**REPORT**

**ON**

# RAG-Enhanced Global Assistant for Agricultural Pest and Disease Management

*Submitted to*

**DEPARTMENT**

**of**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

| | |
|---|---|
| **Gangula Venkata Raja Vineela** | **245322733143** |
| **Gunti Sai Pranitha** | **245322733148** |
| **Rishit Senapati** | **245322733176** |

**Under the guidance**

**Of**

**Mrs. K. J . Archana**

**Assistant Professor**

**Department of Computer Science and Engineering**

**NEIL GOGTE INSTITUTE OF TECHNOLOGY**

Kachavanisingaram Village, Hyderabad, Telangana 500058.

**2025-2026**

# ACKNOWLEDGEMENT

We are also thankful to our Faculty Supervisor, Mrs. K.J. Archana, for his valuable guidance and encouragement given to us throughout the project work

We are also thankful to Dr. T. Srinivas, Project Coordinator, for supporting us from project selection till the submission to complete this project within the scheduled time.

We are also thankful to Dr. P. Vaishali, Heads of the Department for providing us with time to make this project a success within the given schedule.

We take this opportunity to thank all the people who have rendered their full support to our project work. We render our thanks Prof. R. Shyam Sundar, Principal, who encouraged us to do the Project.

We would like to thank the entire CSE Department faculty, who helped us directly or indirectly in the completion of the project.

We sincerely thank our friends and family for their constant motivation during the project work.

# ABSTRACT

Agricultural productivity is often hindered by crop pests and diseases, which require timely and accurate identification to prevent yield losses. Recent advancements in large multimodal models (LMMs), such as Agri-LLaVA, have shown potential in automating pest and disease detection through image and text-based analysis. However, Agri-LLaVA faces two key challenges: misinformation due to hallucinations in model outputs and limited generalization caused by an imbalanced and insufficient dataset.

To address these issues, we propose an enhanced agricultural assistant that integrates Retrieval-Augmented Generation (RAG) to ground model responses in verified agricultural knowledge bases, thereby reducing misinformation and improving reliability. Additionally, we expand and diversify the dataset by incorporating real-world images of crops, pests, and diseases from multiple sources and environmental conditions. This ensures better representation and coverage of rare cases while improving model accuracy.

Our approach aims to create a robust, scalable, and trustworthy system capable of accurately diagnosing crop diseases and suggesting actionable solutions. Experimental results demonstrate improved performance compared to the baseline Agri-LLaVA model, highlighting the effectiveness of RAG and data enrichment. This work contributes to the development of reliable AI-driven agricultural tools to assist farmers, agricultural experts, and policymakers in disease management and sustainable crop production.

**Key Words**: Agricultural AI, Pest and disease detection, Multimodal model, Retrieval-Augmented Generation (RAG), Dataset expansion, Knowledge grounding, Crop disease diagnosis, Deep learning, Hallucination reduction, Sustainable agriculture, Image-textanalysis, Precision farming, Large Language Model (LLM), Plant health monitoring, Computer vision in agriculture.

# INTRODUCTION

Agriculture is a critical pillar of global food security, yet it remains highly vulnerable to crop pests and diseases. These biotic stresses are responsible for substantial yield losses worldwide, with some crops losing up to 40% of their annual production due to undetected or poorly managed infestations. This challenge is further intensified in regions dominated by smallholder farmers, where limited access to agricultural experts, linguistic diversity, and highly variable environmental conditions hinder timely intervention. Early, accurate, and accessible diagnosis—regardless of language, literacy level, or technological familiarity—is therefore essential for promoting sustainable crop health management.

Recent advances in artificial intelligence, particularly Large Multimodal Models (LMMs), have shown promising potential in automating crop pest and disease recognition through the integration of image understanding and text-based reasoning. Models such as LLaVA and its agricultural variant Agri-LLaVA represent important milestones in this direction. Agri-LLaVA, designed specifically for agriculture, leverages a multimodal dataset of approximately 400,000 image–text pairs combined with instruction tuning to deliver capabilities such as symptom recognition, conversational diagnosis, and visual question answering.

