Assisted Few-Shot Learning for Vision-Language Models in Agricultural Stress Phenotype Identification

Muhammad Arbab Arshad
Iowa State University
USA

Talukder Zaki Jubery
Iowa State University
USA

Asheesh K. Singh
Iowa State University
USA

Arti Singh
Iowa State University
USA

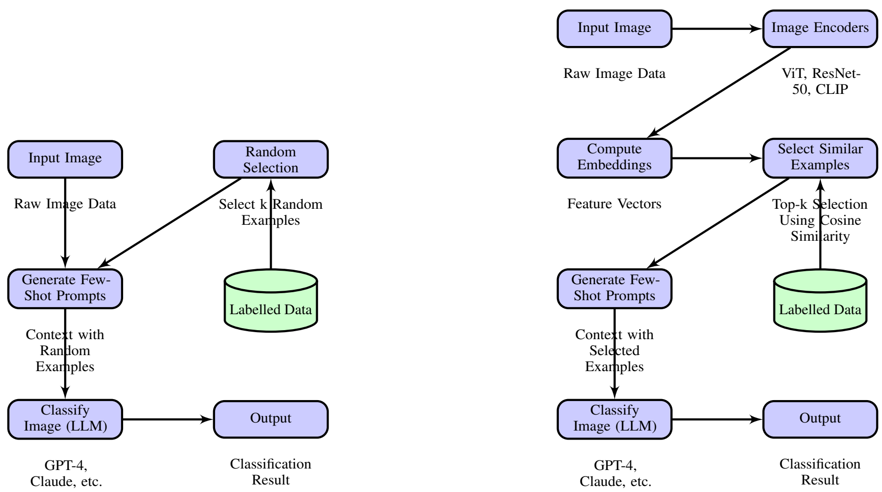
Chinmay Hegde
New York University
USA

Baskar Ganapathysubramanian
Iowa State University
USA

Aditya Balu
Iowa State University
USA

Adarsh Krishnamurthy
Iowa State University
USA

Soumik Sarkar*
Iowa State University
USA



(a) Traditional Few-Shot Learning

(b) Adaptive Few-Shot Learning

Figure 1: Comparison of Traditional and Assisted Few-Shot Learning approaches for image classification. (a) Traditional Few-Shot Learning randomly selects examples from the training set. (b) Assisted Few-Shot Learning uses multiple image encoders (ViT, ResNet-50, CLIP) to compute embeddings and select the most similar examples using cosine similarity. Both approaches use a large language model (e.g., GPT-4, Claude) for final classification.

Abstract

In the agricultural sector, labeled data for crop diseases and stresses are often scarce due to high annotation costs. We propose an Assisted Few-Shot Learning approach to enhance vision-language models (VLMs) for image classification tasks with limited annotated data by optimizing the selection of input examples. Our method employs one image encoder at a time—Vision Transformer (ViT), ResNet-50, or

*Corresponding author: soumiks@iastate.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

CLIP—to retrieve contextually similar examples using cosine similarity of embeddings, thereby providing relevant few-shot prompts to VLMs. We evaluate our approach on the agricultural benchmark for VLMs, focusing on stress phenotyping, where proposed method improves performance in 6 out of 7 tasks. Experimental results demonstrate that, using the ViT encoder, the average F1 score across seven agricultural classification tasks increased from 68.68% to 80.45%, highlighting the effectiveness of our method in improving model performance with limited data.

1 Introduction

In the agricultural sector, obtaining large annotated datasets for specific crop diseases or stresses is often expensive and time-consuming[Ghosal et al., 2018]. This scarcity of labeled data poses a significant challenge for developing accurate and robust classification models. Vision-language models (VLMs) have emerged as a promising solution[Chen et al., 2023a], offering the ability to learn from just a few examples through in-context learning and few-shot techniques. This paper builds on Arshad et al. [2024]'s AgEval benchmark, which evaluates zero-shot[Feuer et al., 2024] and few-shot plant stress phenotyping using multimodal LLMs, as done in general computer vision tasks [Bitton et al., 2023]. While their work showed promising results in agricultural applications, the selection of few-shot examples significantly impacts model performance. Therefore, optimizing the selection of input examples is crucial for enhancing accuracy on agricultural classification tasks.

