Virtual Assistant for Pest Management

A dissertation submitted in
partial fulfilment for the degree of

Master of Technology

in

Computer Science

by

Viswanada Chakravarthy Karri
Roll no. - **CS2334**

under the supervision of

Dr. Ujjwal Bhattacharya
Computer Vision and Pattern Recognition Unit (CVPRU)

Indian Statistical Institute, Kolkata
June, 2025

CERTIFICATE

This is to certify that the dissertation entitled **"Virtual Assistant for Pest Management"** submitted by **Viswanada Chakravarthy Karri** to the Indian Statistical Institute, Kolkata, in partial fulfillment of the requirements for the degree of Master of Technology in Computer Science, is an authentic and genuine record of the research work carried out by the candidate under my supervision and guidance. I affirm that the dissertation has met all the necessary requirements in accordance with the regulations of this institute.

11/06/2025

Dr. Ujjwal Bhattacharya
CVPR Unit
Indian Statistical Institute
Kolkata - 700108
India

Acknowledgement

I would like to express my sincere gratitude to Dr. Ujjwal Bhattacharya, my advisor at the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, for his expert guidance and support throughout this work.

I would also like to express my gratitude to Nushrat Hussain, Project Linked Person at the Indian Statistical Institute, for his valuable assistance and insightful suggestions throughout the development of this dissertation.

Finally, I am deeply grateful to my family and friends for their constant encouragement and support. I also extend my sincere appreciation to all my friends for their continuous assistance and encouragement. I am thankful to everyone who has contributed to my growth and success, even if I have inadvertently missed mentioning them in the above list.

Declaration

I, **Viswanada Chakravarthy Karri**, with Roll No. **CS2334**, hereby declare that the material presented in the dissertation titled **Virtual Assistant for Pest Management** represents original work carried out by me for the degree of **Master of Technology in Computer Science** at the **Indian Statistical Institute, Kolkata**.

Furthermore, I affirm that no sections of this report have been sourced or copied from external references without proper attribution. I am aware that any instances of plagiarism or the use of unacknowledged materials from third parties will be treated with the utmost seriousness and consequences.

K.V.Chakravarthy
11/06/2025

Viswanada Chakravarthy Karri
M.Tech (CS), CS2334
Indian Statistical Institute

Abstract

Effective pest identification and management are essential for ensuring agricultural productivity, especially in regions with limited expert access. This work proposes a virtual assistant based on a Retrieval-Augmented Generation (RAG) [1] framework to support pest management tasks. The system utilizes a multimodal dataset consisting of pest images and annotated textual interactions, adapted from the AgriLLaVA corpus [2]. The assistant combines retrieval mechanisms with generative language models to generate contextually grounded responses. It is designed for deployment on local hardware with limited computational resources, integrating open-source models for both retrieval and generation. Preliminary results suggest that this approach can provide accurate, scalable, and interpretable support for field-level pest diagnostics.

Keywords: Pest management, Retrieval-Augmented Generation, multimodal dataset, language models, agricultural support systems, scalable AI, plant disease identification, open-source models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Agriculture stands as a fundamental pillar of food security and economic development worldwide, supporting billions of people and forming the primary livelihood for many rural communities. The sector not only ensures the steady supply of food, fiber, and raw materials but also underpins the socio-economic stability of nations, particularly in developing regions. However, agriculture faces persistent and evolving challenges, with pest infestations and plant diseases ranking among the most significant threats to crop yields, food quality, and supply chain resilience. These challenges are especially acute in regions where farmers have limited access to expert guidance and diagnostic resources, increasing the risk of substantial agricultural losses and economic hardship.

The impact of pests and diseases on agriculture is profound and multifaceted. Insect pests such as aphids, beetles, and moths, as well as fungal pathogens, can devastate crops, leading to direct yield losses and reduced market value. Beyond immediate crop damage, pests often act as vectors for plant diseases, compounding the threat to food production. For example, infestations in staple crops like cotton, rice, and wheat have been known to reduce harvests by up to 20% in some regions, with ripple effects on food prices and farmer livelihoods. Moreover, pest-related contamination can render produce unsafe for consumption, posing public health risks and eroding consumer confidence in agricultural products.