However, despite these advancements, significant challenges persist. Agricultural images often exhibit high variability due to inconsistent lighting, field conditions, occluded leaves, mixed infections, and natural morphological differences. Existing datasets remain scarce, imbalanced, and insufficiently diverse, limiting the ability of models to generalize across real-world scenarios. Furthermore, general-purpose LMMs—including Agri-LLaVA—are prone to hallucinations: confidently producing incorrect or fabricated outputs, especially when agricultural knowledge is incomplete or ambiguous. Such misinformation poses substantial risks in agricultural decision-making, where inaccurate disease identification or improper pesticide recommendations can lead to severe ecological and economic consequences.

A major factor contributing to these hallucinations is the absence of grounding in verified agricultural knowledge. Current systems rely solely on static learned representations without dynamic retrieval from trusted, up-to-date sources, leading to outdated or misleading advice. Additionally, most existing systems are designed primarily for English-speaking, text-literate users. They lack multilingual capabilities and offer no support for voice-based interaction—an essential feature for many farmers who rely on local languages or prefer speech over typed queries. This results in a technological barrier that limits adoption and effectiveness in low-literacy or linguistically diverse agricultural communities.

To address these limitations, this research proposes an enhanced agricultural assistant that integrates Retrieval-Augmented Generation (RAG) with multimodal disease detection, supported by audio-based interaction and multilingual processing. By grounding model outputs in reliable agricultural knowledge bases, RAG significantly reduces hallucinations and improves factual consistency. Simultaneously, dataset expansion through diverse real-world images strengthens the model's generalization under practical field conditions. The inclusion of Speech-to-Text, Text-to-Speech, and multilingual translation modules ensures accessibility for farmers regardless of language or literacy level.

The objective of this work is to develop a more accurate, accessible, and scalable AI-driven system for crop pest and disease diagnosis. The proposed approach not only enhances diagnostic precision through multimodal understanding but also ensures trustworthiness through knowledge grounding and inclusivity through multilingual and audio-enabled interfaces. By combining LMM capabilities, RAG-based factual alignment, expanded agricultural datasets, and user-friendly interaction modalities, this research aims to overcome the key shortcomings of existing systems and contribute meaningfully to the advancement of sustainable, technology-supported agriculture.

# EXISTING SYSTEM

Agri-LLaVA is the first large-scale, agriculture-specific multimodal model designed to recognize pests and diseases through image–text understanding. It is built upon the LLaVA architecture and leverages a large agricultural dataset comprising approximately 400,000 image–text pairs for feature alignment and 6,000 multimodal conversation samples for instruction tuning. This system aims to bridge the gap between general-purpose Large Multimodal Models (LMMs) and the specialized requirements of agricultural diagnostics.

The Agri-LLaVA workflow consists of two major training phases:

**1. Feature Alignment Pretraining (Stage 1)**

- The model is pretrained on ~391,785 agricultural pest and disease images.
- Images are paired with expert-derived knowledge descriptions.
- The visual encoder and LLM are frozen; only the projection layer is trained.
- Objective: align image features with pest/disease categories and symptom descriptions.

This stage enables Agri-LLaVA to learn visual recognition of crop diseases and pests.

**2. Instruction Tuning (Stage 2)**

- The model is fine-tuned on 6,000 high-quality, GPT-4-generated agricultural conversations.
- Visual encoder is frozen; LLM + projection layer are updated.
- Data covers symptoms, pathogens, control methods, transmission, risk factors, etc.
- Objective: enhance multimodal communication and domain-specific reasoning.

This step teaches Agri-LLaVA to engage in conversational diagnosis and answer complex user queries.

**System Capabilities:**

- Identifies crop type from images.
- Classifies pest/disease categories.
- Describes symptoms and causes.
- Provides prevention and treatment suggestions.
- Performs visual reasoning through VQA.

- Conducts multi-turn agricultural dialogue.

Compared to base LLaVA, Agri-LLaVA demonstrates significantly stronger agricultural knowledge, especially in identifying diseases and describing management practices.

