

Contents lists available at ScienceDirect

### Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



# CDIP-ChatGLM3: A dual-model approach integrating computer vision and language modeling for crop disease identification and prescription

Changqing Yan<sup>a</sup>, Zeyun Liang<sup>a,1</sup>, Han Cheng<sup>a,1</sup>, Shuyang Li<sup>a</sup>, Guangpeng Yang<sup>a</sup>, Zhiwei Li<sup>a</sup>, Ling Yin<sup>b</sup>, Junjie Qu<sup>b</sup>, Jing Wang<sup>c</sup>, Genghong Wu<sup>d</sup>, Qi Tian<sup>d</sup>, Qiang Yu<sup>d</sup>, Gang Zhao<sup>d,\*</sup>

<sup>a</sup> College of Intelligent Equipment, Shandong University of Science and Technology, Taian 271019, China

<sup>b</sup> Guangxi Crop Genetic Improvement and Biotechnology Key Lab, Guangxi Academy of Agricultural Sciences, Nanning 530007, China

<sup>c</sup> College of Resources and Environmental Sciences, China Agricultural University, Beijing 100193, China

<sup>d</sup> College of Soil and Water Conservation Science and Engineering, Northwest A&F University, Yangling 712100, China

#### ARTICLE INFO

Keywords: Crop disease Large language model Fine-tuning ChatGLM3 Deep learning Disease identification Crop protection

#### ABSTRACT

Deep learning (DL) models have shown exceptional accuracy in plant disease identification, yet their practical utility for farmers remains limited due to a lack of professional and actionable guidance. To bridge this gap, we developed CDIP-ChatGLM3, an innovative framework that synergizes a state-of-the-art DL-based computer vision model with a fine-tuned large language model (LLM), designed specifically for Crop Disease Identification and Prescription (CDIP). EfficientNet-B2, evaluated among 10 DL models across 48 diseases and 13 crops, achieved top performance with 97.97 %  $\pm$  0.16 % accuracy at a 95 % confidence level. Building on this, we fine-tuned the widely used ChatGLM3-6B LLM using Low-Rank Adaptation (LoRA) and Freeze-tuning, optimizing its ability to deliver precise disease management prescriptions. We compared two training strategies-multi-task learning (MTL) and Dual-stage Mixed Fine-Tuning (DMT)-using a different combination of domain-specific and general datasets. Freeze-tuning with DMT led to substantial performance gains, achieving a 33.16 % improvement in BLEU-4 and a 27.04 % increase in the Average ROUGE F-score, surpassing the original model and state-of-the-art competitors such as Qwen-max, Llama-3.1-405B-Instruct, and GPT-4o. The dual-model architecture of CDIP-ChatGLM3 leverages the complementary strengths of computer vision for image-based disease detection and LLMs for contextualized, domain-specific text generation, offering unmatched specialization, interpretability, and scalability. Unlike resource-intensive multimodal models that blend modalities, our dual-model approach maintains efficiency while achieving superior performance in both disease identification and actionable prescription generation.

#### 1. Introduction

Crop pests and diseases present significant challenges to global agricultural production. Many farmers resort to indiscriminate pesticide use to manage these issues, which exacerbates environmental pollution and compromises food safety (Devi et al., 2022; Tudi et al., 2021). Traditionally, the identification of crop pests and diseases has relied on the expertise and experience of farmers and agricultural experts (Pathak et al., 2022). However, this approach is prone to errors and inconsistencies, especially in large-scale plantations where the vast crop areas require substantial efforts, making it easy to miss the optimal treatment window (Anderson et al., 2004; Brasier, 2008; Miller et al.,

2009; Spence et al., 2019; Strange et al., 2005). Accurate disease identification is key to effective disease management, as it enables targeted treatments, minimizes crop loss, and reduces unnecessary pesticide usages, ultimately promoting healthier crops and more sustainable agricultural practices (Carvajal-Yepes et al., 2019; Ristaino et al., 2021).

The continuous development of deep learning (DL) technology provides a promising solution for high accuracy detection of crop diseases (Abade et al., 2021; Fang et al., 2015; Roy et al, 2021; Saleem et al., 2019; Shoaib et al., 2023). Compared to traditional manual methods that rely on farmers' personal experience, knowledge, and consultations with experts, DL-based image classification models offer significantly higher accuracy in disease diagnosis (Rahman et al., 2020). Various

\* Corresponding author.

https://doi.org/10.1016/j.compag.2025.110442

Received 16 November 2024; Received in revised form 26 March 2025; Accepted 18 April 2025 Available online 25 April 2025 0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

E-mail address: gang.zhao@nwafu.edu.cn (G. Zhao).

<sup>&</sup>lt;sup>1</sup> These authors contributed equally to this work.

deep learning techniques have been developed and applied to achieve high accuracy in identifying and diagnosing plant diseases (Tian et al., 2024). For example, Too et al. (2019) evaluated the performance of state-of-the-art deep convolutional neural networks including Dense-Nets, VGGNet, Inception V4, and ResNet (with 50, 101, and 152 layers) on plant disease leaf identification and classification, and found that DenseNets achieved the highest accuracy of 99.75 %. Borhani et al. (2022) compared Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and their combination across various datasets, concluding that ViTs generally outperformed CNNs, particularly in complex scenarios. While these results demonstrate the high maturity and effectiveness of DL models like CNNs and ViTs in controlled settings, it is crucial to consider the variability in performance under real-world conditions. Factors such as lighting, disease severity, and crop diversity can affect model accuracy, underscoring the need for further research into robust, field-deployable models (Tian et al., 2024). Additionally, the practical application of these technologies by farmers remains a challenge. Even with accurate disease identification, many farmers, particularly those with limited agronomic knowledge, struggle to make informed decisions regarding the timing and choice of crop protection products (Oing et al., 2023). These farmers require clear, interactive guidance that not only identifies the disease but also provides actionable recommendations on which fungicide to use, when to apply it, and the correct dosage. Addressing this gap is essential to maximizing the benefits of artificial intelligence (AI) in agriculture, ensuring effective crop protection while minimizing resource waste and environmental impact.

Large language models (LLMs) are well-suited to address issues like providing clear, interactive guidance to farmers because they excel in understanding and generating human-like text (Raiaan et al., 2024; Zhao et al., 2023). They enable users to obtain timely, readable responses through direct, interactive questioning. Research has demonstrated the effectiveness of LLMs in agriculture, particularly in crop management (Raiaan et al., 2024; Yenduri et al., 2024). For example, based on Chat Generative Pre-trained Transformer (ChatGPT), a LLM developed by OpenAI, Wang et al. (2022) developed a question answering system for crop pest control and prevention, achieving higher scores than the other six state-of-the-art methods. Silva et al. (2023) evaluated popular large language models as Llama, GPT-3.5 and GPT-4 on their ability to answer agriculture-related questions and the results showed that GPT-4 achieved the highest performance answering 93 % of the questions correctly. Qing et al. (2023) proposed a deep learning vision model YOLOPC and ChatGPT-4 combined method to accurately identify pests and disease of citrus and provide diagnostic report, achieving high identification and reasoning accuracy. Additionally, multimodal LLMs are also employed to solve domain specific problems, demonstrating superior performance on some tasks (Kerdvibulvech, 2025). While these LLMs excel in general language tasks, they often struggle with domain-specific questions, such as those related to plant disease management, due to a lack of specialized knowledge in their training data (Yang et al., 2024). To address these limitations and enhance LLM performance in specialized fields like plant disease management, researchers have explored strategies for incorporating domainspecific knowledge into these models. Retrieval-Augmented Generation (RAG) and Fine-tuning are two common ways that are used to incorporate proprietary and domain-specific data (Gupta et al., 2024). RAG is particularly effective in contexts where relevant data is readily available, such as farm data interpretation, producing concise and contextually accurate responses (Lewis et al., 2020; Swenor, 2022). However, its reliance on external data retrieval systems can lead to inconsistent performance in scenarios with limited or incomplete data (S. Xu et al., 2024). In contrast, Fine-tuning enables the model to internalize domainspecific knowledge by retraining it on curated datasets, enhancing its ability to generate precise and contextually rich responses without dependency on external retrieval systems, although it involves a significant upfront cost due to the extensive data preparation and computational resources required (Xue et al., 2021).

This study aims to enhance the practical application of deep learning in crop disease management by introducing a novel dual-model integration approach that combines specialized disease identification models with LLMs for providing tailored prescriptions. In contrast to previous multi-modal methods that attempt to handle image recognition and text generation within a single framework, our dual-model approach harnesses the strengths of two independently optimized systems. First, we evaluate 10 computer vision models to identify 48 diseases across 13 crops. Next, we fine-tune the large language model ChatGLM3-6B using two methods to improve its domain-specific expertise in crop disease management. Finally, we develop CDIP-ChatGLM3 by combining the best-performing disease identification model with the fine-tuned ChatGLM3-6B to assist relevant growers in Crop Disease Identification and Prescription (CDIP). This dual-model approach not only improves the overall accuracy of diagnosis and recommendations but also offers greater flexibility and scalability, allowing for independent upgrades and improvements to each component. As a result, farmers can seamlessly transition from accurate disease diagnosis to precise treatment recommendations, reducing costs, minimizing environmental impact, and preventing pesticide overuse.

#### 2. Materials and methods

Fig. 1 presents the overall workflow for developing and deploying CDIP-ChatGLM3, an integrated system that combines a disease identification model with a fine-tuned large language model (LLM) for crop disease identification and prescription (CDIP). The identification model development begins with collecting an extensive image dataset containing 48 diseases across 13 crops as well as the healthy leaves for each of the 13 crops. This dataset is used to train and evaluate 10 deep learning models, including CNNs and Transformers, for identifying the health status of crops. After evaluation, the best-performing model is selected to classify crop images as either healthy or infected with a specific disease. Meanwhile, for LLM, disease management information is extracted from 13 well-established books, encompassing the diseases and crops in the dataset, and used to construct a Supervised Fine-Tuning (SFT) instruction-following dataset. This dataset was then used to finetune the LLM to enhance its expertise in crop disease management and language proficiency. The fine-tuned model is further evaluated and compared with the original models and three popular models on specialized dataset. The disease identification model is then integrated with the fine-tuned LLM, enabling an interactive system for crop disease diagnosis and prescription.

#### 2.1. Disease identification models

#### 2.1.1. Dataset collection

The dataset used in this experiment consists of 100,989 images covering 48 diseases across 13 different crops, including apples, cherries, citrus, corn, grapes, peaches, peppers, potatoes, rice, soybeans, strawberries, tomatoes, and wheat, all of which are publicly available (Table 1). Both healthy and diseased images for each crop were included to ensure comprehensive coverage for high-precision disease diagnosis. The images were collected under diverse conditions, including both laboratory setups and real-world field environments, to enhance the robustness of the model (Tian et al., 2024).

The images were sourced from various reputable datasets, including those from Northwest A&F University, and GitHub repositories (Acharya, 2020; Cruz et al., 2019; Feng et al., 2022; Mohanty et al., 2016; Thapa et al., 2020; Xu et al., 2024; Yuan et al., 2023). We carefully cleaned the data by removing any corrupt or low-quality images. After this cleaning process, each image was labeled according to its corresponding disease category, enabling supervised training. We then employed a stratified random sampling approach to partition the labeled dataset into training (60 %), validation (15 %), and test sets (25 %), ensuring that the distribution of the disease categories in each subset



Fig. 1. The overall technical workflow for developing and deploying CDIP-ChatGLM3. It includes, comparison of computer vision models for crop disease identification, fine-tuning large language models, and the integration of both models for interactive prescription of disease management.

remained consistent with that of the overall dataset. Subsequently, data augmentation and preprocessing techniques were applied to the training subset, including random resizing, random cropping, horizontal flipping, and mean–standard deviation normalization. Regarding the validation and test subsets, resizing, center cropping, and normalization in a consistent manner were applied to ensure comparability prior to their utilization. Thirteen representative disease images, one for each of the 13 crops, from the training dataset are displayed in Fig. 2.

