning tasks, with the Disease-P-M-K and Pest-P-K tasks achieving only 59.33% and 56.38%, respectively. These results highlight substantial room for improvement, particularly in complex reasoning.

**Open-source MLLMs gradually outperform proprietary models.** The top-performing open-source model, InternVL-2.5-78B, achieved 65.49% on the test set, outperforming the best proprietary model, Gemini-1.5-Pro, by 0.73%. This result suggests that the application potential of open-source MLLMs in agricultural tasks is steadily increasing, and in some cases, they have already surpassed proprietary models. It also highlights the importance of open-source technologies in the future of agricultural intelligence.

**Optimal MLLMs performance still below expert level.** We evaluated the performance of MLLMs and agricultural experts on the validation set. As shown in Table 1 and Figure 2 (right), although the optimal MLLM has demonstrated strong per-

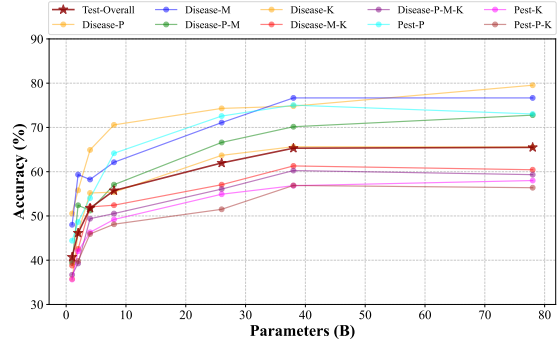


Figure 4: Scaling law of model size in Agri-CM<sup>3</sup>. We use the InternVL-2.5 series as examples.

formance on certain tasks, there remains a performance gap of approximately 8% compared to experts. This indicates that while current MLLMs have made notable progress in agricultural tasks, further optimization is needed to integrate high-level reasoning and domain-specific knowledge.

Besides, the experts' performance did not meet the expected level of excellence. Although agricultural experts have in-depth knowledge of specific crops or pests, their performance decreases when dealing with a broader range of crops and diseases due to the specialized nature of their research. This suggests that while experts have substantial advantages in specific areas, both their performance and that of the models are limited when faced with broader agricultural scenarios.

## 4.3 Further Analysis

**Larger models improve performance, but with diminishing returns.** We conducted experiments using InternVL-2.5 with different parameter sizes, as shown in Figure 4. The results indicate that as the model size increases, overall performance improves significantly. For tasks across different reasoning levels, larger models consistently outperform smaller ones, with an average performance gap of 24.78% between the largest and smallest models. However, as the parameter size further increases, the rate of performance improvement gradually slows, exhibiting diminishing returns. This suggests that although larger models enhance accuracy, the benefits of increasing model size become less pronounced.

**Mixed-reasoning tasks are the key factor limiting model performance.** To explore the impact of different reasoning levels on model performance, we use Disease-P-M-K as a case study. By successively specifying entities in the question: starting

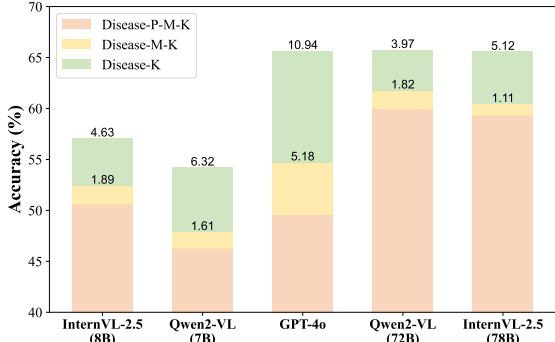


Figure 5: Performance of different models at different reasoning levels.

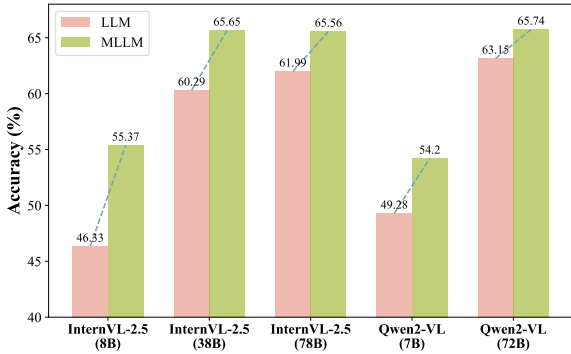


Figure 6: Performance of different MLLMs and their corresponding LLMs on Disease-K (text modality).

from Disease-P-M-K, then specifying crop to create Disease-M-K, and further specifying disease to obtain Disease-K, we investigate how different reasoning levels influence model performance. The experimental results are shown in Figure 5. The results indicate that eliminating perception reasoning (Disease-M-K) leads to only a modest performance improvement of approximately 2.32% on average. However, when mixed perception-cognition reasoning is further removed (Disease-K), the performance increase becomes more significant, with an average improvement of 6.00%. This suggests that mixed-reasoning tasks are more challenging for the models, and improving their ability in this area is key to further enhancing model performance.

**Multi-modal training boosts performance in single-text tasks.** In this benchmark, the Disease-K and Pest-K tasks focus on evaluating pest and disease knowledge in a single-text modality. We compared the performance of MLLMs and their corresponding LLMs on these tasks, as shown in Figure 6 and 12. The results demonstrate that, after multi-modal training, MLLMs exhibit significant improvements in single-text modality tasks compared to their corresponding LLMs, with an

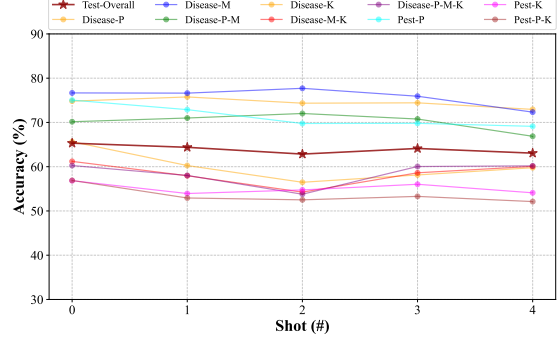


Figure 7: In-Context Learning performance of InternVL-2.5-38B under different shots.

average performance increase of 5.06% and 4.01%. This phenomenon indicates that multi-modal training not only enhances models' ability for joint reasoning over images and text but also improves their adaptability and generalization to complex contexts through image information processing, align with the conclusions of Ji et al. (2024).

#### 4.4 Exploration

**The structured reasoning process enabled by CoT leads to superior performance.** We applied the CoT (Wei et al., 2022) method to enable step-by-step thinking and assessed its impact on performance. As shown in Table 2, GPT-4o demonstrated significant improvement, particularly in complex multi-step reasoning tasks. Compared to other models, GPT-4o effectively follows the multi-step process required by the benchmark: first identifying the crop, diagnosing the disease, and then performing knowledge integration and application based on the options. Other models, lacking such structured reasoning, tend to provide answers without step-by-step logic, resulting in less pronounced CoT performance. A detailed case analysis is provided in Appendix D.1.

**In-context learning cannot stably improve performance.** In this experiment, we conducted in-context learning with different shot counts using InternVL-2.5-38B, as shown in Figure 7. The results indicate that, despite some fluctuations in accuracy across different shot counts, there is no clear upward trend or consistent performance improvement. The main reason for this may be that as the number of images increases, the model could become confused due to the additional information, leading to potential interference. Therefore, this issue is likely related to the selection of examples in in-context learning. Future work could further

Model	CoT	Validation Total	Test Total	Disease						Pest		
				P	M	P-M	K	M-K	P-M-K	P	K	P-K
InternVL-2.5-8B	✗	59.67	55.73	70.58	62.13	57.06	55.37	52.43	50.54	64.16	49.20	48.13
	✓	58.22 ↓1.5	55.30 ↓0.4	69.32 ↓1.3	61.62 ↓0.5	59.59 ↑2.5	50.89 ↓4.5	52.83 ↑0.4	51.09 ↑0.6	64.71 ↑0.5	46.78 ↓2.4	48.53 ↑0.4
InternVL-2.5-38B	✗	68.11	65.31	74.81	76.67	70.16	65.65	61.30	60.26	75.04	56.84	56.89
	✓	64.56 ↓3.5	59.89 ↓5.4	73.29 ↓1.5	72.02 ↓4.7	64.07 ↓6.1	58.05 ↓7.6	55.97 ↓5.3	53.34 ↓6.9	70.15 ↓4.9	52.86 ↓4	49.45 ↓7.4
InternVL-2.5-78B	✗	69.89	65.49	79.54	76.67	72.78	65.56	60.44	59.33	73.04	58.00	56.38
	✓	67.56 ↓2.3	61.71 ↓3.8	79.37 ↓0.2	74.22 ↓2.5	69.23 ↓3.5	61.81 ↓3.8	55.42 ↓5.0	55.52 ↓3.8	69.87 ↓3.2	53.86 ↓4.1	50.70 ↓5.7
Qwen2-VL-7B	✗	60.00	56.32	85.88	58.92	57.82	54.20	47.88	46.27	70.34	47.43	46.26
	✓	60.00 ↓0.0	55.87 ↓0.5	79.71 ↓6.2	62.55 ↓3.6	61.12 ↓3.3	51.61 ↓2.6	48.04 ↓0.2	47.59 ↓1.3	69.41 ↓0.9	45.90 ↓1.5	46.08 ↓0.2
Qwen2-VL-72B	✗	72.00	64.81	80.98	73.29	72.36	65.74	61.77	59.95	72.07	55.07	56.27
	✓	67.44 ↓4.6	62.19 ↓2.6	79.46 ↓1.5	71.17 ↓2.1	66.53 ↓5.8	65.21 ↓0.5	55.57 ↓6.2	55.44 ↓4.5	70.42 ↓1.6	54.83 ↓0.2	52.97 ↓3.3
GPT-4o-mini	✗	63.00	58.21	79.88	69.99	67.12	53.49	52.28	48.37	68.78	48.87	47.80
	✓	61.78 ↓1.2	56.14 ↓2.1	82.25 ↑2.4	68.30 ↓1.7	66.78 ↓0.3	52.06 ↓1.4	48.12 ↓4.2	46.03 ↓2.3	64.67 ↓4.1	46.66 ↓2.2	45.75 ↓2.0
GPT-4o	✗	67.33	63.64	83.85	69.48	68.30	65.65	54.71	49.53	80.09	56.84	51.10
	✓	70.56 ↑3.2	65.43 ↑1.8	87.91 ↑4.1	78.95 ↑9.5	75.23 ↑6.9	61.45 ↓4.2	58.48 ↑3.8	54.04 ↓4.5	80.13 ↑0.0	53.94 ↓2.9	52.49 ↑1.4

Table 2: Performance comparison of different MLLMs with and without CoT.

