ICCV HONOLULU HAWAII
OCT 19-23, 2025

AgroBench: Vision-Language Model Benchmark in Agriculture

Risa Shinoda [1,3]

Nakamasa Inoue [2,3]

Hirokatsu Kataoka [3,4]

Masaki Onishi [3]

Yoshitaka Ushiku [5]

1 The University of Osaka

2 Institute of SCIENCE TOKYO

3 AIST

4 VGG University of Oxford

5 SINICX

*Agronomist AI Benchmark designed to comprehensively evaluate **682 disease categories** across **203 crop types** for 7 vision-language question-answer tasks.*

Disease Management

Question: What should be done to minimize the spread of the disease shown in the image?

Answer: Remove the male buds (bell) immediately using a forked stick.

Weed Detection

Question: What species of weed does the highlighted bounding box represent in the image?

Answer: White goosefoot (Chenopodium album)

Disease Identification

Question: What disease is affecting the plant?

Answer: Bacterial canker.

Crop Management

Question: Is the crop shown in the image ready for harvest based on its current condition?

Answer: Yes, the leaves and stems have turned yellow, and the seed pods are brown to black.

AGROBENCH

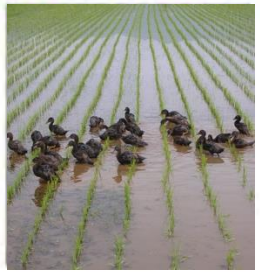
Agronomist AI Benchmark

Pest Identification

Question: What pest is affecting the plant?

Answer: Squash bug.

Traditional Methods



Question: What is one of the main benefits the process shown in the image?

Answer: Weed control and insect pest reduction.

Machine Usage

Question: What is the main purpose of using the machine shown in the image?

Answer: To control weeds in the field without relying on chemical herbicides.

- Crop categories: 203
- Disearse categories: 684
- Expert annotation

Motivation

Agricultural automation is evolving through image recognition.

Disease Identification

Ripeness Recognition

Weed Identification

- Yet, existing models are usually task-specific and crop-specific.
- Complex setups limit their use in real fields.

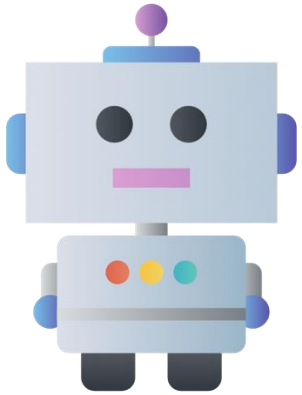
Vision-Language Models（VLM）

- ✓ Capable of solving diverse tasks using broad knowledge
- ✓ Interacts via natural language
- ✓ No advanced computer skills required

Farmers

Natural language interaction

No expertise required



VLMs

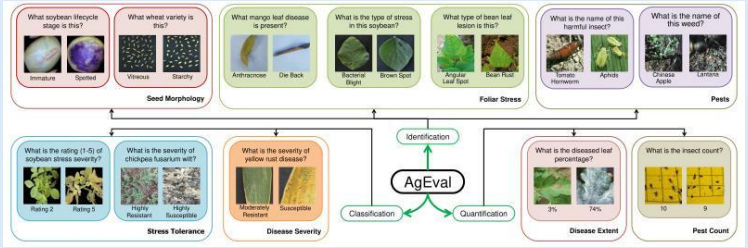
An ideal assistant for farmers!

Related Works

- Research on agricultural VLMs is rapidly growing.
- However, no standardized benchmark exists for fair comparison.
- Recent datasets (e.g., AgMMU, AgEval, Agrostruct, CDDM) cover tasks like disease, weed, and pest recognition, but remain limited in diversity.

Agrostruct (Awais+, WACV'25)

CDDM (Liu+, ECCV'24)



AgEval (Arshad,WACV'25)

AgMMU (Aruna+, arXiv'25)

We Proposed Comprehensive Evaluation Dataset!

Disease Management

Question: What should be done to minimize the spread of the disease shown in the image?

Answer: Remove the male buds (bell) immediately using a forked stick.

Weed Detection

Question: What species of weed does the highlighted bounding box represent in the image?

