Leveraging Vision Language Models for Specialized Agricultural Tasks

Muhammad Arbab Arshad[1] Talukder Zaki Jubery[1] Tirtho Roy[1] Rim Nassiri[1]
Asheesh K. Singh[1] Arti Singh[1] Chinmay Hegde[2] Baskar Ganapathysubramanian[1]
Aditya Balu[1] Adarsh Krishnamurthy[1] Soumik Sarkar[1,*]

[1]Iowa State University, USA
[2]New York University, USA
*Corresponding author: soumiks@iastate.edu

Abstract

As Vision Language Models (VLMs) become increasingly accessible to farmers and agricultural experts, there is a growing need to evaluate their potential in specialized tasks. We present AgEval, a comprehensive benchmark for assessing VLMs' capabilities in plant stress phenotyping, offering a solution to the challenge of limited annotated data in agriculture. Our study explores how general-purpose VLMs can be leveraged for domain-specific tasks with only a few annotated examples, providing insights into their behavior and adaptability. AgEval encompasses 12 diverse plant stress phenotyping tasks, evaluating zero-shot and few-shot in-context learning performance of state-of-the-art models including Claude, GPT, Gemini, and LLaVA. Our results demonstrate VLMs' rapid adaptability to specialized tasks, with the best-performing model showing an increase in F1 scores from 46.24% to 73.37% in 8-shot identification. To quantify performance disparities across classes, we introduce metrics such as the coefficient of variation (CV), revealing that VLMs' training impacts classes differently, with CV ranging from 26.02% to 58.03%. We also find that strategic example selection enhances model reliability, with exact category examples improving F1 scores by 15.38% on average. AgEval establishes a framework for assessing VLMs in agricultural applications, offering valuable benchmarks for future evaluations. Our findings suggest that VLMs, with minimal few-shot examples, show promise as a viable alternative to traditional specialized models in plant stress phenotyping, while also highlighting areas for further refinement. Results and benchmark details are available at: https://github.com/arbab-ml/AgEval

1. Introduction

Food security is a critical global challenge requiring sustainable improvements in agricultural productivity [1]. Agricultural research increasingly utilizes computer vision and AI to optimize crop management and enhance yield, profitability, and sustainability [2]. Phenotyping involves visually inspecting plants to extract agronomically relevant features. While traditional methods are time-consuming and labor-intensive, recent advancements in computer vision offer promising solutions to improve efficiency and scalability [3, 4]. These innovations present opportunities to revolutionize plant stress phenotyping, potentially leading to more objective, rapid, and large-scale assessments that could significantly boost agricultural productivity.

Plant stress phenotyping tasks primarily fall into three categories: identification, classification, and quantification. Our study adopts a comprehensive view of plant stress phenotyping, encompassing traditional stresses, pest infestations, and seed quality issues. Identification detects stress presence (e.g., drought, nutrient deficiency, pathogens, pests). Classification categorizes stress into expert-defined classes, while quantification measures stress severity or extent, including seed quality impacts. Each task requires sophisticated analytical methods [5].

Recent computer vision and machine learning advances offer new opportunities for automating plant stress phenotyping [6]. However, developing specialized models for agricultural applications faces challenges, primarily due to the need for high-quality annotated data. Expert knowledge in plant pathology, entomology, and agronomy is required for annotation, making it costly and time-consuming, thus creating a bottleneck in model development.

Researchers have explored techniques to develop effective models with limited annotated data [7], including transfer learning from large-scale datasets [8], self-supervised learning on unlabeled plant images [7], and leveraging vi-

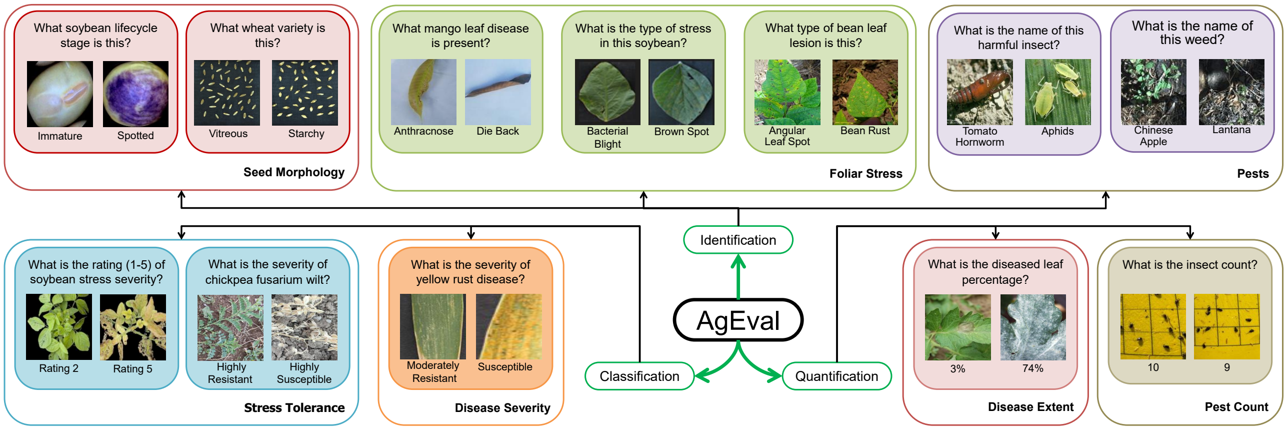

What soybean lifecycle stage is this?
Immature
Spotted
What wheat variety is this?
Vitreous
Starchy
Seed Morphology
What mango leaf disease is present?
Anthracnose
Die Back
What is the type of stress in this soybean?
Bacterial Blight
Brown Spot
What type of bean leaf lesion is this?
Angular Leaf Spot
Bean Rust
Foliar Stress
What is the name of this harmful insect?
Tomato Hornworm
Aphids
What is the name of this weed?
Chinese Apple
Lantana
Pests
Identification
AgEval
Classification
Quantification
What is the rating (1-5) of soybean stress severity?
Rating 2
Rating 5
What is the severity of chickpea fusarium wilt?
Highly Resistant
Highly Susceptible
Stress Tolerance
What is the severity of yellow rust disease?
Moderately Resistant
Susceptible
Disease Severity
What is the diseased leaf percentage?
3%
74%
Disease Extent
What is the insect count?
10
9
Pest Count

Figure 1. *Overview of the AgEval benchmark. The figure showcases sample images across different types of tasks and specific problems, representing diverse plant stress phenotyping challenges in agriculture.*

sion foundation models [9]. Despite progress, challenges remain due to domain differences between general and agricultural imagery, and the need for specialized visual understanding in plant stress phenotyping. The fine-grained nature of agricultural tasks often requires more nuanced feature recognition than general pretrained models offer, highlighting the need for adaptive, domain-specific approaches.

With traditional specialized model development, obtaining sufficient annotated data becomes increasingly challenging for each new use case. Typically, thousands of annotated examples are required to build effective models for specific agricultural tasks. As shown in Table S2, researchers have employed various specialized techniques to address this issue, including transfer learning, hybrid models, and custom architectures tailored to specific agricultural tasks. While these specialized models achieve impressive performance scores (e.g., 94% accuracy on Soybean Disease classification and 91% F1 score on DeepWeeds weed identification), they require significant domain expertise and computational resources to implement effectively. The high performance of these traditional approaches demonstrates their effectiveness for specific tasks, though their development and deployment costs remain substantial challenges.

Recent advances in vision language models (VLMs) have shown promising results in image-text understanding tasks [10, 11]. These models demonstrate remarkable capabilities in processing and understanding visual and textual information jointly [12–14]. Additionally, VLMs can perform few-shot in-context learning (learning from a small number of examples without model updates), adapting to new tasks with just a few examples in the prompt [15]. VLMs often require only a couple of examples to learn and perform well on new tasks, potentially offering a more efficient solution to the data scarcity problem in agricultural applications.

Our study explores VLMs' potential for plant stress phenotyping tasks, hypothesizing that their broad visual-textual understanding could benefit scenarios with limited annotated data. We present a comprehensive evaluation of state-of-the-art VLMs on identification, classification, and quantification of plant stresses, using a curated benchmark dataset reflecting real-world agricultural scenarios. As these models become increasingly accessible to agricultural stakeholders, our study assesses their reliability and effectiveness in specialized tasks, providing insights into their practical potential for plant stress phenotyping.

This study makes several key **contributions** to plant stress phenotyping using VLMs. First, it evaluates these models on plant stress phenotyping tasks, providing insights into their potential to overcome traditional approach limitations. Second, it introduces a curated benchmark dataset mirroring real-world agricultural scenarios. Third, the study analyzes few-shot in-context learning performance on specialized agricultural tasks. Fourth, it presents a comparative analysis of various state-of-the-art VLMs in plant stress phenotyping. Finally, the study provides a quantitative assessment of how example relevance impacts few-shot in-context learning, advancing our understanding of these models in agricultural applications.

2. Methodology

This section details our approach to evaluating vision language models (VLMs) for plant stress phenotyping tasks.

2.1. Task Formulation

We adopt the following taxonomy for plant stress phenotyping tasks, focusing on:

Identification (I): Determining the specific type of stress from predefined options (e.g., identifying bacterial blight in wheat, or identifying a specific weed).

Classification (C): Categorizing stress into distinct severity classes (e.g., classifying iron deficiency chlorosis in soybean leaves into low, medium, or high levels).

Quantification (Q): Measuring the extent or severity of stress numerically (e.g., percentage of leaf area affected by disease).

2.2. AgEval Benchmark Dataset Curation

We curated a diverse dataset comprising 12 subsets, each targeting specific plant stress phenotyping tasks. This collection, compiled from open-source resources, covers three main categories: Identification, Classification, and Quantification. These categories encompass various aspects of plant stress, from seed quality to pest infestations, reflecting the diverse challenges in agricultural stress assessment. The details of these datasets are provided in Table S1. We sampled 100 images from each dataset, evenly distributed across classes. Further details provided in Figure S1.