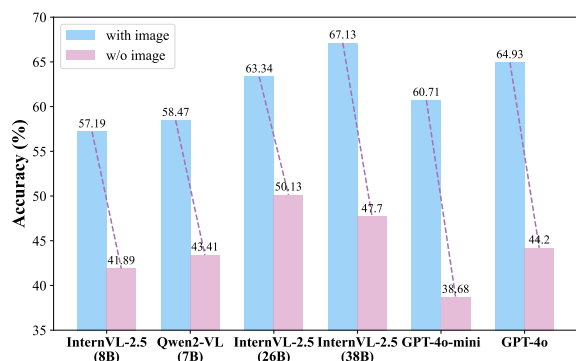


Figure 8: The average performance of different models on seven image-dependent tasks, comparing results with and without the use of images.

explore the selection of examples.

**The dataset relies heavily on visual input.** To assess the dataset’s dependency on the visual modality, we conducted experiments on datasets requiring image input for responses, comparing model performance with and without image input. Specifically, we evaluated MLLMs of different scales under both settings. The results, shown in Figure 8, indicate a significant performance drop across all models when visual information is removed. This finding demonstrates that the dataset is highly dependent on visual input, as models require image information for practical reasoning.

## 5 Error Analysis and Future Work

In this section, we analyze the errors of GPT-4o CoT predictions by randomly sampling 200 incorrect predictions from the most challenging pest and disease reasoning tasks. These errors were annotated and categorized by agricultural experts, and the distribution is shown in Figure 9. Analysis reveals that the primary cause of errors is the lack of

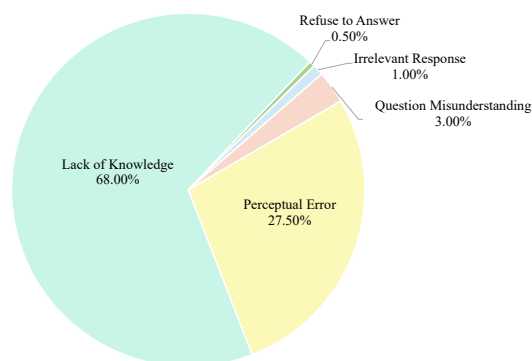


Figure 9: Overall error distribution for 200 annotated GPT-4o errors.

domain-specific knowledge, particularly in disease reasoning tasks. Additionally, in pest reasoning tasks, the model’s insufficient expertise leads to inaccurate pest identification and noticeable perceptual errors, highlighting the limitations of current models: *insufficient training on specialized agricultural knowledge*. To address these issues, future work should focus on enhancing domain knowledge. Overcoming these challenges will improve the performance of multi-modal models in agricultural disease and pest reasoning tasks. Detailed case studies and analyses are provided in the Appendix D.2.

## 6 Conclusion

In this study, we introduce Agri-CM<sup>3</sup>, an innovative multi-level benchmark designed to comprehensively evaluate the complex reasoning capabilities of MLLMs in agricultural pest and disease management, filling a gap in the field. This multi-level design not only evaluates the model’s ability in individual sub-reasoning tasks but also tests its performance in combined reasoning tasks and complete



reasoning chains, offering clear direction for future model optimization. Additionally, Agri-CM<sup>3</sup> provides valuable insights for designing and evaluating MLLMs. Our study emphasizes the importance of integrating visual and language modalities and validates the crucial role of domain-specific knowledge in vertical tasks.

## Limitations

In this study, we introduced Agri-CM<sup>3</sup>, a multi-level benchmark designed to evaluate the complex reasoning abilities of MLLMs in agricultural pest and disease management. Although we have considered various factors, there are still some limitations. Firstly, the current benchmark is based on the Chinese language, which limits its applicability to multilingual tasks. To enhance its usability across language communities, we employed GPT-4o to translate the dataset into English, enabling broader participation from the global research community. However, potential translation biases and semantic shifts introduced by the model may affect a small portion of the content. Secondly, agricultural experts typically specialize in specific areas, which may impact the accuracy of annotations for tasks outside their area of expertise. This specialization bias could lead to annotation inconsistencies, especially for tasks involving different crops or pests. To address this, future efforts may incorporate multi-domain experts or cross-domain knowledge to enhance annotation accuracy and consistency.

## Ethical Considerations

In constructing the Agri-CM<sup>3</sup> benchmark, we leveraged GPT-4o to assist in both translating Chinese content into English and generating synthetic question-answer pairs based on agricultural images and structured domain knowledge. While this approach improves scalability and consistency, it also introduces potential ethical concerns.

First, GPT-4o may encode biases from its pre-training data as a proprietary model, particularly regarding domain-specific or culturally nuanced content. Such biases may affect translation fidelity, domain framing, or answer correctness. This is particularly relevant in agriculture, where regional terminology and knowledge may be underrepresented. Second, although all generated content was validated by human experts, the filtering and editing of LLM outputs remain subject to human

limitations. As a result, some edge-case errors or subtle biases may persist. Third, since the benchmark was originally constructed in Chinese and then translated, there is a risk of semantic drift or misinterpretation during translation. Lastly, all image data were sourced from public datasets or used under fair academic use, with no personal or sensitive information involved. Nonetheless, we acknowledge the importance of ongoing ethical reflection on data provenance, model transparency, and equitable access.

Future iterations of Agri-CM<sup>3</sup> will incorporate multilingual sources, fairness metrics, and explicit documentation of observed biases to improve transparency and inclusiveness.

## Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work was supported in part by the Science and Technology Innovation Award of Heilongjiang Province, China [JD2023GJ01], the Key Research and Development Program of Heilongjiang Province, China [2024ZX01A07], and the National Natural Science Foundation of China [72293584, 72431004]. Jingchi Jiang is the corresponding author of this work.

## References

- Petchiammal A, Briskline Kiruba S, Murugan D, and Pandarasamy Arjunan. 2022. [Paddy doctor: A visual image dataset for automated paddy disease classification and benchmarking](#).
- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. [Yi: Open foundation models by 01.ai](#). *Preprint, arXiv:2403.04652*.

- Anthropic. 2025. [Claude 3.5 sonnet](#). Accessed: 2025-02-07.
- David Arthur and Sergei Vassilvitskii. 2007. `k-means++`: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07*, page 1027–1035, USA. Society for Industrial and Applied Mathematics.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.
- Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. 2024a. [GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.
- Fei Gao, Xiaoli Wang, Tingting Liu, Zhuang Li, and Rui Man. 2020. Spectral and imaging datasets of apple leaf disease and insect pests in china in 2015. *Journal of Agricultural Big Data*, 2(4):120–124.
- Hawi Getachew. 2021. [Wheat leaf dataset](#). Mendeley Data, V1. Accessed: 2025-02-14.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Lakshay Goyal, Chandra Mani Sharma, Anupam Singh, and Pradeep Kumar Singh. 2021. [Leaf and spike wheat disease detection classification using an improved deep convolutional architecture](#). *Informatics in Medicine Unlocked*, 25:100642.
- Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, and Yaodong Yang. 2024. [Align anything: Training all-modality models to follow instructions with language feedback](#).
- Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. 2025. [MMAD: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection](#). In *The Thirteenth International Conference on Learning Representations*.
- Matthew Abu John, Ibukunoluwa Bankole, Oluwatayo Ajayi-Moses, Tofunmi Ijila, Oluwatimilehin Jeje, and Patil Lalit. 2023. Relevance of advanced plant disease detection techniques in disease and pest management for ensuring food security and their implication: A review. *American Journal of Plant Sciences*, 14(11):1260–1295.
- Matheus Thomas Kuska, Mirwaes Wahabzada, and Stefan Paulus. 2024. [AI for crop production – where can large language models \(llms\) provide substantial value?](#) *Computers and Electronics in Agriculture*, 221:108924.
- Jillian Lenné and David Wood. 2024. Crop diversity in agroecosystems for pest management and food production. *Plants*, 13(8):1164.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yanzhong Li. 2023. [Atlas of Alfalfa Diseases, Insect Pests, and Natural Enemies and Dataset of Major Alfalfa Diseases in Heilongjiang Province \(2015-2016\)](#).
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Md Mafiul Hasan Matin Mafi, Mst. Arifa Anjum Ava, Md. Golam Moazzam Moazzam, and Mohammad Shorif Uddin. 2023. [Eggplant disease recognition dataset](#). Mendeley Data, V2. Accessed: 2025-02-14.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. [Med-flamingo: a multimodal medical few-shot learner](#). In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.
- OpenAI. 2023. [Introducing GPT-3.5 & GPT-4](#). Accessed: 2025-02-04.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PMLR.
- Jitesh Shah, Harshadkumar Prajapati, and Vipul Dabhi. 2017. [Rice Leaf Diseases](#). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5R013>.
- Amrita Singh, Nikita Shraogi, Rahul Verma, Joel Saji, Aditya Kumar Kar, Sarita Tehlan, Debabrata Ghosh, and Satyakam Patnaik. 2024. [Challenges in current pest management practices: Navigating problems and a way forward by integrating controlled release system approach](#). *Chemical Engineering Journal*, 498:154989.
- 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, CoDS COMAD 2020*, page 249–253, New York, NY, USA. Association for Computing Machinery.
- Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. 2025. [Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology](#). In *Computer Vision – ECCV 2024*, pages 56–73, Cham. Springer Nature Switzerland.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. [Cambrian-1: A fully open, vision-centric exploration of multimodal llms](#). *arXiv preprint arXiv:2406.16860*.
- Liqiong Wang, Teng Jin, Jinyu Yang, Ales Leonardis, Fangyi Wang, and Feng Zheng. 2024a. [Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases](#). *arXiv preprint arXiv:2412.02158*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *arXiv preprint arXiv:2409.12191*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Tyr Wiesner-Hanks, Ethan L Stewart, Nicholas Kaczmar, Chad DeChant, Harvey Wu, Rebecca J Nelson, Hod Lipson, and Michael A Gore. 2018. [Image set for deep learning: field images of maize annotated with disease symptoms](#). *BMC research notes*, 11:1–3.
- Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. 2019. [Ip102: A large-scale benchmark dataset for insect pest recognition](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8787–8796.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). *arXiv preprint arXiv:2412.10302*.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2025. [A comprehensive survey of large language models and multimodal large language models in medicine](#). *Information Fusion*, 117:102888.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, et al. 2024. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. [MM-LLMs: Recent advances in MultiModal large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. 2024b. Cmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Xianghe Zhang, Xiaoli Wang, Tingting Liu, Lin Hu, and Jingchao Fan. 2022. Image data set of six common orchard pests such as *bactrocera dorsalis*. *Journal of Agricultural Big Data*, 4(1):114–118.

Yutong Zhou and Masahiro Ryo. 2024. Agribench: A hierarchical agriculture benchmark for multimodal large language models. *arXiv preprint arXiv:2412.00465*.

Hongyan Zhu, Shuai Qin, Min Su, Chengzhi Lin, Anjie Li, and Junfeng Gao. 2024. Harnessing large vision and language models in agriculture: A review. *arXiv preprint arXiv:2407.19679*.