#### 2.1.2. Model training and comparison

To identify a lightweight yet highly accurate model for disease identification, we chose five well-established computer vision model architectures, including ResNet (He et al., 2016), MobileNet (Howard et al., 2019), EfficientNet (Tan et al., 2019), Swin-transformer (Liu et al., 2021), FasterNet (Jierun Chen et al., 2023). High performance models with different layer numbers from these architectures were selected for training, with their features compared in Table 2. Additionally, the running performance information, including the approximate GPU memory usage and training time, is presented in Table 2 to facilitate the replication of our experiments. These measurements were obtained using a single NVIDIA RTX 4090, with a batch size of 8 and 8 workers.

The workflow for training, testing, and selecting the most accurate computer vision model is illustrated in Fig. 3. The dataset was first cleaned to remove low-quality images and then split into training (60%), validation (15%), and test sets (25%). Each model was trained on the training set, with performance monitored on the validation set to optimize hyperparameters and prevent overfitting. After training, the models were evaluated on the test set using four key metrics: accuracy, F1-score, recall, and precision. The models' performance was compared across these metrics to identify the most effective model for accurate crop disease identification. It is important to note that this training process was repeated three times to calculate a 95% confidence interval, thereby quantifying the uncertainty of the results.

Table 3 outlines the hyperparameters used in training the disease identification models, including learning rate (Lr), optimizer, and the number of epochs. The learning rate (Lr) controls the step size during gradient descent, which directly impacts the speed and stability of convergence. The optimizers such as Adam or SGD (Stochastic Gradient Descent)—is responsible for adjusting the model weights based on the computed gradients. Each model was trained for a fixed number of epochs—50 in this case—to balance performance and computational efficiency.

The number of collected diseases	or healthy image	e data for each crop.	Conditions include Field (	F), laboratory	(L) and mixed (M).
	2 0	1			

Crop	Disease	Pathogen	No. of images	Condition	Data source
Apple	Alternaria blotch	Alternaria mali	277	F	https://doi.org/10.11922/sciencedb.01627
nppie	Black rot	Botryosphaeria obtusa	428	L	https://autorg/10111922/scienceab.0102/
	Brown rot	Monilinia fructigena	4217	I	https://doi.org/10.11022/sciencedb.01627
	Crow mot	Dostalationsis ann	100	E	https://doi.org/10.11022/sciencedb.01627
	Grey spot	Pestalollopsis spp	199	F	https://doi.org/10.11922/sciencedb.0162/
	Healthy		1812	M	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
					https://www.kaggle.com/c/plant-pathology-2021-tgvc8/data
	Mosaic virus	Apple mosaic virus	156	F	https://doi.org/10.11922/sciencedb.01627
	Powdery mildew	Podosphaera leucotricha	438	F	https://www.kaggle.com/datasets/aibuzz/apple-leaf-disease-powdery-mildew
	Rust	Gymnosporangium juniperi-	5866	M	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
		virginianae			https://www.kaggle.com/c/plant-pathology-2021-fgvc8/data
Cherry	Healthy		2183	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Powdery Mildew	Podosphaera clandestina	193	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Citrus	Healthy	*	677	F	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Huanglonghing	Candidatus Liberibacter spp	5521	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Corn	Grev leaf spot	Cercospora zeae maydis	2114	ī	https://github.com/spMohanty/PlantVillage Dataset
Com	Hoolthy	Cereospora zeae-mayurs	2000	T	https://github.com/spinonancy/riancy/nage-bataset
	Nexthere less	E	2000	L	https://www.kaggle.com/datasets/qrankiisinia/com-teal-infection-dataset
	Northern lear	Exseroniium turcicum	2337	L	https://github.com/spMonanty/Plantvillage-Dataset/tree/master/raw/color
	blight			_	
	Rust	Puccinia sorghi	1946	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Grape	Black measles	Phomopsis viticola	1918	M	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Black rot	Guignardia bidwellii	2030	F	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Downy mildew	Plasmopara viticola	5061	L	The Guangxi Academy of Agricultural Sciences
	•	-			https://www.scidb.cn/en/detail?dataSetId=76b39c9c435d4035b5076412c2ddc
					b61&version=V2
	Leaf blight	Botryosphaeria spp	962	т	https://github.com/cpMohanty/PlantVillage_Dataset/tree/master/raw/color
	Hoolthy	bou yospilacita spp	1450	E	https://github.com/opMohanty/FlantVillage_Dataset/tree/master/raw/color
		o · · · · · ·	1430	r	
	Mosaic virus	Grapevine fanleaf virus	127	L	https://www.scidb.cn/en/detail?dataSetId=76b39c9c435d4035b5076412c2ddc
					b61&version=V2
	Powdery mildew	Erysiphe necator	1085	M	https://www.scidb.cn/en/detail?dataSetId=76b39c9c435d4035b5076412c2ddc
					b61&version=V2
	Yellows	Bois noir	1293	L	https://github.com/DrAlbertCruz/Salento-Grapevine-Yellows-Dataset/tree/mast
					er
Peach	Bacterial spot	Xanthomonas arboricola	1805	М	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Healthy		2160	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Penner	Bell bacterial spot	Xanthomonas spp	1879	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
repper	Healthy	Automonias spp.	2078	I	https://github.com/spMohanty/FlantVillage Dataset/tree/master/raw/color
	Geeb	The state of the s	2078	L	https://glulub.com/spholality/Plantvinage-Dataset/nee/master/raw/color
<b>_</b>	Scab	Elsinoe spp	588	L	https://gitnub.com/spMonanty/PlantVillage-Dataset/tree/master/raw/color
Potato	Healthy		1317	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Late blight	Phytophthora infestans	808	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Rice	Bacterial blight	Xanthomonas oryzae pv.	1023	F	https://www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images
		oryzae			https://doi.org/10.1016/j.biosystemseng.2020.03.020
	Brown spot	Bipolaris oryzae	1694	M	https://www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images
					https://doi.org/10.1016/j.biosystemseng.2020.03.020
	Healthy		1552	М	https://www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images
					https://doi.org/10.1016/i.biosystemseng.2020.03.020
	Hisna	Dicladisna armigera	495	м	https://www.kaggle.com/datasets/nizoroghezuode/rice-leaf-images
	mopu	Dicialispa annigera	150		https://doi.org/10.1016/i.biogystemsong 2020.03.020
	T C 1-1+	Management	0010		https://doi.org/10.1010/J.biosystemseng.2020.03.020
	Lear Diast	Magnaportne oryzae	2219	IVI	https://www.kaggle.com/datasets/mzorogbezuode/rice-teal-images
	_			_	https://doi.org/10.1016/j.biosystemseng.2020.03.020
	Tungro	Rice tungro virus	1248	F	https://www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images
Soybean	Angular leaf spot	Pseudomonas syringae pv.	214	F	https://aistudio.baidu.com/datasetdetail/253471
		syringae			
	Bacterial blight	Pseudomonas syringae pv.	381	F	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
		glycinea			
	Cercospora leaf	Cercospora kikuchii	1255	F	https://datadrvad.org/stash/dataset/doi:10.5061/drvad.41ps1rpi3
	blight	Gereospora kikueliii	1200	1	https://duudiydd.org/sdash/dudset/doi.10.0001/diydd. (histingo
	Downy mildow	Doronocooro monchurico	E10	т	https://datadmind.org/stash/datasat/daii10_E061/dmind_41ps1mi2
	Downy mindew	Concernance and the second sec	1010	L	https://datadropolog/stash/dataset/doi:10.5001/dropol/distantinj5
	Frogeye	Cercospora sojina	1210	F	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
	Healthy		2425	F	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
	Potassium		812	F	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
	deficiency				
	Rust	Phakopsora pachyrhiz	1627	F	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
	Target spot	Corynespora cassiicola	671	L	https://datadryad.org/stash/dataset/doi:10.5061/dryad.41ns1rnj3
Strawbe-	Healthy		2668	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
rrv	Leaf scorch	Diplocarpon earlianum	1740	I.	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Tomato	Bacterial spot	Yanthomonas spp	1671	L	https://github.com/spMohanty/PlantVillage_Dataset/tree/master/raw/color
romato	Farly Blight	Alternaria solani	15/1	I	https://aithub.com/cpMohanty/DaptVillage_Dataset/tros/master/raw/c0101
	Lally Digil		1000	L	https://grundb.com/spivionanty/riantymage-Dataset/free/inaster/raw/color
	rieaitny		1998	L *	https://giulub.com/spivionanty/Plantvillage-Dataset/tree/master/raw/color
	Late blight	Phytophthora infestans	1818	L	nttps://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Leaf mold	Cladosporium fulvum	1848	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Mosaic virus	Tomato mosaic virus	1215	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
	Septoria leaf spot	Septoria lycopersici	1714	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color

(continued on next page)

#### Table 1 (continued)

Crop	Disease	Pathogen	No. of images	Condition	Data source
	Yellow leaf curl virus	Tomato yellow leaf curl virus	1926	L	https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color
Wheat	Brown Rust	Puccinia triticina	2631	F	https://www.kaggle.
					com/datasets/shadabhussain/cgiar-computer-vision-for-crop-disease
	Healthy		2697	F	https://www.kaggle.
					com/datasets/shadabhussain/cgiar-computer-vision-for-crop-disease
	Loose smut	Ustilago tritici	939	М	https://www.kaggle.com/datasets/amankumar2004/wheat-disease-dataset
	Septoria	Septoria tritici	446	F	https://www.kaggle.com/datasets/amankumar2004/wheat-disease-dataset
	Yellow rust	Puccinia striiformis f. sp. tritici	1395	F	https://www.kaggle.com/datasets/amankumar2004/wheat-disease-dataset















Fig. 2. Thirteen representative disease images from the collected disease image dataset, with one for each crop.

#### 2.1.3. Evaluation metrics

We used four widely used evaluation metrics: accuracy, precision, recall, and F1-score. Accuracy represents the proportion of correctly classified samples out of the total number of samples in the test set. It is calculated using the following equation:

Grape downy mildew

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
(1)

where True Positives (TP) and True Negatives (TN) are the correct predictions, and False Positives (FP) and False Negatives (FN) are the incorrect predictions. Recall, also known as sensitivity, measures the model's ability to correctly identify positive instances. It is defined as:

$$Recall = \frac{TP}{TP + FN}$$
(2)

Precision is the proportion of true positive predictions among all the samples that were predicted to be positive. The formula for precision calculation is as follows:

$$Precision = \frac{TP}{TP + FP}$$
(3)

The F1-score is the harmonic mean of precision and recall, which considers both precision and recall, making it especially useful for datasets with imbalanced classes. It can be calculated as follows:

Summary of the features of the ten deep learning vision models used for crop disease identification.

Model	Parameter Size	Description	Training Time	GPU Memory
ResNet-34	21.3 M	Deep residual network with 34 layers for efficient training	~4.5 h	~3.7G
ResNet-50	23.6 M	A residual learning- based CNN known for handling deeper networks effectively	~4.1 h	~4.1G
MobileNetV3- Small	1.58 M	A lightweight neural network designed for mobile and edge devices.	~4.1 h	~3.3G
EfficientNet-B2	7.79 M	A scalable CNN that balances accuracy and efficiency.	~4.5 h	~4.3G
EfficientNet-B1	6.6 M	Smaller version of EfficientNet-B2 with improved accuracy.	~6.12 h	~4.3G
EfficientNet-B0	4.1 M	The baseline EfficientNet model, optimized for speed and parameter efficiency.	~4.6 h	~4.0G
EfficientNetV2- S	20 M	The upgrade of EfficientNetV1 aims to improve the accuracy of the model while also increasing the number of parameters.	~4.5 h	~3.5G
Swin- transformer- Tiny	27.6 M	A resource-efficient model for mobile vision tasks, with hierarchical and windowed attention mechanisms.	~5.5 h	~4.6G
FasterNet-T0	2.7 M	A streamlined CNN that prioritizes speed without compromising on accuracy.	~4.2 h	~3.3G
FasterNet-T1	6.4 M	A sophisticated CNN that maintains a balance between accuracy and efficiency.	~4.1 h	~3.3G

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$
(4)

#### 2.2. Large language model and fine-tuning

#### 2.2.1. Dataset preparation

Six datasets, including the fine-tuning specialized 13-Crop instruction-following dataset, general dataset Alpaca (Taori et al., 2023), general language abilities benchmark test dataset CMMLU (Li et al., 2023), the customized specialized dataset for model expertise test, the Image Understanding Test dataset for multimodal performance test, and the Image Dataset introduced in Section 2.1.1, were utilized for training, testing, validation and benchmarking in the CDIP-ChatGLM3 development process. The key characteristics of these datasets are presented in Table 4. The acquisition process of each dataset will be described in detail in the following sections.