Identification: We considered datasets addressing the identification of **seed morphology** variations due to stresses, **foliar diseases**, and **pests** including weeds and insects that cause stresses. The Durum Wheat Dataset [16] and Soybean Seeds dataset [17] support seed morphology tasks, which involve identifying stress-induced changes in seed characteristics. The Mango Leaf Disease Dataset [18, 19], Bean Leaf Lesions Classification dataset [20], and Soybean Diseases dataset [21] enable foliar stress tasks, focusing on identifying diverse plant stresses affecting leaves, including diseases and adverse environmental conditions. The DeepWeeds dataset [22] and Dangerous Farm Insects dataset [23] facilitate pest identification tasks, which involve recognizing weeds and insects [24] that cause plant stress. These datasets and their associated tasks collectively contribute to assessing stress impacts on seed quality, disease management, and pest control strategies in agriculture.

Classification: We considered datasets for classification of **disease severity** and **stress tolerance** into expert-defined classes. The YELLOW-RUST-19 dataset [25–27] and Fusarium Wilt Disease in Chickpea dataset [28–30] support disease severity tasks, classifying disease stages caused by pathogens based on color and shape changes. The Iron Deficiency Chlorosis (IDC) Soybean Dataset [31] enables stress tolerance tasks, classifying abiotic stress stages caused by factors like nutrient deficiency or drought.

Quantification: We considered datasets addressing quantification of **pest** populations and **disease** extent. The InsectCount dataset [32] supports pest quantification tasks involving counting insects in field images to assess infestation levels and inform pest management decisions. The PlantDoc dataset [33, 34] enables disease quantification tasks, measuring plant stress by segmenting diseased areas in leaf images and quantifying the percentage of affected areas to assess severity and spread.

2.3. Model Selection and Evaluation

We evaluated six vision language models (VLMs): three state-of-the-art models (GPT-4o [35], Claude 3.5 Sonnet [36], and Gemini 1.5 Pro [37]), two budget-friendly options (Claude 3 Haiku [38] and Gemini 1.5 Flash [37]), and one open-source alternative (LLaVA v1.6 34B [39]). This selection encompasses a range of commercially available and open-source options to provide a comprehensive evaluation.

We evaluate VLM performance using both zero-shot and few-shot approaches. Zero-shot testing reveals inherent model capabilities, while few-shot testing (with 1, 2, 4, and 8 examples) demonstrates the models' ability to adapt to new tasks with minimal examples. For few-shot evaluations, we randomly select examples from the dataset. It's important to note that our few-shot learning refers to in-context learning capability with a few examples, not fine-tuning in a few-shot manner [40].

2.4. Performance Metrics

We evaluated model performance using task-specific metrics: F1-score for Identification (I) tasks, and Normalized Mean Absolute Error (NMAE) for both Classification (C) and Quantification (Q) tasks. For identification tasks, we used the weighted F1-score to account for potential class imbalance:

$$\text{F1-score} = \frac{\sum_{i=1}^{c} 2w_i \cdot \text{precision}_i \cdot \text{recall}i}{\sum_{i=1}^{c} w_i(\text{precision}_i + \text{recall}_i)} \tag{1}$$

where c is the number of classes and w_i is the weight of the i-th class, proportional to the number of samples in that class. For classification and quantification tasks, we calculated NMAE as:

$$\text{NMAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\max(y) - \min(y)} \cdot 100\% \tag{2}$$

where n is the number of samples, y_i is the true value, $\hat{y}_i$ is the predicted value, and $\max(y)$ and $\min(y)$ are the maximum and minimum labels in the dataset. Note that the labels for classification problem are mapped to their corresponding ordinal values to be able to calculating NMAE. To handle out-of-vocabulary predictions from vision language models in classification tasks, we assigned the worst possible score to unseen labels in the ordinal mapping. This approach ensures consistency in evaluating predictions across all models.

In addition to these task-specific metrics, we employ Mean Reciprocal Rank (MRR) to compare model performance across datasets. MRR is calculated separately for Identification (ID) and Classification/Quantification (CQ) datasets:

$$\text{MRR}k = \frac{1}{|M|}\sum_{j=1}^{|M|}\frac{1}{|D_k|}\sum_{i=1}^{|D_k|}\frac{1}{r_{k,i,j}} \tag{3}$$

where: $k \in \text{ID}, \text{CQ}$; $|M|$ is the number of models being compared; $|D_k|$ is the number of datasets in category k; and $r_{k,i,j}$ is the rank of model j for the i-th dataset in category k. Ranks are determined as:

$$r_{\text{ID},i,j} = \text{rank}(\text{F1-score}_{i,j}, \text{descending}) \tag{4}$$

$$r_{\text{CQ},i,j} = \text{rank}(\text{NMAE}_{i,j}, \text{ascending}) \tag{5}$$

MRR is a comparative metric that indicates relative performance among models. It is calculated separately for zero-shot ($s = 0$) and 8-shot ($s = 8$) settings:

$$\text{MRR}k, s = \frac{1}{|M|}\sum_{j=1}^{|M|}\frac{1}{|D_k|}\sum_{i=1}^{|D_k|}\frac{1}{r_{k,i,j,s}} \tag{6}$$

2.4.1 Relevance of Examples in Few-Shot Learning

This analysis explores the impact of example relevance in few-shot learning for identification tasks. We investigate how examples from the same category versus diverse examples affect vision language models' (VLMs) performance in predicting image labels. This study aims to understand whether examples from the same dataset (or in a real-world scenario, examples related to the farmer's actual input) positively influence predictions, and whether related information steers the model towards more accurate category-specific predictions.

Our analysis utilizes data from previous experiment runs, where few-shot examples and their labels were logged for each input across different shot settings.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the dataset, where x_i is an input and y_i its true label. For a k-shot setting, let $\mathcal{E}_k(x_i) = \{(e_j, l_j)\}_{j=1}^{k}$ be the set of examples provided in the prompt for input x_i, where e_j is an example and l_j its label, such that $e_j \neq x_i$ for all j.

Definition 1 (Bullseye Shot) *A k-shot prompt for input x_i is considered a bullseye shot if:*

$$\exists j \in \{1, \ldots, k\} : l_j = y_i \tag{7}$$

Intuitively, a bullseye shot occurs when at least one of the k example images provided in the prompt belongs to the same class as the target image being evaluated.
Note: A bullseye shot requires at least one example to match the true label, not necessarily all k examples.

This definition helps us quantify the effectiveness of providing relevant examples, offering insights into how VLMs can rapidly adapt to specialized agricultural tasks with minimal context, and how related information might enhance prediction accuracy for specific categories.

For each dataset and shot number $k \in \{1, 2, 4, 8\}$, we partition $\mathcal{D}$ into bullseye ($\mathcal{B}_k$) and non-bullseye ($\mathcal{N}_k$) subsets:

$$\mathcal{B}_k = \{(x_i, y_i) \in \mathcal{D} : \exists j \in \{1, \ldots, k\}, l_j = y_i \text{ in } \mathcal{E}_k(x_i)\} \tag{8}$$

$$\mathcal{N}_k = \mathcal{D} \setminus \mathcal{B}_k \tag{9}$$

We evaluate the impact using the performance delta from 0-shot ($\Delta F1$):

$$\Delta F1_{\mathcal{B}_k} = F1(\mathcal{B}_k) - F1_0 \tag{10}$$

$$\Delta F1_{\mathcal{N}_k} = F1(\mathcal{N}_k) - F1_0 \tag{11}$$

where $F1_0$ is the 0-shot F1-score and $F1(\cdot)$ is the F1-score on the respective subset. The average impact across all evaluated shot numbers is calculated as:

$$\overline{\Delta F1}_{\mathcal{B}} = \frac{1}{|\mathcal{K}|}\sum_{k \in \mathcal{K}} \Delta F1_{\mathcal{B}_k} \tag{12}$$

$$\overline{\Delta F1}_{\mathcal{N}} = \frac{1}{|\mathcal{K}|}\sum_{k \in \mathcal{K}} \Delta F1_{\mathcal{N}_k} \tag{13}$$

where $\mathcal{K}$ is the set of all shot numbers evaluated. This analysis quantifies the performance impact of having relevant examples in few-shot prompts.

2.4.2 Intra-task Uniformity

To evaluate the performance consistency of Vision Language Models (VLMs) across different classes within individual (identification) tasks, we employ the Coefficient of Variation (CV) of all classes' F1 scores. This analysis is crucial for understanding the robustness and reliability of VLMs in domain-specific applications, not just in agriculture but across various specialized fields. By quantifying performance variability, we can identify potential biases or gaps in model training that may impact real-world deployment.

For each model-dataset combination, we calculate the CV as follows:

$$CV = \frac{\sigma}{\mu} \cdot 100\% \tag{14}$$

where σ is the standard deviation of F1 scores across classes, and μ is the mean F1 score. The CV provides a normalized measure of dispersion, allowing for comparison across datasets with different scales.

For each identification dataset d and model m, we calculate:

$$CV_{d,m} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(F1_{i,d,m} - \overline{F1_{d,m}})^2}}{\overline{F1_{d,m}}} \cdot 100\% \tag{15}$$

Table 1. *0-shot Performance of VLMs on AgEval Benchmark, Models Sorted by Average Performance (Highest to Lowest)*

(a) *Identification - Metric: F1 Score (Higher is Better).* *Highest* *Second Highest*

Model	Seed Morphology		Foliar Stress			Pests	
	Durum Wheat	Soybean	Mango Leaf Disease	Bean Leaf Lesions	Soybean Diseases	Dangerous Insects	Weeds
Gemini-pro-1.5	55.56	26.24	42.91	77.22	21.78	82.67	46.83
GPT-4o	55.1	19.0	58.41	65.92	3.7	82.79	38.77
Claude-3.5-sonnet	55.56	38.7	49.82	68.65	8.54	82.02	18.85
Gemini-flash-1.5	53.64	24.58	42.85	70.61	14.41	80.38	32.83
Claude-3-haiku	36.06	31.24	29.83	55.26	12.69	51.28	13.86
LLaVA v1.6 34B	40.56	13.74	13.63	44.03	8.54	18.54	8.68

(b) *Classification and Quantification - Metric: NMAE (Lower is Better).* *Lowest* *Second Lowest*

Model	Disease Severity		Stress Tolerance	Pest	Disease
	Yellow Rust 19	FUSARIUM 22	IDC	InsectCount	PlantDoc
Claude-3.5-sonnet	22.29	18.25	26.28	16.25	15.59
GPT-4o	17.19	37.0	18.88	15.8	18.14
Gemini-flash-1.5	31.25	24.0	19.39	16.32	21.22
Gemini-pro-1.5	26.25	33.0	30.87	29.0	9.57
Claude-3-haiku	37.08	25.75	22.86	28.34	22.14
LLaVA v1.6 34B	35.94	30.6	25.51	26.19	41.72

where n is the number of classes in dataset d, $F1_{i,d,m}$ is the F1 score for class i, and $\overline{F1_{d,m}}$ is the mean F1 score across all classes for dataset d and model m.