## A Dataset Agri-CM<sup>3</sup> Details

### A.1 Data Collection

To ensure high-quality image data, we collected a large number of images from existing agricultural pest and disease datasets and related websites, resulting in 21 distinct pest and disease datasets (see

Table 3). We filtered the images based on resolution and excluded those with fewer instances. Additionally, we applied K-means clustering to eliminate highly similar images, ensuring dataset diversity. In the end, we retained 5,727 high-quality images. To further validate the images, two agricultural experts were hired to review and filter out the following categories of unusable images:

- i. Images containing crop, disease, pest names, or symptom-related terms.
- ii. Images with large watermarks obscuring key features of pests and diseases.
- iii. Images unrelated to pests and diseases.

After filtering, we obtained 3,939 high-quality agricultural pest and disease images with crop and pest labels. These images formed the foundation for generating subsequent question-answer data.

To ensure the quality of the generated content and prevent hallucinations, we gathered domain-specific background knowledge related to pests and diseases. The sources included professional agricultural books, Baidu Baike<sup>1</sup>, and Wikipedia<sup>2</sup>, totaling 5.2 million words of specialized content. This knowledge was structured into a database for reference by GPT-4o during data generation, providing a solid foundation for generating precise and reliable question-answer data.

### A.2 Data Generation

Firstly, we use GPT-4o to generate high-level complex reasoning data based on agricultural pest and disease images, crop, pest, and disease captions, along with relevant agricultural background knowledge. Based on this high-level data, we progressively derive low-level complex reasoning data, with the derivation sequence shown in Table 4 and the prompts provided in Prompt A.1, A.2 A.3, and A.4. To ensure the accuracy of the generated data, we provide an example to guide the model towards generating more accurate results.

It is important to note that the pest recognition task only involves the perception stage, and therefore, the mixture of perception-cognition reasoning stage is not included in the generation of high-level pest data.

### A.3 Expert Validation

To ensure the quality of the generated data, we incorporated external domain-specific background

<sup>1</sup><https://baike.baidu.com/>

<sup>2</sup><https://zh.wikipedia.org/>

Dataset Name	Dataset Type	# Crop	# Disease/Pest	Link
PlantDoc-Dataset (Singh et al., 2020)	Disease	13	28	<a href="https://github.com/pratikayal/PlantDoc-Dataset">https://github.com/pratikayal/PlantDoc-Dataset</a>
Rice disease image classification	Disease	1	5	<a href="https://www.kaggle.com/nurnob101/rice-disease">https://www.kaggle.com/nurnob101/rice-disease</a>
Rice Diseases	Disease, Pest	1	4	<a href="https://www.kaggle.com/renzmarrionperea/ricedisease">https://www.kaggle.com/renzmarrionperea/ricedisease</a>
ai_challenger_pdr2018	Disease	10	40	<a href="https://aistudio.baidu.com/datasetdetail/76075">https://aistudio.baidu.com/datasetdetail/76075</a>
Agricultural Pest Research Base	Disease, Pest	8	40	<a href="http://www.icgroupcas.cn/website_bchtk/fuwu.html">http://www.icgroupcas.cn/website_bchtk/fuwu.html</a>
CGIAR Computer Vision for Crop Disease	Disease	1	3	<a href="https://www.kaggle.com/datasets/shadabhussain/cgiar-computer-vision-for-crop-disease">https://www.kaggle.com/datasets/shadabhussain/cgiar-computer-vision-for-crop-disease</a>
IP102 v1.1 (Wu et al., 2019)	Pest	8	102	<a href="https://github.com/xpwu95/IP102">https://github.com/xpwu95/IP102</a>
Wheat Leaf Dataset (Getachew, 2021)	Disease	1	3	<a href="https://data.mendeley.com/datasets/wgd66f8n6h/1">https://data.mendeley.com/datasets/wgd66f8n6h/1</a>
Large Wheat Disease Classification Dataset (Goyal et al., 2021)	Disease	1	4	<a href="https://drive.google.com/drive/folders/10HKtwD1UrdmhqxrPQEeF_X_pqKotxRGD">https://drive.google.com/drive/folders/10HKtwD1UrdmhqxrPQEeF_X_pqKotxRGD</a>
Spectral and image dataset of apple leaf diseases and insect pests in China in 2015 (Gao et al., 2020)	Disease, Pest	1	3	<a href="https://www.agridata.cn/data.html#/datadetail?id=4363">https://www.agridata.cn/data.html#/datadetail?id=4363</a>
Image dataset of six common orchard pests including <i>Bactrocera dorsalis</i> (Zhang et al., 2022)	Pest	0	6	<a href="https://www.agridata.cn/data.html#/datadetail?id=286640">https://www.agridata.cn/data.html#/datadetail?id=286640</a>
Corn Disease Image Database	Disease	1	5	<a href="https://www.agridata.cn/data.html#/datadetail?id=288848">https://www.agridata.cn/data.html#/datadetail?id=288848</a>
Field images of maize annotated with disease symptoms (Wiesner-Hanks et al., 2018)	Disease	1	1	<a href="https://osf.io/p67rz/">https://osf.io/p67rz/</a>
Database of rice pests and diseases in the field in South China and Southwest China from 2018 to 2020	Disease, Pest	1	20	<a href="https://www.agridata.cn/data.html#/datadetail?id=291593">https://www.agridata.cn/data.html#/datadetail?id=291593</a>
Data on the occurrence of insect pests in major crop fields in China from 2013 to 2018	Pest	5	53	<a href="https://www.agridata.cn/data.html#/datadetail?id=4437">https://www.agridata.cn/data.html#/datadetail?id=4437</a>
Data on the occurrence of major crop field diseases in China from 2013 to 2018	Disease	5	34	<a href="https://www.agridata.cn/data.html#/datadetail?id=4436">https://www.agridata.cn/data.html#/datadetail?id=4436</a>
Atlas of Alfalfa Diseases, Insect Pests, and Natural Enemies and Dataset of Major Alfalfa Diseases in Heilongjiang Province (2015-2016) (Li, 2023)	Pest	1	203	<a href="https://www.scidb.cn/en/detail?dataSetId=DS_455c0351d04b4a7e912b6d01a66ee5f7">https://www.scidb.cn/en/detail?dataSetId=DS_455c0351d04b4a7e912b6d01a66ee5f7</a>
Rice Leaf Diseases (Shah et al., 2017)	Disease	1	3	<a href="https://archive.ics.uci.edu/dataset/486/rice+leaf+diseases">https://archive.ics.uci.edu/dataset/486/rice+leaf+diseases</a>
Eggplant Disease Recognition Dataset (Mafi et al., 2023)	Disease, Pest	1	13	<a href="https://data.mendeley.com/datasets/r3tb5mzn4d/2">https://data.mendeley.com/datasets/r3tb5mzn4d/2</a>
paddy-doctor-diseases-medium (A) (et al., 2022)	Disease, Pest	1	13	<a href="https://ieee-dataport.org/competitions/design-classifier-classify-diseases-paddy-based-leaf-color">https://ieee-dataport.org/competitions/design-classifier-classify-diseases-paddy-based-leaf-color</a>
Baidu Encyclopedia Data	Pest	1	5	<a href="https://baike.baidu.com/">https://baike.baidu.com/</a>

Table 3: Sources for agricultural pest and disease image data collection.

Derivation Sequence
Disease-P-M-K $\xrightarrow{\text{Specified Crop}}$ Disease-M-K $\xrightarrow{\text{Specified Disease}}$ Disease-K
Disease-P-M $\xrightarrow{\text{Specified Crop}}$ Disease-M
Pest-P-K $\xrightarrow{\text{Specified Pest}}$ Pest-K
Disease-P
Disease-P

Table 4: Derivation sequence of high to low level complex reasoning data.

knowledge during the generation process. However, due to the inherent hallucination issue of MLLMs, errors in the data are unavoidable. Therefore, we hired two agricultural experts to annotate the data and filter out erroneous entries. To ensure annotation consistency, the experts underwent training, followed by the annotation of 200 samples for preliminary labeling. The annotation consistency was calculated, and when it fell below 90%, further training was provided to address discrepancies,

followed by reannotation of another 200 samples. The detailed process is shown in Figure 10. After two training rounds, the experts achieved over 90% consistency, with the results in Table 5.

Metrics	Round 1	Round 2
# Consistency of Question	92.5	94.0
# Consistency of Options	90.5	93.0
# Consistency of Answer	87.5	92.5
# Consistency of Answer Explanation	87.5	93.0

Table 5: Consistency among experts in different labeling rounds.

#### A.4 Data Statistics

Through rigorous expert review, we obtained 3,939 high-quality images and 15,901 multiple-choice questions. A total of 900 high-resolution data entries were selected from each category, with 100

### Prompt A.1: High-Level Data Generation

【指令】你是一名农学领域的教师，你需要依据作物疾病某个方面的背景知识生成考察知识的推理逻辑图，然后依据图片和逻辑图再生成多项单选题。通过逻辑图生成问题是为了引导学生从识别图中的作物开始，逐步深入到判断是该作物的哪种疾病、该疾病的相关特性是什么，从全面考察他们在农学领域的知识运用与逻辑思维能力。

【示例】

{Example}

【信息】

作物: {Crop Name}

疾病: {Disease or Pest Name}

背景知识: {Background Knowledge}

【要求】

1. 逻辑图的前两个节点固定为"作物", "疾病", 逻辑图至少包含4个节点。
2. 根据图片和逻辑图生成问题和选项, 保证问题需要根据图片才能回答。
3. 问题中不应出现作物名称, 疾病名称、害虫名称、致病因子的名称、致病因子的类型 (如孢子、细菌等)。
4. 每个选项由一个词或者一个短语组成。
5. 答案中给出正确选项的字母。
6. 根据逻辑图和背景知识给出解题逻辑, 每个步骤对应逻辑图中的一个节点, 每一步中要有推理的结果。
7. 对于每个选择题输出生成选择题用到的背景知识。
8. 输出选项解析, 逐个分析选项。
9. 生成3个多项选择题。

Metric	Validation Overall	Test Overall	Disease						Pest		
			P	M	P-M	K	M-K	P-M-K	P	K	P-K
Avg. Question Length	19.43	19.40	12.61	18.56	22.70	17.63	24.80	22.59	13.37	21.36	21.95
Avg. Option Length	3.66	3.55	2.03	5.11	3.55	5.11	3.54	3.54	3.39	3.28	3.28
# Question / # Images	900 / 200	15001 / 3767	1283 / 1283	1283 / 1283	1283 / 1283	1218 / -	1375 / 1160	1389 / 1160	2656 / 2656	2586 / -	2828 / 2332
Avg. Image Resolution	2081 × 1642	1815 × 1489	2128 × 1783	2128 × 1783	2128 × 1783	-	2171 × 1808	2186 × 1819	1444 × 1140	-	1449 × 1139

Table 6: Detailed statistics on the Agri-CM<sup>3</sup> benchmark.

entries per category, to form the **Validation** set. The remaining 15,001 entries were used as the **Test** set. The dataset includes 11 crops, 45 diseases, and 107 pests. Figure 14 shows data cases for different tasks, and Figure 13 illustrates the label distribution across the global, disease, and pest levels, indicating a balanced distribution across all labels.