(a) Sample image

Category	Subcategory	Task
Identification (I)	Foliar Stress	Bean Leaf Lesions
Question: What type of bean leaf lesion is this? **Ground Truth:** Bean Rust		

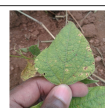
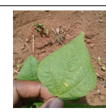
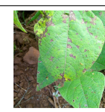
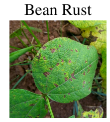
Method	Examples			
VIT-based	 Bean Rust	 Bean Rust	 Bean Rust	 Angular Leaf Spot
Traditional	 Angular Leaf Spot	 Angular Leaf Spot	 Angular Leaf Spot	 Healthy

Figure 2: Analysis of Adaptive few-shot example selection for Bean Leaf Lesions Task

Recent approaches to enhance few-shot learning capabilities in image classification leverage vision-language pre-trained models and employ techniques such as implicit knowledge distillation [Peng et al., 2024], contrastive losses [Kato et al., 2024], and fine-tuning attention pooling layers [Zhu et al., 2024]. While these methods have shown promising results, their complex adaptation processes and additional training steps may limit applicability in domains like agriculture, where labeled data and computational resources are limited[Sarkar et al., 2024]. This underscores the need for more efficient and adaptable few-shot learning techniques that maximize the utility of the limited available data without extensive modifications.

Few-shot learning techniques have been applied to specialized domains beyond general image classification, including action recognition [Wang et al., 2024], species recognition [Liu et al., 2024], and remote sensing applications [Chen et al., 2023b]. These applications demonstrate the adaptability of vision-language pre-trained models to domain-specific challenges. However, the necessity for domain-specific adaptations indicates that existing methods may not always generalize well across different fields, suggesting the need for approaches that require fewer modifications when applied to new domains.

In this work, we address the gap in the application of VLMs to agricultural problems by proposing an assisted few-shot learning approach that optimizes the selection of input examples. Our method intelligently selects the most relevant examples for each input image through simple similarity-based image retrieval, maximizing the utility of the limited available data. We demonstrate that this approach enhances few-shot learning performance in agriculture, improving accuracy on agricultural classification tasks while minimizing computational requirements and domain-specific adjustments. This advancement represents a step forward in adapting machine learning techniques to the challenges of agricultural image classification and offers a generalizable approach applicable across various domains.

2 Methodology

2.1 Foundation and Dataset Preparation

Our methodology builds upon the AgEval benchmark [Arshad et al., 2024], a framework designed to evaluate vision-language models on specialized agricultural tasks. We focus specifically on the identification subset of the benchmark, which encompasses challenges in plant stress phenotyping, disease detection, and crop variety classification. AgEval demonstrated the potential of multimodal large language models in addressing complex agricultural challenges, particularly in scenarios with limited labeled data. The AgEval benchmark includes tasks that test vision-language models' capabilities in agricultural contexts, such as identifying various plant stresses, diseases, and crop varieties from images, often with limited examples. These tasks reflect real-world agricultural scenarios where expert knowledge is crucial but labeled data may be scarce.

While the original study utilized various vision-language models, we concentrate on (GPT-4o) for our experiments. This choice is motivated by GPT-4o's superior performance across AgEval tasks[Arshad et al., 2024], particularly its significant improvement in few-shot learning scenarios, 8-shots, with F1 scores increasing from 46.24 to 73.37.

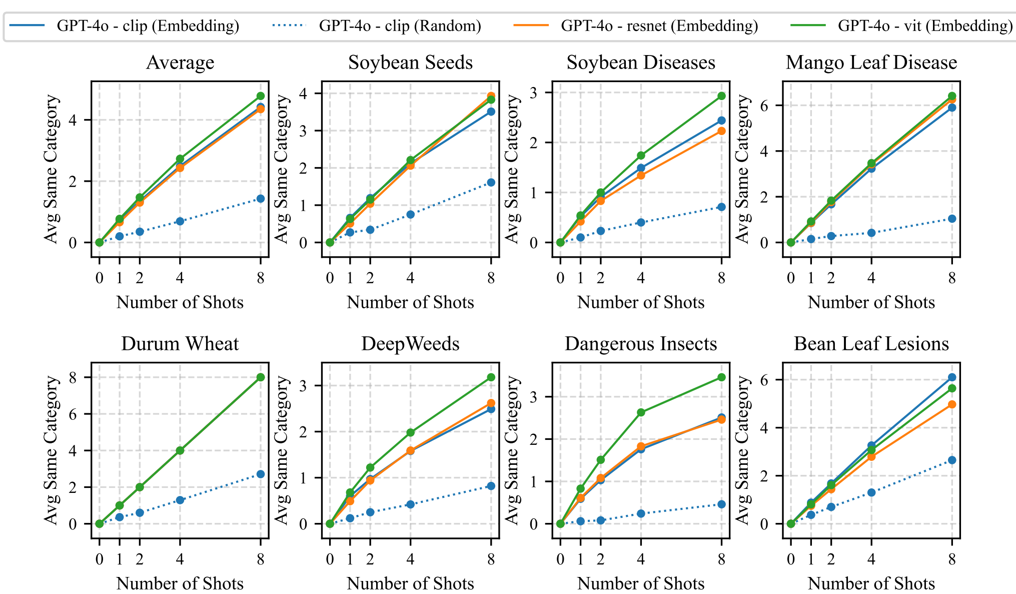


Figure 3: Comparison of average same category performance between assisted few-shot learning and baseline approaches on the AgEval benchmark.

2.2 Assisted Few-Shot Learning Approach

We employ an assisted few-shot learning strategy that utilizes a general-purpose prompt for identification tasks, detailed in the supplementary materials. To evaluate the model's performance under varying levels of supervision, we implement shot variations with 0, 1, 2, 4, and 8 examples.

To enhance the selection of examples, we integrate multiple image encoders—specifically CLIP, ResNet-50, and ViT—to compute image embeddings. An illustration provded in Figure 2 The implementation details of these encoders are provided in the supplementary section. For a given input image I, the encoder E generates an embedding $\mathbf{e}_I$:

$$\mathbf{e}_I = E(I).$$

We compute pre-embeddings for all candidate images in the dataset. The similarity between the input image embedding $\mathbf{e}_I$ and a candidate example embedding $\mathbf{e}_j$ is calculated using cosine similarity:

$$\text{sim}(\mathbf{e}_I, \mathbf{e}_j) = \frac{\mathbf{e}_I \cdot \mathbf{e}_j}{\|\mathbf{e}_I\|\|\mathbf{e}_j\|}.$$

The top k examples with the highest similarity scores are selected to form the few-shot examples, where k corresponds to the number of shots. This selection process can be formalized as:

$$\mathcal{E}_k = \underset{j \neq i}{\arg\,\text{top}_k} \left(\text{sim}(\mathbf{e}_I, \mathbf{e}_j) \right).$$

This similarity-based selection is integrated into the inference pipeline, allowing the model to utilize contextually relevant examples dynamically. We evaluate performance using the F1 score. By incorporating the most similar examples, the model can better generalize from limited data.

3 Results

We evaluate the effectiveness of various encoders in retrieving relevant examples and their impact on classification performance.