We then compute average CVs across datasets and models:

$$\overline{CV_d} = \frac{1}{|M|} \sum_{m \in M} CV_{d,m} \tag{16}$$

$$\overline{CV_m} = \frac{1}{|D|} \sum_{d \in D} CV_{d,m} \tag{17}$$

where M is the set of all models and D is the set of all datasets.

These metrics allow us to quantify and compare the consistency of model performance across classes, highlighting areas where VLMs excel and identifying potential challenges in specific classes or model behaviors. This analysis provides valuable insights into the adaptability and robustness of VLMs in handling diverse tasks, which is essential for their effective application in specialized domains beyond agriculture, such as medical imaging, industrial quality control, or environmental monitoring.

2.5. Prompt Engineering

We designed task-specific prompts to guide the VLMs in performing the ICQ tasks. The prompts were structured to provide clear instructions and ensure consistent output formatting across all models.

```
Given the image, identify the class. Use the
↪  following list of possible classes for your
↪  prediction It should be one of the :
↪  {expected_classes}. Be attentive to subtle
↪  details as some classes may appear similar.
↪  Provide your answer in the following JSON
↪  format:
{"prediction": "class_name"}...
```

For identification and classification tasks, we used a universal prompt template shown above (unless otherwise stated). Specialized prompts used for quantification tasks are given in the Supplement. For few-shot scenarios, we prepend examples to these prompts, maintaining the same structure and format across all shot counts to ensured consistency in inputs given to multiple model.

3. Results

3.1. Zero-shot Performance

The zero-shot performance of Vision Language Models (VLMs) on the AgEval benchmark reveals interesting patterns (Figure 2). In identification tasks, Gemini-pro-1.5 demonstrates the strongest performance with an MRR of 0.69. For classification and quantification tasks, GPT-4o emerges as the top performer with an MRR of 0.70, closely followed by Claude-3.5-sonnet at 0.54. This highlights the varying strengths of different models across task types, emphasizing the importance of model selection based on specific requirements within the agricultural domain.

The strong performance of larger models like GPT-4o and Gemini-pro-1.5 suggests that general-purpose training translates well to domain-specific performance, underscoring the value of comprehensive pretraining in enabling VLMs to adapt to specialized agricultural tasks without additional fine-tuning.

Identification Tasks (F1 Score): In zero-shot identification tasks, Gemini-pro-1.5 leads with an average F1 score of 50.45 across the 7 tasks, outperforming others in 4 out of 7 tasks. GPT-4o follows with 46.24 F1, showing strength in Mango Leaf Disease identification (58.41). Claude-3.5-sonnet performs competitively (46.02), particularly in Soybean Seeds identification (38.70).

Classification and Quantification Tasks (NMAE): For

Table 2. *8-shot Performance of VLMs on AgEval Benchmark, Models Sorted by Average Performance (Highest to Lowest)*

(a) *Identification - Metric: F1 Score (Higher is Better).* *Highest* *Second Highest*

Model	Seed Morphology		Foliar Stress			Pests	
	Durum Wheat	Soybean	Mango Leaf Disease	Bean Leaf Lesions	Soybean Diseases	Dangerous Insects	Weeds
GPT-4o	95.94	48.29	80.96	86.9	62.96	82.56	56.03
Gemini-pro-1.5	79.66	52.19	71.68	78.17	24.41	82.98	49.96
Claude-3.5-sonnet	89.66	51.17	61.68	84.78	11.07	81.89	27.17
Gemini-flash-1.5	83.7	48.09	64.66	73.42	23.67	82.72	41.89
Claude-3-haiku	53.29	38.02	38.92	46.42	8.81	45.08	15.34
LLaVA v1.6 34B	46.8	23.1	22.84	48.5	10.53	12.08	13.23

(b) *Classification and Quantification - Metric: NMAE (Lower is Better).* *Lowest* *Second Lowest*

Model	Disease Severity		Stress Tolerance	Pest	Disease
	Yellow Rust 19	FUSARIUM 22	IDC	InsectCount	PlantDoc
Claude-3.5-sonnet	16.04	14.0	16.84	5.75	11.31
Gemini-pro-1.5	17.08	17.0	12.04	9.57	13.04
Gemini-flash-1.5	20.83	17.5	15.56	6.11	12.92
GPT-4o	15.83	19.75	60.82	6.84	10.93
Claude-3-haiku	25.69	21.75	23.06	19.16	17.57
LLaVA v1.6 34B	30.56	60.0	60.82	13.18	26.28

Table 3. *Few Shot Learning: Impact of having at least 1 example with same category as ground truth (Bullseye example).* *Highest* *Lowest* *across 1, 2, 4, and 8-shot settings for both Bullseye and Non-Bullseye. Average Impact values are in* ***bold****.*

	Baseline	Bullseye Shots					Non-Bullseye Shots				
	0-shot	1-shot	2-shot	4-shot	8-shot	Avg. Impact	1-shot	2-shot	4-shot	8-shot	Avg. Impact
Durum Wheat	51.18	+31.10	+23.74	+28.02	+30.67	**+28.38**	-01.74	+01.03	-04.39	-11.85	**-04.24**
Soybean Seeds	27.95	+27.10	+15.54	+13.79	+21.77	**+19.55**	-01.48	+05.57	+05.44	+09.20	**+04.68**
Mango Leaf Disease	44.76	+22.61	+22.90	+17.34	+28.30	**+22.79**	-02.02	-02.96	-00.40	+03.31	**-00.52**
Bean Leaf Lesions	67.53	+11.46	+05.01	+07.67	+06.90	**+07.76**	-06.08	-05.61	-09.65	-07.87	**-07.30**
Soybean Diseases	12.23	+49.05	+17.71	+09.26	+06.54	**+20.64**	+00.26	+06.98	+04.95	+11.73	**+05.98**
Dangerous Insects	75.83	-12.05	-10.92	-12.67	-04.15	**-09.95**	-00.92	+02.04	+03.37	+01.90	**+01.60**
DeepWeeds	30.23	+22.37	+24.79	+13.85	+12.86	**+18.47**	-02.03	-03.32	-00.35	+01.55	**-01.04**
Average	44.24	+21.66	+14.11	+11.04	+14.70	**+15.38**	-02.00	+00.53	-00.15	+01.14	**-00.12**

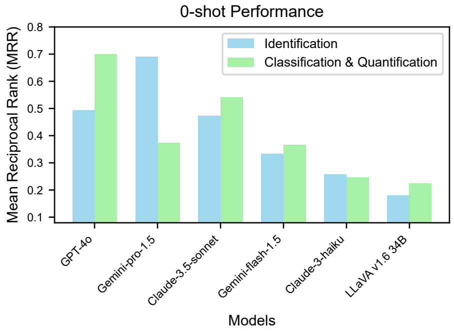


Figure 2. *Zero-shot comparative performance of VLMs.*

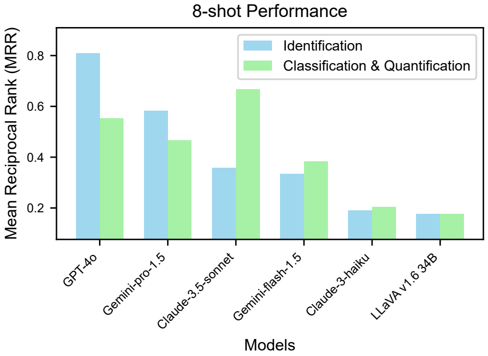


Figure 3. *8-shot comparative performance of VLMs.*

zero-shot classification and quantification, Claude-3.5-sonnet leads with the lowest average NMAE of 19.73 across 5 tasks. GPT-4o follows closely (21.40), outperforming in 3 out of 5 tasks. These results reinforce that larger VLMs can effectively leverage their general-purpose training to perform well on specialized agricultural tasks, even without domain-specific fine-tuning.