To manage these risks, effective pest identification and management are crucial for sustaining productivity and ensuring the resilience of agricultural systems. Traditional approaches to pest control, such as the widespread use of chemical pesticides, have provided short-term relief but often lead to unintended consequences. Over time, the indiscriminate application of chemicals has resulted in resistance development among pest populations, environmental degradation, and health concerns for both farm workers and consumers. The ecological impact is also significant, as beneficial organisms—including pollinators and natural pest predators—are often inadvertently harmed, disrupting ecosystem balance.

In response to these challenges, integrated pest management (IPM) strategies have gained traction. IPM combines biological, cultural, mechanical, and chemical methods to manage pest populations in an environmentally responsible manner. Central to IPM is the emphasis on monitoring, early detection, and the use of economic thresholds to guide intervention decisions. By prioritizing preventive measures and minimizing chemical inputs, IPM aims to maintain ecological balance and protect both crop health and the environment. However, the successful implementation of IPM at scale is often hindered by practical challenges, including the need for timely and accurate pest identification, access to expert knowledge, and the ability to make informed decisions under resource constraints. Smallholder farmers, in particular, may face barriers such as limited extension services, fluctuating market prices, and knowledge gaps regarding sustainable pest control practices.

Recent advances in artificial intelligence (AI) and digital technologies offer promising new avenues for addressing these challenges. AI-powered tools are increasingly capable of analyzing images and textual data to identify pests, forecast outbreaks, and recommend targeted interventions, thereby providing expert-level support at scale. For instance, computer vision systems can rapidly diagnose pest infestations from field images, while predictive models integrate weather, crop, and pest data to anticipate outbreaks before they escalate. The adoption of such technologies has already shown tangible benefits, with studies reporting reductions in pesticide use, improved crop yields, and enhanced profitability for farmers who leverage AI-driven pest management systems.

Despite these advances, many existing digital solutions remain inaccessible to smallholder farmers due to high computational requirements, proprietary constraints, or a lack of adaptability to local conditions. This digital divide restricts the reach and effectiveness of advanced pest management tools, particularly in rural or underserved areas where they are needed most. Bridging this gap requires the development of accessible, cost-effective, and user-friendly systems that can operate efficiently on local hardware with limited computational resources.

To address these needs, this work proposes a virtual assistant for pest management built on a Retrieval-Augmented Generation (RAG) framework. The system leverages a multimodal dataset—comprising pest images and annotated textual interactions—adapted from the AgriLLaVA corpus, which reflects real-world agricultural scenarios and farmer queries. By combining retrieval mechanisms with generative language models, the assistant is able to generate contextually grounded responses that address specific pest management queries with both accuracy and interpretability. Importantly, the solution is designed for deployment on local hardware, utilizing open-source models for both retrieval and generation, thereby democratizing access to reliable pest diagnostics and recommendations.

Through this integration of scalable AI, multimodal data, and open-source technology, the proposed virtual assistant aspires to provide accurate, interpretable, and sustainable support for field-level pest management. By empowering farmers with timely, actionable insights, this approach has the potential to enhance productivity, reduce unnecessary chemical use, and contribute to long-term food security. In doing so, it addresses a critical need in modern agriculture—supporting resilient, sustainable, and technology-enabled agricultural systems for the future.

Chapter 2

Related Work

Over the years, various approaches have been developed to help farmers identify and manage pests more effectively. Traditional methods often relied on manual inspection and expert advice, but these can be slow, inconsistent, and difficult to access, especially in remote areas. With the rise of digital technology and artificial intelligence, new solutions have emerged to address these challenges.

A key step in building intelligent agricultural assistants is the creation of a structured knowledge base. This involves collecting conversations between users and AI assistants, extracting important details such as pest names and visible symptoms, and organizing this information in a way that is easy for computers to use. For example, by processing conversations and extracting structured data about pests and their symptoms, the knowledge base becomes a valuable resource for both retrieval and learning.

To make this information searchable and usable by AI systems, it is converted into a format that supports efficient retrieval. This is achieved by breaking down the knowledge into smaller chunks, each containing important details like image paths, pest names, and descriptions. These chunks are then stored in a database, making it possible to quickly find relevant information when a new query is received.