## Disadvantages Of Existing System

### Hallucinations and Factual Inconsistency:

The existing system relies solely on internally learned representations without access to real-time authoritative validation. As a result, it may generate *highly confident yet incorrect outputs*, often referred to as hallucinations. For example, the model may correctly localize a leaf lesion but mistakenly attribute it to an incorrect pathogen whose features appear more frequently in the training corpus. Such factual inconsistencies pose significant risks when the system is used for agricultural decision-making, where incorrect diagnoses can lead to crop loss or misuse of pesticides.

### Static and Non-Adaptive Knowledge Base:

The knowledge embedded in the model is fixed at the time of training, making it inherently outdated as soon as new pest variants, emerging diseases, or revised treatment guidelines appear. Updating this knowledge requires extensive retraining, which is resource-intensive and infeasible for rapid adaptation. Consequently, the model lacks responsiveness to evolving agricultural threats and updated best practices.

### Insufficient Generalization to Real-World Conditions:

Although trained on a sizable dataset, the model's performance degrades when confronted with images captured under diverse real-world conditions such as variable lighting, shadows, occluded leaves, different crop growth stages, mixed infections, or environmental noise. This indicates insufficient robustness and limited generalization beyond the curated datasets used for training, reducing its reliability for field-level adoption.

### Lack of Explainability and Source Verification:

The system can provide descriptive answers but cannot cite specific, up-to-date scientific sources or field-verified references for its recommendations. This lack of transparent reasoning undermines user trust and makes it difficult for farmers, agronomists, or policymakers to validate or cross-check the advice provided. Without traceable evidence, the system’s outputs cannot be confidently used in high-stakes agricultural decision-making.

**Dataset Imbalance and Limited Diversity:**

The underlying dataset exhibits imbalance across crop types, diseases, and pest classes. Rare diseases and region-specific infestations are underrepresented, leading the model to overfit to dominant patterns and underperform on minority or unseen cases. Limited environmental diversity in the dataset further constrains the system’s ability to handle real-world variation.

**Absence of Retrieval or Knowledge Grounding:**

The model operates purely on pre-trained parameters without the ability to retrieve updated agronomic knowledge, scientific reports, or validated disease management protocols. This absence of retrieval-augmented reasoning contributes directly to hallucinations and outdated recommendations.

**Limited Support for Complex or Multi-Modal Agricultural Scenarios:**

Current systems mainly operate on single images and text queries. They lack integration with other data modalities such as weather conditions, soil parameters, satellite imagery, or multi-stage plant monitoring, thereby restricting their applicability in comprehensive precision farming workflows.

**Lack of Multimodal and Multilingual Accessibility:**

Existing agricultural multimodal systems such as Agri-LLaVA primarily rely on text and image inputs and often support only a single language, typically English. This limits accessibility for farmers who prefer voice-based interaction or who speak regional languages. The absence of Speech-to-Text (STT), Text-to-Speech (TTS), and multilingual processing features makes these systems unsuitable for rural, low-literacy user groups, thereby restricting their real-world usability and adoption.

# LITERATURE SURVEY

**Research Gap:**

Despite the availability of numerous CNN-based plant disease detection models and general-purpose Large Multimodal Models (LMMs) such as GPT-4V, a significant gap remains in developing an agricultural-specific multimodal system that is both context-aware and factually grounded. Existing models either function as “black-box” classifiers that provide predictions without reasoning or as conversational agents that are prone to hallucinations—particularly when agricultural domain knowledge is sparse or ambiguous. None of these systems sufficiently incorporate Retrieval-Augmented Generation (RAG) to dynamically access verified agricultural knowledge bases, making them vulnerable to misinformation and unsafe recommendations.