The 13-Crop Instruction-Following Dataset was derived from 13 books covering crop cultivation and disease management, with one book dedicated to each crop (see STable 1). Since a high-quality fine-tuning dataset is critical to the supervised fine-tuning method adopted in the following section, a three-phase based high quality dataset generation method was employed (Dong et al., 2023). Compared with the two-phase based method SELF-QA (Zhang and Yang, 2023), this method can generate more instructions with greater conciseness and responses with more comprehensive details, making it easier and more accurate for farmers to follow. The workflow is illustrated in Fig. 4.

We first collected the specialized pdf textbooks on disease management for 13 crops (see Supplementary Materials STable1). Thereafter, optical character recognition (OCR) technology was utilized to extract the textual content into an editable format. Subsequently, the extracted text was manually segmented—by chapters, paragraphs, or disease categories—to produce multiple context-specific text files with complete semantic logic, with each segment saved as an individual .txt file. Each txt file served as contextual data fed to the Llama3.1-405b-instruct model, an instruction-tuned generative language model developed by Meta (Dubey et al., 2024).

With the context data prepared, the generation of fine-tuning datasets proceeded in three phases (Fig. 4). Firstly, Llama3.1-405b-instruct

#### Table 3

Overview of the hyperparameters used in disease identification models architecture.

Model architecture	Learning Rate (Lr)	Epochs	Optimizer
ResNet	0.01	50	SGD
EfficientNet	0.0001	50	Adam
EfficientNetV2	0.01	50	SGD
MobileNetv3	0.0001	50	Adam
Swin-Transformer	0.0001	50	AdamW
FasterNet	0.0001	50	Adam



diseases, 13 Crops)

Fig. 3. The workflow for crop disease data preparation, model training, validation, testing, and selection.

Key statistics of all the six datasets used for training, validation, and benchmarking in the CDIP-ChatGLM3 development.

Dataset	Crops	Diseases	Samples	Data Type	Task	Source	Purpose
Image Dataset	13	48	100,989	Image	Classification	The crop datasets listed in Table 1	Disease identification models training, validation, and testing.
13-Crop Instruction- Following Dataset	13	48+	2500	Text	Fine-tuning	Extracted from the selected high quality crop protection books and constructed via Llama 3.1 APIs	LLM model fine-tuning to enhance its expertise for application in agricultural domain.
Alpaca General Dataset	N/A	N/A	520,000	Text	Fine-tuning	Standard Alpaca Dataset https://huggingface.co/datasets/tatsu-lab/alpaca	Alleviating the catastrophic forgetting issue during the fine- tuning process for retaining general capabilities.
CMMLU Test Dataset	N/A	N/A	11,528	Text	Benchmarking	Standard CMMLU dataset https://github.com/haonan-li/CMMLU	Testing the general language abilities of LLMs.
Specialized Abilities Dataset	13	30	1000	Text	Benchmarking	200 disease prevention and control instructions were selected from the 13-Crop Instruction- Following Dataset, and the remaining 800 variants were generated via the Llama 3.1 API according to the selected 200 instructions.	Evaluating LLMs' expertise in agricultural disease control.
Image Understanding Test Dataset	13	48	610	Image	Classification	10 images were randomly selected for each crop disease and its healthy counterpart from the Image Dataset.	Evaluating LLMs' abilities to interpret and classify disease images.



Fig. 4. Workflow for generating a 13-Crop instruction-following dataset from the collected specialized crop pdf books.

model generated instructions by incorporating contextual information from the partitioned text according to Prompt1 (as shown in STable2). Secondly, these instructions were executed to produce an initial output according to Prompt2 (as shown in STable 3). Finally, the Llama3.1-405b-instruct model refined the initial output by integrating it with the contextual data according to Prompt3 (as shown in STable 4), thereby ensuring logical coherence and key insights extraction. To further clarify this process, an illustrative example of the instructionfollowing data generation process for professional disease prevention and control information on rice sheath blight is provided in STable 5. After generation, all instruction-following data were subjected to manual inspection, data cleaning, and integration to construct a comprehensive and high-quality instruction-following dataset for disease management across the 13 crops. The number of generated instruction-following data instances for each crop is shown in Fig. 5. The instance counts for each crop category were calibrated to achieve an even distribution, yielding a total of 2,500 instances across categories, with each comprising between 6.7 % and 9.6 % of the total. This dataset



**Fig. 5.** Distribution of the 13-Crop instruction-following dataset across different crops. Each section represents the proportion of dataset allocated to a specific crop.

was used for fine-tuning the large language model, with 10 % randomly sampled as a validation set during training to ensure a consistent distribution of crop categories.

Due to its extensive diversity and comprehensive domain coverage (Table 4), the publicly available Alpaca dataset, comprising a wealth of diverse, high-quality, and open-ended general questions covering various fields such as mathematics and coding, was selected to mitigate catastrophic forgetting. Catastrophic forgetting is quite common in domain-specific large language models (Kirkpatrick et al., 2017), especially when they learn specialized knowledge and terminology. Acquiring specialized knowledge often results in the loss of general knowledge, thereby weakening their language abilities. A portion of general data mixed with specialized data can help the model "recall" the lost general knowledge, alleviating the issue of catastrophic forgetting during the training on specialized knowledge (Hayes et al., 2020). Alpaca served as the general dataset to be combined with our highquality instruction-following specialized dataset on 13 different crops for fine-tuning training (Dong et al., 2023; Taori et al., 2023). By setting a mixing ratio between the specialized dataset and Alpaca, catastrophic forgetting issue is anticipated to be substantially mitigated.

Chinese Massive Multitask Language Understanding (CMMLU) benchmark dataset (Li et al., 2023) was used to test the model's reasoning abilities after fine-tuning. CMMLU is a comprehensive benchmark designed to evaluate large language models' knowledge and reasoning abilities within the context of Chinese language and culture. It spans 67 subjects, including topics in science, technology, engineering, and mathematics (STEM) fields, social sciences, and humanities, with a particular focus on tasks that are specific to China, such as traditional Chinese medicine and Chinese driving rules. The dataset consists of single-choice questions, with both a few-shot development set and a more extensive test set for each subject. The questions and their answer choices are designed to test models' understanding of complex and domain-specific knowledge, making it an important tool for assessing Chinese-language models like ChatGLM in diverse areas.

To test the fine-tuned model's specialized abilities, 200 frequently encountered or representative crop disease prevention and control questions, along with their original responses serving as the standard answers, were selected from the fine-tuning specialized dataset to create the specialized abilities test dataset. Based on these selected questions, four variants of each question, expressing the same meaning as the original, were obtained using Llama3.1-405b-instruct (S. Chen et al., 2023) (the variants obtaining process is provided in Supplementary Material STable 6). The original question and the standard answer as well as the four variants comprised a group. The created expertise test dataset comprised 200 groups in total. When testing the model's specialized abilities, these groups of questions were input into the fine-tuned models, the original models, and three state-of-the-art models including GPT-40 (Achiam et al., 2023), Llama3.1-405b-instruct and Qwen-max (Bai et al., 2023) and the responses were generated for evaluation of their expertise abilities.

To compare the classification performance of our model and the SOTA multimodal language models on disease images, we randomly selected 10 images from each disease category to construct the Image Understanding Test Dataset.

#### 2.2.2. ChatGLM3-6B as the base large language model

ChatGLM3-6B is an advanced version of the ChatGLM series designed specifically for Chinese language tasks, making it an ideal baseline model for fine-tuning in specialized domains, such as crop disease management, where less educated Chinese farmers are involved (GLM et al., 2024). Additionally, ChatGLM3-6B achieves an optimal balance between overall performance and computational efficiency. With 6 billion parameters, it provides robust general-purpose language understanding while remaining more accessible for fine-tuning on available GPUs, especially compared to larger models with tens of billions of parameters. This moderate size allows for efficient model training and finetuning within standard hardware constraints, making it a feasible choice for organizations or researchers with limited access to extensive computational resources. Moreover, ChatGLM3-6B maintains robust performance across a variety of tasks, ensuring it can handle both general language processing and domain-specific fine-tuning, such as crop disease management (GLM et al., 2024). Its ability to integrate specialized knowledge while preserving general language capabilities, as well as the widespread availability compared to the popular LLMs as ChatGPT, makes it a versatile and practical choice for fine-tuning.

#### 2.2.3. Fine-tuning ChatGLM3-6B for crop disease management

Despite the high proficiency of ChatGLM3-6B in natural language processing tasks, fine-tuning is necessary to meet the specific needs of farmers in addressing crop diseases effectively (Gupta et al., 2024). By tailoring its output to agricultural requirements in crop protection, finetuning enhances the model's responsiveness in disease prevention and control area. Commonly used fine-tuning methods should be carefully assessed and selected to maximize capability enhancement in these areas. To achieve this, the model fine-tuning procedures were conducted using the LLaMA-Factory operational platform (Zheng et al., 2024). Two widely used fine-tuning methods, Low-Rank Adaptation (LoRA) and Freeze-tuning method (Freeze) (Houlsby et al., 2019; Hu et al., 2021; Zheng et al., 2024), were compared. Following this, comprehensive finetuning of ChatGLM3-6B was performed to enhance its expertise while retaining its general language abilities.

LoRA is a fine-tuning method designed to adapt pre-trained language models efficiently with a minimal number of trainable parameters. It works by introducing low-rank adapters into the layers of the pretrained model, training only the dimensionality reduction matrix and the dimensionality enhancement matrix, while keeping the core model parameters fixed (Hu et al., 2021). This allows for a more focused and computationally efficient training process. The process involves three key steps: loading the pre-trained model, inserting LoRA adapters, and training the adapters.

On the other hand, the Freeze tuning method selectively freezes specific layers or most parameters in the pre-trained model, allowing only a small subset of decoder layer parameters to be updated during fine-tuning (Houlsby et al., 2019; Levine et al., 2022). This method helps retain the learned features from the pre-training phase, thus preserving the core knowledge while adapting the model to task-specific needs. The fine-tuning process for the Freeze-tuning method involves four steps:

loading the pre-trained model, freezing the desired layers, modifying the model as necessary, and training the model.

Different fine-tuning methods may yield varying performance levels across different tasks (Dettmers et al., 2024; Jeong, 2024). In this study, we evaluated both fine-tuning methods on ChatGLM3-6B under the parameters outlined in Table 5, which includes details such as batch size, learning rate, and epoch variations. These fine-tuning methods were tested across multiple epochs (3, 5, 10, 15, 20 and 30), with key distinctions in trainable layers for the Freeze method and specific values like LoRA alpha (16) and rank (8) for LoRA, as well as the choice of AdamW as the optimizer and a cosine learning rate scheduler. Additionally, the running performance for fine-tuning ChatGML3 on a single NVIDIA 4090 GPU is also provided in the last two rows of Table 5. The comparison of the fine-tuning methods helps identify the most effective method for improving the model's disease prevention and control capabilities while adhering to computational constraints.

As aforementioned, fine-tuning could cause catastrophic forgetting and weaken the general language abilities. To fine-tune the model for enhanced expertise while preserving general language abilities, two training strategies, including multi-task learning(MTL) and Dual-stage Mixed Fine-Tuning (DMT) (Dong et al., 2023), were introduced to determine the optimal effectiveness of various combination of professional and general dataset, and their performance was systematically compared. MTL directly mixes the expertise and general data for the fine-tuning. In contrast, DMT first executes fine-tuning on specialized dataset, and then utilizes varying proportions of specialized and general data for further fine-tuning to enhanced model performance. A workflow was designed to illustrate the entire fine-tuning process (Fig. 6).