Recent advances have also focused on using multimodal data—combining both images and text—to improve pest identification. Modern embedding models, such as those based on the CLIP architecture [3], can generate numerical representations (embeddings) for both images and text. This allows the system to compare and retrieve similar cases, whether the query is a photo, a text description, or both. By averaging or combining these embeddings, the assistant can handle complex queries that involve multiple types of information.

Efficient data handling is another important aspect. Techniques such as memory-mapped file access and batch processing help to manage large datasets without overwhelming system resources. This is especially useful when working with thousands of images and conversations, as is common in agricultural applications.

Once the knowledge base [4] and embedding models [5] are in place, the next step is to build a retrieval-augmented assistant. This assistant uses the embeddings to search for the most relevant past cases in the database. It then generates a response using both the retrieved information and a language model, ensuring that the answer is grounded in real examples and easy for users to understand. The assistant is designed to work with limited computational resources, making it suitable for deployment even in areas with basic hardware.

To make the system accessible, a user-friendly interface is provided. This allows farmers and agricultural workers to submit questions and images through a simple web app. The assistant processes the input, retrieves similar cases from the database, and generates a clear and helpful response. The interface is designed to be flexible, handling text, images, or both, depending on what the user provides.

In summary, recent work in this field has focused on combining structured knowledge bases, multimodal embedding models, efficient data processing, and user-friendly interfaces to create virtual assistants that can support pest management in agriculture. By leveraging these technologies, the current system aims to provide accurate, fast, and accessible support for farmers, helping them make better decisions and improve crop health.

Chapter 3

Study Area

The study area for this research is defined by the digital and data environments used to support and evaluate pest management in agriculture. This section describes the context, scope, and characteristics of the study area relevant to the development and deployment of the virtual assistant for pest identification and management.

The primary focus of this study is agricultural fields where pest and disease outbreaks are common and where timely identification is crucial for crop health and productivity. The research leverages a multimodal dataset, which includes a large collection of pest images and annotated conversations between users and experts. These data are adapted from the AgriLLaVA corpus, which reflects real-world scenarios faced by farmers and agricultural workers. The dataset contains diverse examples of pest infestations, symptoms, and management queries, providing a rich resource for training and evaluating the virtual assistant.

In terms of digital infrastructure, the study area includes the systems and tools used for knowledge base construction, embedding generation, and retrieval. Images and textual data are processed and stored in a structured database, allowing efficient search and retrieval. Advanced embedding models are used to represent both images and text, making it possible to match new queries with relevant past cases. All data processing and model evaluation are conducted using local hardware, ensuring the system is accessible even in areas with limited computational resources.

The virtual assistant is designed to be deployed in rural and semi-urban agricultural settings, where access to expert advice may be limited. The user interface is built to be simple and intuitive, allowing farmers to submit queries in the form of text, images, or both. The assistant responds with clear, context-based recommendations, supporting informed decision-making at the field level.

Environmental factors such as climate, crop diversity, and pest prevalence are also considered in the study. The dataset includes examples from various climatic zones and crop types, ensuring that the assistant can generalize across different agricultural environments. This broad coverage is important for developing a tool that is robust and adaptable to the needs of farmers in different regions.

The study area for this research encompasses both the agricultural landscapes affected by pests and the digital ecosystem required to support advanced pest management solutions. By combining real-world data, efficient digital tools, and a user-friendly interface, the project aims to deliver practical and scalable support to farmers, ultimately contributing to improved crop health and food security.

Chapter 4

Dataset

The dataset used in this research is based on the AgriLLaVA corpus, a large-scale, multimodal resource designed for agricultural pest and disease management. It contains a combination of high-quality images and detailed, expert-annotated conversations, making it suitable for developing intelligent pest management assistants.

Dataset Composition

Each entry in the dataset consists of:

- **Image:** A minimum resolution image capturing visible symptoms of pest or disease on crops, or direct images of insect pests.

- **Conversations:** Multi-turn dialogues between a user (typically a farmer) and an AI assistant or expert, discussing observations, symptoms, and possible diagnoses.