3.2. Full (8)-shot Performance

The introduction of eight examples per task (8-shot learning) leads to significant improvements in model performance (Figure 3), achieving results that traditionally require thousands of annotated examples. For identification tasks, GPT-4o achieves the highest MRR of 0.81, a substantial increase from its zero-shot performance. In classification and

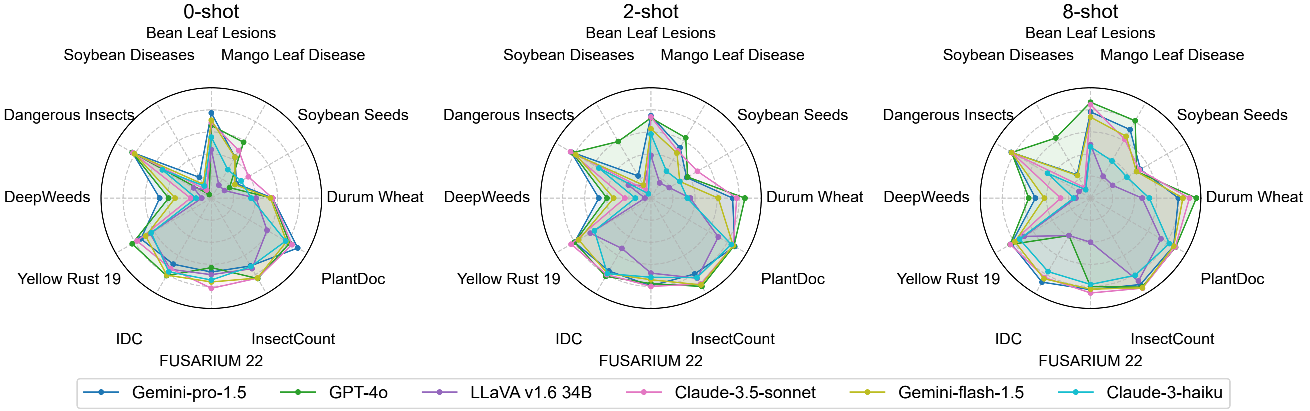


Figure 4. *Performance comparison of models across 0-shot, 2-shot, and 8-shot settings on various datasets. F1 scores are shown directly, while NMAE is inverted (100 - NMAE) for consistent visualization, with higher values indicating better performance*

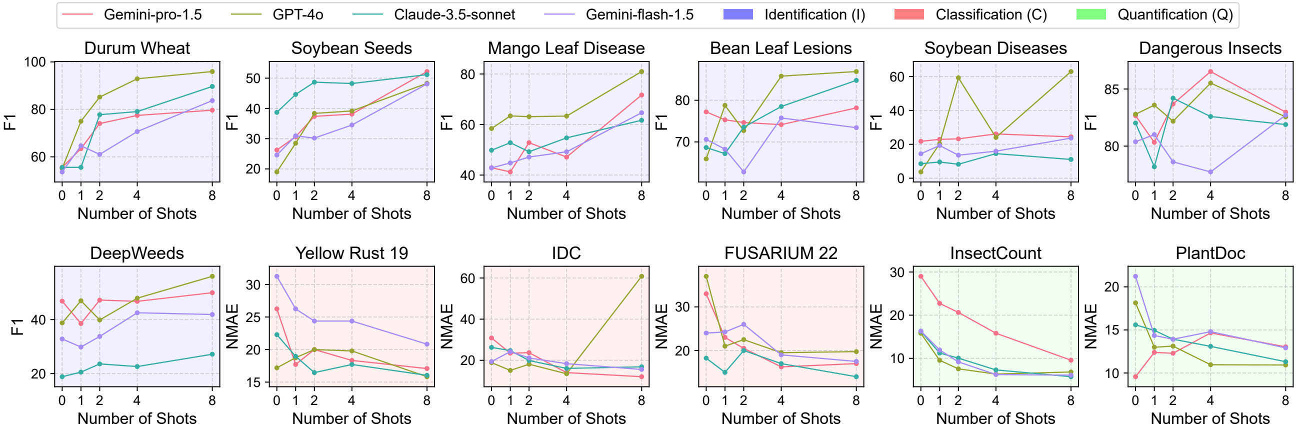


Figure 5. *Performance comparison on individual tasks of the AgEval benchmark across different shot settings (0 to 8 shots) for top-4 performing VLMs.*

quantification tasks, Claude-3.5-sonnet emerges as the top performer with an MRR of 0.66, while GPT-4o maintains strong performance with an MRR of 0.55.

Most models show improved performance with 8-shot learning compared to zero-shot, particularly in complex tasks, highlighting the efficiency of VLMs in learning from limited data.

Identification Tasks (F1 Score): In 8-shot identification, GPT-4o achieves an average F1 score of 73.37, outperforming in 5 out of 7 tasks. It excels in Durum Wheat (95.94) and Bean Leaf Lesions (86.90) identification. Gemini-pro-1.5 maintains strong performance (62.72), leading in DeepWeeds identification (49.96).

Classification and Quantification Tasks (NMAE): For 8-shot classification and quantification, Claude-3.5-sonnet leads with an average NMAE of 12.78, excelling in 3 out of 5 tasks. Gemini-pro-1.5 shows substantial improvement (13.74), particularly in IDC classification (12.04).

LLaVA's performance unexpectedly deteriorates with increased shot numbers, suggesting potential limitations in its in-context learning capabilities for AgEval tasks. These results underscore the varying impacts of few-shot learning across models and tasks, highlighting the importance of model selection based on specific agricultural task requirements.

3.3. Relevance of Examples in Few-Shot Learning

The relevance of examples in few-shot learning significantly influences model performance across various identification tasks. As shown in Table 3, the presence of exact category examples (Bullseye) consistently improves F1 scores, with an average increase of 15.38% across all shot settings. This impact is most pronounced in the 1-shot scenario (+21.66%). The absence of exact category examples (Non-Bullseye) has a minimal overall impact (-0.12% on average), suggesting VLM robustness to less relevant examples, especially as the number of shots increases.

The benefits of relevant examples persist across different shot counts, albeit with diminishing returns. The average Bullseye impact decreases from +21.66% in 1-shot to +14.70% in 8-shot settings. Dataset-specific variations are substantial: Soybean Diseases shows the highest Bullseye impact (+49.05% in 1-shot), while Bean Leaf Lesions exhibits more modest improvements (+11.46% in 1-shot). The Dangerous Insects dataset presents an interesting case, showing slight negative impacts even with Bullseye examples (-9.95% average impact), which may indicate domain-specific nuances. These findings underscore the importance of example selection in few-shot learning, particularly in low-shot scenarios. Please note that Llava was excluded from this analysis due to challenges in few-shot learning for our benchmark (see Figure 5).

3.4. Intra-task Uniformity

Among VLMs, GPT-4o demonstrated the most consistent performance (Figure 6) with the lowest average CV (26%), while Claude-3-haiku showed higher variability (CV=58%). Gemini-pro-1.5 and Gemini-flash-1.5 exhibited moderate consistency (CV≈39%), with Claude-3.5-sonnet performing slightly better (CV=32%).

Regarding datasets, Soybean Diseases exhibited the highest average CV (81.29%), indicating variability in model performance across its classes, potentially due to its low image resolution. Conversely, Durum Wheat showed the lowest average CV (14.28%), implying more uniform performance. DeepWeeds and Mango Leaf Disease also demonstrated higher variability (CV >40%), highlighting areas for targeted improvements in VLM training to enhance performance uniformity.

These findings underscore the importance of considering not just overall accuracy, but also consistency across classes when selecting models for agricultural identification tasks. Detailed plots for each dataset's classes are provided in the supplement.

3.5. Key Findings

Here are 4 key findings from our evaluation:

(1) GPT-4o demonstrates strong adaptability across AgEval tasks, showing significant improvement with few-shot learning (F1 score increase from 46.24 to 73.37).

(2) Example relevance significantly impacts few-shot learning: On average, exact category examples (bullseyes) improve F1 scores by 15.38%, while related examples from different classes have minimal impact.

(3) VLM performance within datasets shows some variation: Coefficient of Variation ranges from 26.02% (GPT-4o) to 58.03% (Claude-3-haiku), indicating opportunities for further refinement in achieving consistent accuracy across all classes in plant stress phenotyping tasks.

(4) Different models exhibit complementary strengths: Gemini-pro-1.5 excels in zero-shot identification (MRR 0.69), while GPT-4o leads in zero-shot classification/quantification (MRR 0.70). Claude Sonnet-3.5 consistently performs well in classification/quantification tasks, showcasing the diverse capabilities of VLMs in agricultural applications.

4. Conclusions

This study introduces AgEval, a benchmark for evaluating Vision Language Models (VLMs) on plant stress phenotyping tasks. We assembled diverse tasks across crops and stress types. Our evaluation examines zero-shot and few-shot performance, example relevance impact, and performance consistency using Coefficient of Variation. Results show VLMs' potential in addressing plant stress phenotyping challenges, with complementary strengths across models and tasks. This work establishes a baseline for future VLMs in agricultural contexts.

VLMs demonstrate adaptability to specialized agricultural tasks, with improvements in few-shot learning. Intra-task uniformity variation highlights refinement opportunities. These findings show VLMs are scalable solutions for plant stress phenotyping with minimal context. While VLMs may not match specialized models' peak performance, they offer valuable flexibility and reduced data requirements for practical agricultural applications

Future research could expand to broader agricultural tasks beyond plant stress phenotyping. Exploring increased shot counts could provide further insights. Fine-tuning models could enhance non-bullseye example performance and improve intra-task uniformity. Assessing practical deployment aspects in real-world settings, including computational requirements, data update strategies, integration with existing systems, and environmental and privacy considerations, will be crucial.

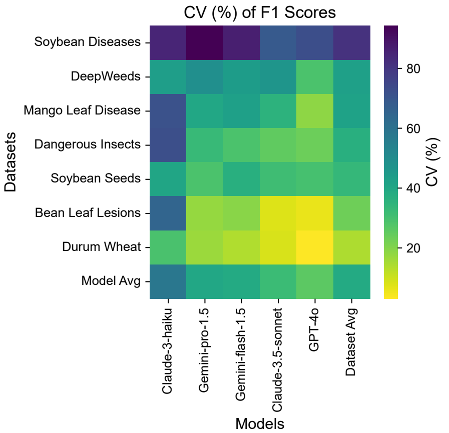


Figure 6. *Heatmap of Coefficient of Variation (CV) for F1 scores (lower is better) across models and identification datasets.*

References

[1] Pamela K Anderson, Andrew A Cunningham, Nikkita G Patel, Francisco J Morales, Paul R Epstein, and Peter Daszak. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in ecology & evolution*, 19(10):535–544, 2004. 1

[2] Soumik Sarkar, Baskar Ganapathysubramanian, Arti Singh, Fateme Fotouhi, Soumyashree Kar, Koushik Nagasubramanian, Girish Chowdhary, Sajal K Das, George Kantor, Adarsh Krishnamurthy, Nirav Merchant, and Asheesh K. Singh. Cyber-agricultural systems for crop breeding and sustainable production. *Trends in Plant Science*, 2023. 1

[3] Noah Fahlgren, Malia A Gehan, and Ivan Baxter. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology*, 24:93–99, 2015. 1

[4] José Luis Araus and Jill E Cairns. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in plant science*, 19(1):52–61, 2014. 1