Answer: White goosefoot (Chenopodium album)

Disease Identification

Question: What disease is affecting the plant?

Answer: Bacterial canker.

Crop Management

Question: Is the crop shown in the image ready for harvest based on its current condition?

Answer: Yes, the leaves and stems have turned yellow, and the seed pods are brown to black.

AGROBENCH

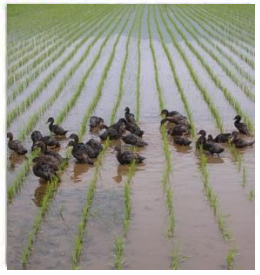
Agronomist AI Benchmark

Pest Identification

Question: What pest is affecting the plant?

Answer: Squash bug.

Traditional Methods



Question: What is one of the main benefits the process shown in the image?

Answer: Weed control and insect pest reduction.

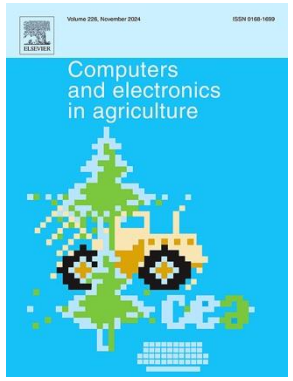
Machine Usage

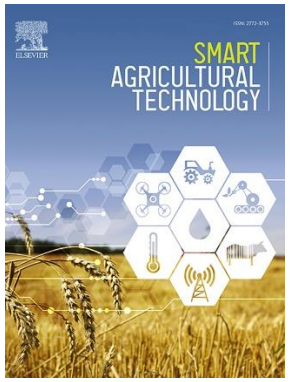
Question: What is the main purpose of using the machine shown in the image?

Answer: To control weeds in the field without relying on chemical herbicides.

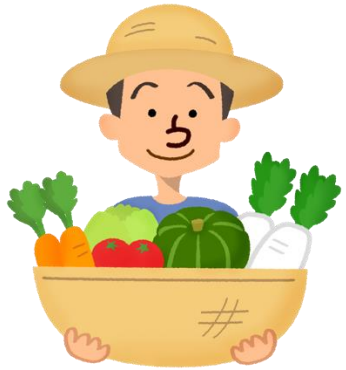
- Crop categories: 203
- Disearse categories: 684
- Expert annotation