Moreover, widely used datasets such as IP102 and PlantVillage lack environmental diversity, covering mostly clean, well-lit, laboratory-style images that do not reflect real-world farming conditions. This limits the generalization capability of current models under variable field environments involving occlusions, mixed infections, irregular lighting, and naturally aged foliage. Additionally, existing systems provide limited support for multilingual interaction or voice-based queries, which restricts accessibility for farmers who rely on local languages or voice input due to low literacy or device limitations.

| Sr No. | Title of Paper | Author | Year | Methodology | Dataset | Research Gap | Performance |
|---|---|---|---|---|---|---|---|
| 1 | Flamingo: A visual language model for few-shot learning | Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. | 2022 | Visual encoder + frozen LM with cross-attention | MultiModal MassiveWeb, ALIGN | Lacks specific fine-tuning for agricultural nuances (pests/diseases). | Strong few-shot reasoning |
| 2 | Qwen-vl: A versatile vision-language model | J Bai, S Bai, S Yang, S Wang, et al. | 2023 | Versatile vision-language understanding backbone | COCO, LVIS | Not optimized for biological taxonomy or symptom recognition. | Robust multi-task ability |
| 3 | Sparks of AGI: Early experiments with gpt-4 | Sebastien Bubeck, Varun Chandrasekaran, et al. | 2023 | Evaluations across reasoning and vision | Proprietary / Not Disclosed | Closed source; hallucinations in specialized scientific fields. | Early AGI behaviors |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **4** | X-llm: Bootstrapping advanced LLMs | Feilong Chen, Minglun Han, Haozhi Zhao, et al. | 2023 | Multimodal pretraining via text-language mapping | Specific Dataset Not Disclosed | Treats modalities as foreign languages; complex to adapt for pest data. | Flexible cross-modality transfer |
| **5** | Allava: Harnessing gpt4v-synthesized data | Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, et al. | 2024 | Synthetic GPT-4V captions for training lightweight models | LAION, Vision FLAN | Relies heavily on synthetic data which may contain hallucinations. | Efficient with small compute |
| **6** | Sharegpt4v: Improving large multi-modal models | Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, et al. | 2023 | Enhanced caption datasets for fine-tuning | COCO, LAION, CC 3M | General captions lack the scientific precision needed for agronomy. | Better caption alignment |
| **7** | PaLM: Scaling language modeling with pathways | Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. | 2023 | Transformer-based, large distributed training | Massive web corpus | Unimodal (text-only focus in base); requires massive compute resources. | Strong reasoning |
| **8** | Instructblip: Towards general-purpose vision-language models | Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng, et al. | 2024 | Instruction-aware Query Transformer | 26 public VL datasets | General instruction following is good, but fails in subtle disease differentiation. | Strong instruction following |
| **9** | Llama-adapter v2: Parameter-efficient visual instruction model | Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, et al. | 2023 | Parameter Efficient visual instruction tuning | COCO Caption | Lightweight but may lack depth for complex biological reasoning. | Easy domain adaptation |
| **10** | Multimodal-gpt: A vision and language model for dialogue | Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, et al. | 2023 | GPT backbone + image encoder | VQA v2.0, OKVQA | Struggles with fine-grained visual details necessary for early pest detection. | Visual-aware chat |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **11** | A probabilistic interpretation of precision recall and f-score | Cyril Goutte and Eric Gaussier | 2005 | Probabilistic evaluation theory | Benchmark IR datasets | Theoretical metric paper; not a model implementation. | Better metric interpretation |
| **12** | Leaf and spike wheat disease detection | Lakshay Goyal, Chandra Mani Sharma, Anupam Singh, et al. | 2021 | Deep CNN for classification with SVM | LWDCD2020 | Classification only; no conversational or explanation capability. | 97.88% accuracy |
| **13** | An open access repository of images on plant health | David Hughes, Marcel Salathe, et al. | 2015 | Data collection and curation | PlantVillage | Dataset is lab-controlled; lacks background complexity of real fields. | Core training source |
| **14** | Lisa: Reasoning segmentation via large language model | Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, et al. | 2023 | Embedding-as-mask to enable segmentation | COCO, ADE20K | Focuses on segmentation, not biological diagnosis or treatment advice. | Strong zero-shot segmentation |
| **15** | Bloom: A 176b-parameter open-access model | Teven Le Scao, Angela Fan, Christopher Akiki, et al. | 2023 | Transformer trained on multilingual corpora | ROOTS corpus | Text-only; requires adaptation for visual inputs. | Competitive performance |
| **16** | Llava-med: Training a large language-and-vision assistant | Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, et al. | 2024 | Fine-tuning on biomedical data (PubMed) | PubMed Vision | Medical focus; proves domain tuning works but not directly usable for crops. | Improved reasoning |