3.1 Evaluation of encoder effectiveness in retrieving contextually relevant examples for input images

Our analysis reveals that all three encoders—CLIP, ResNet-50, and ViT—significantly outperform random selection in retrieving contextually relevant examples. For the 8-shot scenario, the encoders retrieved on average 4.35 to 4.78 same-category images, compared to 1.42 for random selection. This trend persists across all shot variations (1, 2, 4, and 8), with ViT generally exhibiting the highest retrieval accuracy, followed closely by CLIP and ResNet-50. Figure 3 illustrates this comparison across different shot scenarios.

3.2 Impact on classification performance

The enhanced retrieval of relevant examples translates to improved classification performance across agricultural tasks. For the 8-shot scenario, assisted few-shot learning consistently outperforms random selection. The average F1 score across all tasks increased from 68.68% (random selection) to 77.16%, 79.16%, and 80.45% for CLIP, ResNet-50, and ViT encoders, respectively. This improvement is also evident in the 2-shot scenario, where the average F1 score rose from 60.45% to 69.19%, 69.07%, and 72.33% for the respective encoders. These results demonstrate the effectiveness of our approach in leveraging contextually relevant examples for improved classification performance. Figure 4 provides a comprehensive view of the performance improvements across different shot scenarios and encoders. Notably, in 6 out of 7 tasks, improvement occurs, highlighting the broad applicability of our approach across various agricultural classification tasks. Detailed results are provided in Table 1.

4 Discussion

The assisted few-shot learning method presented shows improvements over traditional approaches in agricultural classification tasks. By utilizing image encoders such as ViT, ResNet-50, and CLIP to select contextually relevant examples, the model achieved higher F1 scores, increasing from 68.68%

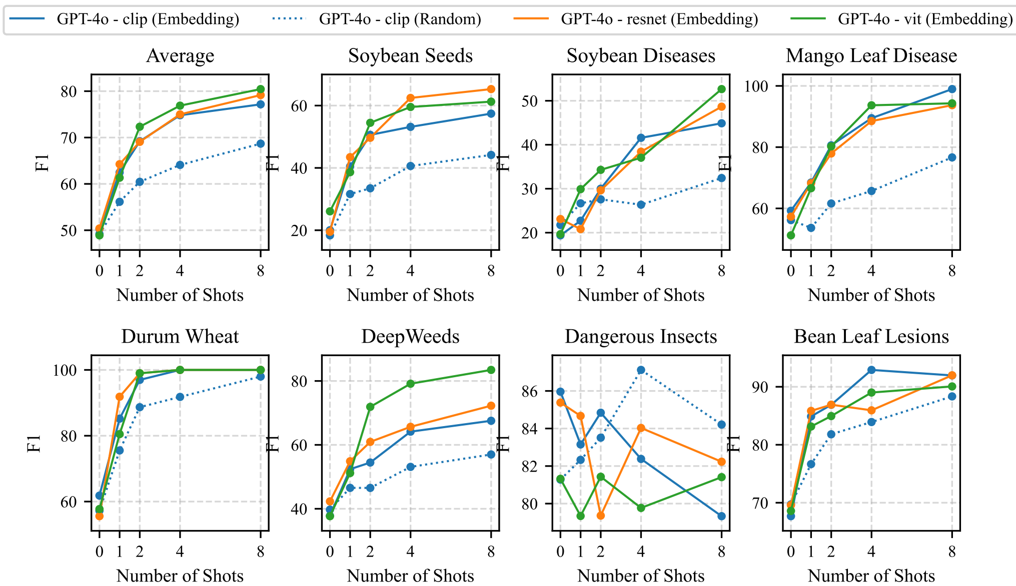


Figure 4: Adaptive few-shot learning vs baseline few-shot learning on AgEval benchmark.