[5] Jordan R Ubbens and Ian Stavness. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Frontiers in plant science*, 8:1190, 2017. 1

[6] Michael P Pound, Jonathan A Atkinson, Alexandra J Townsend, Michael H Wilson, Marcus Griffiths, Aaron S Jackson, Adrian Bulat, Georgios Tzimiropoulos, Darren M Wells, Erik H Murchie, et al. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience*, 6(10):gix083, 2017. 1

[7] Franklin C Ogidi, Mark G Eramian, and Ian Stavness. Benchmarking self-supervised contrastive learning methods for image-based plant phenotyping. *Plant Phenomics*, 5: 0037, 2023. 1

[8] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 1

[9] Feng Chen, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Adapting vision foundation models for plant phenotyping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 604–613, 2023. 2

[10] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[11] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023. 2

[12] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[13] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

[14] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 2

[15] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024. 2

[16] Esra Kaya and İsmail Saritas. Towards a real-time sorting system: Identification of vitreous durum wheat kernels using ann based on their morphological, colour, wavelet and gaborlet features. *Computers and Electronics in Agriculture*, 166:105016, 2019. 3, S2

[17] Warcoder. Soyabean seeds dataset. https://www.kaggle.com/datasets/warcoder/soyabean-seeds, Accessed: 2024-07-15. 3, S2

[18] Arya Shah. Mango leaf disease dataset. https://www.kaggle.com/datasets/aryashah2k/mango-leaf-disease-dataset, Accessed: 2024-07-15. 3, S2

[19] Sarder Iftekhar Ahmed, Muhammad Ibrahim, Md Nadim, Md Mizanur Rahman, Maria Mehjabin Shejunti, Taskeed Jabid, and Md Sawkat Ali. Mangoleafbd: A comprehensive image dataset to classify diseased and healthy mango leaves. *Data in Brief*, 47:108941, 2023. 3, S2

[20] Marquis03. Bean leaf lesions classification. https://www.kaggle.com/datasets/marquis03/bean-leaf-lesions-classification, Accessed: 2024-07-15. 3, S2

[21] Sambuddha Ghosal, David Blystone, Asheesh K Singh, Baskar Ganapathysubramanian, Arti Singh, and Soumik Sarkar. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18):4613–4618, 2018. 3, S2

[22] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019. 3, S2

[23] Tarun Dalal. Dangerous insects dataset. https://www.kaggle.com/datasets/tarundalal/dangerous-insects-dataset, Accessed: 2024-07-15. 3, S2

[24] Benjamin Feuer, Ameya Joshi, Minsu Cho, Shivani Chiranjeevi, Zi Kang Deng, Aditya Balu, Asheesh K Singh, Soumik Sarkar, Nirav Merchant, Arti Singh, et al. Zero-shot insect detection via weak language supervision. *The Plant Phenome Journal*, 7(1):e20107, 2024. 3

[25] Tolgahayit. Yellow rust disease in wheat. https://www.kaggle.com/datasets/tolgahayit/yellowrust19-yellow-rust-disease-in-wheat, Accessed: 2024-07-15. 3, S2

[26] Tolga Hayit, Hasan Erbay, Fatih Varçın, Fatma Hayit, and Nilüfer Akci. Determination of the severity level of yellow rust disease in wheat by using convolutional neural networks. *Journal of plant pathology*, 103(3):923–934, 2021.

[27] Tolga Hayıt, Hasan Erbay, Fatih Varçın, Fatma Hayıt, and Nilüfer Akci. The classification of wheat yellow rust disease based on a combination of textural and deep features. *Multimedia Tools and Applications*, 82(30):47405–47423, 2023. 3, S2

[28] Tolgahayit. Fusarium wilt disease in chickpea dataset. https://www.kaggle.com/datasets/tolgahayit/fusarium-wilt-disease-in-chickpea-dataset, Accessed: 2024-07-15. 3, S2

[29] Tolga Hayit, Ali Endes, and Fatma Hayit. Knn-based approach for the classification of fusarium wilt disease in chickpea based on color and texture features. *European Journal of Plant Pathology*, 168(4):665–681, 2024.

[30] Tolga Hayit, Ali Endes, and Fatma Hayit. The severity level classification of fusarium wilt of chickpea by pre-trained deep learning models. *Journal of Plant Pathology*, 106(1): 93–105, 2024. 3, S2

[31] Hsiang Sing Naik, Jiaoping Zhang, Alec Lofquist, Teshale Assefa, Soumik Sarkar, David Ackerman, Arti Singh, Asheesh K Singh, and Baskar Ganapathysubramanian. A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant methods*, 13:1–12, 2017. 3, S2

[32] AT Nieuwenhuizen, J Hemming, Dirk Janssen, HK Suh, L Bosmans, V Sluydts, N Brenard, E Rodríguez, and MDM Tellez. Raw data from yellow sticky traps with insects for training of deep learning convolutional neural network for object detection. *Wageningen University & Research*, 2019. 3, S2

[33] Fakhrealam9537. Leaf disease segmentation dataset. https://www.kaggle.com/datasets/fakhrealam9537/leaf-disease-segmentation-dataset, Accessed: 2024-07-15. 3, S2

[34] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. 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, 2020. Association for Computing Machinery. ISBN 9781450377386. doi: 10.1145/3371158.3371196. URL https://doi.org/10.1145/3371158.3371196. 3, S2

[35] openai. GPT-4o(ision) System Card. https://cdn.openai.com/gpt-4o-system-card.pdf, Accessed 15-07-2024. 3

[36] Anthropic. Claude 3.5 sonnet model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf, Accessed 15-07-2024. 3

[37] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3

[38] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf, Accessed 15-07-2024. 3

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[40] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*, 2023. 3

[41] Murat Koklu. Durum wheat dataset. https://www.kaggle.com/datasets/muratkokludataset/durum-wheat-dataset, Accessed: 2024-07-15. S2

[42] Imsparsh. Deepweeds. https://www.kaggle.com/datasets/imsparsh/deepweeds, Accessed: 2024-07-15. S2

S1. Supplementary Material

This section provides additional details about the datasets used in this study, including their names, links, and the classes they contain.

S1.1. Prompts

For identification tasks, we used a universal prompt template, which was provided in the prompt engineering section, asking models to identify the class from a given list and provide the answer in JSON format. For classification and quantification tasks, we employed specialized prompts tailored to each dataset's requirements. These prompts included specific instructions on rating scales or counting methods relevant to the task at hand.

IDC Dataset:

```
Analyze this image of a soybean canopy to
↪  determine the iron deficiency chlorosis
↪  (IDC) severity rating. The images are of
↪  soybean plants exhibiting various levels of
↪  IDC symptoms, ranging from healthy green
↪  plants to those with severe chlorosis and
↪  necrosis. Evaluate the extent of yellowing
↪  and browning in the canopy. Provide your
↪  answer in the following JSON format:
{{"prediction": "number"}}
Replace "number" with your best estimate of the
↪  IDC severity rating based on your analysis
↪  of the image.
The number should be entered exactly as a whole
↪  number (without any symbols) in a range of
↪  {expected_classes}. Higher value means more
↪  severity.
The response should start with {{ and contain
↪  only a JSON object (as specified above) and
↪  no other text.
```

Insect Count:

```
Analyze this image of a yellow sticky insect
↪  trap. Count the total number of visible
↪  insects caught on the trap. Only look for
↪  insects which are easily visible to naked
↪  eye and look bigger compared to the other
↪  background artifacts. Provide your answer in
↪  the following JSON format:
{"prediction": "number"}
Replace "number" with your best estimate of the
↪  total insect count based on your analysis of
↪  the image. The number should be entered
↪  exactly as a whole number (without any
↪  symbols) in a range of {expected_classes}
↪  The response should start with { and contain
↪  only a JSON object (as specified above) and
↪  no other text.
```

PlantDoc *(Disease Quantification)*

```
Analyze this image of a leaf to get the total
↪  percentage of affected leaf. The images are
↪  of several plant leaf-like Apple Scab Leaf,
↪  Apple rust leaf, Bell_pepper leaf spot, Corn
↪  leaf blight, Potato leaf early blight, etc.
↪  The affected area is: diseased leaf area /
↪  total image area. Provide your answer in the
↪  following JSON format:
{"prediction": "number"}
Replace "number" with your best estimate of the
↪  percent on your analysis of the image. The
↪  number should be entered exactly as a whole
↪  number (without any symbols) in a range of
↪  {expected_classes} The response should start
↪  with { and contain only a JSON object (as
↪  specified above) and no other text.
only a JSON object (as specified above) and no
↪  other text.
```

S1.2. Additional dataset details

Table S1 provides a comprehensive overview of the datasets used in the AgEval benchmark. It categorizes each dataset based on its primary task (Identification, Classification, or Quantification) and subcategory (e.g., Seed Morphology, Foliar Stress, Pests). The table includes key information such as the number of images, classes, year of creation, geographical location, and the evaluation metric used. This diverse collection of datasets covers various aspects of plant stress phenotyping, ranging from seed quality assessment to disease severity classification across different crops and regions. Table S2 provides a comparison of the performance of traditional models on these datasets.

Figure S1 provides a treemap visualization of the AgEval benchmark datasets, illustrating the distribution and hierarchy of tasks, subcategories, and individual classes. This comprehensive view highlights the diverse range of plant stress-related challenges addressed by AgEval, for all the AgEval benchmark. The size of each rectangle corresponds to the number of instances in that class, offering insights into the dataset composition and balance. We sampled 100 images in total from each dataset and the size corresponds to the resulting number of instances per class in each dataset used to build AgEval.