| **17** | Mini-gemini: Mining the potential of multi-modality | Yanwei Li, Yuechen Zhang, Chengyao Wang, et al. | 2024 | Dual vision encoders for high-res embeddings | Synthetic multimodal data | High resolution is good, but requires specific agri-data alignment. | Zero-shot benchmarks |
| **18** | Improved Baselines with Visual Instruction Tuning | Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee | 2024 | LLaVA improved with MLP connector | COCO, VQA | The baseline for Agri-LLaVA, but lacks external knowledge retrieval. | Improved efficiency |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **19** | Visual Instruction Tuning (LLaVA) | Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee | 2024 | GPT-4 generates multimodal instructions | COCO | Hallucinates facts when domain knowledge is missing. | 85.1% on synthetic data |
| **20** | Efficient vision language instruction tuning | Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, et al. | 2024 | MMA connects encoders and LLMs via adapters | COCO, VQAv2 | Focus on efficiency over depth of knowledge. | Low training cost |
| **21** | Macaw-llm: Multi-modal language modeling | Chenyang Lyu, Minghao Wu, Longyue Wang, et al. | 2023 | Cross-modal fusion (Image, Audio, Video) | Multi-modal datasets | Integration of video is promising but complex for simple diagnostic apps. | Multi-sensory understanding |
| **22** | Video-chatgpt: Towards detailed video understanding | Muhammad Maaz, Hanoona Rasheed, Salman Khan, et al. | 2023 | Frame encoder + GPT backbone | VideoQA datasets | Video focused; higher compute latency for real-time field use. | Detailed temporal reasoning |
| **23** | Language models are few-shot learners | Ben Mann, N Ryder, M Subbiah, J Kaplan, et al. | 2020 | Transformer pretraining (GPT-3) | WebText, Books | Text only; outdated compared to GPT-4 based architectures. | Few-shot capabilities |
| **24** | GPT-4 (Vision) System Card | OpenAI | 2023 | Multimodal architecture | Internal OpenAI data | Closed source; API costs are prohibitive for small farmers. | Powerful visual reasoning |
| **25** | Detgpt: Detect what you need via reasoning | Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, et al. | 2023 | Visual reasoning + LLM | Detection datasets | Good for locating pests, but weak on treatment recommendation. | Detects objects by reasoning |
| **26** | Pandagpt: One model to instruction-follow them all | Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, et al. | 2023 | Multimodal instruction alignment | Visual-language data | General-purpose instruction following; lacks specialized agricultural ontology. | Universal instruction model |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **27** | Stanford alpaca: An instruction-following llama model | Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. | 2023 | Fine-tuned LLaMA with self-instruct data | Text instruction data | Text-only model; requires integration with vision encoders for crop diagnosis. | Strong alignment |
| **28** | Open and efficient foundation language models | Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. | 2023 | Optimized transformer training | Web, books, code | Base foundation model; no visual capabilities or pest knowledge. | Strong open model |
| **29** | To see is to believe: Prompting gpt-4v | Junke Wang, Lingchen Meng, Zejia Weng, Bo He, et al. | 2023 | Prompt engineering study for visual tasks | GPT-4V image tasks | Focuses on prompting techniques rather than architectural changes. | Improved visual responses |
| **30** | Cogvlm: Visual expert for pretrained language models | Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, et al. | 2023 | Vision-aware fine-tuning | Image-caption datasets | High computational cost; general visual expert may miss subtle disease patterns. | Strong visual understanding |
| **31** | Visionllm: Large language model is also an open-ended decoder | Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, et al. | 2024 | Unified vision decoder model | Visual benchmarks | Good for segmentation, but lacks the reasoning depth for complex biology. | General visual task solver |
| **32** | Finetuned language models are zero-shot learners | Jason Wei, Maarten Bosma, Vincent Y Zhao, et al. | 2021 | Fine-tuned GPT models via tuning | Text corpora | Zero-shot learning is often insufficient for high-stakes crop identification. | Improves zero-shot learning |
| **33** | Ip102: A large-scale benchmark dataset | Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, et al. | 2019 | Insect pest recognition benchmark | 102 insect classes | A dataset paper, not a solution model; data is lab-based. | Benchmark accuracy >90% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **34** | Multiinstruct: Improving multi-modal zero-shot learning | Zhiyang Xu, Ying Shen, and Lifu Huang | 2022 | Instruction-tuned multi-modal data | COCO, VQA | General domain instruction tuning; lacks "agricultural logic". | Stronger cross-modal zero-shot |
| **35** | mplug-owl: Modularization empowers large language models | Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, et al. | 2023 | Modular plug-in vision encoders | Visual QA datasets | Modular design is efficient but visual encoders lack plant pathology features. | Flexible multi-modal LLM |