Task Definition





Agricultural Engineering Journals



Domain Knowledge Required on the Farm

7 Core Tasks Identified (with Subcategories)

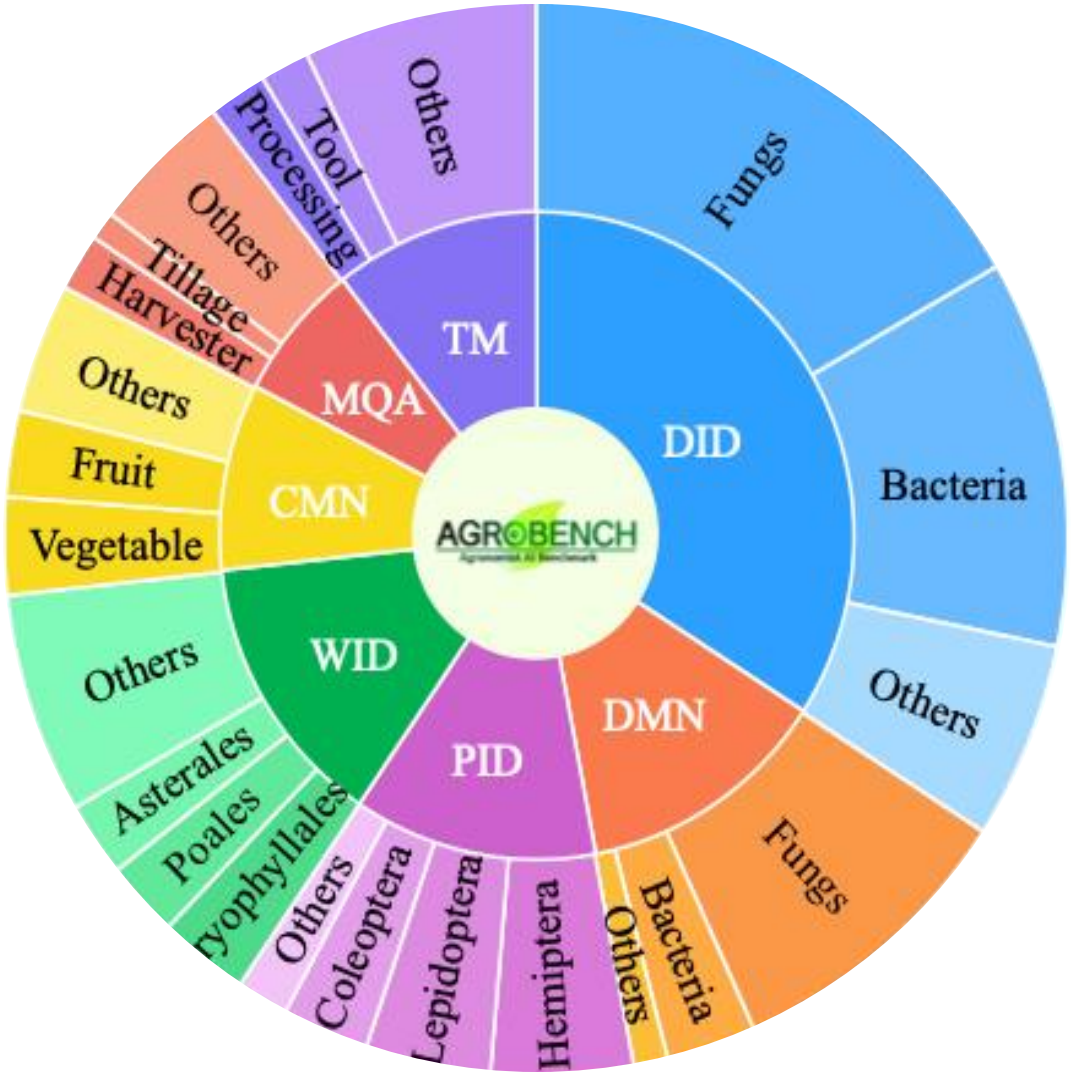


Image and Label Collection

- Remove blurry or unclear images
- Prioritize images taken in real-world agricultural settings (not in lab environments)

50,000
Image-Label pairs

Annotator
Ph.D. in Agriculture

4,000
Image-Lavel pairs

QA Annotation

Detection tasks: Automatically generated from labels

Other tasks (e.g., pest control, crop management, machinery, traditional farming): Manually annotated

Images were carefully reviewed

QA pairs were created based on reliable sources

Question: Is the crop shown in the image ready for harvest based on its current condition?

Answer: Yes, the leaves and stems have turned yellow, and the seed pods are brown to black.

Annotators
Ph.D. in Agriculture

+ also reviewed by students and professors in Agriculture

Image-QA pairs

Task Definition
Disease Identification（DID）

- 682 plant and disease categories
- 1,502 images
- Covering a wide range of plant parts

DID task Question: Can you identify the disease affecting the crop?

A. Blister spot
B. Flyspeck
C. Fire blight
D. Apple proliferation
E. Cedar apple rust

Ground truth: D Labels: Apple proliferation

Task Definition
Disease Management（DMN）

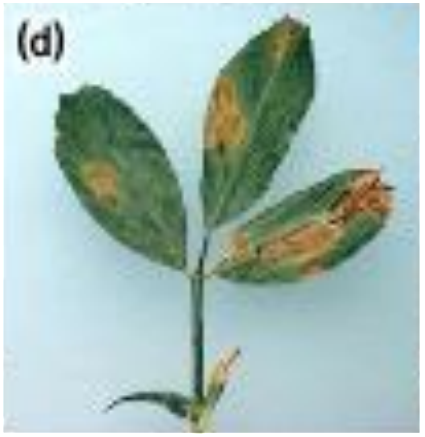
- 141 crop–disease categories, 569 QAs
- Based on disease severity, we provide different answer options and define different correct answers.

Initial Stage

DMN task Question: Based on the disease level of this plant, what should be done to maintain this condition?
A. Immediately cut and harvest the entire field.
B. Continue monitoring every one to four weeks and remove affected leaves.
C. Stop examining plants until the disease becomes more severe.
D. Avoid removing any leaves, as this could stress the plant further.
E. Increase irrigation and fertilization to boost plant health.