S1.3. Additional details on intra-task uniformity

Figure S2 provides a detailed examination of intra-task uniformity across different datasets in the AgEval benchmark. Each subfigure represents a specific dataset, showcasing the F1 scores for the highest, median, and lowest performing classes based on 0-shot performance. The visualization for each class displays both the 0-shot F1 score (solid bars) and the additional gain in F1 score achieved with 8-shot learning (hatched bars) for all six evaluated models. This comprehensive view highlights the significant performance disparities among classes within each task, supporting our finding that the coefficient of variance

Table S1. *Classification of Agricultural Image Datasets. Categories: I (Identification), C (Classification), Q (Quantification)*

Dataset	Category	Subcategory	Description	# of Classes	Year	Location	Metric
Durum Wheat [16, 41]	I	Seed Morphology	Wheat variety identification	3	2019	Turkey	F1
Soybean Seeds [17]	I	Seed Morphology	Soybean quality prediction	5	N/A	N/A	F1
Mango Leaf Disease [18, 19]	I	Foliar Stress	Mango leaf disease classification	8	2022	Bangladesh	F1
Bean Leaf Lesions [20]	I	Foliar Stress	Bean leaf lesion type classification	3	N/A	N/A	F1
Soybean Diseases [21]	I	Foliar Stress	Soybean stress identification	9	2016	United States	F1
Dangerous Insects [23]	I	Pests	Harmful insects identification	15	N/A	N/A	F1
DeepWeeds [22, 42]	I	Pests	Weeds species identification	9	2019	Australia	F1
Yellow Rust 19 [25–27]	C	Disease Severity	Wheat yellow rust severity	6	2021	Turkey	NMAE
FUSARIUM 22 [28–30]	C	Disease Severity	Chickpea fusarium wilt severity	5	2023	Turkey	NMAE
IDC [31]	C	Stress Tolerance	Soybean stress severity	5	2015	United States	NMAE
InsectCount [32]	Q	Pest Count	Insect count in images	-	2021-2022	N/A	NMAE
PlantDoc [33, 34]	Q	Disease	Percentage of the leaf that is diseased	-	N/A	N/A	NMAE

Table S2. *Performance of the Traditional Models on Agricultural Image Datasets*

Dataset	Method/Approach Used	Reported Metric (Score)	Train	Validation	Test
Durum Wheat [16, 41]	Transfer learning with EfficientNetB3	F1 Score (100)	227(70%)	49 (15%)	49 (15%)
Soybean Seeds [17]	Transfer learning with ResNet50	Accuracy (89)	4410(80%)	-	1103(20%)
Mango Leaf Disease [18, 19]	Transfer learning with EfficientNetB3	Accuracy (100)	3200(80%)	480(12%)	320(8%)
Bean Leaf Lesions [20]	Hybrid Model (ViT, SVM)	F1 score (91)	974 (84%)	133 (11%)	60 (5%)
Soybean Diseases [21]	Convolutional neural network	Accuracy (94)	53266(81%)	5918(9%)	6576(10%)
Dangerous Insects [23]	Transfer learning with Xception	Accuracy (77)	1272 (80%)	287 (18%)	32 (2%)
DeepWeeds [22, 42]	Transfer learning with ResNet26	F1 score (91)	11205(64%)	2801(16%)	3501(20%)
Yellow Rust 19 [25–27]	CNN-CGLCM with SVM	Accuracy (92)	13500 (90%)	-	1500 (10%)
FUSARIUM 22 [28–30]	Hybrid Classifier (ViT,CatBoost)	F1 score (75)	2950(68%)	521 (12%)	868(20%)
IDC [31]	Hierarchical classification	Accuracy (96)	1479(75%)	-	493 (25%)
InsectCount [32]	Internal dataset	No baseline published			
PlantDoc [33, 34]	No baseline exists on this data for this task				

(CV) ranges from 26.02% to 58.03% across models. The stark differences between the highest and lowest performing classes underscore the need for subject matter expertise to achieve reliable performance, especially for "difficult" classes.

S1.4. Anecdotal Samples from Each Task:

Two samples and their corresponding predictions with respect to 0 and 8 shot are provided later. Please note that the questions are for illustration and actual prompts provided are in Section S1.1

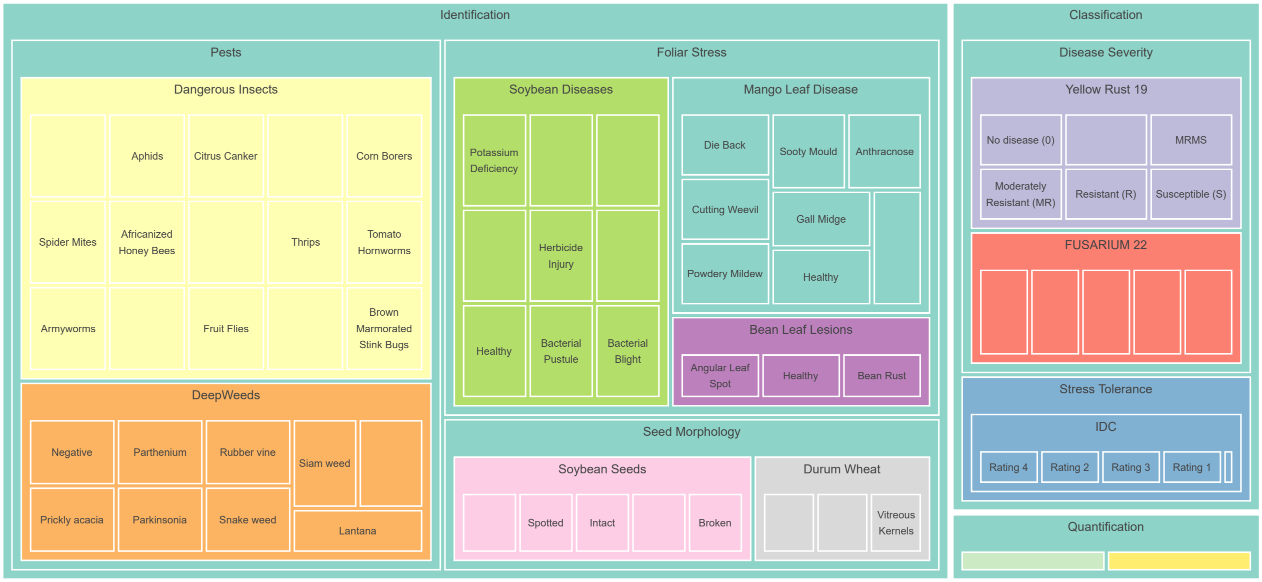
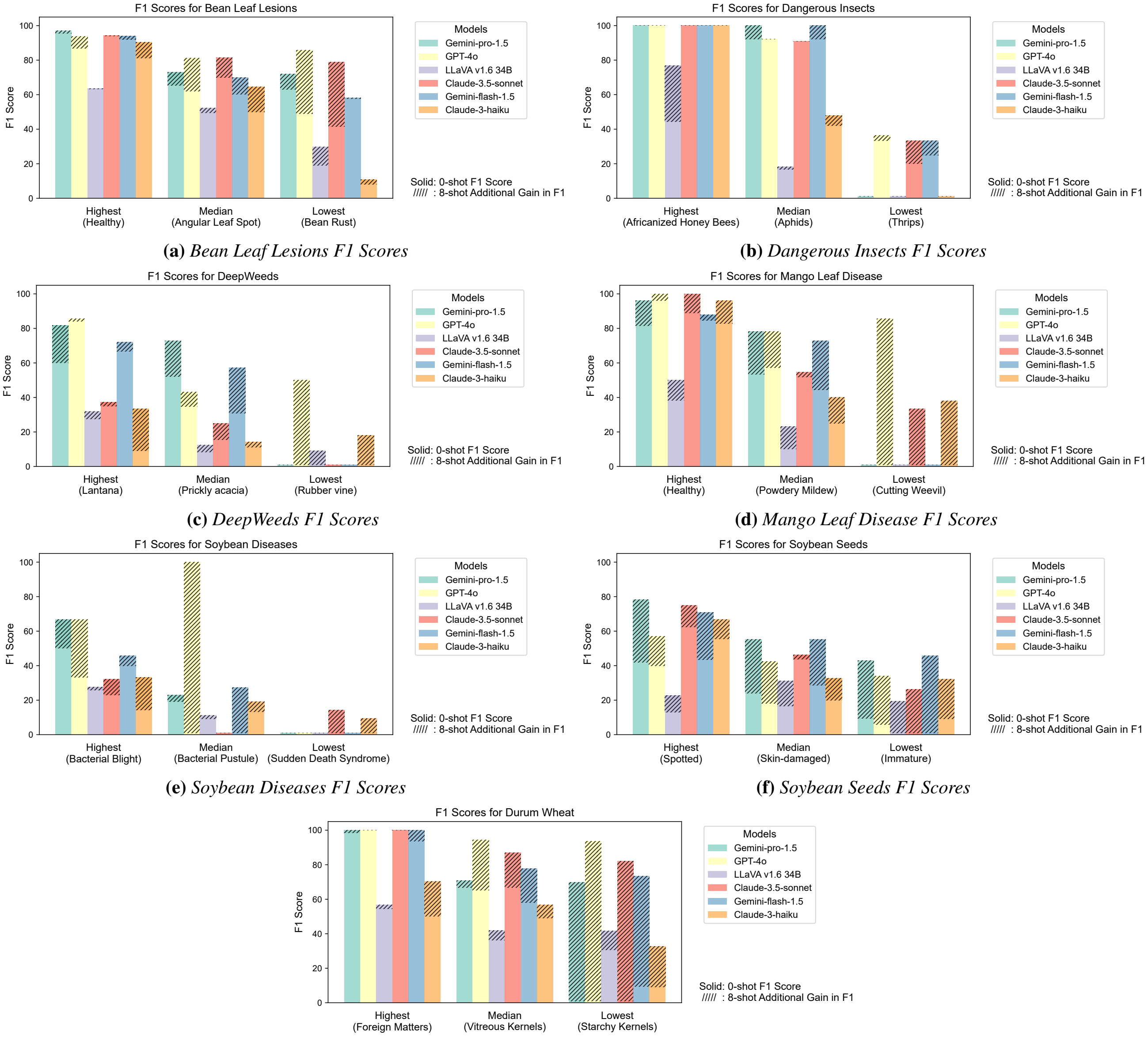


Figure S1. *Visualization of AgEval Benchmark Dataset - This treemap illustrates the distribution of datasets used in AgEval for plant stress identification, classification, and quantification. It contains subcategories, dataset names, and specific class names. Each rectangle represents a unique class name, with its size proportional to the count of instances. The visualization demonstrates the diversity of plant stress-related tasks covered by the AgEval framework across various crops and conditions.*



(a) *Bean Leaf Lesions F1 Scores*

(b) *Dangerous Insects F1 Scores*

(c) *DeepWeeds F1 Scores*

(d) *Mango Leaf Disease F1 Scores*

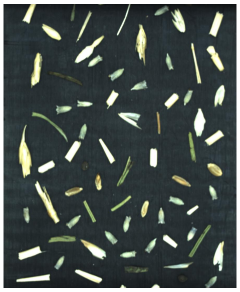
(e) *Soybean Diseases F1 Scores*

(f) *Soybean Seeds F1 Scores*

(g) *Durum Wheat F1 Scores*

Figure S2. *Comparison of F1 Scores for classes within datasets for Highest, Medium and Lowest performing class.*

What wheat variety is this?