Table 1: Assisted Few-shot Performance Comparison with Baseline (8-shot)

Task	Baseline	clip	resnet	vit
Bean Leaf Lesions	88.34	91.96 (+3.62)	91.98 (+3.64)	90.06 (+1.72)
Dangerous Insects	84.21	79.33 (-4.88)	82.23 (-1.98)	81.41 (-2.80)
DeepWeeds	56.99	67.54 (+10.55)	72.26 (+15.27)	83.46 (+26.47)
Durum Wheat	97.98	100.00 (+2.02)	100.00 (+2.02)	100.00 (+2.02)
Mango Leaf Disease	76.65	98.96 (+22.31)	93.71 (+17.06)	94.31 (+17.66)
Soybean Diseases	32.43	44.88 (+12.45)	48.66 (+16.23)	52.66 (+20.23)
Soybean Seeds	44.16	57.42 (+13.26)	65.29 (+21.13)	61.26 (+17.10)
Average	68.68	77.16 (+8.48)	79.16 (+10.48)	80.45 (+11.77)

to 80.45% in the 8-shot scenario with the ViT encoder. This suggests that incorporating similar examples enhances the model's performance, particularly in settings with limited annotated data.

Future work will aim to extend this methodology to a wider range of VLMs, including smaller and more efficient models, to assess scalability and resource efficiency. Additionally, applying this assisted few-shot learning approach across various tasks and domains may contribute to the development of a generalized framework. Enhanced agricultural image classification has the potential to support farmers in effective crop monitoring and management.

References

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.

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, 2023a.

Muhammad Arbab Arshad, Talukder Zaki Jubery, Tirtho Roy, Rim Nassiri, Asheesh K Singh, Arti Singh, Chinmay Hegde, Baskar Ganapathysubramanian, Aditya Balu, Adarsh Krishnamurthy, et al. Ageval: A benchmark for zero-shot and few-shot plant stress phenotyping with multimodal llms. *arXiv preprint arXiv:2407.19617*, 2024.

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.

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

Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE TRANSACTIONS ON MULTIMEDIA*, 26:3469–3480, 2024. ISSN 1520-9210. doi: 10.1109/TMM.2023.3311646.

Naoki Kato, Yoshiki Nota, and Yoshimitsu Aoki. Proto-adapter: Efficient training-free clip-adapter for few-shot image classification. *SENSORS*, 24(11), JUN 2024. doi: 10.3390/s24113624.

Yao Zhu, Yuefeng Chen, Xiaofeng Mao, Xiu Yan, Yue Wang, Wang Lu, Jindong Wang, and Xiangyang Ji. Enhancing few-shot clip with semantic-aware fine-tuning. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 2024 AUG 27 2024. ISSN 2162-237X. doi: 10.1109/TNNLS.2024.3443394.

Soumik Sarkar, Baskar Ganapathysubramanian, Arti Singh, Fateme Fotouhi, Soumyashree Kar, Koushik Nagasubramanian, Girish Chowdhary, Sajal K Das, George Kantor, Adarsh Krishnamurthy, et al. Cyber-agricultural systems for crop breeding and sustainable production. *Trends in Plant Science*, 29(2):130–149, 2024.

Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 132(6):1899–1912, JUN 2024. ISSN 0920-5691. doi: 10.1007/s11263-023-01917-4.

Lei Liu, Linzhe Yang, Feng Yang, Feixiang Chen, and Fu Xu. Clip-driven few-shot species-recognition method for integrating geographic information. *REMOTE SENSING*, 16(12), JUN 2024. doi: 10.3390/rs16122238.

Deliang Chen, Jianbo Xiao, Kyle Gao, Yanyan Lu, Sarah Fatholahi, and Jonathan Li. Natural language aided remote sensing image few-shot classification. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6298–6301. IEEE, 2023b.

A Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix. All such materials **SHOULD be included in the main submission.**

```
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"}
Replace "class_name" with the appropriate class from
↪  the list above based on your analysis of the
↪  image. The labels should be entered exactly as
↪  they are in the list above i.e.,
↪  {expected_classes}. The response should start
↪  with { and contain only a JSON object (as
↪  specified above) and no other text.
```

Figure 5: General purpose prompt for identification tasks in the AgEval benchmark. This prompt is utilized in both baseline and assisted few-shot learning approaches, remaining consistent across all experiments.