Ground truth: B Labels: Alfalfa Bacterial leaf spot

Severe



DMN task Question: Based on the disease level of this plant, what action should be taken?
A. Reduce sunlight exposure to slow disease progression.
B. Increase watering and apply additional fertilizers immediately.
C. Plant additional alfalfa to compensate for the loss.
D. Stop all cutting and harvesting until symptoms decrease.
E. Prompt cutting of the alfalfa to reduce loss in yield and quality.

Ground truth: E Labels: Alfalfa Bacterial leaf spot

Task Definition
Pest Identification（PID）

- 134 categories, 544 images
- Covering various life stages, from larva to adult

PID task Question: Can you identify the disease affecting the crop?

A. Large Cabbage White
B. Cabbage Looper
C. White Grub
D. White Scale Insects
E. Cassava Whitefly

Ground truth: A Labels: Large Cabbage White

Task Definition
Weed Identification (WID)

- 108 categories, 609 images
- Collaboration with the authors of the CropAndWeed dataset [Steininger et.al.]

WID task Question: What species of weed does the highlighted bounding box represent in the image ?

A. Common dandelion (Taraxacum officinale)
B. Mollugo verticillata (Carpetweed)
C. Bromus sterilis (Barren Brome)
D. Polygonum aviculare (Prostrate Knotweed)
E. Solanum nigrum (Black Nightshade)

Ground truth: B Labels: Mollugo verticillata (Carpetweed)

Task Definition
Crop Management（CMN）

- 411 QA pairs, covering diverse topics.

Facility

Q: What is the purpose of the method shown in the image?

A: Allowing some spears to grow into full plants.

Harvest

Q: Is the crop shown in the image ready for harvest based on its current condition?

A: Yes, the leaves and stems have turned yellow.

Storage

Q: What is the recommended action for the crop shown in the image harvested?

A: Dry them before storage

Environment

Q:What soil conditions are ideal for the plant shown in the image to grow well?

A: Highly acidic soil with pH 4.5–5.8

Tool

Q: What is the purpose of using the equipment shown in the image?

A: To grow uniform seedlings

Task Definition
Machine QA (MQA)

- 98 machinery types, 303 images
- Covers not just identification, but also usage and operation.

Aquatic weed harvester

Smooth Roller

Flail Mower

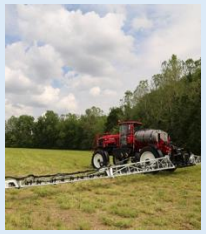
Land Imprinter

Transplanting Machine

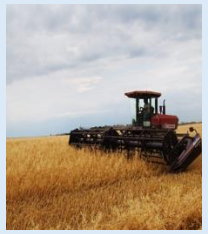
Seed Drill

Mower

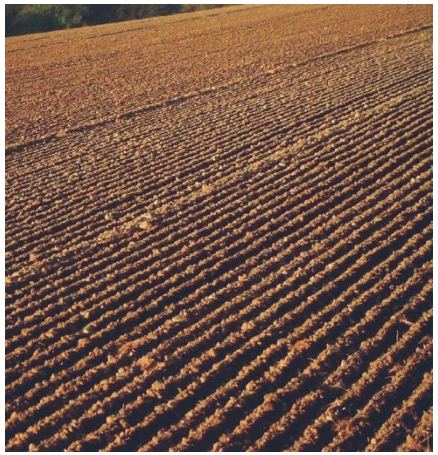
Cultivator


Self-Propelled Sprayer

Strip-Trill Implement


Swather

Reversible Plough



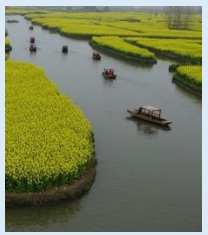
MQA task Question: What machine was used to achieve the current state of the field shown in the image?

A. Disc plough
B. Rotary tiller
C. Spring-tooth harrow
D. Cambridge roller
E. Subsoiler

Ground truth: D Labels: Cambridge roller

Task Definition
Traditional Method（TM）

- 77 traditional farming methods, 404 QAs
- Covers farming methods shaped by diverse regional traditions.