# PROPOSED ARCHITECTURE

To overcome the limitations identified in existing agricultural multimodal systems such as Agri-LLaVA, we propose an enhanced architecture integrating Retrieval-Augmented Generation (RAG), a Data Expansion Module, and additional support for audio-based interaction and multilingual processing. These extensions significantly improve factual reliability, accessibility, and real-world usability, enabling the system to serve a broader population of farmers, including those who prefer voice interaction or local languages.

## 1. Data Acquisition and Expansion

A comprehensive dataset is constructed by gathering real-world agricultural images from farms, extension centers, research stations, and publicly available repositories. The images include variations in lighting, crop maturity, weather conditions, occlusions, and background complexity. Each sample is annotated with crop species, disease or pest class, and corresponding symptom descriptions. This expanded dataset augments existing collections such as IP102, improving domain coverage, reducing overfitting, and enhancing the model's generalization to field-level conditions.

## 2. Knowledge Base Construction

A trusted agricultural knowledge repository is developed using validated sources including research publications, government guidelines, pesticide regulatory manuals, and crop-specific diagnostic protocols. These documents are converted into high-dimensional vector embeddings using domain-optimized embedding models and stored in a vector database (e.g., FAISS, Milvus, or Pinecone). This enables efficient similarity search and ensures the availability of up-to-date, scientifically verified information during inference.

## 3. User Input Interface (Image, Text, and Audio)

The system accepts three forms of input from users such as farmers, field technicians, or agricultural advisors:

- Image Input: A photo of the affected plant or leaf.
- Text Input: A typed question regarding visible symptoms.
- Audio Input: A spoken query, processed through a Speech-to-Text (STT) module (e.g., Whisper) to convert audio into text.

This multimodal input capability improves usability, especially in rural or low-literacy settings.

### 4. Multilingual Processing Module

To support diverse language users, the system incorporates a multilingual translation module.

- Incoming text (or transcribed audio) is translated into a standard internal language (e.g., English).
- Outgoing responses are translated back into the user’s preferred language using models such as NLLB, M2M100, or Opus-MT.

This ensures seamless accessibility for speakers of local and regional languages.

### 5. Visual Feature Extraction

The uploaded plant image is processed using a transformer-based vision encoder (e.g., CLIP, SigLIP, or ViT variants). The encoder extracts high-level visual features representing disease patterns, lesion textures, shape irregularities, and pigmentation anomalies. These visual embeddings are used alongside textual embeddings during retrieval and reasoning.

### 6. Retrieval Augmentation (RAG Module)

The translated user query and visual embeddings are fed into the Retrieval-Augmented Generation (RAG) module. The module performs semantic search over the vector knowledge base to identify relevant documents containing disease characteristics, treatment methods, regulatory information, and management protocols. Retrieved results are ranked and returned as contextual evidence to support grounded inference.

### 7. Contextual Fusion and Multimodal Reasoning

The retrieved knowledge snippets are combined with the user’s query and the extracted visual features to form an enriched, context-aware prompt. This augmented input is forwarded to the Large Multimodal Model (LMM), which performs grounded reasoning by aligning image-based insights with validated knowledge. This reduces hallucinations and improves diagnostic accuracy.

### 8. Response Generation and Recommendation

The model produces a final, evidence-based output containing:

- Accurate disease or pest identification
- Explanation of visible symptoms
- Scientifically validated treatment procedures
- Preventive recommendations
- Optional citations from retrieved sources

**9. