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.
```

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

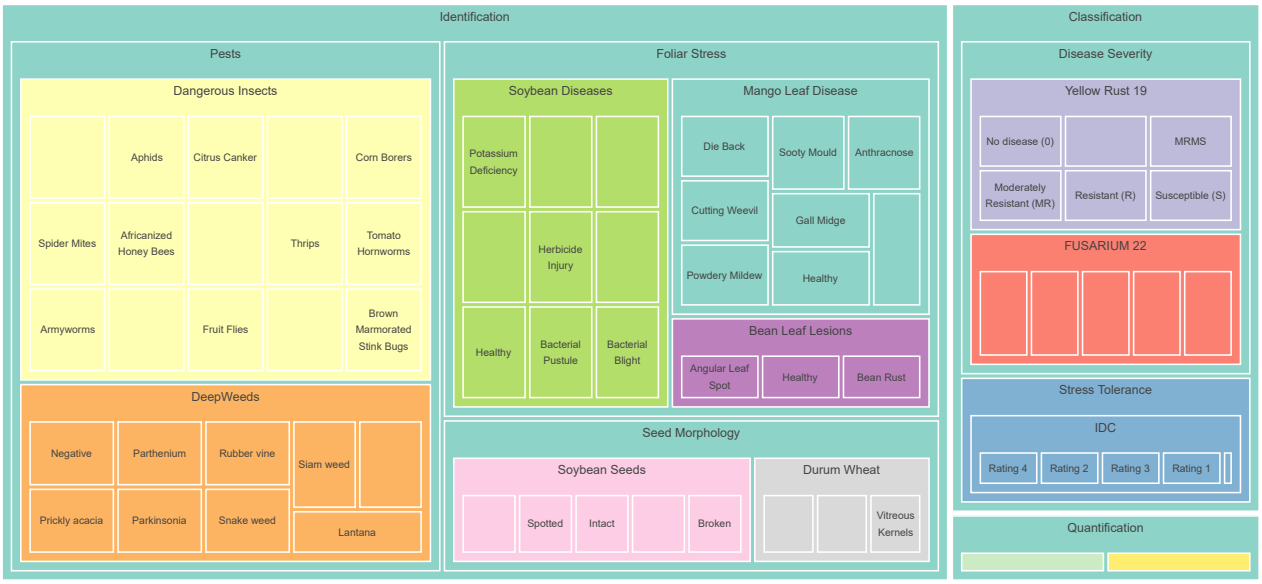
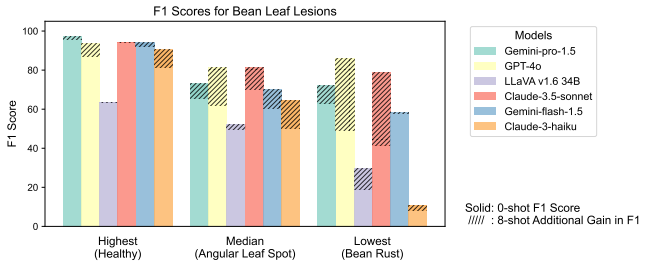
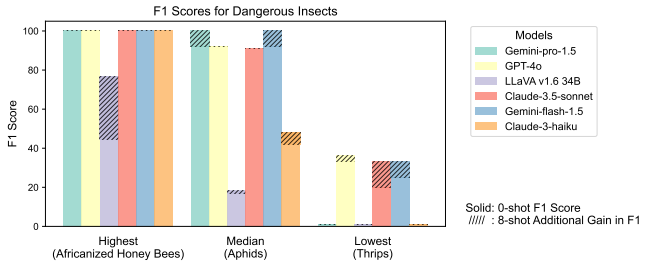


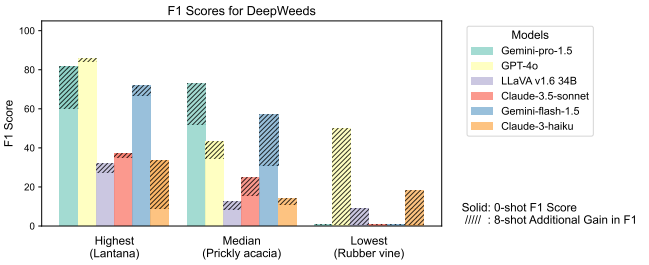
Figure S1. *Visualization of AgEval Benchmark Dataset - This treemap illustrates the distribution of datasets used in AgEval for plant stress identification, classification, and quantification. It contains subcategories, dataset names, and specific class names. Each rectangle represents a unique class 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.*



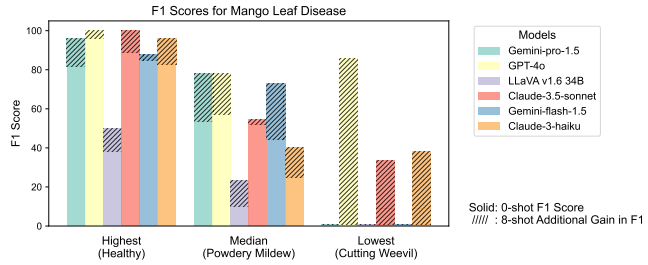
(a) *Bean Leaf Lesions F1 Scores*



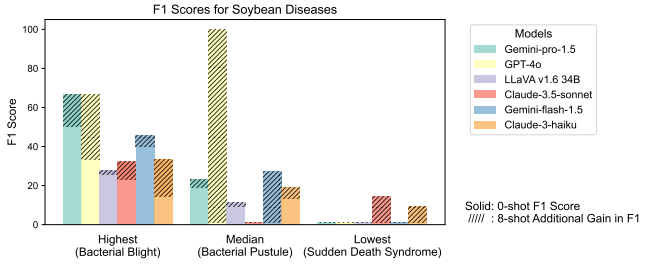
(b) *Dangerous Insects F1 Scores*



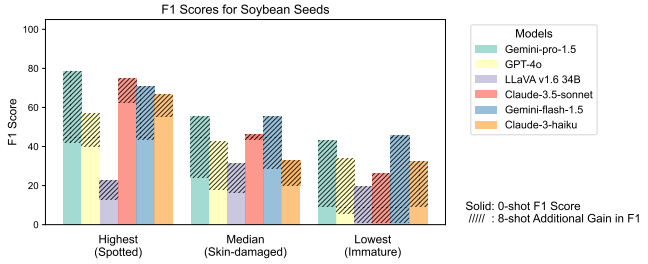
(c) *DeepWeeds F1 Scores*



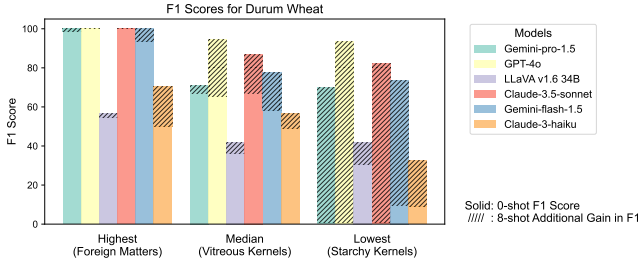
(d) *Mango Leaf Disease F1 Scores*



(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?

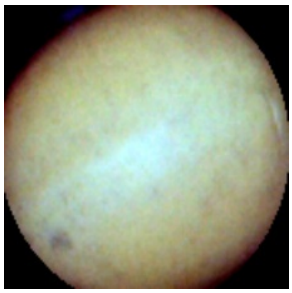
Category	Subcategory	Task
Identification (I)	Seed Morphology	Durum Wheat

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 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 wheat variety is this?

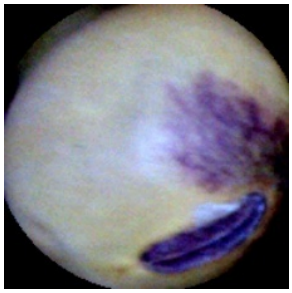