TM task Question: What is the agricultural method shown in the picture?

A. Terrace farming
B. Hydroponics
C. Drip irrigation
D. Zai pits
E. Crop rotation

Ground truth: D Labels: Zai pits

Reasoning Annotation

We also have reasoning annotation for the clarity and further investigation.

Question
At what stage of growth is the method shown in the image typically performed?

Options
A. During the early growth stages after the rice seeds sprout
B. When the rice plants start bearing grains
C. After the rice grains have matured
D. Before the rice plants reach full height
E. Right before the rice is ready for harvest

Answer A. During the early growth stages after the rice seeds sprout

Reasoning
The method shown in the image is duck rice farming, implemented during the early stages to allow ducks to help control weeds and pests naturally, without damaging the rice grains.

Category Duck rice method

Experiment Setup

💭 Models

- 4 closed (API-based) models
- 8 open-source models

🧑 Human Evaluation （for reference）

- 28 graduates from the Faculty/Graduate School of Agriculture
- Each answered 20 questions, selected to fit their academic
- Books were allowed, but internet searches were prohibited

📊 Evaluation Method

- Accuracy measured based on 5-choice questions

Result (Main)

The most difficult task!

Model	DID	DMN	PID	WID	CMN	MQA	TM	Overall (all)	Overall (subset)
Random Choice	21.77	15.64	20.40	17.90	16.06	22.11	19.31	19.03	19.11
Human	25.00	22.50	45.00	20.00	36.25	57.50	51.25	-	36.79
Closed-Source Vision Language Models (VLMs)									
GPT-4o mini [33]	53.60	80.67	60.04	35.14	64.23	70.96	69.80	62.06	69.65
GPT-4o [32]	64.18	87.35	77.76	44.17	75.43	82.84	82.43	73.45	79.26
Gemini1.5-Flash [46]	55.06	79.96	70.04	50.90	64.72	78.22	73.27	67.45	68.82
Gemini1.5-Pro [46]	62.92	81.55	74.45	55.17	71.05	82.84	77.72	72.24	69.74
Open-Source Vision Language Models (VLMs)									
EMU2Chat [44]	42.01	48.33	43.75	23.81	40.39	37.62	47.77	40.53	33.84
LLaVA-Next-8B [19]	45.47	72.58	43.01	30.05	54.26	56.11	57.46	51.28	57.84
LLaVA-Next-72B [19]	54.95	80.00	49.81	26.98	66.92	66.11	70.38	59.31	64.36
QwenVLM-7B [5]	51.26	80.49	63.97	33.17	66.42	76.24	77.48	64.15	66.41
QwenVLM-72B [5]	57.99	87.87	73.35	34.48	75.91	80.86	84.16	70.66	72.45
CogVLM-19B [50]	29.16	53.78	52.39	25.45	54.01	71.62	66.09	50.36	44.27
LLaVa-7B [18]	36.02	62.74	38.79	24.79	53.77	46.53	55.20	45.41	46.14
LLaVA-13B [18]	40.21	68.89	44.49	24.79	59.37	54.13	58.42	50.04	55.31

Result (Main)

There is still room for perceptual improvement.