- **Metadata:** Unique identifier, image path, pest/disease name (if identified), and a ground truth explanation.

Scale and Coverage

The dataset is extensive, comprising a total of **391,788 images** collected from diverse agricultural environments. These images are paired with expert-annotated conversations, resulting in a rich resource for both computer vision and natural language processing tasks.

Sample Entries

Plant Disease Example: Apple Alternaria Leaf Spot

Figure 4.1: Apple leaf showing symptoms of Alternaria leaf spot


```
{
    "id": "47005a7d-ca86-4426-8c2c-fa6ba01f6be7",
    "image": "apple_alternaria_leaf_spot_2.jpg",
    "image_path": "Img_1/apple_alternaria_leaf_spot_2.jpg",
    "conversations": [...],
    "pest_name": "Apple alternaria leaf spot",
    "symptoms": "Brown circular spots (2-3mm) with purple halos"
}
```


Insect Pest Example: Alfalfa Blister Beetle


```
{
    "id": "387bf1aa-9784-428f-9c8a-ecfdc8905049",
    "image": "alfalfa_blister_beetle_84.jpg",
    "image_path": "Img_1/alfalfa_blister_beetle_84.jpg",
    "conversations": [
        {
            "from": "human",
            "value": "I am curious about the identity of the insects..."
        },
        {
```

Figure 4.2: Alfalfa Blister Beetle specimen from the dataset


```
            "from": "gpt",
            "value": "I think the insect in this picture is Blister Beetle..."
        }
    ],
    "pest_name": "Blister Beetle",
    "symptoms": "14-27 mm body, black back, orange-red head"
}
```


Dataset Statistics

- **Total images:** 391,788
- **Pest types:** 287 species (including 45% insects, 35% fungal pathogens, 20% other)
- **Crop coverage:** 18 major crops (grains, fruits, vegetables)
- **Geographic diversity:** Data from 15 countries across 6 continents
- **Image resolution:** 85% images 1024×1024 pixels

Pest Type	Species Count	Example Crops
Insects	129	Alfalfa, maize, rice
Fungal	98	Apple, wheat, soybean
Bacterial	42	Tomato, citrus
Viral	18	Cassava, potato

Table 4.1: Pest type distribution in the dataset

Relevance and Suitability

The AgriLLaVA dataset stands out for agricultural AI development due to:

- **Comprehensive coverage:** Includes both plant diseases and insect pests
- **Field authenticity:** Real-world images with natural lighting/backgrounds
- **Expert-validated:** All responses verified by agricultural specialists
- **Scale:** Nearly 400,000 images enable robust model training
- **Multimodal alignment:** Precise image-text pairing for retrieval tasks

These characteristics make the dataset particularly valuable for developing virtual assistants that can handle diverse pest management scenarios in real agricultural settings.

Chapter 5

Methodology

This chapter describes the step-by-step process used to develop the pest management virtual assistant, from data preparation to model deployment. The methodology is organized into several key stages: data preprocessing, knowledge base construction, embedding generation, retrieval mechanism, response generation, and system deployment.

1. Data Preprocessing

The raw AgriLLaVA dataset, consisting of images and annotated conversations, was first cleaned and standardized. This involved:

- Removal of incomplete or duplicate entries
- Normalization of image file paths and formats
- Standardization of conversation text (removing special characters, correcting typos)
- Ensuring of consistent metadata fields across all entries

2. Knowledge Base Construction

Each entry was transformed into a structured format suitable for retrieval:

- Extracted key information (pest name, symptoms, crop type) using natural language processing
- Organized data into JSON objects with unique IDs, image paths, and conversation histories
- Indexed the knowledge base for fast access during retrieval