Further statistical analysis of the dataset is provided in Table 6, demonstrating the high image quality of Agri-CM<sup>3</sup>. The average resolution is 1800 × 1500 pixels, ensuring high clarity and detail. Additionally, homogenized images were filtered out automatically by the model, ensuring diversity. All images were also manually annotated and validated by agricultural experts to guarantee data reliability and usability.

#### A.5 Diversity Analysis of Agri-CM<sup>3</sup>

**Image Diversity Analysis.** Figure 11 illustrates the spatial representations of images encoded using CLIP (Radford et al., 2021). The wide distribution of image representations across the space indicates rich semantic content and intense coverage, suggesting that the dataset has a high diversity and comprehensive representation. This diversity en-

ables a comprehensive evaluation of the model’s ability across various agricultural pest and disease management tasks.

**Text Diversity Analysis.** Further analysis of phrase frequencies in questions and options is presented in the word clouds in Figure 15. In the question word cloud, terms like “害虫 (Pest)”, “病害 (Disease)”, “特征 (Characteristic)” and “防治 (Prevention)” dominate, reflecting the benchmark’s focus on agricultural pest and disease management, aligning with our objectives. In the option word cloud, terms like “水稻 (Rice)”, “小麦 (Wheat)” and “稻瘟病 (Rice Blast)” are prominent, validating the targeted approach of our dataset in reflecting key agricultural issues.

#### A.6 Multilingual Dataset

Given the strong machine translation capabilities of current large language models—particularly proprietary models like GPT-4o—we utilized GPT-4o to translate our Chinese dataset into English. This not only ensures high-quality translations but also enhances reproducibility and facilitates community-driven extensions.

### Prompt A.2: High-Level Data Generation (Translation)

[Instruction] You are a teacher in the field of agronomy. Based on the background knowledge of a crop disease in a specific aspect, you are required to generate a reasoning logic diagram to assess the knowledge. Then, based on the diagram, generate multiple-choice questions. The purpose of generating questions from the logic diagram is to guide students from identifying the crop to gradually delving into determining the type of disease affecting the crop, its related characteristics, and comprehensively assess their knowledge application and logical thinking abilities in the field of agronomy.

【Example】

{Example}

【Information】

Crop: {Crop Name}

Disease: {Disease or Pest Name}

Background Knowledge: {Background Knowledge}

【Requirements】

1. The first two nodes of the logic diagram should be fixed as “Crop” and “Disease,” and the diagram should include at least four nodes.
2. Generate questions and options based on the image and logic diagram, ensuring that the questions can only be answered by referring to the diagram.
3. The question should not mention the crop name, disease name, pest name, pathogen name, or the type of pathogen (such as spores, bacteria, etc.).
4. Each option should be a word or a short phrase.
5. Provide the correct option in the answer key.
6. Based on the logic diagram and background knowledge, provide the reasoning for solving the question, where each step corresponds to a node in the diagram. Each step should have a reasoning outcome.
7. Provide the background knowledge used to generate the multiple-choice questions.
8. Output option analysis, analyzing each option individually.
9. Generate three multiple-choice questions.

### Prompt A.3: Low-Level Data Derivation

【指令】你是一名农学老师，你现在要修改一个农学多项选择题的问题，将原先没有明确作物和疾病名称、结合图片进行的提问，改为在题目中给出作物名称、结合图片进行提问。

【示例】

{Example}

【信息】

作物: {Crop Name}

需要修改的问题: {Question}

## B Evaluation

For closed-source models, we directly call the corresponding APIs. For open-source models, if the model supports vLLM<sup>3</sup>, we deploy it in the OpenAI API format and test it according to the API call procedure. If the model does not support vLLM, we use the official inference method provided by the model developers to perform reasoning. During the evaluation, to ensure the stability of the results, we set the temperature parameter to 0.0 and the maximum output length to 1024.

See Prompt B.1, B.3, B.5 for the prompts in the experiment. The English Prompt after translation is B.2, B.4, and B.6.

<sup>3</sup><https://github.com/vllm-project/vllm>

### B.1 Evaluation Metric

Our benchmark consists entirely of multiple-choice questions. For models that directly respond with an option, we evaluate their performance by matching the answer string with the correct option. For CoT methods, we consider the final answer provided after the reasoning process. The evaluation metric used is accuracy (ACC), calculated using the formula:

$$ACC = \frac{n}{N}. \quad (1)$$

where  $n$  denotes the amount of data answered correctly and  $N$  denotes the amount of all data.

### B.2 Evaluation Models

In this study, we evaluated 45 MLLMs and 5 LLMs on Agri-CM<sup>3</sup>. The specific parameters of these

#### Prompt A.4: Low-Level Data Derivation (Translation)

【Instruction】 You are an agronomy teacher. You are now required to modify a multiple-choice question in agronomy by revising the original question, which previously did not specify the crop and disease names and was based on images, to a version where the crop name is provided in the question along with the image for reference.

【Example】  
{Example}

【Information】  
Crop: {Crop Name}  
Question to be modified: {Question}

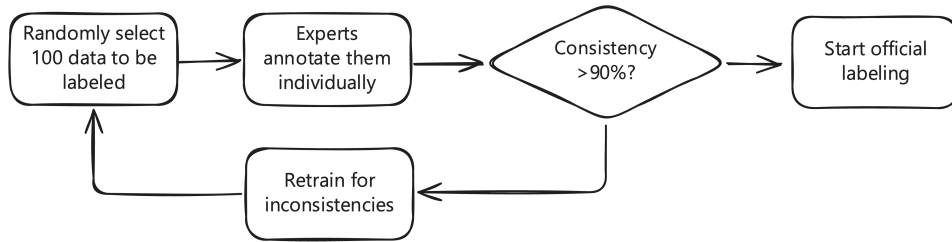


Figure 10: Expert annotation training process. To ensure annotation quality and consistency, formal annotation begins only when expert agreement exceeds 90%.

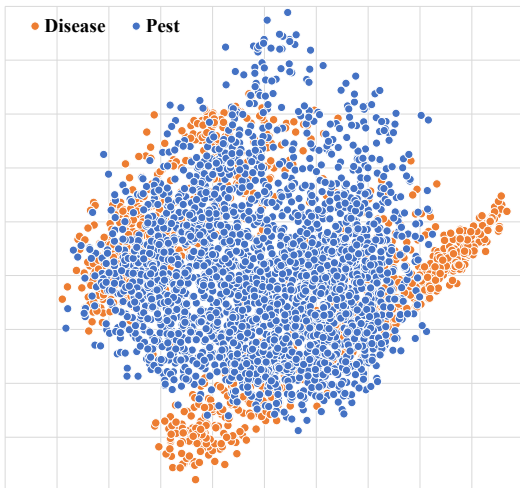


Figure 11: Image spatial representation via CLIP encoding and PCA dimensionality reduction.

models are listed in Table 7.

## C Supplementary Results

In this section, we present the complete experimental results for all models on the benchmark in Table 10, including 45 MLLMs and 5 LLMs. Additionally, Table 11 provides the full CoT experimental results for 11 models of varying scales, covering both open-source and proprietary models.

**Expert Performance Analysis.** To systematically evaluate expert performance across tasks, we present each expert’s scores on individual tasks in

#### Prompt A.5: Translate Chinese Dataset to English

You are a translation expert in the field of agriculture. Below is a Chinese agricultural exam question. Please translate it into English and follow the requirements below: 1. Ensure that the translation maintains professionalism, academic rigor, and accuracy. 2. The output should be in JSON format.

Question: question Options:

A. {Option A}

B. {Option B}

C. {Option C}

D. {Option D}

Options Analysis: {Options Analysis}

Translation Result:

Table 8. Experimental results indicate that current agricultural experts typically possess deep expertise in specific crops or pest and disease categories, with their knowledge being relatively domain-focused rather than broadly distributed. As a result, when confronted with the diverse crop and pest categories included in our benchmark, their overall performance exhibits noticeable variability, highlighting limitations in cross-domain generalization.

**Mixed reasoning poses a greater challenge than perceptual complexity for current models.**

To systematically evaluate deficiencies in different model capabilities, we introduce two diagnostic metrics using the Disease-P-M-K task as a case