The workflow begins with the generated specialized crop disease control dataset, referred to as the 13-Crop instruction-following dataset, followed by fine-tuning using Freeze and LoRA methods. After finetuning, an evaluation is conducted to select the optimal fine-tuning method and the best-performing model. Based on the selected method and model, further fine-tuning with MTL and DMT is performed, and the fine-tuning results are compared to identify the model that achieves the highest expertise performance while retaining the general language abilities.

#### Table 5

Overview of parameters set in fine-tuning ChatGLM3-6B using two different methods, Freeze-tuning and LoRA. The table outlines key configurations for both methods, including batch size, epochs, model-specific parameters (trainable layers for Freeze-tuning and alpha/rank for LoRA), learning rate, optimization techniques, and running performance.

Fine-tuning	Parameter	Freeze	LoRA
Training Setup	Train batch size	2	2
	Gradient	8	8
	accumulation		
	steps		
	Stage	Supervised Fine-	Supervised Fine-
		Tuning	Tuning
Epochs	Epoch	3, 5, 10, 15, 20, 30	3, 5, 10, 15, 20, 30
Validation	Validation size	0.1	0.1
	Validation batch	1	1
	size		
Model-Specific	Freeze trainable	2	-
	layers		
	LoRA alpha	-	16
	LoRA rank	-	8
Optimization	Learning rate	5.0E-05	1.0E-04
	Learning rate	Cosine	Cosine
	scheduler type		
	Optimizer	AdamW	AdamW
Running	GPU Memory	~22 GB	~17 GB
performance	Training Time	0.4 h (3 epochs) –	0.3 h (3 epochs) –
		4 h (30 epochs)	2.7 h (30 epochs)

#### 2.2.4. Metrics for evaluating fine-tuned LLM

Four commonly used metrics, including Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), CMMLU accuracy, and output error rate, were employed to evaluate the effectiveness of the two fine-tuning methods and the two training strategies.

BLEU is utilized to assess the quality of machine translation outputs (Papineni et al., 2002). It primarily measures the similarity between the machine translation output and the reference translation, with a particular emphasis on sentence accuracy and precise matching. BLEU evaluates the precision of machine translations by calculating the degree of match for n-grams (sequences of N consecutive words). The underlying principle of BLEU is that the closer the machine translation output is to professional human translation, the better the quality. A higher score of BLEU indicates superior machine translation performance. The ROUGE metric, commonly used for evaluating text summarization, compares the overlap between automatically generated text and reference summaries (Lin, 2004). We use three of its variations, including ROUGE-N, ROUGE-L and Average ROUGE F-score (Zhang et al., 2019). This metric is frequently reported to assess the performance of automatic summarization systems, machine translation, or any task involving the generation of natural language text. A higher Average ROUGE F-score indicates better alignment between the generated text and the reference text, suggesting that the model produces more coherent, relevant, and accurate output.

Both BLEU and ROUGE provide crucial insights into the quality of text generated by models like ChatGLM3-6B. While BLEU focuses on the precision and fluency of machine translations by evaluating n-gram matches, ROUGE evaluates the content retention and coherence of automatically generated text in comparison to human reference summaries. Together, these metrics help gauge how well the fine-tuned model captures both accuracy and meaning in specific domain tasks such as crop disease management.

To evaluate general language abilities of the models, CMMLU accuracy was employed to measure the ratio of the correctly answered questions to the total number of questions, which can be calculated as follows:

$$CMMLUaccuracy = \frac{Q_{correct}}{Q_{total}} \times 100\%$$
(5)

where  $Q_{correct}$  denotes the number of correctly answered questions and  $Q_{total}$  denotes the number of all the questions.

Additionally, during the fine-tuning process for generating answers, we observed that the model produces repetitively loops in varying probabilities across different epochs, resulting in excessively long responses. These repetitive loops could greatly destroy the robustness of the model. To assess this possibility, we introduced the output error rate (referred to as error rate) to quantify the likelihood of such errors occurring, which can be calculated as follows:

$$Error Rate = \frac{R_{token > 5000}}{R_{total}} \times 100\%$$
(6)

where  $R_{token}$  denotes the number of responses with more than 5000 tokens and  $R_{total}$  denotes the total number of generated responses.

Moreover, to evaluate the computational resource requirements of LLM models, the following formula was employ to estimate their memory usage (Anthony et al., 2023):

$$M \approx \frac{P \times Q \times 10^9}{8 \times (1024)^3} \times 1.2 \tag{7}$$

where M denotes the approximate GPU memory requirement, providing a simplified estimate of the Video Random Access Memory (VRAM) necessary to execute the model. P is the number of parameters in the model. Q specifies the numerical precision of each parameter, measured

I



Fig. 6. The workflow for fine-tuning ChatGLM3-6B. MTL represents multi-task learning and DMT for Dual-stage mixed fine-tuning.

in bits.

### 2.3. CDIP-ChatGLM3: Integration of identification models and fine-tuned LLM

By integrating the selected best crop disease identification model (CDIP) with the fine-tuned high-performance ChatGLM3-6B, we created CDIP-ChatGLM3, a practical tool for crop disease management. The integration was accomplished according to the following workflow. First, disease images are captured and uploaded, followed by analysis using the crop disease identification model. Once the identification result is generated, it is converted into a Chinese keyword, which is then embedded into a query or consultation, forming a disease-specific consultation. This consultation is then submitted to the fine-tuned ChatGLM3-6B. To obtain an accurate response, the LLM is instructed as follows:

"You are now a crop disease prevention assistant. Please answer questions in Chinese, describe the specific details of each question thoroughly, and display the content in a structured format."

Once the instruction is executed, the LLM outputs a response, embedding the disease-related keyword into a predefined template:

"The state of the leaf is {keyword}. If you have any questions, please feel free to ask."

Then farmers can obtain responses regarding the identification results or interact directly with the systems to access the relevant knowledge or any information they seek.

This integration enables seamless combination between the disease identification model and the LLM. With the selected deep learning vision model for disease identification and the fine-tuned ChatGLM3-6B for prescription, we developed and deployed our CDIP-ChatGLM3 based application. When a disease image is uploaded, CDIP-ChatGLM3 provides reliable and accurate responses for prevention and control in a timely, interactive manner. With CDIP-ChatGLM3, the efficiency of disease prevention and control could be significantly improved.

#### 2.4. Experiments design

2.4.1. Experiments on the models' expertise and general language abilities changes during fine-tuning

To select the best fine-tuning method and assess both disease-specific expertise and general language abilities of the fine-tuned model, we conducted fine-tuning experiments on both domain-specific and the combination of domain-specific and general dataset, using both LoRA and Freeze. Fine-tuning was conducted across 30 epochs. The finetuning method with the highest performance was chosen for further fine-tuning with MTL and DMT training strategies on a combined specialized and general dataset, with the goal of maintaining high expertise capabilities while retaining general language proficiency and low error rate. The Chinese Massive Multitask Language Understanding (CMMLU) benchmark (Li et al., 2023) was employed to evaluate the fine-tuned models' general language abilities in Chinese cultural context with the Zero-shot and Five-shot learning, respectively. Here, Zero-shot learning directly inputs the prompt to the LLM, and Five-shot learning presents a set of high-quality demonstrations on the target task, each consisting of both input and desired output. For Five-shot learning, the model is first exposed to high-quality examples, enabling it to better understand human intentions and the criteria for determining desirable answers.

### 2.4.2. Evaluation of the fine-tuned model's performance on the original domain-specific materials and its variants

To further test if the fine-tuned model can learn, understand, and apply the knowledge from the fine-tuning specialized dataset, we conducted assessments on the expertise abilities test dataset. Using the original ChatGLM3-6B, the fine-tuned ChatGLM3-6B, and three state-ofthe-art LLM models, including GPT-4o, Llama-3.1-405B, and Qwen-max, we generated responses to each group of five questions in the dataset. Subsequently, we compared the responses generated to the standard answers (i.e., the answers from the original training materials). BLEU-4 and average ROUGE F-scores were used to assess the professional capability of the five large language models in answering disease-related questions.

#### 2.4.3. Comparative evaluation of CDIP-ChatGLM3 and multimodal models

To evaluate the performance of our proposed dual-model CDIP-ChatGLM3 and multimodal models in disease image classification and understanding, we designed and conducted a comparative experiment. Several state-of-the-art universal multimodal language models, including LLaVA-v1.6-7B (Li et al., 2024), MiniCPM-V-2.6 (Yao et al., 2024), Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024), GLM-4v-Plus-20250111 (GLM et al., 2024), Qwen2.5-VL-72B-Instruct (A. Yang et al., 2024), DeepSeek-VL2 (Wu et al., 2024), and InternVL2.5-78B (Chen et al., 2024) are selected for comparison. We utilized either the APIs provided by these models or their publicly available pretrained versions without conducting any fine-tuning or training. Specifically, LLaVA-v1.6-7B, MiniCPM-V-2.6, and Llama-3.2-11B-Vision-Instruct were run locally via the Ollama inference framework in INT4 or Q4 K M formats, whereas Qwen2.5-VL-72B-Instruct, DeepSeek-VL2, and InternVL2.5-78B were accessed through external APIs in BF16 quantization. GLM-4v-Plus-20250111 is closed-source, providing no corresponding model information. Meanwhile, CDIP-ChatGLM3 used FP16 (16-bit) quantization. The memory calculation formula (Eq. (7)) was employed to estimate the GPU memory requirements of these models. The constructed image understanding test dataset was fed into these models. To ensure a fair comparison, particularly with the specialized dual-model CDIP-ChatGLM3, we standardized the prompt for all multimodal language models as follows:

"What disease is in the picture? Please select ONE and ONLY ONE option from the following list and output EXACTLY that option without any extra words: [48 crop disease categories + 13 healthy crop categories]."

Finally, Accuracy (Acc) and F1-score were employed as evaluation metrics to comprehensively quantify the models' capability in disease image classification and understanding.

Moreover, the same method outlined in Section 2.4.2 was applied to assess expertise abilities and evaluate the professional competence of our method and the multimodal models in the agricultural domain, using the Specialized Abilities Test Dataset.

#### 2.4.4. Computing environment

The experiments were conducted on a system equipped with an Intel (R) Core(TM) i9-14900 K CPU, 128 GB of RAM, and an NVIDIA(R) RTX 4090 graphics card with 24 GB of video memory. The software environment comprised Windows 11, CUDA version 12.6, Python version 3.10.14, and PyTorch version 2.4.0. This configuration enabled effective fine-tuning and evaluation of the models, ensuring their ability to deliver real-time, accurate disease management recommendations for farmers.

#### 3. Results

#### 3.1. Performance of disease identification models

The experimental results of the 10 classification models for classifying the diseases are shown in Fig. 7. The models with EfficientNet achitecture achieved both the highest and lowest performance in terms of all the metrics, among which EfficientNetB2 ranked the first, and EfficientNetV2-S ranked the last. ResNet, MobileNet, and Swintransformer achieved approximately and moderately high performance. FasterNets performed only lower than the three highest EfficientNets. For most of the architecture, the models and their variants of different sizes showed only slight differences in performance. Only the EfficientNetV2-S of the EfficientNets had a significant inferior performance compared to other EfficientNets, while two FasterNets had a mild performance difference. Among the 10 models, EfficientNet-B2 achieves the highest accuracy (97.97 %  $\pm$  0.16 %), recall (97.17 %  $\pm$  0.46 %), precision (97.18 %  $\pm$  0.59 %), and F1-score (97.12 %  $\pm$  0.10 %) at a 95 % confidence level.