Table 2: Summary of encoder models and specific library functions used from the `transformers` library

Encoder Type	Model Name	Library Functions Used
CLIP	`openai/clip-vit-base-patch16`	`CLIPProcessor.from_pretrained`, `CLIPModel.from_pretrained`
ViT	`google/vit-base-patch16-224-in21k`	`AutoImageProcessor.from_pretrained`, `ViTModel.from_pretrained`
ResNet	`microsoft/resnet-50`	`AutoFeatureExtractor.from_pretrained`, `ResNetModel.from_pretrained`

(a) Sample image

Category	**Subcategory**	**Task**
Identification (I)	Invasive Species	Dangerous Insects
Question: What is the name of this harmful insect? **Ground Truth:** Colorado Potato Beetles		

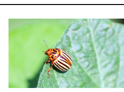
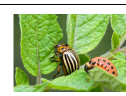
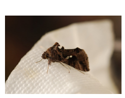
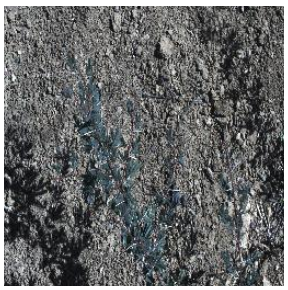
Method	**Examples**			
VIT-based	 Colorado Potato Beetles	 Colorado Potato Beetles	 Colorado Potato Beetles	 Colorado Potato Beetles
Traditional	 Cabbage Loopers	 Fall Armyworms	 Armyworms	 Cabbage Loopers

Figure 6: Analysis of Assisted few-shot example selection for Dangerous Insects Task



(a) Sample image

Category	**Subcategory**	**Task**
Identification (I)	Invasive Species	DeepWeeds
Question: What is the name of this weed? **Ground Truth:** Prickly acacia		

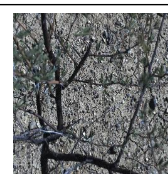
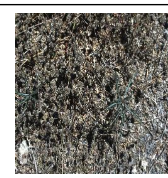
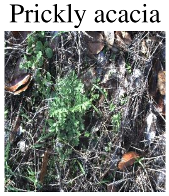
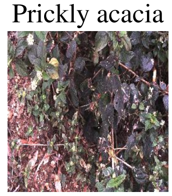
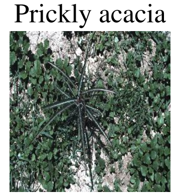
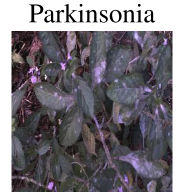
Method	**Examples**			
VIT-based	 Prickly acacia	 Prickly acacia	 Prickly acacia	 Parkinsonia
Traditional	 Parthenium	 Lantana	 Parkinsonia	 Snake weed

Figure 7: Analysis of Assisted few-shot example selection for DeepWeeds Task

(a) Sample image

Category	Subcategory	Task
Identification (I)	Seed Morphology	Durum Wheat

Question: What wheat variety is this?
Ground Truth: Starchy Kernels

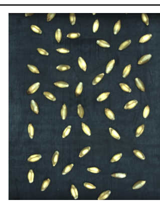
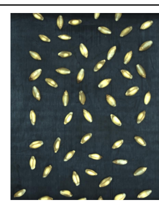
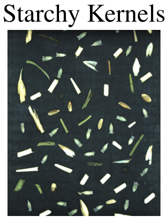
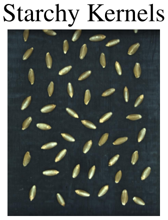
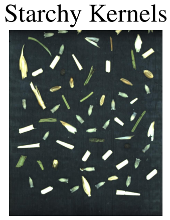
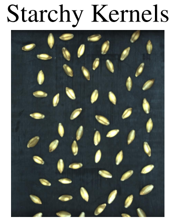
Method	Examples			
VIT-based	 Starchy Kernels	 Starchy Kernels	 Starchy Kernels	 Starchy Kernels
Traditional	 Foreign Matters	 Vitreous Kernels	 Foreign Matters	 Starchy Kernels