Models	Size	Vision Encoder	LLM	Parameter Link
LLaVA1.5-7B	7.06B	CLIP ViT-L/14	Vicuna-v1.5-7B	<a href="https://huggingface.co/llava-hf/llava-1.5-7b-hf">https://huggingface.co/llava-hf/llava-1.5-7b-hf</a>
LLaVA1.5-13B	13.4B	CLIP ViT-L/14	Vicuna-v1.5-13B	<a href="https://huggingface.co/llava-hf/llava-1.5-13b-hf">https://huggingface.co/llava-hf/llava-1.5-13b-hf</a>
LLaVA-1.6-7B	7.57B	CLIP ViT-L/14	Mistral-7B	<a href="https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf">https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf</a>
LLaVA-1.6-13B	13.4B	CLIP ViT-L/14	Vicuna-v1.5-13B	<a href="https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf">https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf</a>
LLaVA-1.6-34B	34.8B	CLIP ViT-L/14	Nous-Hermes-2-Yi-34B	<a href="https://huggingface.co/llava-hf/llava-v1.6-34b-hf">https://huggingface.co/llava-hf/llava-v1.6-34b-hf</a>
LLaVA-OneVision-7B	8.03B	SigLIP-SO400M	Qwen2-7B	<a href="https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf">https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf</a>
LLaVA-OneVision-72B	73.2B	SigLIP-SO400M	Qwen2-72B	<a href="https://huggingface.co/llava-hf/llava-onevision-qwen2-72b-ov-hf">https://huggingface.co/llava-hf/llava-onevision-qwen2-72b-ov-hf</a>
Phi-3.5-vision-instruct	4.15B	CLIP ViT-L/14	Phi-3 Mini	<a href="https://huggingface.co/microsoft/Phi-3.5-vision-instruct">https://huggingface.co/microsoft/Phi-3.5-vision-instruct</a>
Qwen-VL-Chat	9.6B	ViT-bigG	Qwen-7B	<a href="https://huggingface.co/Qwen/Qwen-VL-Chat">https://huggingface.co/Qwen/Qwen-VL-Chat</a>
Qwen2-VL-7B-Instruct	8.29B	ViT-bigG	Qwen2-7B	<a href="https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct">https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct</a>
Qwen2-VL-72B-Instruct	73.4B	ViT-bigG	Qwen2-72B	<a href="https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct">https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct</a>
MiniCPM-Llama3-V2.5	8.54B	SigLip-400M	Llama3-8B-Instruct	<a href="https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5">https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5</a>
MiniCPM-V2.6	8.1B	SigLip-400M	Qwen2-7B	<a href="https://huggingface.co/openbmb/MiniCPM-V-2_6">https://huggingface.co/openbmb/MiniCPM-V-2_6</a>
MiniCPM-o-2.6	8.67B	SigLip-400M et al.	Qwen2.5-7B	<a href="https://huggingface.co/openbmb/MiniCPM-o-2_6">https://huggingface.co/openbmb/MiniCPM-o-2_6</a>
CogVLM2	19.5B	EVA-CLIP-E	LLaMA3-8B	<a href="https://huggingface.co/THUDM/cogvlm2-llama3-chat-19B">https://huggingface.co/THUDM/cogvlm2-llama3-chat-19B</a>
GLM4-9B	13.9B	EVA-CLIP-E	GLM4-9B	<a href="https://huggingface.co/THUDM/glm-4-9b">https://huggingface.co/THUDM/glm-4-9b</a>
InternVL2-1B	938M	InternViT-300M-448px	Qwen2-0.5B-Instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2-1B">https://huggingface.co/OpenGVLab/InternVL2-1B</a>
InternVL2-2B	2.21B	InternViT-300M-448px	internlm2-chat-1.8b	<a href="https://huggingface.co/OpenGVLab/InternVL2-2B">https://huggingface.co/OpenGVLab/InternVL2-2B</a>
InternVL2-4B	4.15B	InternViT-300M-448px	Phi-3-mini-128k-instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2-4B">https://huggingface.co/OpenGVLab/InternVL2-4B</a>
InternVL2-8B	8.08B	InternViT-300M-448px	internlm2.5-7b-chat	<a href="https://huggingface.co/OpenGVLab/InternVL2-8B">https://huggingface.co/OpenGVLab/InternVL2-8B</a>
InternVL2-26B	25.5B	InternViT-6B-448px-V1-5	internlm2-chat-20b	<a href="https://huggingface.co/OpenGVLab/InternVL2-26B">https://huggingface.co/OpenGVLab/InternVL2-26B</a>
InternVL2.5-1B	938M	InternViT-300M-448px-V2.5	Qwen2.5-0.5B-Instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-1B">https://huggingface.co/OpenGVLab/InternVL2_5-1B</a>
InternVL2.5-2B	2.21B	InternViT-300M-448px-V2.5	internlm2.5-1.8b-chat	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-2B">https://huggingface.co/OpenGVLab/InternVL2_5-2B</a>
InternVL2.5-4B	3.71B	InternViT-300M-448px-V2.5	Qwen2.5-3B-Instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-4B">https://huggingface.co/OpenGVLab/InternVL2_5-4B</a>
InternVL2.5-8B	8.08B	InternViT-300M-448px-V2.5	internlm2.5-7b-chat	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-8B">https://huggingface.co/OpenGVLab/InternVL2_5-8B</a>
InternVL2.5-26B	25.5B	InternViT-6B-448px-V2.5	internlm2.5-20b-chat	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-26B">https://huggingface.co/OpenGVLab/InternVL2_5-26B</a>
InternVL2.5-38B	38.4B	InternViT-6B-448px-V2.5	Qwen2.5-32B-Instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-38B">https://huggingface.co/OpenGVLab/InternVL2_5-38B</a>
InternVL2.5-78B	78.4B	InternViT-6B-448px-V2.5	Qwen2.5-72B-Instruct	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-78B">https://huggingface.co/OpenGVLab/InternVL2_5-78B</a>
Yi-VL-6B	6.6B	CLIP ViT-H/14	Yi-6B-Chat	<a href="https://huggingface.co/01-ai/Yi-VL-6B">https://huggingface.co/01-ai/Yi-VL-6B</a>
Yi-VL-34B	34B	CLIP ViT-H/14	Yi-34B-Chat	<a href="https://huggingface.co/01-ai/Yi-VL-34B">https://huggingface.co/01-ai/Yi-VL-34B</a>
Deepseek-v1.2-tiny	1.0B/3.37B	SigLIP-SO400M	DeepSeek-MoE	<a href="https://huggingface.co/deepseek-ai/deepseek-v1.2-tiny">https://huggingface.co/deepseek-ai/deepseek-v1.2-tiny</a>
Deepseek-v1.2-small	2.8B/16.1B	SigLIP-SO400M	DeepSeek-MoE	<a href="https://huggingface.co/deepseek-ai/deepseek-v1.2-small">https://huggingface.co/deepseek-ai/deepseek-v1.2-small</a>
Deepseek-v1.2	4.05B/27.5B	SigLIP-SO400M	DeepSeek-MoE	<a href="https://huggingface.co/deepseek-ai/deepseek-v1.2">https://huggingface.co/deepseek-ai/deepseek-v1.2</a>
Molmo-7B	8.02B	CLIP ViT-L/14	Qwen2-7B	<a href="https://huggingface.co/allenai/Molmo-7B-D-0924">https://huggingface.co/allenai/Molmo-7B-D-0924</a>
Molmo-72B	73.3B	CLIP ViT-L/14	Qwen2-72B	<a href="https://huggingface.co/allenai/Molmo-72B-0924">https://huggingface.co/allenai/Molmo-72B-0924</a>
Cambrian-8B	8.33B	SVA	Llama3-8B-Instruct	<a href="https://huggingface.co/nyu-visionx/cambrian-8b">https://huggingface.co/nyu-visionx/cambrian-8b</a>
Cambrian-13B	13.4B	SVA	Vicuna-v1.5-13B	<a href="https://huggingface.co/nyu-visionx/cambrian-13b">https://huggingface.co/nyu-visionx/cambrian-13b</a>
Cambrian-34B	34.8B	SVA	Hermes2-Yi-34B	<a href="https://huggingface.co/nyu-visionx/cambrian-34b">https://huggingface.co/nyu-visionx/cambrian-34b</a>
-----				
GPT-4o	-	-	-	-
GPT-4o-mini	-	-	-	-
Gemini-1.5-flash	-	-	-	-
Gemini-1.5-pro	-	-	-	-
Yi-Vision	-	-	-	-
Qwen-VL-PLUS	-	-	-	-
Qwen-VL-MAX	-	-	-	-
-----				
GLM4-9B	9.4B	-	-	<a href="https://huggingface.co/THUDM/glm-4-9b-chat">https://huggingface.co/THUDM/glm-4-9b-chat</a>
Qwen2.5-7B	7.62B	-	-	<a href="https://huggingface.co/Qwen/Qwen2.5-7B-Instruct">https://huggingface.co/Qwen/Qwen2.5-7B-Instruct</a>
InternLM-2.5-7B	7.74B	-	-	<a href="https://huggingface.co/internlm/internlm2_5-7b-chat">https://huggingface.co/internlm/internlm2_5-7b-chat</a>
GPT-3.5-Turbo	-	-	-	-
GPT-4-Turbo	-	-	-	-

Table 7: The model architecture of 45 MLLMs and 5 LLMs evaluated on Agri-CM<sup>3</sup>.

#### Prompt B.1: Zero-Shot Normal

以下是一道农学单项选择题，不需要做任何分析和解释，直接输出答案。

问题: {question}  
选项: {options\_str}  
答案:

#### Prompt B.2: Zero-Shot Normal (Translation)

The following is a single-choice question in Agronomy. No analysis or explanation is required, just output the answer.

Question: {question}  
Options: {options\_str}  
Answer:

study:

- **Perception Drop:** The percentage of samples where the model correctly answers both Disease-K and Disease-M-K, but fails on Disease-P-M-K. This reflects performance degradation due to increased perceptual complexity.
- **Mixed Reasoning Drop:** The percentage of samples where the model performs well on Disease-K but fails on Disease-M-K, capturing

the impact of added reasoning complexity.

We compute these metrics across a range of representative open-source and proprietary models. As shown in Table 9, Mixed Reasoning Drop consistently exceeds Perception Drop, highlighting mixed reasoning as a more significant challenge for current models.

Expert	Validation	Disease						Pest		
		P	M	P-M	K	M-K	P-M-K	P	K	P-K
1	80.00	100.00	85.00	83.00	66.00	71.00	69.00	95.00	76.00	75.00
2	79.11	100.00	73.00	77.00	73.00	73.00	71.00	94.00	76.00	75.00
Average	79.56	100.00	79.00	80.00	69.50	72.00	70.00	94.50	76.00	75.00

Table 8: Expert performance across agricultural tasks.

#### Prompt B.3: Zero-Shot CoT

以下是一道农学单项选择题，请一步一步地思考，并在最后给出答案。

问题: {question}  
选项: {options\_str}  
答案:

#### Prompt B.4: Zero-Shot CoT (Translation)

The following is a single-choice question in Agronomy. Please think step by step and provide the answer at the end.

Question: {question}  
Options: {options\_str}  
Answer:

## D Case Study

### D.1 GPT-4o CoT Case Study

Through analyzing the CoT responses of GPT-4o and InternVL-2.5-78B, we observed that GPT-4o effectively follows the multi-level reasoning process required by the benchmark, as shown in Figure 16. Specifically, GPT-4o diagnoses the disease, analyzes key features, and performs knowledge integration and application, ultimately selecting the correct answer through option analysis. In contrast, InternVL-2.5-78B lacks this structured reasoning process. These findings demonstrate that GPT-4o’s structured multi-step reasoning significantly enhances CoT performance. This highlights the importance of multi-level structured reasoning in improving model performance and provides valuable insights for future model optimization.

### D.2 Error Analysis and Case Study

We randomly selected 200 erroneous predictions from GPT-4o’s CoT responses in the most challenging disease and pest reasoning tasks. These errors were annotated by two agricultural experts and categorized into five types:

#### Prompt B.5: Few-Shot Normal

以下是一道农学单项选择题，不需要做任何分析和解释，直接输出答案。

以下是一些示例:

{examples}  
现在，请根据以下题目输出答案:  
问题: {question}  
选项: {options\_str}  
答案:

#### Prompt B.6: Few-Shot Normal (Translation)

The following is a single-choice question in Agronomy. No analysis or explanation is required, just output the answer.

Here are some examples:  
{examples}

Now, please output the answer based on the following question:  
Question: {question}  
Options: {options\_str}  
Answer:

- **Question Misunderstanding:** Errors caused by incorrect comprehension of the question.
- **Perceptual Error:** Errors in interpreting visual information from images.
- **Lack of Knowledge:** Errors due to incorrect or insufficient application of domain-specific knowledge.
- **Irrelevant Response:** Model responses unrelated to the question.
- **Refuse to Answer:** The model refuses to respond due to privacy or safety concerns.