Fig. 8 shows the classification accuracy of the best-performing EfficientNet-B2 model across all diseases, with cherry, citrus, corn, grape, peach, strawberry, and tomato crops exhibiting high classification accuracy. For these crops, most disease categories exhibited accuracies exceed 95 %, with a significant portion achieving 100 % accuracy. In contrast, the classification accuracy for diseases in apple, pepper, potato, rice, and wheat crops was comparatively lower, with substantial variation in accuracy across disease categories within these crops. Notably, cherry and strawberry crops had the highest average classification accuracy, reaching 100 %, while rice had the lowest average accuracy, with rice hispa disease classification accuracy at only 70.18 %. Further details on the disease identification model can be found in the Supplementary Material. Specifically, SFig. 5 highlights the top 40 misclassification pairs, while SFig. 6 presents the model's classification performance at different levels.

### 3.2. The performance of CDIP-ChatGLM3 in disease management prescription

Fig. 9 presents the fine-tuning results of ChatGLM3-6B on specialized dataset using LoRA and Freeze methods (Fig. 9a and Fig. 9b), and the test results of the fine-tuned models' general capability measured on the benchmark CMMLU (Fig. 9c and Fig. 9d). It was observed that both BLEU-4 and ROUGE scores increased with the fine-tuning epochs for both fine-tuning methods, achieving improved performance over the original ChatGLM3-6B. Notably, the Freeze method significantly outperformed LoRA, achieving a BLEU-4 index of nearly 350 at the 10th epoch. Meanwhile, the general language abilities exhibited fluctuations during the fine-tuning process and achieved lower performance metrics compared to the original ChatGLM3-6B. Additionally, the LoRA fine-



Fig. 7. Disease classification performance of 10 models in terms of accuracy, precision, recall and F1-score, based on the average results from three experimental runs. The highest values achieved are indicated with stars.



Fig. 8. Classification accuracy across all diseases using best EfficientNet-B2, with different colors representing various crops.



**Fig. 9.** Performance changes of two fine-tuning methods across epochs. (a) Changes in specialized abilities during LoRA based fine-tuning measured by BLEU-4 Index, Average ROUGE F-score, and error rate. (b) Specialized abilities changes during Freeze based fine-tuning measured by BLEU-4 Index, Average ROUGE Fscore, and error rate. (c) General language abilities changes during LoRA based fine-tuning measured by CMMLU accuracy changes. (d) General language abilities changes during Freezing based fine-tuning measured by CMMLU accuracy changes. Thick short vertical line segments represent the range of accuracy change for three tests. Zero-shot Learning directly inputs the prompt to the LLM. Few-shot Learning presents a set of high-quality demonstrations, each consisting of both input and desired output, on the target task.

tuning method exhibited a higher error rate, characterized by an initial sharp increase followed by a rapid decline. In contrast, the Freeze method maintained a lower error rate, with fluctuations observed across different epochs. At the 10th epoch, a relatively high expertise ability metric as well as a low error rate was achieved. Despite a slight decline in general language abilities compared to the original ChatGLM3-6B, it remained the best model when considering all factors. In contrast, although the better expertise ability could be achieved at the 20th epoch, the general language ability and the error rate decreased obviously, leading to poor robustness and accuracy. These results indicate the superior effectiveness of the Freeze method for fine-tuning on the specialized dataset, leading to the adoption of fine-tuned model at epoch 10 using Freeze (Freeze10) in subsequent experiments.

Based on the selected fine-tuned model by Freeze at the 10th epoch

(Freeze10) and Freeze method, the results of further fine-tuning using MTL and DMT training strategies on the dataset combining specialized and general language abilities are presented in Fig. 10. When MTL was used (Fig. 10a), the expertise ability first improved gradually with the increase of general data, reaching the highest BLEU-4 and ROUGE when the general data was four times size of special data. Meanwhile, the error rate first changed smoothly before sharply increasing, reaching the highest value at the same point as the expertise ability peaked. In contrast, when DMT was used (Fig. 10b), the expertise ability increased rapidly with the addition of specialized dataset until the maximum amount of specialized dataset were reached, followed by a slower incline before the general data was increased to twice the amount of the specialized dataset. Meanwhile, the error rate fluctuated, achieving its lowest value upon the inclusion of the maximum specialized dataset



**Fig. 10.** The performance changes of Freeze method with different training strategies and dataset ratios. (a) Specialized abilities changes during MTL strategy based fine-tuning measured by BLEU-4 Index, Average ROUGE F-score, and error rate. (b) Specialized abilities changes during DMT strategy based fine-tuning measured by BLEU-4 Index, Average ROUGE F-score, and error rate. (c) General language abilities changes during MTL strategy based fine-tuning measured by CMMLU accuracy. (d) General language abilities changes during DMT strategy based fine-tuning measured by CMMLU accuracy, and DMT represents dual-stage mixed training strategy. Thick short vertical line segments represent the range of accuracy change for three tests. Freeze10 (S:2.5 K) represents the model fine-tuned for 10 epochs using Freeze on the specialized dataset (the size is 2.5 K). Freeze10-MTL(S:2.5 K + G1.25 K) represents the model fine-tuned for 10 epochs using Freeze on the specialized dataset of 1.25 K using Freeze for 10 epochs. Freeze10(S:2.5 K) - G1.25 K) represents the model that first was fine-tuned for 10 epochs using Freeze on the specialized ataset of 1.25 K using Freeze for 10 epochs. Freeze10(S:2.5 K) - G1.25 K) represents the model that first was fine-tuned for 10 epochs using Freeze on the specialized ataset (the size is 2.5 K), followed by applying the DMT strategy for 10 epochs of Freeze-based fine-tuning on a combination of 2.5 K specialized dataset and 1.25 K general dataset.

(Fig. 10b). On the other hand, the general language abilities of the models reached near-peak levels at the same position where the expertise ability peaked, closely approaching the performance of original ChatGLM3-6B (Fig. 10d). Taking all these factors into consideration, DMT method achieved the best performance when the maximum amount of specialized dataset was included, combined with an equal amount of general data.

As for the general language abilities, there was a difference between Zero-shot and Five-shot method. When a 0-shot approach was used, high accuracy was obtained, showing fluctuations as the proportion of general data increased (Fig. 10c and Fig. 10d), or epochs increased (Fig. 9c and Fig. 9d). In contrast, with a Five-shot approach, most models demonstrated lower accuracy compared to the Zero-shot method, exhibiting similar trends in performance as Zero-shot. No obvious change on both Zero-shot and Five-shot were found.

Based on the experimental results above, we ultimately selected the Freeze10(S:2.5 K)-DMT(S:2.5 K + G2.5 K) model as the fine-tuned model for developing CDIP-ChatGLM3.

The test results for the expertise abilities of our fine-tuned models and other comparative models are shown in Fig. 11. Our fine-tuned model significantly outperformed all other models in terms of both BLEU-4 Index and Average ROUGE F-score when applied to the domain



Fig. 11. Test results for the expertise capabilities of the original ChatGLM3-6B, Qwen-Max, Llama-3.1-405B-Instruct, GPT-4, and our fine-tuned model, CDIP-ChatGLM3 with metrics of BLEU-4 and Average ROUGE F-score. Bold diagonal lines indicate broken axes.

of disease control and prevention. The original ChatGLM3-6B exhibited the lowest BLEU-4 Index value among the models evaluated. Llama-3.1-405B-Instruct ranked second, but significantly lower than our fine-tuned ChatGLM3-6B (CDIP-ChatGLM3). In contrast, GPT-4o achieved a BLEU-4 Index comparable to that of the ChatGLM3-6B, which was the lowest. Additionally, the Qwen-max achieved a BLEU-4 Index only slightly better than the lowest one. However, the ranking of these models varied when sorted by Average ROUGE F-score or BLEU-4, with the exception that our CDIP-ChatGLM3 consistently ranked first for both metrics. Specifically, GPT-4o and Qwen-max achieved similar Average ROUGE F-score, both of which was noticeably lower than the highest score. ChatGLM3-6B achieved an Average ROUGE F-score slightly lower than that of Qwen-max. While ranking second in terms of BLEU-4 Index, Llama-3.1-405B-Instruct achieved the lowest Average ROUGE F-score.

The experimental results comparing the image classification abilities as well as expertise abilities of our model with the popular universal multimodal models are presented in Table 6. Among all the multimodal models, Qwen2.5-VL-72B-Instruct, InternVL2.5-78B, and GLM-4v-Plus-20250111 achieved the top three scores on both image identification and expertise abilities, while exhibiting significantly higher memory consumption compared to other models. In contrast, LLaVA-v1.6-7B and MiniCPM-V-2.6 exhibited moderate domain-specific expertise while demonstrating the lowest performance in image identification and minimal memory requirements. Compared to these two models, DeepSeek-VL2 showed modest performance in image identification and ranked near the bottom in expertise. Although Llama-3.2-11B-Vision-Instruct required less memory than most multimodal models, its domain-specific expertise was among the lowest observed. These results indicate that, the pretrained versions or API access of universal multimodal models, could only achieved limited performance on our dataset (disease images and disease instruction Q&A), particularly in image classification and specialized expertise tasks (Table 6). In contrast, our dual-model CDIP-ChatGLM3, which underwent vision training and finetuning, achieved significantly better performance in both aspects. Notably, our model has the smallest size and reduced inference memory consumption compared to most multimodal models, enabling it's training and deployment with comparatively limited resources.

Additionally, a comprehensive performance summary of all tested

#### Table 6

Disease identification and expertise abilities compariso	n of multimodal models and the dual-model CDIP-ChatGLM3.
--	--

Model	Туре	Params	Language Model	Vision Model	Loading Precision (Q)	Inference Memory (G)	Image Understanding Test Metric		Specialized Abilities Test Metric	
							Acc	F1- score	BLEU-4 Index (‰)	Average ROUGE F-score (‰)
LLaVA-v1.6-7B	Multi- modal	7B	Llama-7B	CLIP-300 M	INT4 (4-bit)	~5	0.04	0.04	0.87	292.40
MiniCPM-V-2.6		8B	Qwen2-7B	CLIP-500 M	INT4 (4-bit)	~6	0.04	0.04	2.03	228.38
Llama-3.2-11B- Vision-Instruct		11B	Mllama-9.8B	Mllama-895 M	Q4_K_M (~4.5bit)	~8	0.04	0.04	0.68	72.06
GLM-4v-Plus- 20250111		N/A	N/A	N/A	N/A	N/A	0.12	0.13	0.90	309.86
Qwen2.5-VL-72B- Instruct		73.4B	Qwen2.5-72B	Qwen-ViT	BF16 (16-bit)	~164	0.21	0.21	1.45	304.36
DeepSeek-VL2		27.5B	DeepSeekMoE- 27B	SigLIP-400 M	BF16 (16-bit)	~62	0.11	0.12	0.11	108.91
InternVL2.5-78B		78B	Qwen-2.5-72B	InternViT- 6B-v2.5	BF16 (16-bit)	~173	0.19	0.2	0.45	304.38
CDIP-ChatGLM3	Dual- model	6B	FT-ChatGLM3- 6B	EfficientNet- b2	FP16 (16-bit)	~13	0.98	0.98	106.86	463.13

Comprehensive performance summary of all tested vision and language models.

Туре	Model	Image Dataset Test (Avg)			Image Understanding Test				Specialized Abilities Test		
		Acc	Р	R	F1	Acc	Р	R	F1	BLEU-4 Index(‰)	Avg Rouge F-score(‰)
CV	EfficientNet-B0	0.97	0.96	0.96	0.96	N/A				N/A	
	EfficientNet-B0	0.97	0.96	0.95	0.95						
	EfficientNet-B2	0.98	0.97	0.97	0.97						
	EfficientNetV2-S	0.89	0.90	0.89	0.89						
	FasterNet-T0	0.98	0.97	0.96	0.96						
	FasterNet-T1	0.98	0.96	0.96	0.96						
	MobileNetV3-Small	0.96	0.96	0.94	0.94						
	ResNet-34	0.96	0.95	0.95	0.95						
	ResNet-50	0.97	0.95	0.95	0.95						
	Swin-transformer-Tiny	0.97	0.97	0.96	0.96						
MLLM	LLaVA-v1.6-7B	N/A				0.04	0.05	0.04	0.04	0.87	292.40
	MiniCPM-V-2.6					0.04	0.04	0.04	0.04	2.03	228.38
	Llama-3.2-11B-Vision-Instruct					0.04	0.04	0.04	0.04	0.68	72.06
	GLM-4v-Plus-20250111					0.12	0.14	0.12	0.13	0.90	309.86
	Qwen2.5-VL-72B-Instruct					0.21	0.21	0.21	0.21	1.45	304.36
	DeepSeek-VL2					0.11	0.14	0.11	0.12	0.11	108.91
	InternVL2.5-78B					0.19	0.21	0.19	0.2	0.45	304.38
LLM	ChatGLM3-6B					N/A				0.55	286.06
	Qwen-max									0.87	316.67
	Llama-3.1-405B-Instruct									3.71	101.55
	GPT-40									0.61	327.48
Dual-model	CDIP-ChatGLM3	0.98	0.97	0.97	0.97	0.98	0.98	0.98	0.98	106.86	463.13

vision and language models is also presented in Table 7 to facilitate the readers.