Category	Subcategory	Task
Identification (I)	Seed Morphology	Durum Wheat

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: 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?

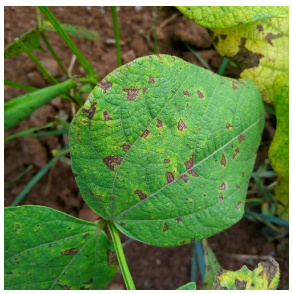
Category	Subcategory	Task
Identification (I)	Foliar Stress	Mango Leaf Disease

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 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 mango leaf disease is present?

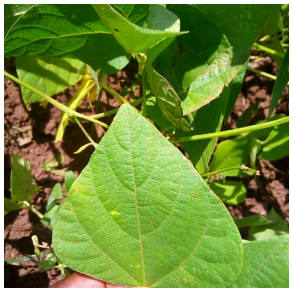
Category	Subcategory	Task
Identification (I)	Foliar Stress	Mango Leaf Disease

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: 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?

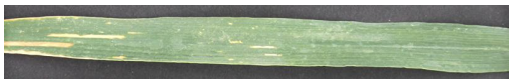
Category	Subcategory	Task
Identification (I)	Invasive Species	DeepWeeds

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?

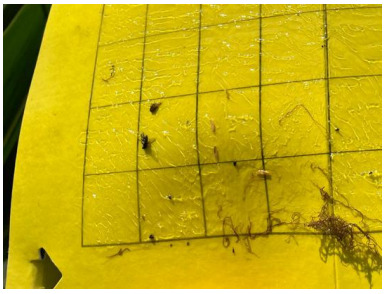
Category	Subcategory	Task
Classification (C)	Stress Tolerance	FUSARIUM 22

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 diseased leaf percentage?

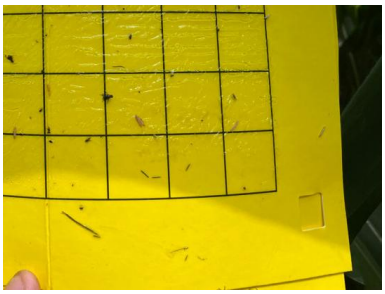
Category	Subcategory	Task
Quantification (Q)	Disease	PlantDoc

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 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: 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