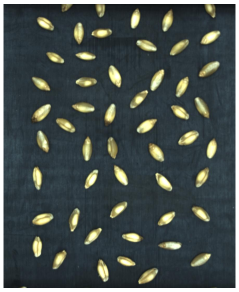
Category	Subcategory	Task
Identification (I)	Seed Morphology	Durum Wheat

Ground Truth: Foreign Matters

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Foreign Matters	Foreign Matters
GPT-4o	Foreign Matters	Foreign Matters
LLaVA v1.6 34B	Starchy Kernels	Vitreous Kernels
Claude-3.5-sonnet	Foreign Matters	Foreign Matters
Gemini-flash-1.5	Foreign Matters	Foreign Matters
Claude-3-haiku	Vitreous Kernels	Vitreous Kernels

What wheat variety is this?



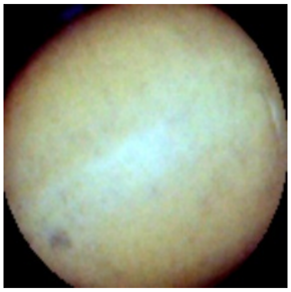
Category	Subcategory	Task
Identification (I)	Seed Morphology	Durum Wheat

Ground Truth: Starchy Kernels

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Vitreous Kernels	Starchy Kernels
GPT-4o	Vitreous Kernels	Starchy Kernels
LLaVA v1.6 34B	Vitreous Kernels	Vitreous Kernels
Claude-3.5-sonnet	Vitreous Kernels	Starchy Kernels
Gemini-flash-1.5	Vitreous Kernels	Vitreous Kernels
Claude-3-haiku	Vitreous Kernels	Vitreous Kernels

What is the quality of the soybean seed?



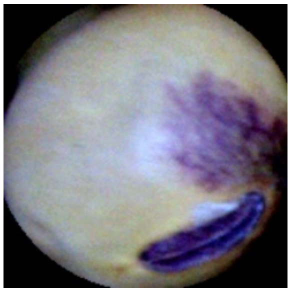
Category	Subcategory	Task
Identification (I)	Seed Morphology	Soybean Seeds

Ground Truth: Intact

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Spotted	Intact
GPT-4o	Spotted	Intact
LLaVA v1.6 34B	Immature	Intact
Claude-3.5-sonnet	Spotted	Intact
Gemini-flash-1.5	Intact	Spotted
Claude-3-haiku	Intact	Intact

What is the quality of the soybean seed?



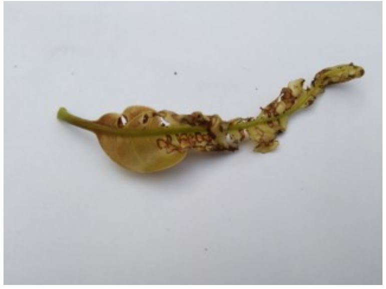
Category	Subcategory	Task
Identification (I)	Seed Morphology	Soybean Seeds

Ground Truth: Spotted

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Skin-damaged	Spotted
GPT-4o	Skin-damaged	Skin-damaged
LLaVA v1.6 34B	Skin-damaged	nan
Claude-3.5-sonnet	Spotted	Spotted
Gemini-flash-1.5	Skin-damaged	Spotted
Claude-3-haiku	Skin-damaged	Spotted

What mango leaf disease is present?



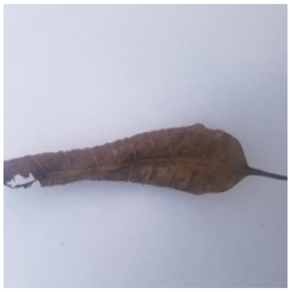
Category	Subcategory	Task
Identification (I)	Foliar Stress	Mango Leaf Disease

Ground Truth: Anthracnose

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Cutting Weevil	Bacterial Canker
GPT-4o	Cutting Weevil	Gall Midge
LLaVA v1.6 34B	Other	Anthracnose
Claude-3.5-sonnet	Cutting Weevil	Die Back
Gemini-flash-1.5	nan	Anthracnose
Claude-3-haiku	Die Back	Anthracnose

What mango leaf disease is present?



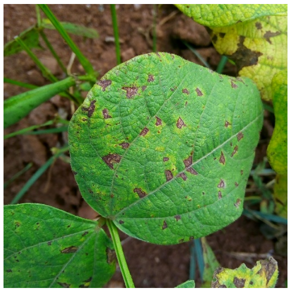
Category	Subcategory	Task
Identification (I)	Foliar Stress	Mango Leaf Disease

Ground Truth: Die Back

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	nan	nan
GPT-4o	Die Back	Die Back
LLaVA v1.6 34B	Die Back	Bacterial Canker
Claude-3.5-sonnet	Die Back	Die Back
Gemini-flash-1.5	nan	nan
Claude-3-haiku	Die Back	Die Back

What type of bean leaf lesion is this?



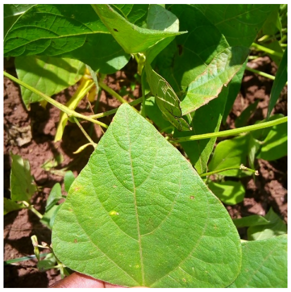
Category	Subcategory	Task
Identification (I)	Foliar Stress	Bean Leaf Lesions

Ground Truth: Angular Leaf Spot

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Angular Leaf Spot	Bean Rust
GPT-4o	Angular Leaf Spot	Angular Leaf Spot
LLaVA v1.6 34B	Bean Rust	Angular Leaf Spot
Claude-3.5-sonnet	Angular Leaf Spot	Angular Leaf Spot
Gemini-flash-1.5	Angular Leaf Spot	Angular Leaf Spot
Claude-3-haiku	Angular Leaf Spot	Angular Leaf Spot

What type of bean leaf lesion is this?



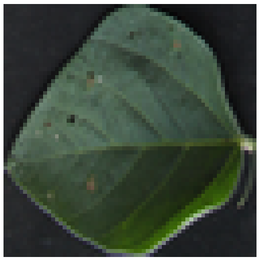
Category	Subcategory	Task
Identification (I)	Foliar Stress	Bean Leaf Lesions

Ground Truth: Healthy

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Healthy	Healthy
GPT-4o	Healthy	Healthy
LLaVA v1.6 34B	Healthy	Angular Leaf Spot
Claude-3.5-sonnet	Healthy	Bean Rust
Gemini-flash-1.5	Healthy	Healthy
Claude-3-haiku	Healthy	Healthy

What is the type of stress in this soybean?



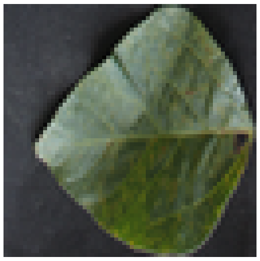
Category	Subcategory	Task
Identification (I)	Foliar Stress	Soybean Diseases

Ground Truth: Frogeye Leaf Spot

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Healthy	Potassium Deficiency
GPT-4o	Bacterial Pustule	Bacterial Blight
LLaVA v1.6 34B	Healthy	Iron Deficiency Chlorosis
Claude-3.5-sonnet	Healthy	Healthy
Gemini-flash-1.5	Healthy	Healthy
Claude-3-haiku	Potassium Deficiency	Potassium Deficiency

What is the type of stress in this soybean?



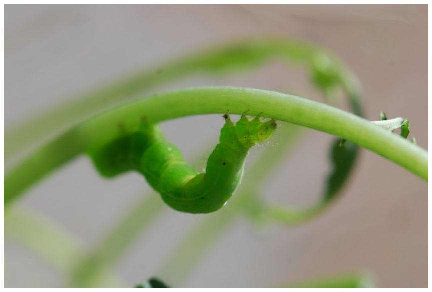
Category	Subcategory	Task
Identification (I)	Foliar Stress	Soybean Diseases

Ground Truth: Bacterial Pustule

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Herbicide Injury	Iron Deficiency Chlorosis
GPT-4o	Healthy	Bacterial Pustule
LLaVA v1.6 34B	Healthy	Healthy
Claude-3.5-sonnet	Healthy	Healthy
Gemini-flash-1.5	Iron Deficiency Chlorosis	Healthy
Claude-3-haiku	Frogeye Leaf Spot	Sudden Death Syndrome

What is the name of this harmful insect?



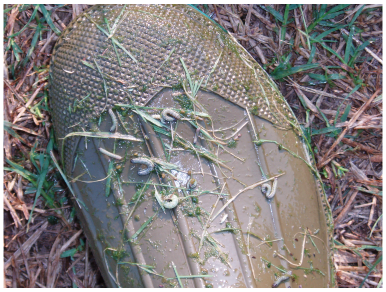
Category	Subcategory	Task
Identification (I)	Invasive Species	Dangerous Insects

Ground Truth: Cabbage Loopers

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Cabbage Loopers	Cabbage Loopers
GPT-4o	Cabbage Loopers	Cabbage Loopers
LLaVA v1.6 34B	Cabbage Loopers	nan
Claude-3.5-sonnet	Cabbage Loopers	Cabbage Loopers
Gemini-flash-1.5	Cabbage Loopers	Cabbage Loopers
Claude-3-haiku	Aphids	Tomato Hornworms

What is the name of this harmful insect?