To further evaluate the performance of CDIP-ChatGLM3 for practical application, a sample prescription on grapevine downy mildew disease management, generated by both the original ChatGLM3-6B and the fine-tuned model CDIP-ChatGLM3, respectively, is presented in Table 8. In Q1 question and its responses, the underlined text demonstrates that the CDIP-ChatGLM3 provided more specific and actionable responses to the terms '*what is the typical time frame*', and '*chemical control*' compared to the original ChatGLM3-6B. The detailed information on grapevine growth stages, as well as specific chemical agent names and usage methods, comprised practical guidance for farmers to follow for efficient disease control. In the follow-up question Q2, the fine-tuned model offered more comprehensive, logically structured, and actionable recommendations for integrated disease management than the original model. In contrast, the responses from the original ChatGLM3-6B were overly general, limiting their practical application for farmers.

Furthermore, responses from the three comparative large-scale LLMs—GPT-40, Qwen-max, and Llama-3.1-405B-Instruct, are presented in Supplementary Materials STable 7. Compared to these models, CDIP-ChatGLM3 delivered more pragmatic and readily implementable recommendations, exhibiting superior precision in domain-specific decision-making contexts. However, for the qualitative inquiries, the other large language models generated more comprehensive and extended responses than CDIP-ChatGLM3. For more examples, the readers could refer to Supplementary STable 9 and STable 10 for rice blast and wheat leaf blight prescriptions.

#### 4. Discussion

Crop disease management challenges require effective identification and prescription solutions to increase the adoption of advanced technologies among farmers. Current DL models achieve high accuracy in identifying plant diseases but fail to provide actionable guidance, limiting their practical impact. This study addresses these challenges by developing CDIP-ChatGLM3, a dual-model framework that integrates the precision of EfficientNet\_b2 for disease identification with the contextual expertise of a fine-tuned ChatGLM3-6B. By combining accurate detection with tailored prescriptions, this approach bridges the gap between technical capabilities and practical usability, promoting the adoption of AI-driven solutions in agriculture.

#### 4.1. The performance of disease identification models

Extensive comparative experiments involving ten widely used deep learning models, evaluated on a comprehensive dataset of lab and field images, yield two important findings. Firstly, within the same architecture, model size has only a minimal impact on performance. While a larger model size may yield slightly improved results, the differences remain modest. Compared to the training data, the impact of architecture on the accuracy of the models could be almost ignorable in most cases, which aligns with Tian et al. (2024). Secondly, most of the EfficientNets outperform other deep learning architectures, with EfficientNet-B2 achieving the highest performance with an accuracy of 97.97 %  $\pm$  0.16 % at a 95 % confidence level. Similar to our work, Tian et al. (2024) compared ResNet, MobileNet, and Efficient-BO, finding that EfficientB0 achieved the highest accuracy. Despite their comparison only involves Efficient-B0, their findings are consistent with ours. Additionally, the comparison between ResNet and EfficientNet in Atila et al. (2021) is also consistent with our findings. Although higher accuracy was achieved in their research using EfficientNet-B2, it was based solely on the PlantVillage dataset, with no validation conducted on a dataset as extensive as the one used in our research. Zhang et al. (2024) achieved the highest accuracy of 87.51 percent using ResNet50 only on maize of Plant Village dataset, achieving lower accuracy than our results.

The superiority of EfficientNets and EfficientNet-B2 could be attributed to the optimal compound scaling method, which combines the depth, width and resolution (Tan et al., 2019), which could potentially reduce expenses while meeting essential requirements for disease identification. The high accuracy and lightweight EfficientNet-B2 could generate accurate disease identification result, ensuring that the large language model can provide the fast and accurate prescription.

#### 4.2. The impact of fine-tuning on the abilities of language models

The fine-tuning experiments reveal two key insights. First, when fine-tuning with specialized dataset, large language models (LLMs) show a significant enhancement in their expertise abilities. However, this improvement in specialized capabilities comes at the cost of a gradual decline in their general language abilities. This effect may be attributed to the fine-tuning process, where parameters related to domain-specific knowledge are adjusted, while parameters governing general knowledge

Sample 1: grapevine downy mildew consultation and the corresponding prescription responses from both ChatGLM3-6B and the fine-tuned model. The original content is in Chinese and is provided in STable 8.

### Q&A by different LLMs

Q1: What is the typical time frame for implementing chemical control measures against grapevine downy mildew?

ChatGLM3-6B: Grape downy mildew is a prevalent plant disease commonly found in warm, humid environments. The following outlines the typical periods for chemical management of grape downy mildew:

Chemical control is generally advised during the following stages:

In conclusion, chemical management of grape downy mildew should be adapted based on disease development and weather conditions, as timely treatments are essential for effective disease control.

- 1. Preventive Measures: During grape cultivation, downy mildew can be mitigated by selecting resistant grape varieties, maintaining appropriate planting density, performing timely pruning, and implementing effective vineyard management practices.
- Treatment Measures: Upon the appearance of downy mildew, chemical control can be achieved through the application of fungicides targeting the downy mildew pathogen, as well as biological control agents. Early intervention during initial stages of infection is essential for effective disease management.
- 1. Pre-Bud Break in Spring: This is a crucial period for vineyard management, allowing for soil disinfection and pruning to reduce disease risk.
- 2. Early Fruit Development: Downy mildew commonly occurs during this period, necessitating regular monitoring and treatment to prevent disease progression.
- 3. Pre-Harvest (Prior to Fruit Maturity): This period is vital for maintaining fruit quality, where protection against disease is crucial for optimal fruit quality.

CDIP-ChatGLM3:

- 1. Winter Vineyard Sanitation: Apply 500-fold diluted 77 % copper sulfate-calcium mixture for preventive treatment.
- 2. Pre-High-Risk Period: Apply 600-fold diluted 80 % mancozeb wettable powder or 600-fold diluted 72 % metalaxyl-mancozeb wettable powder for preventive control.
- 3. High-Risk/Infection Period: Use 1500-fold diluted 20 % cymoxanil suspension, 2000-fold diluted 687.5 g/L fluopicolide-propamocarb suspension, 2000-fold diluted 50 % oxathiapiprolin wettable powder, or 2000-fold diluted 20 % cymoxanil suspension for effective control.
- Emergency Control During Bloom and Post-Bloom Stages (when downy mildew infects flower clusters and young fruit): Apply 2000-fold diluted 687.5 g/L fluopicolide-propamocarb suspension or 2000-fold diluted 50 % oxathiapiprolin wettable powder for rapid intervention.

Q2: Are there alternative methods for controlling it?

ChatGLM3-6B: In addition to chemical control, there are other methods for managing grape downy mildew:

- 1. Biological Control: Utilize biological control methods, including natural predators, antagonistic pathogens, and biocontrol agents. These approaches can effectively reduce the occurrence and spread of grape downy mildew.
- 2. Physical Control: Apply physical control measures, such as black nets and shade nets, to limit the incidence and transmission of the disease.
- Cultural Control: Implement appropriate vineyard management practices, such as pruning, thinning, and maintaining optimal planting density, to decrease disease incidence and spread.
   Integrated Control: Combine multiple control strategies in an integrated approach to effectively manage grape downy mildew.
- CDIP-ChatGLM3:

Apart from chemical control, common methods for managing grape downy mildew include: 1. Cultural Control:

- Select resistant grape varieties.
- Perform proper pruning and remove diseased tissue.
- Apply appropriate irrigation to avoid excessive moisture.
- Balance fertilization, avoiding over-application of nitrogen.
- **Biological Control:**
- Introduce natural predators, such as antagonistic fungi targeting downy mildew pathogens.
- Use biological fungicides, like Bacillus subtilis.
- 3. Physical Control:
- Utilize UV lamps to kill pathogens.
- Cover the vineyard with white plastic film to inhibit pathogen growth.
- 4. Integrated Control:
- · Combine cultural, biological, and chemical control methods for comprehensive management.
- Select appropriate control periods based on disease development.
- · Conduct timely interventions at critical stages, such as early infection and pre- and post-bloom phases.
- Practice winter vineyard sanitation by removing infected tissue to reduce pathogen sources.
- When applying chemicals, choose low-toxicity, low-residue options and prioritize safe usage.

must also adapt to the updated domain-specific parameters. As a result, the general knowledge parameters can deviate from their intended functions, leading to a decline in the model's overall versatility. This finding is consistent with Dong et al. (2023), who highlighted the adversarial interaction between specialized and general language abilities during fine-tuning, confirming that fine-tuning for specialized tasks can inadvertently hinder the model's broader linguistic capabilities. Wang et al. (2024) and Tariq et al. (2024) also verified the effectiveness of fine-tuning on specialized abilities. Furthermore, while the comparative results of models such as Llama, ChatGPT, and ChatGLM may differ from ours—potentially due to differences in research focus and experimental setup—Wang et al. (2023) underscore the effectiveness of fine-tuning in agricultural technology knowledge question-answering applications.

Fine-tuning a LLM with a balanced combination of specialized and general data can enhance its domain-specific expertise while preserving general language capabilities. Fine-tuning solely on specialized dataset often degrades general abilities, necessitating strategies to restore them without compromising specialization. While combining specialized and general data for fine-tuning appears to address this, simply merging the datasets offers only marginal gains in specialization and weakens general abilities. A two-stage fine-tuning approach-first on specialized dataset, followed by fine-tuning on a mixed dataset—has proven more effective. This method restores general abilities while retaining improved specialization, as corroborated by findings from Cook et al., (2024) and Dong et al., (2023). Similarly, Junying Chen et al. (2023) demonstrated the importance of proper fine-tuning order and data composition in developing the HuatuoGPT-II model for traditional Chinese medicine. Furthermore, Tinn et al. (2023) highlighted that finetuning strategies, such as freezing layers, can stabilize baseline models and improve outcomes in domain-specific applications, underscoring the need for task-specific approaches tailored to model size and

#### objectives.

Additionally, the prescription experiments demonstrated CDIP-ChatGLM3 provided superior specific and actionable suggestions over other models. This further confirmed the efficacy of the fine-tuning process, in agreement with the quantitative metric comparisons. Meanwhile, it underscores the importance of targeted training and specialized dataset. In contrast, the comparative models offered more comprehensive and extended responses than CDIP-ChatGLM3, demonstrating superior general language abilities. One possible reason is that the relatively smaller model size of CDIP-ChatGLM3 results in diminished general language abilities compared to LLMs with larger parameter counts and a more extensive general knowledge base (Zhao et al., 2023).

Therefore, it can be concluded that, both the fine-tuning on mixed expertise and general dataset, and the inherent language abilities, which are largely dependent on the model size, jointly determine the final specialized performance.

## 4.3. Integration of identification and prescription for crop disease diagnosis

The integration of a high efficiency deep learning vision model for disease identification and fine-tuned language model for disease prevention and control knowledge consultation, namely CDIP-ChatGLM3, is developed to provide practitioners with a practical and flexible tool for crop protection. This integration enables the rapid and accurate identification of crop disease and the followed timely and professional prescriptions as well as the interactive and instant communication.