Model	DID	DMN	PID	WID	CMN	MQA	TM	Overall (all)	Overall (subset)
Random Choice	21.77	15.64	20.40	17.90	16.06	22.11	19.31	19.03	19.11
Human	25.00	22.50	45.00	20.00	36.25	57.50	51.25	-	36.79
Closed-Source Vision Language Models (VLMs)									
GPT-4o mini [33]	53.60	80.67	60.04	35.14	64.23	70.96	69.80	62.06	69.65
GPT-4o [32]	64.18	87.35	77.76	44.17	75.43	82.84	82.43	73.45	79.26
Gemini1.5-Flash [46]	55.06	79.96	70.04	50.90	64.72	78.22	73.27	67.45	68.82
Gemini1.5-Pro [46]	62.92	81.55	74.45	55.17	71.05	82.84	77.72	72.24	69.74
Open-Source Vision Language Models (VLMs)									
EMU2Chat [44]	42.01	48.33	43.75	23.81	40.39	37.62	47.77	40.53	33.84
LLaVA-Next-8B [19]	45.47	72.58	43.01	30.05	54.26	56.11	57.46	51.28	57.84
LLaVA-Next-72B [19]	54.95	80.00	49.81	26.98	66.92	66.11	70.38	59.31	64.36
QwenVLM-7B [5]	51.26	80.49	63.97	33.17	66.42	76.24	77.48	64.15	66.41
QwenVLM-72B [5]	57.99	87.87	73.35	34.48	75.91	80.86	84.16	70.66	72.45
CogVLM-19B [50]	29.16	53.78	52.39	25.45	54.01	71.62	66.09	50.36	44.27
LLaVa-7B [18]	36.02	62.74	38.79	24.79	53.77	46.53	55.20	45.41	46.14
LLaVA-13B [18]	40.21	68.89	44.49	24.79	59.37	54.13	58.42	50.04	55.31

Result (Main)

Closed-model is superior to human.

Model	DID	DMN	PID	WID	CMN	MQA	TM	Overall (all)	Overall (subset)
Random Choice	21.77	15.64	20.40	17.90	16.06	22.11	19.31	19.03	19.11
Human	25.00	22.50	45.00	20.00	36.25	57.50	51.25	-	36.79
Closed-Source Vision Language Models (VLMs)									
GPT-4o mini [33]	53.60	80.67	60.04	35.14	64.23	70.96	69.80	62.06	69.65
GPT-4o [32]	64.18	87.35	77.76	44.17	75.43	82.84	82.43	73.45	79.26
Gemini1.5-Flash [46]	55.06	79.96	70.04	50.90	64.72	78.22	73.27	67.45	68.82
Gemini1.5-Pro [46]	62.92	81.55	74.45	55.17	71.05	82.84	77.72	72.24	69.74
Open-Source Vision Language Models (VLMs)									
EMU2Chat [44]	42.01	48.33	43.75	23.81	40.39	37.62	47.77	40.53	33.84
LLaVA-Next-8B [19]	45.47	72.58	43.01	30.05	54.26	56.11	57.46	51.28	57.84
LLaVA-Next-72B [19]	54.95	80.00	49.81	26.98	66.92	66.11	70.38	59.31	64.36
QwenVLM-7B [5]	51.26	80.49	63.97	33.17	66.42	76.24	77.48	64.15	66.41
QwenVLM-72B [5]	57.99	87.87	73.35	34.48	75.91	80.86	84.16	70.66	72.45
CogVLM-19B [50]	29.16	53.78	52.39	25.45	54.01	71.62	66.09	50.36	44.27
LLaVa-7B [18]	36.02	62.74	38.79	24.79	53.77	46.53	55.20	45.41	46.14
LLaVA-13B [18]	40.21	68.89	44.49	24.79	59.37	54.13	58.42	50.04	55.31

Result (Without Image)

Dataset designed for image-based reasoning, but models answer correctly even without images —indicating reliance on text rather than vision.

	DID	DMN	PID	WID	CMN	MQA	TM	Overall
GPT-4o [32]	1.93	72.58	18.75	1.00	40.39	25.08	48.27	29.71
LLaVA-Next-8B [19]	26.10	70.30	21.88	19.70	53.77	30.36	40.35	37.49

Question

Which environmental condition increases the susceptibility of the plants to the disease shown in the image?