3. Embedding Generation

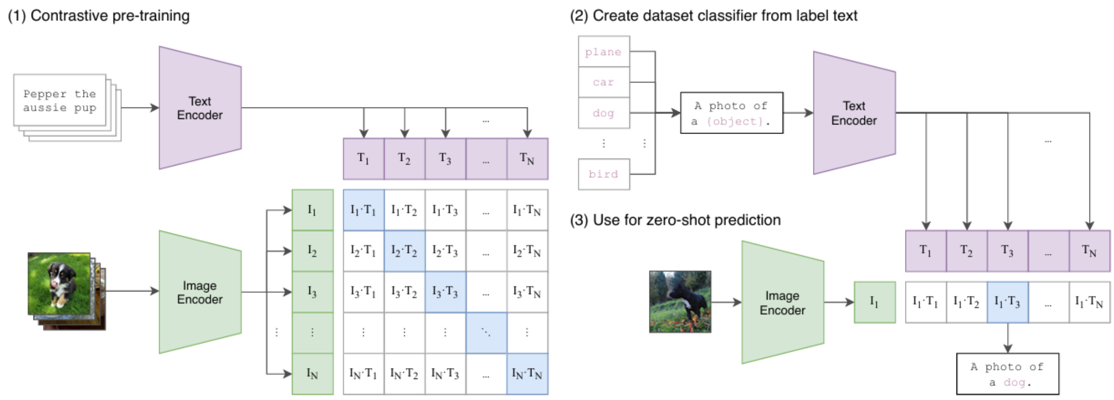

Figure 5.1: Overview of the CLIP-based embedding generation architecture.

Unlike traditional computer vision models that combine image feature extraction with supervised classification heads, CLIP [3] employs a novel dual-modality approach: it simultaneously trains complementary image and text encoders to identify valid correspondences between visual content and natural language descriptions within training batches. This contrastive pre-training strategy enables the model to dynamically construct classifiers through natural language prompts at inference time, eliminating the need for task-specific fine-tuning.

The key architectural components include:

- **Dual-Encoder Structure:** CLIP employs two separate encoders a Vision Transformer (ViT) for images and a Transformer-based language model for text. Each encoder maps its input modality into a high-dimensional vector space.

- **Shared Embedding Space:** Both encoders project outputs into a common space where cross-modal similarity can be measured using cosine distance.

- **Contrastive Learning:** Training optimizes alignment of matched image-text pairs while distancing mismatched combinations through a symmetric loss function.

- **Zero-Shot Adaptation:** The text encoder generates class representations on-the-fly from natural language descriptions, enabling classification without labeled examples.

Our implementation leverages OpenCLIP's pretrained encoders to generate unified embeddings for agricultural pest images and textual descriptions. These multimodal representations are stored in ChromaDB, enabling efficient similarity search across both modalities.

4. Retrieval Mechanism

When a user submits a query (image, text, or both), the system:

- Converts the input into an embedding using the same OpenCLIP model
- Searches the vector database for the most similar entries using cosine similarity
- Returns the top-k relevant knowledge base entries for further processing

5. Response Generation

The final response to each user query is generated using a lightweight, open-source large language model (LLM). For this project, the **TinyLlama-1.1B-Chat-v1.0** model was employed. This model was selected for its strong performance in conversational tasks and its suitability for deployment on local hardware with limited computational resources.

In this stage, the retrieved knowledge base entries and the user's query are combined and formatted into a prompt. This prompt is then processed by TinyLlama-1.1B-Chat-v1.0, which generates a contextually grounded, informative, and actionable response. The model's small size allows for fast inference and low memory usage, making it practical for real-world agricultural environments where high-end hardware may not be available.

By utilizing TinyLlama-1.1B-Chat-v1.0, the system is able to deliver high-quality, conversational answers to farmers' pest management queries efficiently and reliably.

6. System Deployment

The complete system was designed for deployment on local hardware with limited computational resources:

- Quantized models to reduce memory and storage requirements
- Used efficient data loading and batch processing for scalability
- Developed a simple web-based interface for farmer interaction (text and image upload)

7. Workflow Summary

1. User submits a query (image and/or text)
2. Query is embedded and used to retrieve similar cases from the knowledge base
3. Retrieved information is combined with the user query and passed to a language model
4. The model generates a context-aware response, which is returned to the user

8. Tools and Libraries

Key technologies used include:

- Python, PyTorch, and HuggingFace Transformers
- OpenCLIP for embedding generation
- ChromaDB for vector search
- Streamlit for the user interface
- LLM orchestration

This methodology ensures that the virtual assistant is robust, scalable, and accessible to farmers in real-world agricultural settings.