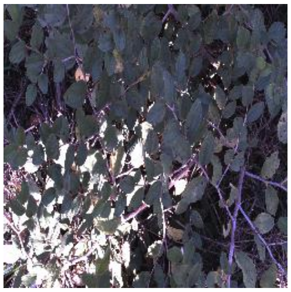
Category	Subcategory	Task
Identification (I)	Invasive Species	Dangerous Insects

Ground Truth: Fall Armyworms

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Armyworms	Armyworms
GPT-4o	Cabbage Loopers	Armyworms
LLaVA v1.6 34B	Cabbage Loopers	nan
Claude-3.5-sonnet	Armyworms	Armyworms
Gemini-flash-1.5	Fall Armyworms	Armyworms
Claude-3-haiku	Armyworms	nan

What is the name of this weed?



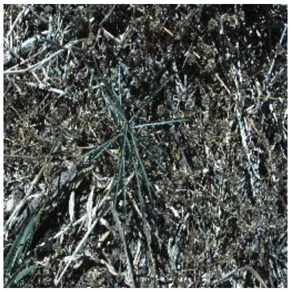
Category	Subcategory	Task
Identification (I)	Invasive Species	DeepWeeds

Ground Truth: Chinee apple

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Chinee apple	Chinee apple
GPT-4o	Chinee apple	Chinee apple
LLaVA v1.6 34B	Parthenium	Parkinsonia
Claude-3.5-sonnet	Lantana	Lantana
Gemini-flash-1.5	Prickly acacia	Chinee apple
Claude-3-haiku	Parthenium	Parthenium

What is the name of this weed?



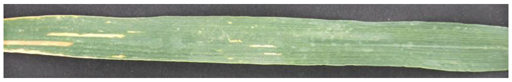
Category	Subcategory	Task
Identification (I)	Invasive Species	DeepWeeds

Ground Truth: Parkinsonia

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	nan	nan
GPT-4o	Parthenium	Negative
LLaVA v1.6 34B	nan	Snake weed
Claude-3.5-sonnet	Snake weed	Parthenium
Gemini-flash-1.5	nan	Siam weed
Claude-3-haiku	Parthenium	Snake weed

What is the severity of yellow rust disease?



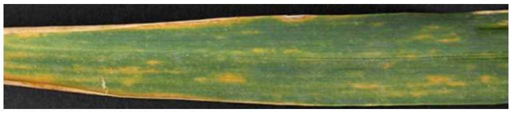
Category	Subcategory	Task
Classification (C)	Disease Severity	Yellow Rust 19

Ground Truth: MRMS

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Moderately Resistant (MR)	Moderately Resistant (MR)
GPT-4o	Moderately Susceptible (MS)	Moderately Resistant (MR)
LLaVA v1.6 34B	Susceptible (S)	No disease (0)
Claude-3.5-sonnet	Moderately Resistant (MR)	No disease (0)
Gemini-flash-1.5	Moderately Resistant (MR)	MRMS
Claude-3-haiku	Susceptible (S)	Moderately Resistant (MR)

What is the severity of yellow rust disease?



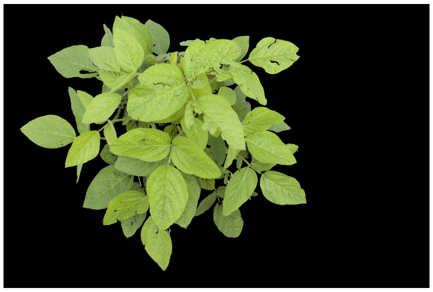
Category	Subcategory	Task
Classification (C)	Disease Severity	Yellow Rust 19

Ground Truth: Resistant (R)

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Moderately Resistant (MR)	Susceptible (S)
GPT-4o	Moderately Resistant (MR)	Moderately Susceptible (MS)
LLaVA v1.6 34B	Susceptible (S)	Moderately Susceptible (MS)
Claude-3.5-sonnet	Moderately Susceptible (MS)	Moderately Susceptible (MS)
Gemini-flash-1.5	Moderately Resistant (MR)	Moderately Susceptible (MS)
Claude-3-haiku	Moderately Susceptible (MS)	MRMS

What is the rating (1-5) of soybean stress severity?



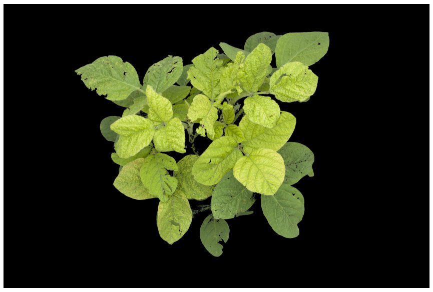
Category	Subcategory	Task
Classification (C)	Stress Tolerance	IDC

Ground Truth: 1

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	1.0	1.0
GPT-4o	3	1
LLaVA v1.6 34B	4.0	nan
Claude-3.5-sonnet	2.0	2
Gemini-flash-1.5	1	2
Claude-3-haiku	1.0	3.0

What is the rating (1-5) of soybean stress severity?



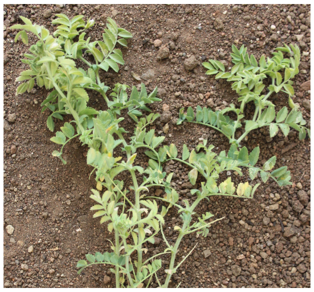
Category	Subcategory	Task
Classification (C)	Stress Tolerance	IDC

Ground Truth: 2

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	2.0	4.0
GPT-4o	4	2
LLaVA v1.6 34B	3.0	nan
Claude-3.5-sonnet	3.0	3
Gemini-flash-1.5	2	3
Claude-3-haiku	1.0	3.0

What is the severity of chickpea fusarium wilt?



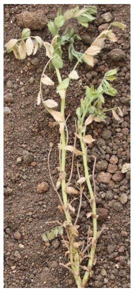
Category	Subcategory	Task
Classification (C)	Stress Tolerance	FUSARIUM 22

Ground Truth: Resistant

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Susceptible	Resistant
GPT-4o	Highly Susceptible	Highly Resistant
LLaVA v1.6 34B	Highly Susceptible	nan
Claude-3.5-sonnet	Susceptible	Moderately Resistant
Gemini-flash-1.5	Moderately Resistant	Moderately Resistant
Claude-3-haiku	Resistant	Resistant

What is the severity of chickpea fusarium wilt?



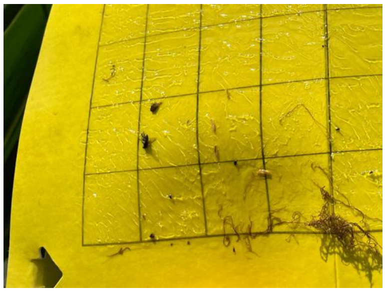
Category	Subcategory	Task
Classification (C)	Stress Tolerance	FUSARIUM 22

Ground Truth: Susceptible

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	Susceptible	Susceptible
GPT-4o	Highly Susceptible	Highly Susceptible
LLaVA v1.6 34B	Resistant	nan
Claude-3.5-sonnet	Susceptible	Highly Susceptible
Gemini-flash-1.5	Highly Susceptible	Highly Susceptible
Claude-3-haiku	Susceptible	Moderately Resistant

What is the insect count?



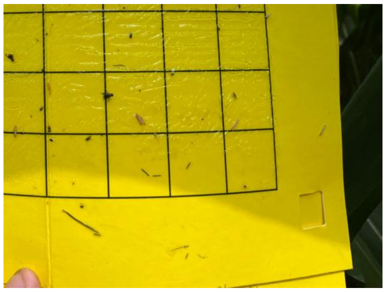
Category	Subcategory	Task
Quantification (Q)	Pest	InsectCount

Ground Truth: 2

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	10	5.0
GPT-4o	6	9
LLaVA v1.6 34B	7.0	4.0
Claude-3.5-sonnet	8	3
Gemini-flash-1.5	4	6
Claude-3-haiku	17.0	17.0

What is the insect count?



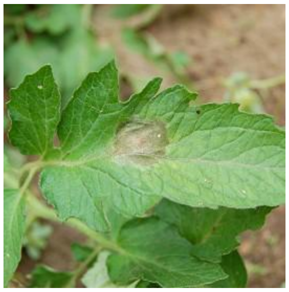
Category	Subcategory	Task
Quantification (Q)	Pest	InsectCount

Ground Truth: 1

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	8	8.0
GPT-4o	9	2
LLaVA v1.6 34B	0.0	11.0
Claude-3.5-sonnet	15	0
Gemini-flash-1.5	1	2
Claude-3-haiku	22.0	3.0

What is the diseased leaf percentage?



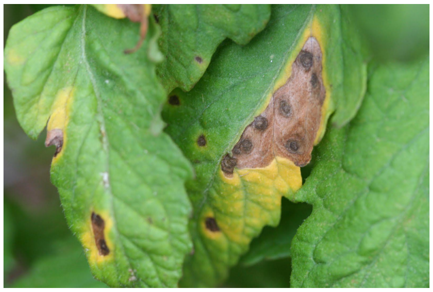
Category	Subcategory	Task
Quantification (Q)	Disease	PlantDoc

Ground Truth: 3

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	10.0	3.0
GPT-4o	7	8
LLaVA v1.6 34B	5.0	7.0
Claude-3.5-sonnet	12	4
Gemini-flash-1.5	5	4
Claude-3-haiku	19.0	3.0

What is the diseased leaf percentage?



Category	Subcategory	Task
Quantification (Q)	Disease	PlantDoc

Ground Truth: 12

Predictions:

Model Name	0 shot	8 shot
Gemini-pro-1.5	10.0	5.0
GPT-4o	18	12
LLaVA v1.6 34B	10.0	nan
Claude-3.5-sonnet	23	15
Gemini-flash-1.5	32	12
Claude-3-haiku	18.0	30.0