Figure 8: Analysis of Assisted few-shot example selection for Durum Wheat Task

(a) Sample image

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

Question: What mango leaf disease is present?
Ground Truth: Powdery Mildew

Method	Examples			
VIT-based	 Powdery Mildew	 Powdery Mildew	 Powdery Mildew	 Powdery Mildew
Traditional	 Healthy	 Powdery Mildew	 Gall Midge	 Cutting Weevil

Figure 9: Analysis of Assisted few-shot example selection for Mango Leaf Disease Task

(a) Sample image

Category	**Subcategory**	**Task**
Identification (I)	Foliar Stress	Soybean Diseases

Question: What is the type of stress in this soybean?
Ground Truth: Iron Deficiency Chlorosis

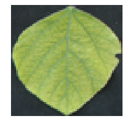
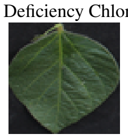
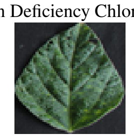
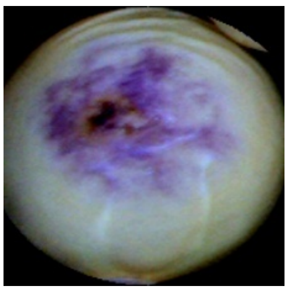
Method	**Examples**			
VIT-based	 Iron Deficiency Chlorosis	 Iron Deficiency Chlorosis	 Iron Deficiency Chlorosis	 Iron Deficiency Chlorosis
Traditional	 Frogeye Leaf Spot	 Healthy	 Bacterial Pustule	 Sudden Death Syndrome

Figure 10: Analysis of Assisted few-shot example selection for Soybean Diseases Task



(a) Sample image

Category	**Subcategory**	**Task**
Identification (I)	Seed Morphology	Soybean Seeds
Question: What soybean lifecycle stage is this? **Ground Truth:** Spotted		

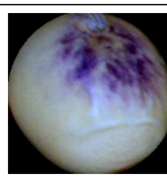
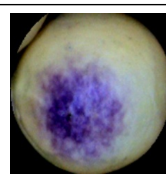
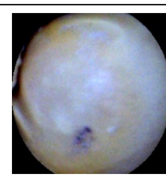
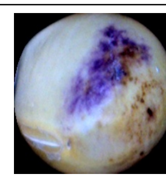
Method	**Examples**			
VIT-based	 Spotted	 Spotted	 Intact	 Spotted
Traditional	 Skin-damaged	 Immature	 Intact	 Skin-damaged

Figure 11: Analysis of Assisted few-shot example selection for Soybean Seeds Task

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) clearly state the paper's main contributions, including the proposed Assisted Few-Shot Learning approach and its performance improvements over traditional methods.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Section 4, including the need for validation on other language models and potential challenges in scaling to larger datasets.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper focuses on empirical results and does not include theoretical proofs or formal theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2 provides detailed information on the methodology, datasets, and experimental setup. Section 3 presents comprehensive results, including performance metrics and comparisons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully. If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We build on the open-source AgEval benchmark, and reasonable details for replication are provided in Sections 2 and 3. While we plan to make our code available soon, it is not open at the time of submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 2 provides comprehensive details on the experimental setup, including datasets, models used, and evaluation metrics. Section 3 further elaborates on specific experimental configurations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not include error bars or statistical significance tests in the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 2 explains that computations are handled via OpenAI API calls, so we draw on publicly available information about the API's requirements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper adheres to ethical guidelines listed in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The discussion section (Section 4) addresses only the positive societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any data or models that pose a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators of all assets used, including the AgEval benchmark, as detailed in Section 3.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects or crowdsourcing, making IRB approval not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.