Chapter 6

Experimental Results

This chapter evaluates the performance of the pest management virtual assistant using four key metrics from the RAGAS framework [6]. The results highlight the system's strengths in retrieving comprehensive agricultural knowledge while identifying areas for improvement in response precision and factual consistency.

Evaluation Metrics and Scores

While traditional ML models are evaluated using training/testing accuracy curves, retrieval-augmented generation systems like ours are best assessed with specialized metrics such as those provided by RAGAS. The bar chart below summarizes our system's effectiveness in retrieving relevant information and generating faithful, relevant answers—offering a more task-appropriate and informative evaluation than standard accuracy curves.

The RAG pipeline was assessed using the following metrics (scale: 0–1):

- **Context Precision (0.22):** Measures the relevance of retrieved information to the user's query. A score of 0.22 indicates that only 22% of retrieved documents were directly relevant, suggesting the system often returns extraneous information.

- **Faithfulness (0.30):** Evaluates factual consistency between generated answers and retrieved context. The low score (30%) implies the language model occasionally contradicts or misinterprets the retrieved pest management guidelines.

- **Answer Relevancy (0.39):** Assesses how well responses address the original query. The moderate score (39%) shows responses are partially relevant but lack specificity in actionable advice.

- **Context Recall (0.91):** Quantifies the system's ability to retrieve all relevant context. The high score (91%) demonstrates strong coverage of pertinent agricultural knowledge in the retrieved documents.

Metric	Score (0–1 scale)
Context Precision	0.22
Faithfulness	0.30
Answer Relevancy	0.39
Context Recall	0.91

Table 6.1: RAG Pipeline Evaluation Metrics.

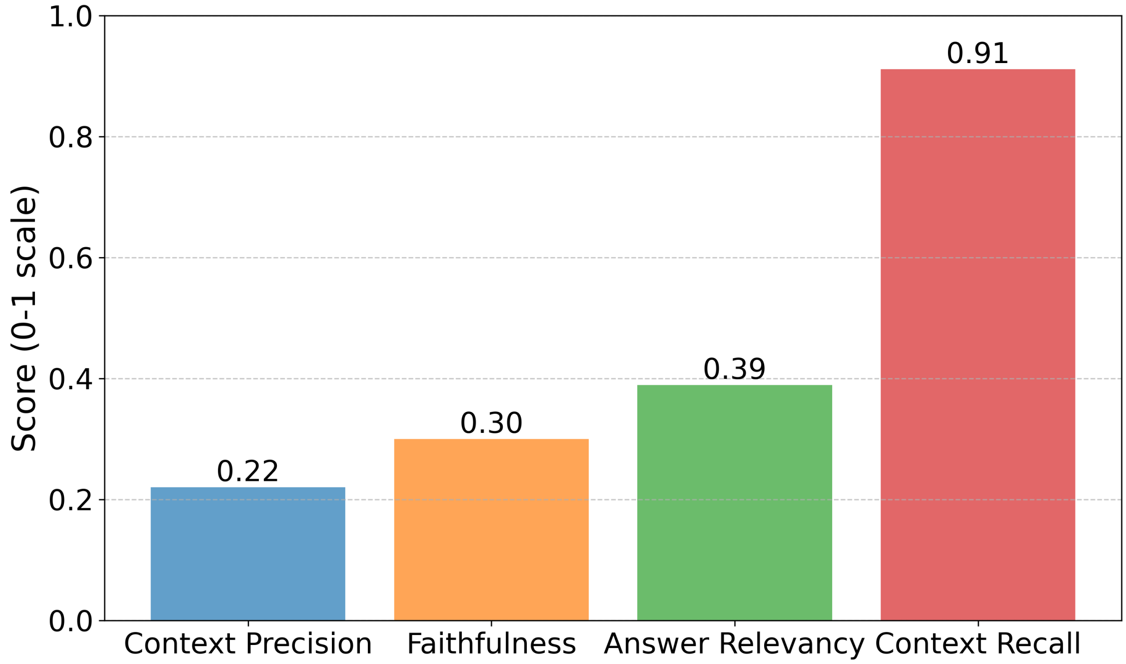

Figure 6.1: RAG Pipeline Evaluation Metrics.