Similar to our research, Qing et al. (2023) combined identification model YOLOv8 and GPT4, aiming to address the challenge of accurately analyzing agricultural pests and diseases images and providing diagnostic reports. However, only two crops and a few diseases were involved in their research. Further, they just used the general GPT4 and did not fine-tune it with sufficient specialized material, resulting in the LLM's inability to provide professional responses to specific crop diseases. Therefore, their system might not satisfy the requirements of the practical use of farmers.

While universal multimodal language models exhibit exceptional performance on general-purpose tasks and demonstrate high versatility, our comparative experiments revealed that their performance on domain-specific tasks as crop disease identification and prescription, remains suboptimal despite a substantially larger model size (Table 6). This finding is consistent with Kerdvibulvech (2025), who demonstrated that the zero-shot multimodal language models excel at simple, general tasks, but suffer from more complex, domain-specific tasks or those with limited training samples. This discrepancy is likely due to the reliance on generalized training data, resulting in a deficiency of domain-specific examples. Furthermore, to enhance their versatility across a broad range of tasks, a trade-off is made, resulting in compromised performance on domain-specific tasks. In contrast, fine-tuned on expertise dataset, CDIP-ChatGLM3 demonstrated high performance in crop disease and prescription tasks.

Additionally, for specific-domain multimodal models, Lu et al. (2024) employed multimodal transformers for agricultural disease detection and question-answering systems. Despite the high performance they achieved, the crops and disease covered in their study are relatively limited and the corpus may not be as comprehensive as ours. Comparatively, based on extensive crops and the corresponding diseases images as well as the fine-tuning on the comprehensive high-quality corpus, our CDIP-ChatGLM3 achieved superior performance on not only disease identification but the prevention and control prescriptions. Liu et al. (2025) also fine-tuned another multimodal model to address the crop diagnosis problem, achieving a high accuracy on both disease identification and question answering. Although more images encompassing more crops and diseases (16 crops, 60 diseases) were employed for training, they achieved lower identification accuracy than ours,

possibly due to the challenge from complexity of information fusion between modalities and the huge data required to train multimodal models. By contrast, our study fine-tuned the identification model and the language model individually on relatively smaller dataset, and then integrated them, gaining a superior performance on both identification accuracy and accurate response. While multimodal models have shown promise in combining image and text data, the computer vision + finetuned LLM approach provides a more specialized, interpretable, scalable, and resource-efficient solution for crop disease identification and prescription. By leveraging the strengths of separate models tailored for each modality, this approach is better suited for real-world agricultural applications, offering superior performance, flexibility, and practicality.

#### 4.4. Deployment and potential usability assessments of CDIP-ChatGLM3

With the smallest size and a lower inference memory consumption compared to the most comparative multimodal models, our model demands considerably fewer computational resources than most of these multimodal LLMs, thereby enhancing the feasibility for deployment on consumer-grade GPUs. Specifically, CDIP-ChatGLM3 requires only around 13 GB of VRAM at FP16 precision for inference, enabling its operation on consumer-grade GPUs and supporting rapid API-based model inference for potential edge or mobile applications. Through additional quantization (e.g., 8-bit or 4-bit) or pruning, the model's memory footprint can be reduced to 6 GB, 3 GB, or even lower, thereby expanding the range of devices capable of hosting it. To validate this feasibility, the model was deployed on a server equipped with a P40 GPU featuring 20 GB of VRAM and 32 GB of system memory. Under these relatively moderate hardware specifications, which is lower than those typically required by most multimodal LLMs, CDIP-ChatGLM3 exhibits stable performance and high accuracy.

While CDIP-ChatGLM3 delivers actionable, domain-specific disease prevention recommendations, it is essential to continuously collect and integrate feedback from end-users—such as farmers, agricultural extension officers, and agronomy researchers—to conduct an objective, comprehensive evaluation of the model under real-world farming conditions. To support this process, we have developed a mobile application that is freely available for download (https://www.united-smartag. com/team/#/servicedetail), thereby facilitating practical access to CDIP-ChatGLM3. With sufficient user data, systematic usability assessments will be performed to comprehensively delineate the model's strengths and identify areas for improvement.

#### 4.5. Limitations and future work

Despite promising results demonstrated in disease recognition accuracy and provision of prevention and control recommendations, several limitations persist. Firstly, although our leaf disease dataset comprises 48 disease types, potential misclassification could happen when confronted with scenarios outside the dataset or in complex realworld production environments. Secondly, due to the limitations of the large language model adopted and the acquired materials, the details for inquiries from farmers may still be provided with some possible errors or flaws occasionally, resulting in a perception of inadequacy. Finally, although the dual-model framework achieves high accuracy on both disease identification and prescription with a high efficiency and facilitates the upgrading of individual models and continuous service provision, the interaction between the vision and language modalities is not fully exploited due to the independent operation of the two models.

To address these issues, future research will focus on three key areas. The vison model will be continuously updated by incorporating the newly collected disease leaf images into the training process, enhancing our identification models. Meanwhile, the large language model should undergo ongoing fine-tuning and iterations using updated datasets related to disease prevention and control, ensuring that the system remains adaptable to evolving production environments. To address the interaction issue between the two modalities, three key strategies will be implemented. First, a vision encoder will be introduced to extract image features, which will then be integrated with text input to the language model, thereby improving the model's capacity for multimodal information utilization. Second, coupled training of vision model and language model will replace the current two steps independent training process, such as end-to-end co-training. This strategy will jointly optimize the vision model and the language model, leading to enhanced overall performance. Finally, an attention mechanism will be incorporated into the vision model to prioritize critical image information, enabling the language model to effectively integrate both image and text data for generating prescription, rather than focusing solely on the keywords of the identification results.

#### 5. Conclusions

In this study, we introduced a dual-model approach that delivered both accurate disease identification and timely, interactive, and precise prescription recommendations. This approach integrates an advanced deep learning vision model with a specialized Chinese language model that had been fine-tuned on high-quality expert materials. Three primary conclusions could be drawn. Firstly, the EfficientNet-B2 based disease identification model we built outperformed all other models, achieving an accuracy of 97.97  $\% \pm 0.16$ %, a recall of 97.17  $\% \pm 0.46$ %, an F1-score of 97.12  $\% \pm 0.10$ %, and a precision of 97.18  $\% \pm 0.59$ % at a 95 % confidence level.

Secondly, the fine-tuned model CDIP-ChatGLM3 using Freeze methods with DMT training strategy, evaluated on different combinations of specialized and general datasets on crop disease prevention and control corpus datasets, outperforms other models on both specialized and general abilities, achieving substantially higher performance in terms of BLEU-4, ROUGE, and CMMLU accuracy metrics, as well as a low error rate.

Finally, the CDIP-ChatGLM3 achieved timely and interactive accurate disease prevention and control with the disease images as input, which substantially facilitates the farmers and ensures secure crop production.

In the future, we will develop identification models with even higher efficiency and accuracy as well as large language models with higher expertise and general abilities to continually update our CDIP-ChatGLM3 model for practical use of farmers.

#### Data avai lability

The data and code used in this study are available on GitHub (https://github.com/SmartAG-Team/CDIP-ChatGLM3).

#### CRediT authorship contribution statement

Changqing Yan: Writing - review & editing, Writing - original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. Zeyun Liang: Writing - review & editing, Writing - original draft, Visualization, Validation, Software, Methodology, Data curation. Han Cheng: Writing - review & editing, Writing - original draft, Software, Methodology, Conceptualization. Shuyang Li: Writing - review & editing, Writing - original draft, Validation, Software, Resources. Guangpeng Yang: Writing - review & editing, Visualization, Validation, Data curation. Zhiwei Li: Writing – review & editing, Resources, Investigation, Data curation. Ling Yin: Writing - review & editing, Project administration, Investigation. Junjie Qu: Writing - review & editing, Supervision, Project administration, Investigation, Formal analysis. Jing Wang: Writing - review & editing, Methodology, Conceptualization. Genghong Wu: Writing - review & editing, Formal analysis. Qi Tian: Writing - review & editing, Resources. Qiang Yu: Writing - review & editing, Project administration, Funding acquisition. Gang Zhao: Writing - review & editing, Writing - original draft, Validation, Supervision, Resources, Project administration, Methodology,

Funding acquisition, Formal analysis, Conceptualization.

#### Funding

This work was supported by the Shaanxi Key R&D Program Project (grant no. 2023-ZDLNY-64).

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.110442.

#### Data availability

Data will be made available on request.

#### References

- Abade, A., Ferreira, P.A., de Barros Vidal, F., 2021. Plant diseases recognition on images using convolutional neural networks: a systematic review. Comput. Electron. Agric. 185, 106125. https://doi.org/10.1016/j.compag.2021.106125.
- Acharya, R., 2020. Corn Leaf Infection Dataset, Version 1.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., others, 2023. GPT-4 Technical Report. arXiv. 2303.08774. doi: 10.48550/arXiv.2303.08774.
- Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., Daszak, P., 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol. Evol. 19, 535–544. https://doi.org/ 10.1016/j.tree.2004.07.021.
- Anthony, Q., Biderman, S., Schoelkopf, H., 2023. Transformer Math 101.
- Atila, Ü., Uçar, M., Akyol, K., Uçar, E., 2021. Plant leaf disease classification using EfficientNet deep learning model. Ecol. Inform. 61, 101182. https://doi.org/ 10.1016/j.ecoinf.2020.101182.
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al., 2023. Qwen Technical Report arXiv. 2309.16609. https://doi.org/10.48550/ arXiv.2309.16609.
- Borhani, Y., Khoramdel, J., Najafi, E., 2022. A deep learning based approach for automated plant disease classification using vision transformer. Sci. Rep. 12, 11554. https://doi.org/10.1038/s41598-022-15163-0.
- Brasier, C.M., 2008. The biosecurity threat to the UK and global environment from international trade in plants. Plant Pathol. 57, 792–808. https://doi.org/10.1111/ j.1365-3059.2008.01886.x.
- Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K.A., Giovani, B., Saunders, D.G., Kamoun, S., Legg, J.P., Verdier, V., Lessel, J., 2019. A global surveillance system for crop diseases. Science 364, 1237–1239. https://doi.org/10.1126/science.aaw1572.
- Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C.-H., Chan, S.-H.-G., 2023a. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12021–12031. https://doi.org/10.1109/CVPR52729.2023.01157.
- Chen, Junying, Wang, X., Ji, K., Gao, A., Jiang, F., Chen, S., Zhang, H., Song, D., Xie, W., Kong, C., others, 2023. HuatuoGPT-II: One-Stage Training for Medical Adaption of LLMs. arXiv. 2311.09774. doi: 10.48550/arXiv.2311.09774.
- Chen, S., Kann, B.H., Foote, M.B., Aerts, H.J., Savova, G.K., Mak, R.H., Bitterman, D.S., 2023c. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol. 9, 1459–1462. https://doi.org/10.1001/jamaoncol.2023.2954.
- Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., others, 2024. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv. 2412.05271. doi: 10.48550/ arXiv.2412.05271.
- Cook, A., Karakuş, O., 2024. LLM-commentator: novel fine-tuning strategies of large language models for automatic commentary generation using football event data. Knowl. -Based Svst. 300, 112219. https://doi.org/10.1016/j.knosvs.2024.112219.
- Cruz, A., Ampatzidis, Y., Pierro, R., Materazzi, A., Panattoni, A., De Bellis, L., Luvisi, A., 2019. Detection of grapevine yellows symptoms in Vitis vinifera L with artificial intelligence. Comput. Electron. Agric. 157, 63–76. https://doi.org/10.1016/j. compag.2018.12.028.
- Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2024. QLoRA: efficient finetuning of quantized LLMs. Adv. Neural Inf. Process. Syst. 36. https://dl.acm.org/ doi/10.5555/3666122.3666563.
- Devi, P.I., Manjula, M., Bhavani, R., 2022. Agrochemicals, environment, and human health. Annu. Rev. Environ. Resour. 47, 399–421. https://doi.org/10.1146/annurevenviron-120920-111015.