Through the analysis of error distributions, we identified the following characteristics: In disease reasoning tasks, which involve a relatively small number of crop types (only 11), perceptual tasks are relatively simple, and most errors are attributed to a lack of domain knowledge. This indicates that

Model	PD	MRD
internvl2.5-8B	3.50	14.33
Internvl2.5-78B	2.76	15.71
Qwen2-VL-7B	4.43	14.27
Qwen2.5-VL-7B	3.17	13.33
Qwen2.5-VL-72B	4.54	16.2
Qwen-VL-MAX	4.99	18.05
Gemini-1.5-flash	3.72	15.71
Gemini-1.5-pro	4.67	17.09
GPT-4o	6.66	21.99
<b>Average</b>	<b>4.21</b>	<b>15.96</b>

Table 9: Comparative analysis of Perception Drop (PD) and Mixed Reasoning Drop (MRD) across models.

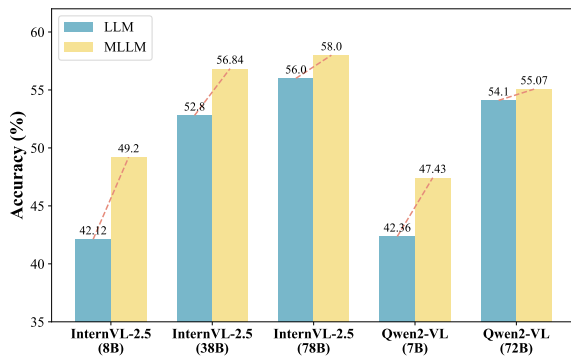


Figure 12: Performance of different MLLMs and their corresponding LLMs on Pest-K (text modality).

models require further enhancement in domain-specific knowledge reasoning to handle disease-related tasks effectively. In contrast, pest reasoning tasks, with 107 pest types, exhibit significantly higher perceptual complexity, making perceptual errors the primary cause of failures. Analysis reveals that the perceptual errors in pest reasoning tasks typically arise from the model's failure to accurately identify pests, which stems from a lack of domain-specific knowledge.

Future efforts should strengthen models' ability to combine visual perception with domain-specific knowledge to improve performance in complex reasoning tasks. Additionally, expanding datasets to cover a broader range of pest and disease types and with improving multi-modal fusion techniques will be crucial for reducing errors and enhancing reasoning accuracy. Detailed case analyses for the five error types are provided in Figures 18, 19, 20, 21, 22, and 23.

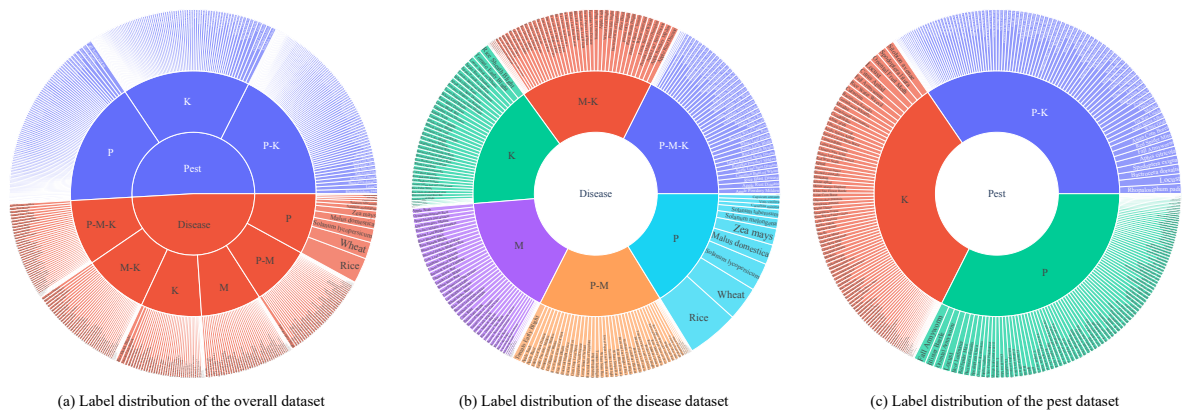


Figure 13: Label distribution across the global, disease, and pest levels.



Figure 14: Data cases in the Agri-CM<sup>3</sup> benchmark. The abbreviation in Agri-CM<sup>3</sup> means P: Perception, M: Mixture of Perception and Cognition, and K: Knowledge Application.

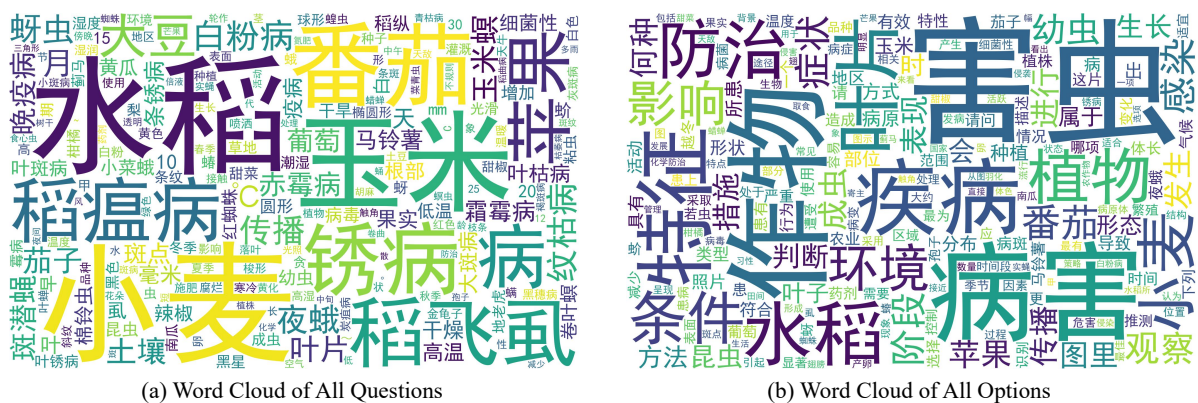


Figure 15: The word frequency information for both questions and options is separately calculated and presented in the form of word clouds.


Model	Validation Total	Test Total	Disease						Pest		
			P	M	P-M	K	M-K	P-M-K	P	K	P-K
Random Choice	23.22	25.14	25.44	26.04	25.44	24.15	24.31	26.61	25.04	24.94	24.89
Expert Performance	79.56	-	-	-	-	-	-	-	-	-	-
<b>Multi-Modal Large Language Model(MLLMs): Text + Image as Input</b>											
LLaVA-1.6-13B (Liu et al., 2024)	33.33	31.43	35.42	36.77	28.74	42.49	30.61	30.17	24.37	35.24	28.12
Cambrian-8B (Tong et al., 2024)	40.11	37.59	41.42	52.75	47.17	40.07	36.63	33.9	30.05	36.56	34.38
LLaVA-1.5-7B (Liu et al., 2023a)	38.00	35.67	27.73	51.99	51.99	40.70	40.66	40.82	22.65	32.94	32.84
Deepseek-vl2-tiny (Wu et al., 2024)	40.22	36.38	28.57	49.2	41.08	38.82	37.33	36.77	23.83	39.5	39.44
LLaVA-1.5-13B (Liu et al., 2023a)	41.33	38.19	43.87	53.00	52.41	40.52	42.54	42.38	23.40	33.31	36.47
Cambrian-13B (Tong et al., 2024)	41.56	38.78	54.86	52.16	45.05	41.86	36.08	32.74	30.59	38.82	33.8
InternVL-2-1B (Chen et al., 2024c)	44.33	39.11	47.00	50.63	43.28	42.31	37.83	37.40	27.50	39.58	39.41
Deepseek-VL2-small (Wu et al., 2024)	44.56	39.42	48.35	45.98	37.28	47.58	44.86	42.13	19.95	43.48	41.02
InternVL-2.5-1B (Chen et al., 2024b)	44.33	40.71	50.55	48.01	39.39	35.78	38.78	36.70	44.41	35.64	39.85
LLaVA-1.6-7B (Liu et al., 2024)	42.11	40.73	45.82	52.41	46.58	40.34	42.62	42.46	33.06	37.97	39.08
Phi-3.5-Vision (Abdin et al., 2024)	46.11	41.69	39.14	57.99	53.85	44.10	44.66	45.18	30.01	38.50	40.29
Deepseek-VL2 (Wu et al., 2024)	46.33	42.91	36.18	47.93	41.5	48.12	48.47	45.23	37.28	44.17	42.56
InternVL-2-2B (Chen et al., 2024c)	47.11	42.96	45.31	51.23	47.17	42.22	40.27	38.96	44.91	39.74	41.09
Qwen-VL (Bai et al., 2023)	46.11	44.26	66.61	45.98	37.70	44.54	40.58	39.66	47.65	41.67	39.63
MiniCPM-V2.5 (Yao et al., 2024)	44.78	44.75	64.33	47.00	46.49	50.09	42.62	42.07	40.14	45.53	38.20
InternVL-2-4B (Chen et al., 2024c)	50.00	45.04	56.72	55.45	50.04	47.85	46.08	44.87	36.03	42.04	42.89
Yi-VL-6B (AI et al., 2024)	49.56	45.45	35.93	49.54	41.67	49.19	47.96	45.72	50.94	43.04	43.66
Yi-VL-34B (AI et al., 2024)	48.78	45.98	49.79	52.83	38.46	51.25	48.43	44.25	42.68	46.50	44.76
InternVL-2.5-2B (Chen et al., 2024b)	50.56	46.14	55.79	59.34	52.41	46.24	42.54	39.27	48.63	42.04	39.81
CogVLM2 (Hong et al., 2024)	48.11	46.37	76.84	45.65	43.11	39.45	39.53	37.24	58.92	38.21	40.91
Cambrian-34B (Tong et al., 2024)	56.00	49.16	74.13	54.35	45.98	53.13	45.41	38.71	53.99	47.91	39.11
MiniCPM-o-2.6 (Yao et al., 2024)	55.78	51.56	65.51	58.07	51.48	52.42	46.75	45.23	60.52	45.37	44.83
InternVL-2-8B (Chen et al., 2024c)	56.33	51.70	56.21	58.41	53.34	54.03	48.98	48.76	57.75	47.02	46.44
InternVL-2.5-4B (Chen et al., 2024b)	55.33	51.76	64.92	58.24	51.23	55.19	52.04	49.38	54.03	46.30	45.93
Qwen2.5-VL-3B (Bai et al., 2025)	54.44	52.08	73.03	51.31	44.46	52.42	47.76	45.23	62.79	46.14	45.53
MiniCPM-V2.6 (Yao et al., 2024)	54.67	52.24	65.34	58.58	54.35	49.02	46.70	45.96	64.05	44.85	45.42
GLM4V-9B (GLM et al., 2024)	56.00	52.65	86.73	55.62	56.04	50.72	47.84	47.71	57.79	43.32	44.17
LLaVA-1.6-34B (Liu et al., 2024)	57.78	52.71	59.09	59.51	54.35	57.60	51.96	50.47	54.58	48.51	47.76
Qwen2.5-VL-7B (Bai et al., 2025)	56.44	53.67	76.16	53.85	50.38	49.46	45.33	43.75	73.51	45.70	44.87
LLaVA-OneVision-7B (Li et al., 2024)	57.00	53.92	70.33	58.66	53.00	51.25	50.78	50.54	64.87	45.86	46.41
Molmo-7B (Deitke et al., 2024)	55.89	54.76	54.69	66.19	59.93	52.50	49.92	49.77	67.41	47.91	47.54
InternVL-2.5-8B (Chen et al., 2024b)	59.67	55.73	70.58	62.13	57.06	55.37	52.43	50.54	64.16	49.20	48.13
Qwen2-VL-7B (Wang et al., 2024b)	60.00	56.32	85.88	58.92	57.82	54.20	47.88	46.27	70.34	47.43	46.26
InternVL-2-26B (Chen et al., 2024c)	61.33	57.01	81.4	64.67	60.27	56.26	49.06	49.92	66.82	50.2	46.08
InternVL-2.5-26B (Chen et al., 2024b)	64.78	61.97	74.30	71.09	66.61	63.69	57.06	56.07	72.57	54.91	51.50
Molmo-72B (Deitke et al., 2024)	66.33	63.13	63.14	73.96	71.01	66.01	60.75	59.56	68.86	58.37	55.61
LLaVA-OneVision-72B (Li et al., 2024)	69.00	64.37	78.61	72.19	68.98	65.38	60.99	58.79	70.07	57.32	57.66
Qwen2.5-VL-72B (Bai et al., 2025)	67.78	64.70	84.02	68.64	67.03	63.33	56.86	54.46	79.11	55.47	53.41
Qwen2-VL-72B (Wang et al., 2024b)	72.00	64.81	80.98	73.29	72.36	65.74	61.77	59.95	72.07	55.07	56.27
InternVL-2.5-38B (Chen et al., 2024b)	68.11	65.31	74.81	76.67	70.16	65.65	61.30	60.26	75.04	56.84	56.89
InternVL-2.5-78B (Chen et al., 2024b)	69.89	65.49	79.54	76.67	72.78	65.56	60.44	59.33	73.04	58.00	56.38
Qwen-VL-PLUS (Wang et al., 2024b)	51.67	51.54	80.56	47.76	48.27	47.14	45.29	42.69	68.39	40.23	45.42
Yi-Vision (AI et al., 2024)	61.56	56.59	60.27	58.66	55.45	60.82	53.61	51.01	70.31	51.85	48.35
Qwen-VL-MAX (Wang et al., 2024b)	59.44	57.97	76.84	60.19	60.19	62.25	52.04	48.83	66.12	53.46	49.67
GPT-4o-mini (Hurst et al., 2024)	63.00	58.21	79.88	69.99	67.12	53.49	52.28	48.37	68.78	48.87	47.80
Gemini-1.5-flash (Team et al., 2024)	64.67	61.73	87.57	68.55	69.74	57.33	52.51	52.72	80.40	49.92	47.73
GPT-4o (Hurst et al., 2024)	67.33	63.64	83.85	69.48	68.30	65.65	54.71	49.53	80.09	56.84	51.10
Gemini-1.5-pro (Team et al., 2024)	67.78	64.73	89.18	69.82	70.33	62.43	57.30	54.20	78.72	57.36	52.49
<b>Large Language Models (LLMs): Only Text as Input</b>											
GLM4-9B (GLM et al., 2024)	41.89	35.37	34.66	38.55	27.98	45.26	40.03	38.72	21.56	41.67	36.91
Qwen2.5-7B (Hui et al., 2024)	40.67	35.37	36.43	44.04	19.78	48.21	40.89	37.48	20.74	42.04	36.69
InternLM2.5-7B (Cai et al., 2024)	40.44	36.55	44.55	39.73	32.12	46.33	39.72	38.34	22.38	42.12	35.48
GPT-3.5-Turbo (OpenAI, 2023)	39.11	37.01	32.04	44.72	32.29	45.80	43.33	39.04	23.28	41.55	39.11
GPT-4-Turbo (Achiam et al., 2023)	47.11	41.84	33.39	43.20	35.08	57.07	52.83	46.27	23.08	50.36	44.21