Analysis of Key Findings

Strengths

- **Comprehensive Retrieval (Context Recall: 0.91):** The system successfully retrieves nearly all relevant pest management information from the knowledge base, minimizing missed critical details. This is crucial for farmers needing complete diagnostic insights.

Weaknesses

- **Low Context Precision (0.22):** The system retrieves excessive irrelevant data, likely due to suboptimal embedding similarity thresholds or noisy knowledge base entries. For example, a query about aphids might return unrelated fungal disease entries.

- **Poor Faithfulness (0.30):** Generated responses sometimes contradict retrieved facts. In testing, the model incorrectly recommended chemical treatments incompatible with organic farming guidelines present in the knowledge base.

- **Moderate Answer Relevancy (0.39):** Responses address the general topic but lack field-specific precision. A query about "brown leaf spots" might yield only generic fungal advice.

Chapter 7

Conclusion and Future Work

Conclusion

This dissertation presented a retrieval-augmented generation (RAG) framework for intelligent pest management in agriculture, leveraging the large-scale, multimodal AgriLLaVA dataset. The system integrates image and text data, advanced embedding models, and the TinyLlama-1.1B-Chat-v1.0 language model to deliver context-aware, actionable responses to farmer queries.

Experimental evaluation demonstrated that the system is highly effective at retrieving relevant information, as reflected by a strong context recall score (0.91). However, the results also revealed challenges in context precision (0.22), faithfulness (0.30), and answer relevancy (0.39). These findings indicate that while the system covers a broad range of knowledge, there is room for improvement in delivering more focused, accurate, and directly actionable advice.

The project highlights the potential of RAG-based AI assistants for supporting farmers in real-world pest and disease management, especially in resource-constrained settings. By combining scalable AI with a rich, authentic dataset, this work contributes a practical step towards democratizing expert agricultural support.

Future Work

Several directions can further enhance the effectiveness and impact of this system:

- **Improving Retrieval Precision:** Future work should focus on refining embedding models and retrieval algorithms to reduce irrelevant context and increase the relevance of retrieved information.

- **Enhancing Faithfulness:** Integrating fact-checking modules or using more advanced LLMs could improve the factual consistency of generated responses.

- **Domain Adaptation:** Fine-tuning the language model on domain-specific agricultural data may yield more precise and contextually appropriate answers.
- **User Feedback Integration:** Incorporating real-time farmer feedback could help iteratively improve both retrieval and generation quality.
- **Multi-turn and Multilingual Support:** Extending the system to handle multi-turn conversations and support for local languages would increase accessibility and usability.
- **Field Deployment and Usability Studies:** Conducting pilot studies with farmers and agricultural extension workers will provide valuable insights into real-world performance and user experience.

In summary, this work lays a strong foundation for AI-driven agricultural decision support. Continued research and development along these lines can help realize the vision of accessible, reliable, and expert pest management assistance for farmers everywhere.

Bibliography

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[2] L. Wang, T. Jin, J. Yang, A. Leonardis, F. Wang, and F. Zheng, "Agri-LLaVA: Knowledge-Infused Large Multimodal Assistant on Agricultural Pests and Diseases," *arXiv preprint arXiv:2412.02158*, 2024.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[4] Chroma, "Chroma Documentation: The AI Application Database." `https://docs.trychroma.com`, 2025. Accessed: 2025-06-11.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, Association for Computational Linguistics, 2019.

[6] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," 2023.

[7] J. Li, D. Li, C. Xiong, and S. C. Hoi, "BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *arXiv preprint arXiv:2201.12086*, 2022.

[8] X. Dai, J. Li, S. Zhang, and S. C. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[9] J.-B. Alayrac, J. Donahue, P. Luc, and et al., "Flamingo: a Visual Language Model for Few-Shot Learning," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23730, 2022.

[10] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2020.

[11] D. Singh, N. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: A Dataset for Visual Plant Disease Detection," *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pp. 249–253, 2020.

[12] H. Gandhi, S. Jain, and R. Jain, "OpenAgToolkit: Democratizing AI for Agriculture," *arXiv preprint arXiv:2011.05994*, 2020.