Dong, G., Yuan, H., Lu, K., Li, C., Xue, M., Liu, D., Wang, W., Yuan, Z., Zhou, C., Zhou, J., 2023. How Abilities in Large Language Models are Affected by Supervised Finetuning Data Composition. arXiv. 2310.05492. doi: 10.48550/arXiv.2310.05492.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., others, 2024. The LLaMA 3 Herd of Models. arXiv. 2407.21783. doi: 10.48550/arXiv.2407.21783.

Fang, Y., Ramasamy, R.P., 2015. Current and prospective methods for plant disease detection. Biosensors 5, 537–561. https://doi.org/10.3390/bios5030537.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., others, 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv. 2406.12793. doi: 10.48550/arXiv.2406.12793.

Gupta, A., Shirgaonkar, A., Balaguer, A. de L., Silva, B., Holstein, D., Li, D., Marsman, J., Nunes, L.O., Rouzbahman, M., Sharp, M., others, 2024. RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv. 2401.08406. doi: 10.48550/arXiv.2401.08406.

Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C., 2020. REMIND Your Neural Network to Prevent Catastrophic Forgetting, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 466–483. doi: 10.1007/978-3-030-58598-3\_28.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-Efficient Transfer Learning for NLP, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, pp. 2790–2799.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q., 2019. Searching for MobileNetV3. In: In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv. 2106.09685. doi: 10.48550/arXiv.2106.09685.

Jeong, C., 2024. Fine-tuning and Utilization Methods of Domain-specific LLMs. arXiv. 2401.02981. doi: 10.13088/jiis.2024.30.1.093.

Jingze Feng, Xiaofei Chao, 2022. Apple Tree Leaf Disease Segmentation Dataset. doi: 10.11922/sciencedb.01627.

Kerdvibulvech, C., 2025. Multimodal AI Model for Zero-Shot Vehicle Brand Identification. Multimedia Tools and Applications. 1–20. https://doi.org/10.1007/ s11042-024-20559-3.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Others, 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 3521–3526. https://doi.org/10.1073/pnas.1611835114.

Levine, Y., Dalmedigos, I., Ram, O., Zeldes, Y., Jannai, D., Muhlgay, D., Osin, Y., Lieber, O., Lenz, B., Shalev-Shwartz, S., others, 2022. Standing on the Shoulders of Giant Frozen Language Models. arXiv. 2204.10019. doi: 10.48550/arXiv.2204.10019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inform. Process. Syst. 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401.

Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C., 2024. LLaVA-NeXTinterleave: tackling multi-image, video, and 3D in large multimodal models. arXiv. 2407.07895. doi: 10.48550/arXiv.2407.07895.

Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., Baldwin, T., 2023. CMMLU: measuring massive multitask language understanding in Chinese. arXiv. 2306.09212. doi: 10.48550/arXiv.2306.09212.

Lin, C.-Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries, in: Annual Meeting of the Association for Computational Linguistics. https://api. semanticscholar.org/CorpusID:964287.

Liu, X., Liu, Z., Hu, H., Chen, Z., Wang, Kohou, Wang, Kai, Lian, S., 2025. A multimodal benchmark dataset and model for crop disease diagnosis, in: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp. 157–170. doi: 10.1007/978-3-031-73016-0 10.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.

Lu, Y., Lu, X., Zheng, L., Sun, M., Chen, S., Chen, B., Wang, T., Yang, J., Lv, C., 2024. Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems. Plants 13, 972. https://doi.org/10.3390/ plants13070972.

Miller, S.A., Beed, F.D., Harmon, C.L., 2009. Plant disease diagnostic capabilities and networks. Annu. Rev. Phytopathol. 47, 15–38. https://doi.org/10.1146/annurevphyto-080508-081743.

Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Front. Plant Sci. 7, 1419. https://doi.org/10.3389/ fpls.2016.01419.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: A Method for Automatic Evaluation of Machine Translation, in: Isabelle, P., Charniak, E., Lin, D. (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. doi: 10.3115/1073083.1073135. Pathak, V.M., Verma, V.K., Rawat, B.S., Kaur, B., Babu, N., Sharma, A., Dewali, S., Yadav, M., Kumari, R., Singh, S., others, 2022. Current status of pesticide effects on environment, human health and its eco-friendly management as bioremediation: a comprehensive review. Front. Microbiol. 13, 962619. https://doi.org/10.3389/ fmicb.2022.962619.

Qing, J., Deng, X., Lan, Y., Li, Z., 2023. GPT-aided diagnosis on agricultural image based on a new light YOLOPC. Comput. Electron. Agric. 213, 108168. https://doi.org/ 10.1016/j.compag.2023.108168.

Rahman, C.R., Arko, P.S., Ali, M.E., Khan, M.A.I., Apon, S.H., Nowrin, F., Wasif, A., 2020. Identification and recognition of rice diseases and pests using convolutional neural networks. Biosyst. Eng. 194, 112–120. https://doi.org/10.1016/j. biosystemseng.2020.03.020.

Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S., 2024. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access 12, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742.

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N. V., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Others, 2021. The persistent threat of emerging plant disease pandemics to global food security. Proc. Natl. Acad. Sci. 118, e2022239118. https://doi.org/10.1073/pnas.2022239118.

Roy, A.M., Bhaduri, J., 2021. A deep learning enabled multi-class plant disease detection model based on computer vision. AI. 2, 413–428. https://doi.org/10.3390/ ai2030026.

Saleem, M.H., Potgieter, J., Arif, K.M., 2019. Plant disease detection and classification by deep learning. Plants. 8, 468. https://doi.org/10.3390/plants8110468.

Shoaib, M., Shah, B., EI-Sappagh, S., B., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., Ali, F., 2023. An Advanced deep learning models-based plant disease detection: a review of recent research. Front. Plant Sci. 14. https://doi.org/10.3389/ fpls.2023.1158933.

Silva, B., Nunes, L., Estevão, R., Aski, V., Chandra, R., 2023. GPT-4 as an Agronomist Assistant? Answering Agriculture Exams Using Large Language Models. arXiv. 2310.06225. doi: 10.48550/arXiv.2310.06225.

Spence, N., Hill, L., Morris, J., 2019. How the global threat of pests and diseases impacts plants, people, and the planet. PLANTS, PEOPLE, PLANET 2, 5–13. https://doi.org/ 10.1002/ppp3.10088.

Strange, R.N., Scott, P.R., 2005. Plant disease: a threat to global food security. Annu. Rev. Phytopathol. 43, 83–116. https://doi.org/10.1146/annurev. phyto.43.113004.133839.

Swenor, A., 2022. Using random perturbations to mitigate adversarial attacks on NLP models. Proceedings of the AAAI Conference on Artificial Intelligence. 36, 13142–13143. https://doi.org/10.1609/aaai.v36i11.21707.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Presented at the International Conference on Machine Learning, pp. 6105–6114. https://doi.org/10.48550/arXiv.1905.11946.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T. B., 2023. Stanford Alpaca: An Instruction-following LLaMA model. GitHub repository.

Tariq, A., Luo, M., Urooj, A., Das, A., Jeong, J., Trivedi, S., Patel, B., Banerjee, I., 2024. Domain-specific LLM Development and Evaluation – A Case-study for Prostate Cancer. medRxiv. 2024.03.15.24304362. doi: 10.1101/2024.03.15.24304362.

Thapa, R., Zhang, K., Snavely, N., Belongie, S., Khan, A., 2020. The plant pathology challenge 2020 data set to classify foliar disease of apples. Appl. Plant Sci. 8, e11390. https://doi.org/10.1002/aps3.11390.

Tian, Q., Zhao, G., Yan, C., Yao, L., Qu, J., Yin, L., Feng, H., Yao, N., Yu, Q., 2024. Enhancing practicality of deep learning for crop disease identification under field conditions: insights from model evaluation and crop-specific approaches. Pest Management Science. 80, 5864–5875. https://doi.org/10.1002/0s.8317.

Management Science. 80, 5864–5875. https://doi.org/10.1002/ps.8317.
 Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2023.
 Fine-tuning large neural language models for biomedical natural language processing. Patterns. 4, 100729. https://doi.org/10.1016/j.patter.2023.100729.

Too, E.C., Yujian, L., Njuki, S., Yingchun, L., 2019. A comparative study of fine-tuning deep learning models for plant disease identification. Comput. Electron. Agric. 161, 272–279. https://doi.org/10.1016/j.compag.2018.03.032.

Tudi, M., Daniel Ruan, H., Wang, L., Lyu, J., Sadler, R., Connell, D., Chu, C., Phung, D.T., 2021. Agriculture development, pesticide application and its impact on the environment. Int. J. Environ. Res. Public Health 18, 1112. https://doi.org/10.3390/ ijerph18031112.

Wang, A., Liu, C., Yang, J., Weng, C., 2024. Fine-tuning large language models for rare disease concept normalization. J. Am. Med. Inform. Assoc. 31, 2076–2083. https:// doi.org/10.1093/jamia/ocae133.

Wang, H., Wu, H., Zhu, H., Miao, Y., Wang, Q., Qiao, S., Zhao, H., Chen, C., Zhang, J., 2022. A residual LSTM and Seq2Seq neural network based on GPT for Chinese ricerelated question and answer system. Agriculture. 12, 813. https://doi.org/10.3390/ agriculture12060813.

Wang, T., Liu, J., Wang, N., Cui, Y., 2023. Agricultural technology knowledge intelligent question-answering system based on large language model. Smart Agriculture 5, 105. https://doi.org/10.12133/j.smartag.SA202311005.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., others, 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv. 2412.10302. doi: 10.48550/ arXiv.2412.10302.

Xu, M., Park, J.-E., Lee, J., Yang, J., Yoon, S., 2024a. Plant disease recognition datasets in the age of deep learning: challenges and opportunities. Front. Plant Sci. 15, 1452551. https://doi.org/10.3389/fpls.2024.1452551.

Xu, S., Pang, L., Yu, M., Meng, F., Shen, H., Cheng, X., Zhou, J., 2024. Unsupervised Information Refinement Training of Large Language Models for RetrievalAugmented Generation, in: Ku, L.-W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp. 133–145. doi: 10.18653/v1/2024.acl-long.9.

- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., others, 2024. Qwen2.5 Technical Report. arXiv. 2412.15115. doi: 10.48550/ arXiv.2412.15115.
- Yang, X., Gao, J., Xue, W., Alexandersson, E., 2024. PLLaMa: An Open-source Large Language Model for Plant Science. arXiv. 2401.01600. doi: 10.48550/ arXiv.2401.01600.
- Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., others, 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv. 2408.01800. doi: 10.48550/arXiv.2408.01800.

- Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., Srivastava, G., Maddikunta, P.K.R., Raj, G.D., Jhaveri, R.H., Prabadevi, B., Wang, W., others, 2024. GPT (generative pretrained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access. 12, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497.
- Yuan Yuan, Lei Chen, 2023. An Image Dataset for IDADP-Grape Disease Identification. doi: 10.11922/sciencedb.j00001.00311.
- Zhang, F., Jin, X., Lin, G., Jiang, J., Wang, M., An, S., Hu, J., Lyu, Q., 2024. Hybrid attention network for citrus disease identification. Comput. Electron. Agric. 220, 108907. https://doi.org/10.1016/j.compag.2024.108907.
- Zhang, H., Gong, Y., Yan, Y., Duan, N., Xu, J., Wang, J., Gong, M., Zhou, M., 2019. Pretraining-based natural language generation for text summarization. arXiv. 1902.09243. doi: 10.48550/arXiv.1902.09243.
- Zhang, X., Yang, Q., 2023. Self-QA: Unsupervised Knowledge Guided Language Model Alignment. arXiv. 2305.11952. doi: 10.48550/arXiv.2305.11952.
- Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., others, 2023. A Survey of Large Language Models. arXiv. 2303.18223. doi: 10.48550/arXiv.2303.18223.
- Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y., 2024. LLaMAFactory: Unified Efficient Fine-Tuning of 100+ Language Models. arXiv. 2403.13372. doi: 10.48550/arXiv.2403.13372.