Table 10: The zero-shot result of 45 MLLMs and 5 LLMs in Agri-CM<sup>3</sup>. The abbreviation in Agri-CM<sup>3</sup> means P: Perception, M: Mixture of Perception and Cognition Reasoning, and K: Knowledge Application. Test total represents the average of all test instances.

Model	CoT	Validation Total	Test Total	Disease						Pest		
				P	M	P-M	K	M-K	P-M-K	P	K	P-K
InternVL-2.5-1B	✗	44.33	40.71	50.55	48.01	39.39	35.78	38.78	36.70	44.41	35.64	39.85
	✓	39.33 ↓5.0	35.01 ↓5.7	39.05 ↓11.5	41.50 ↓6.5	31.87 ↓7.5	34.97 ↓0.8	35.01 ↓3.8	32.89 ↓3.8	36.62 ↓7.8	31.54 ↓4.1	34.49 ↓5.4
InternVL-2.5-2B	✗	50.56	46.14	55.79	59.34	52.41	46.24	42.54	39.27	48.63	42.04	39.81
	✓	46.00 ↓4.6	41.14 ↓5.0	40.41 ↓15.4	53.09 ↓6.2	44.46 ↓7.9	40.07 ↓6.2	40.58 ↓2.0	35.85 ↓3.4	44.48 ↓4.2	37.69 ↓4.4	38.05 ↓1.8
InternVL-2.5-4B	✗	55.33	51.76	64.92	58.24	51.23	55.19	52.04	49.38	54.03	46.30	45.93
	✓	52.78 ↓2.5	48.76 ↓3.0	57.31 ↓7.6	53.34 ↓4.9	44.38 ↓6.8	53.67 ↓1.5	49.84 ↓2.2	45.26 ↓4.1	49.73 ↓4.3	46.54 ↑0.2	45.20 ↓0.7
InternVL-2.5-8B	✗	59.67	55.73	70.58	62.13	57.06	55.37	52.43	50.54	64.16	49.20	48.13
	✓	58.22 ↓1.5	55.30 ↓0.4	69.32 ↓1.3	61.62 ↓0.5	59.59 ↑2.5	50.89 ↓4.5	52.83 ↑0.4	51.09 ↑0.6	64.71 ↑0.5	46.78 ↓2.4	48.53 ↑0.4
InternVL-2.5-26B	✗	64.78	61.97	74.30	71.09	66.61	63.69	57.06	56.07	72.57	54.91	51.50
	✓	62.78 ↓2.0	61.97 ↓0.0	74.30 ↓0.0	73.37 ↑2.3	68.30 ↑1.7	56.89 ↓6.8	55.10 ↓2.0	52.41 ↓3.7	71.67 ↓0.9	50.52 ↓4.4	49.85 ↓1.6
InternVL-2.5-38B	✗	68.11	65.31	74.81	76.67	70.16	65.65	61.30	60.26	75.04	56.84	56.89
	✓	64.56 ↓3.5	59.89 ↓5.4	73.29 ↓1.5	72.02 ↓4.7	64.07 ↓6.1	58.05 ↓7.6	55.97 ↓5.3	53.34 ↓6.9	70.15 ↓4.9	52.86 ↓4	49.45 ↓7.4
InternVL-2.5-78B	✗	69.89	65.49	79.54	76.67	72.78	65.56	60.44	59.33	73.04	58.00	56.38
	✓	67.56 ↓2.3	61.71 ↓3.8	79.37 ↓0.2	74.22 ↓2.5	69.23 ↓3.5	61.81 ↓3.8	55.42 ↓5.0	55.52 ↓3.8	69.87 ↓3.2	53.86 ↓4.1	50.70 ↓5.7
Qwen2-VL-7B	✗	60.00	56.32	85.88	58.92	57.82	54.20	47.88	46.27	70.34	47.43	46.26
	✓	60.00 ↓0.0	55.87 ↓0.5	79.71 ↓6.2	62.55 ↑3.6	61.12 ↑3.3	51.61 ↓2.6	48.04 ↑0.2	47.59 ↑1.3	69.41 ↓0.9	45.90 ↓1.5	46.08 ↓0.2
Qwen2-VL-72B	✗	72.00	64.81	80.98	73.29	72.36	65.74	61.77	59.95	72.07	55.07	56.27
	✓	67.44 ↓4.6	62.19 ↓2.6	79.46 ↓1.5	71.17 ↓2.1	66.53 ↓5.8	65.21 ↓0.5	55.57 ↓6.2	55.44 ↓4.5	70.42 ↓1.6	54.83 ↓0.2	52.97 ↓3.3
GPT-4o-mini	✗	63.00	58.21	79.88	69.99	67.12	53.49	52.28	48.37	68.78	48.87	47.80
	✓	61.78 ↓1.2	56.14 ↓2.1	82.25 ↑2.4	68.30 ↓1.7	66.78 ↓0.3	52.06 ↓1.4	48.12 ↓4.2	46.03 ↓2.3	64.67 ↓4.1	46.66 ↓2.2	45.75 ↓2.0
GPT-4o	✗	67.33	63.64	83.85	69.48	68.30	65.65	54.71	49.53	80.09	56.84	51.10
	✓	70.56 ↑3.2	65.43 ↑1.8	87.91 ↑4.1	78.95 ↑9.5	75.23 ↑6.9	61.45 ↓4.2	58.48 ↑3.8	54.04 ↑4.5	80.13 ↑0.0	53.94 ↓2.9	52.49 ↑1.4

Table 11: All experimental results for performance comparison of different mlms with and without CoT.

**Disease-P-M-K: Wheat Stem Rust**



**Question:** 图片中的作物病害萌发所需的湿度是多少？

**Options:** A: 40% B: 55% C: 75% D: 100%

**Answer:** D

**Background Knowledge:** 小麦秆锈病菌对湿度要求较高，需要水滴或100%湿度才能萌发。

**InternVL-2.5-78B CoT Response:**  
 要确定图片中的作物病害萌发所需的湿度，我们需要分析图片中的植物病害特征。图片中可以看到植物叶片上有明显的锈斑，这通常是锈病的症状。锈病是由真菌引起的，常见于高湿度环境中。锈病的病原菌通常在相对湿度较高时萌发和传播。根据植物病理学的知识，**锈病的病原菌在相对湿度达到75%以上时，会更有利于其孢子的萌发和侵染。**因此，结合图片中的病害特征和锈病的湿度需求，答案是：**C. 75%**

**GPT-4o CoT Response:**  
 从图片中可以看出，该作物的病害为一种锈病。锈病通常由真菌引起，而这些病原体在萌发和传播时需要一定的湿度。  
 分析过程：  
 1. **锈病的特点:** 锈病是由真菌引起的，病菌孢子在萌发和侵入宿主植物时，需要高湿度的环境。  
 2. **真菌孢子萌发的湿度需求:** 一般来说，锈病病菌孢子的萌发通常需要空气相对湿度达到 **95%-100%** 的水平。因此，这样的病害通常在潮湿、雨雾多的环境下容易爆发。  
 3. **选项分析:**  
 • A. 40%: 湿度较低，不可能支持真菌孢子的萌发。  
 • B. 55%: 仍然偏低，也不适合锈病病菌的萌发。  
 • C. 75%: 虽然湿度较高，但对于锈病来说通常不足以确保孢子萌发。  
 • D. 100%: 这是典型的适合锈病病菌萌发的湿度范围。  
**答案: D. 100%**

Figure 16: Comparison of CoT responses between GPT-4o and InternVL-2.5-78B. GPT-4o demonstrates superior performance due to its reasoning logic, which better aligns with the benchmark's evaluation criteria. **Red bold** highlights the wrong content.