Options

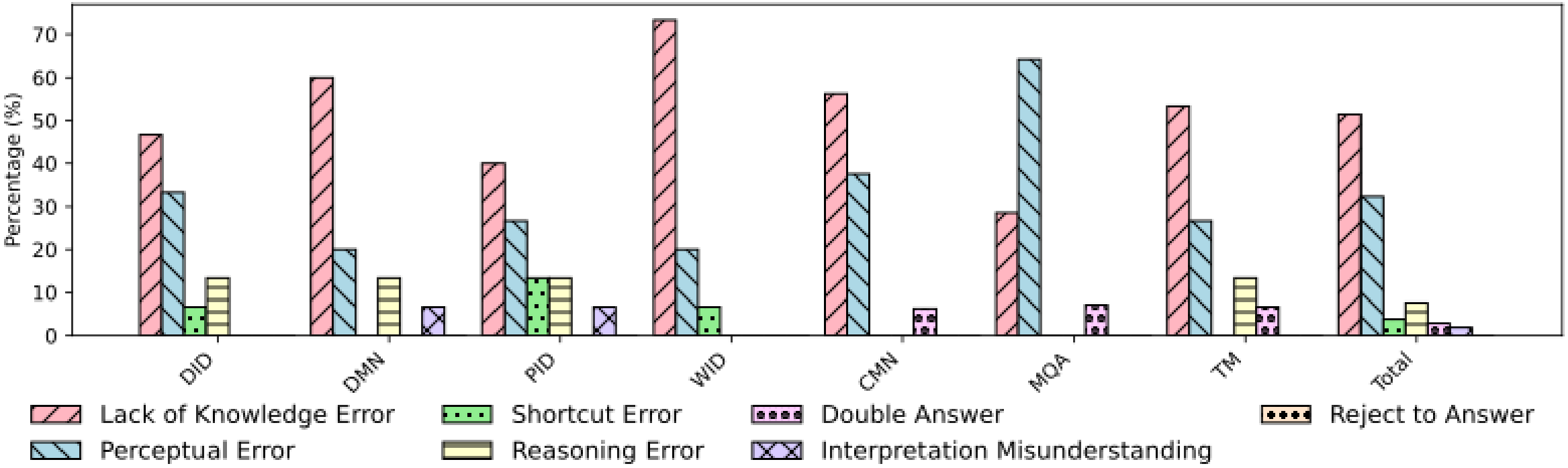
A. Shady conditions with low humidity.
B. Cool, dry conditions throughout the season.
C. Windy conditions with high sunlight exposure.
D. Low soil potassium levels.
E. High soil nitrogen content.

Ground truth: E **Labels**: High soil nitrogen content.

