Leveraging Vision Language Models for Specialized Agricultural Tasks

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

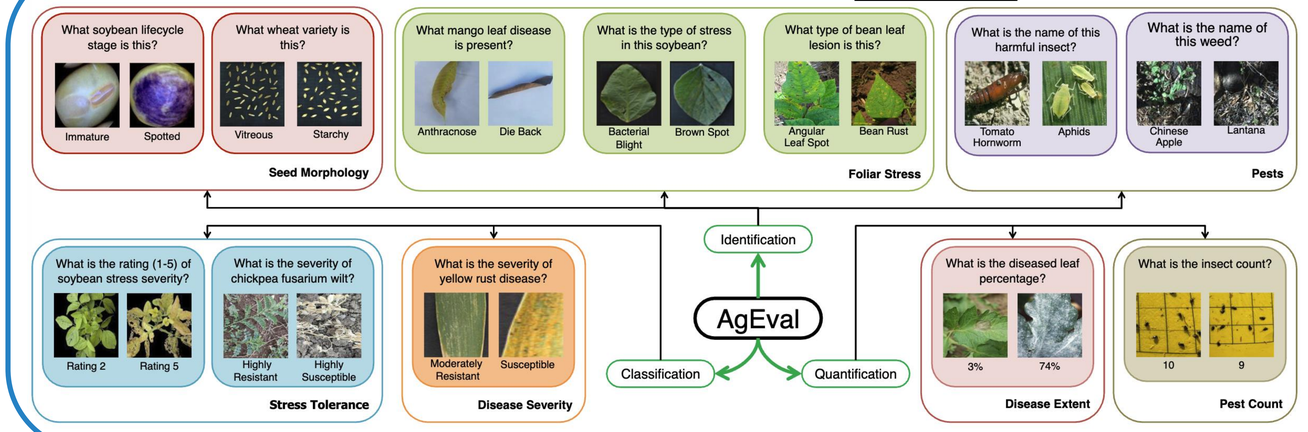
Iowa State Univeristy, New York University

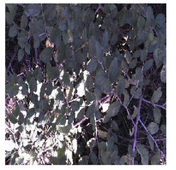
AI Institute for Resilient Agriculture (AIIRA) is supported by the National Science Foundation (NSF) and United States Department of Agriculture - National Institute of Food and Agriculture award #2021-67021-35329

Introduction

- **Current Challenges:**
  - Each agricultural AI model requires thousands of expert-annotated images.
  - Expert annotation is expensive and creates development bottlenecks.
- **Evaluates state-of-the-art multimodal LLMs on plant stress phenotyping tasks.**
  - SOTA: GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro.
  - Budget Friendly: Claude 3 Haiku, and Gemini 1.5 Flash
- **12 diverse tasks:**
  - Diversity: Seed quality, foliar stress, pests, disease severity, and stress tolerance
  - Span: Identification, Classification and Quantification.

Benchmark





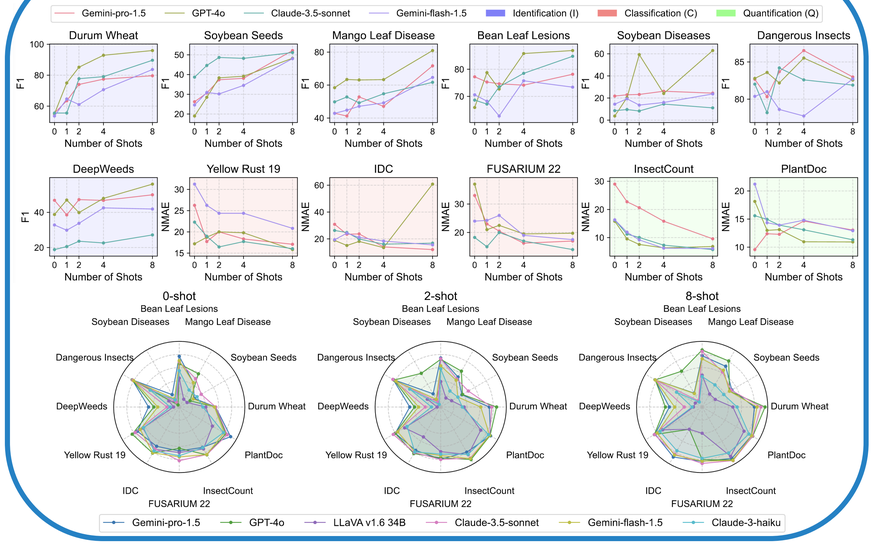
Category	Subcategory	Task
Identification (I)	Invasive Species	DeepWeeds

Ground Truth: Chinee apple

Predictions:

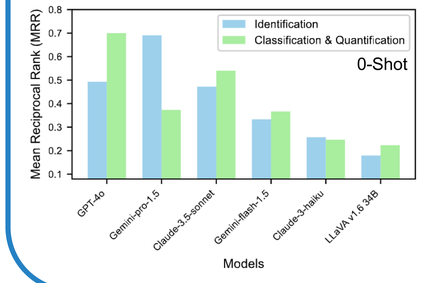
Model Name	0 shot	8 shot
Gemini-pro-1.5	Chinee apple	Chinee apple
GPT-4o	Chinee apple	Chinee apple
LLaVA v1.6 34B	Parthenium	Parkinsonia
Claude-3.5-sonnet	Lantana	Lantana
Gemini-flash-1.5	Prickly acacia	Chinee apple
Claude-3-haiku	Parthenium	Parthenium

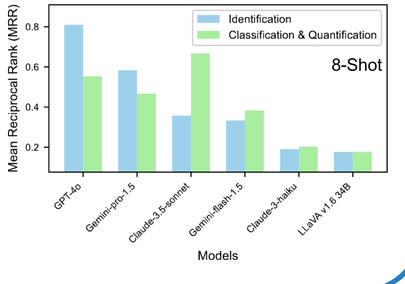
Analysis



Results

- Zero-Shot: Gemini-pro-1.5 leads in zero-shot identification (MRR 0.69), GPT-4o in classification/quantification (MRR 0.70).
- GPT-4o shows superior adaptability, with F1 score increasing from 46.24% to 73.37% in 8-shot identification.
- Bullseye examples (same category) improve F1 scores by 15.38% on average.
- Coefficient of Variation (CV) 26.02% (GPT-4o) to 58.03% (Claude-3-haiku). Subject matter expertise.





Conclusion

- Introduced comprehensive benchmark for agricultural multimodal LLM evaluation.
- Evaluated zero-shot and few-shot performance across diverse agricultural tasks.
- Analyzed example relevance impact and intra-task variability in LLMs.
- Established baselines for future research in agricultural AI applications.

Future work

- Expand scope of Benchmark.
- Increase shot counts for further analysis.
- Adapted Few Shot Learning for smart selection of input examples provided to Improve performance.