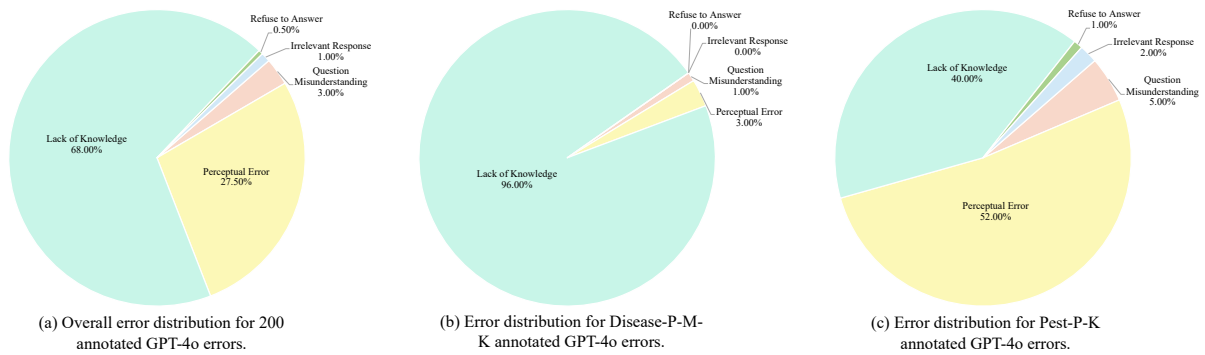



Figure 17: Error distribution on different complex reasoning tasks.

Disease-P-M-K: Lack of Knowledge



**Disease:** 苹果黑星病

**Question:** 这种疾病在什么温度下生长最为理想?

**Options:** A: 30°C B: 10°C C: 20°C D: 25°C

**Answer:** C

**GPT-4o Response:**

从图片看, 这种疾病显然是由真菌病害引起的, **看起来很像霉菌或白粉病**。真菌病害通常在温暖潮湿的环境中生长最为理想。真菌普遍在20°C到25°C的温度下生长得很好。30°C可能会太高, 抑制一些真菌的生长, 10°C则可能太低而不利于活跃生长。因此, 答案可能是在20°C或25°C。**在选择20°C或25°C之间时, 考虑到许多霉菌类真菌在较温暖的环境中活跃, 因此25°C可能更为合理。**

答案: D. 25°C

**Error Analysis:**

模型缺乏专业知识答错题目, 如果准确诊断为**苹果黑星病**, 便可以根据专业知识确定**苹果黑星病病原菌最适生长温度为20°C**。

Figure 18: A lack of knowledge example in Disease-P-M-K. **Red bold** highlights the wrong content.

Pest-P-K: Lack of Knowledge



Pest: 叶蝉科

Question: 图片中展示的害虫通常活跃在哪种环境区域?

Options: A: 亮色区域 B: 阴影处 C: 水面 D: 干燥处

Answer: A

GPT-4o Response:

这是一只**叶蝉科的害虫**，通常以吸食植物汁液为生，并且其颜色鲜艳显眼，如红色和蓝色，这表明它可能有捕食者防御的特性。以下是对问题的逐步分析：

第一步：害虫特征分析

叶蝉类害虫通常生活在植物表面，尤其是叶片表面，这与图片展示的虫子形态和它站在叶片上的姿态是吻合的。

第二步：与环境的关系

叶蝉类害虫一般会选择适宜其生存和繁殖的环境区域，具体特点如下：

- **亮色区域 (A): 叶蝉可能会暴露于阳光下，但这并不是其最偏好的环境，因为阳光直射可能会导致过热。**
- **阴影处 (B): 叶蝉多数活跃在植物叶片的背面或阴影较多的地方，以躲避阳光直射以及捕食者。**
- 水面 (C): 叶蝉并不生活在水面，它们与植物密切相关。
- 干燥处 (D): 干燥的环境并不是叶蝉的主要活动场所，因为它们需要植物汁液以维持生存。

Error Analysis:

模型缺乏专业知识答错题目，根据相关的农学专业可知叶蝉科的成虫具有**强趋光性**，因此相对于阴影处，该害虫更喜欢在**较亮的区域**活动。

Figure 19: A lack of knowledge example in Pest-P-K. **Red bold** highlights the wrong content.

Disease-P-M-K: Perceptual Error



Disease: 茄子病毒病

Question: 这种病害在什么环境条件下更容易发生?

Options: A: 阴凉湿润 B: 短日照 C: 低温多湿 D: 高温干旱

Answer: D

GPT-4o Response:

根据图片中植物叶片上的**白色粉状物质**，这种病害很可能是**白粉病**。白粉病是由子囊菌目真菌引起的一种常见作物病害，通常表现为叶片上出现白色粉状霉层。

白粉病通常在以下环境条件下更容易发生：

- 高湿度：高湿度有利于真菌孢子的萌发和繁殖。
- 温度适中：白粉病在温暖的天气中更为常见。

根据这些信息，以及选项来看：

- A. 阴凉湿润：不完全符合，因为虽然湿润有利于病害发生，但阴凉的环境不太符合白粉病的高发条件。
- B. 短日照：与白粉病的发生没有直接关系。
- C. 低温多湿：低温环境通常不利于白粉病的发展。
- D. 高温干旱：干旱不利于白粉病，但略高的温度可能会促进白粉病在温暖湿润环境中的发展。

虽然题目提供的选项可能不完全匹配白粉病的典型环境条件，但在这些选项中，“阴凉湿润”相对于其他选项来说可能更能引发白粉病的发生，尽管它并不是最完美的描述。因此，选择：

A. 阴凉湿润

Error Analysis:

模型感知图片中病害类型发生错误，由图片中**茄子叶片发生皱缩卷曲、黄化**可以感知、推断可能感染为**茄子病毒病**，但是模型错误感知为白粉病，导致后续判断错误。

Figure 20: An example of perceptual error in Disease-P-M-K. **Red bold** highlights the wrong content.



### Pest-P-K: Perceptual Error



Pest: 豆芫菁

Question: 图片中的昆虫若以幼虫形式度过冬天, 会在哪里越冬呢?

Options: A: 土表越冬 B: 土中越冬 C: 植物越冬 D: 水越冬

Answer: B

#### GPT-4o Response:

**这是一只黑肩绿天牛 (Atanycolus cappaerti), 通常称为“红头黑肩天牛”, 成虫通体黝黑, 头部呈明显红色。**

在思考这道题目时, 重点在于确定这种昆虫的幼虫习性。天牛科 (Cerambycidae) 的昆虫幼虫主要以木材或植物组织为食, 这通常意味着它们会在木质植物中度过一部分或全部幼虫阶段。因此, 这类昆虫的幼虫经常会在植物组织中 (如枯死或衰弱的树干、树枝中) 越冬。

所以, 考虑到黑肩绿天牛属昆虫的一般生活习性, 幼虫最有可能是在植物组织中度过冬天。

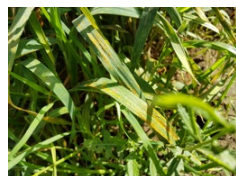
因此, 图片中的昆虫若以幼虫形式度过冬天, 会在哪里越冬呢? 选择C. 植物越冬。

#### Error Analysis:

模型感知错误, 没有正确判断出图中害虫的种类是**豆芫菁**, 因此未能正确推断害虫的越冬场所 (**土中越冬**)。

Figure 21: An example of perceptual error in Pest-P-K. **Red bold** highlights the wrong content.

### Disease-P-M-K: Question Misunderstanding



Disease: 小麦叶锈病

Question: 为减缓图片中病害的扩展, 采用哪种栽培防病策略最为合适?

Options: A: 喷施杀菌剂 B: 及时排水 C: 增施氮肥 D: 深耕灭茬

Answer: B

#### GPT-4o Response:

这张图片中显示的病害是小麦锈病, 其特征是叶片上出现黄色至橙色的小斑点。这种病害是由真菌引起的, 主要通过空气传播, 在较潮湿的条件下容易扩散。

为了减缓这种病害的扩展, 我们需要采取一种有效的**栽培策略**。让我们依次分析各个选项:

- A. 喷施杀菌剂: 这是应对真菌性病害的常见方法, 通过直接杀灭病原体, 可以有效控制病害的扩散。
- B. 及时排水: 虽然过多的潮湿可能促进锈病的发展, 但仅仅排水通常不足以有效控制病害。
- C. 增施氮肥: 过分施用氮肥可能导致植株过于茂盛, 透风性变差, 反而可能加重病害。
- D. 深耕灭茬: 这是一种清除土壤表面作物残留的方法, 可以减少病原菌的越冬体积, 但对于已经发病的植株来说, 短期内效果有限。

综上所述, 对于已经出现的小麦锈病, **喷施杀菌剂是最直接和有效的防治措施**。因此, 正确答案是:

A. 喷施杀菌剂

#### Error Analysis:

题目中的“**栽培防病策略**”是指**农业防治措施**, 而模型给出的答案是“喷施杀菌剂”。尽管**喷施杀菌剂**也能有效防治病害, 但它属于**化学防治措施**, 与题目要求不符, 反映出模型对问题的理解存在偏差。

Figure 22: A question misunderstanding example in Disease-P-M-K. **Red bold** highlights the wrong content.

Pest-P-K: Question Misunderstanding



Pest: 蚜虫

Question: 图中所示害虫特征——腹管——首次出现在什么地质年代?

Options: A: 侏罗纪 B: 二叠纪 C: 三叠纪 D: 白垩纪

Answer: D

GPT-4o Response:

这道题目要求我们确定图中所示的“腹管”这一害虫特征首次出现的地质年代。

- 图中展示的是蚜虫 (Aphid)，这些昆虫广为人知的特征之一就是腹部有突出的腹管。
  - 根据化石记录，蚜虫的祖先被认为出现在二叠纪 (Permian) 时代后期，并在三叠纪 (Triassic) 时期得以发展。
  - **尤其是蚜虫在三叠纪已经出现，它们的化石已在这段时期的地层中被发现。**
- 因此，考虑到问题中问的是“首次出现在什么地质年代”，答案是C. 三叠纪。

Error Analysis:

模型的回复的问题与题目中描述的问题不一致：问题中提问的是蚜虫的“腹管”结构是什么时候出现的，而模型回复的是蚜虫出现的时期。

Figure 23: A question misunderstanding example in Pest-P-K. **Red bold** highlights the wrong content.