GPT-4o: E

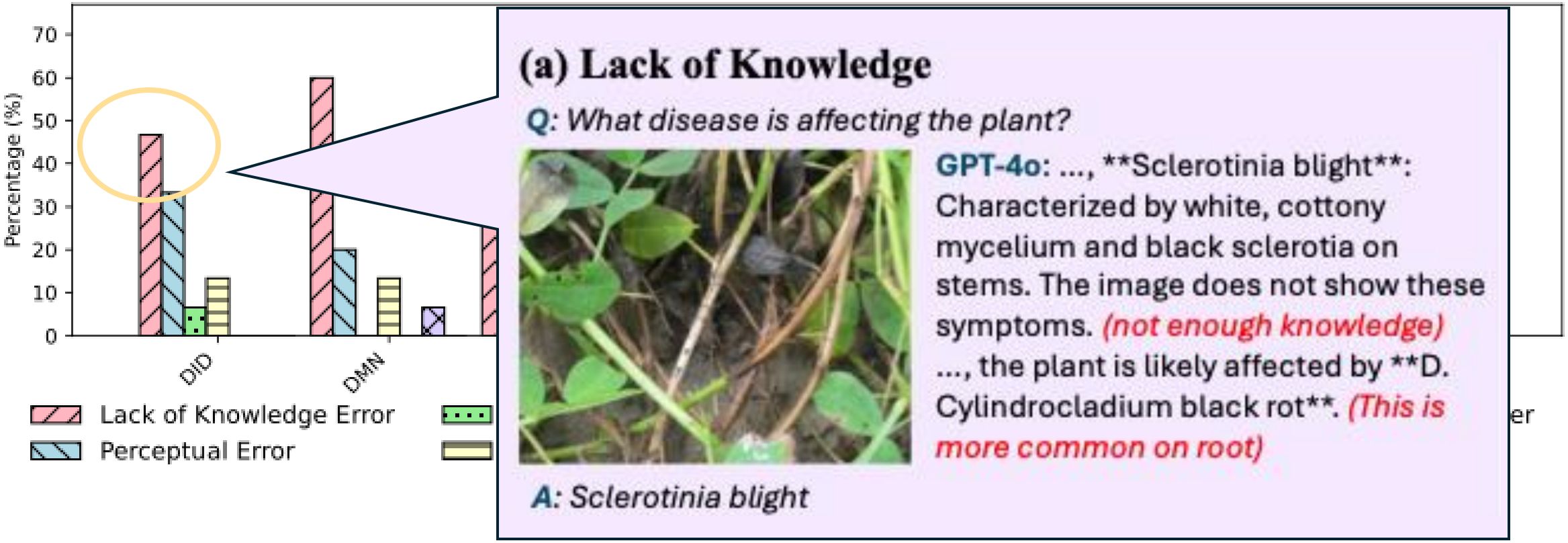
Error Analysis: GPT4o

15 incorrect cases per task were sampled and analyzed for error patterns.



Error Analysis: GPT4o (Lack of Knowledge)

- The model tends to fail in fully describing the traits of the choices due to its lack of knowledge



Error Analysis: GPT4o (Perceptual Error)

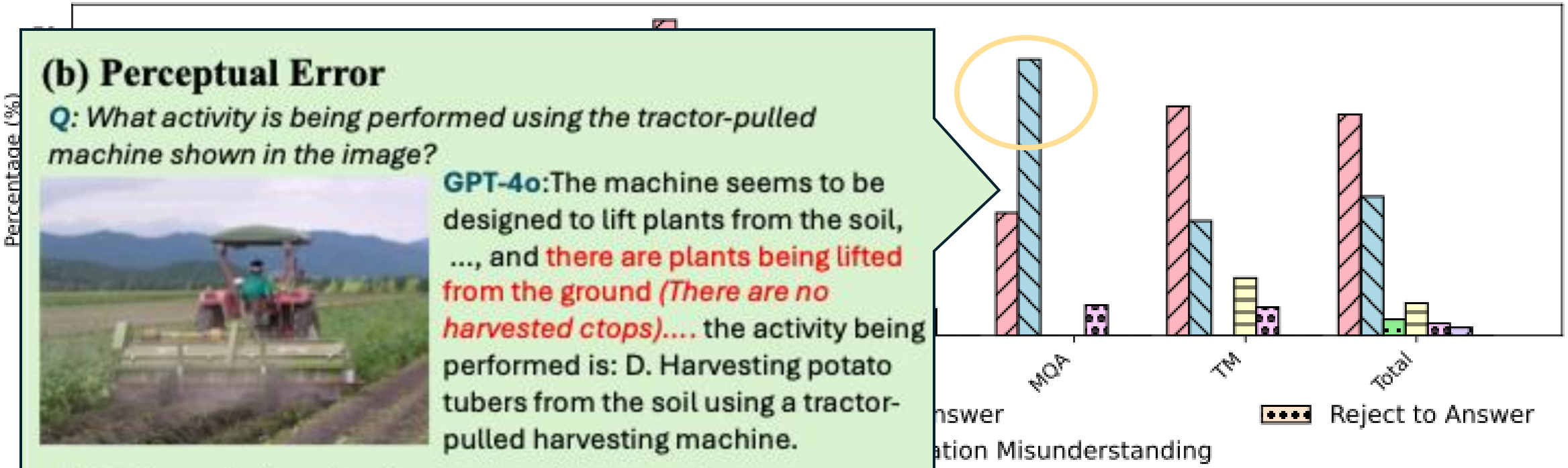
- The model hullusinates the visual elements to support their guessing.

(b) Perceptual Error

Q: What activity is being performed using the tractor-pulled machine shown in the image?

GPT-4o:The machine seems to be designed to lift plants from the soil, ..., and there are plants being lifted from the ground *(There are no harvested ctops)*.... the activity being performed is: D. Harvesting potato tubers from the soil using a tractor-pulled harvesting machine.

A: Cutting and chopping the stems and leaves using a tractor-pulled leaf chopper.



Summery

Project Page

We propose AgroBench, a benchmark for evaluating VLMs on agricultural tasks.

- Covers diverse tasks, plant parts, and symptom severities
- All models show room for improvement, especially in weed and disease identification.
- Errors mainly stem from limited agricultural knowledge.
- AgroBench enables comprehensive and sustainable agricultural AI research.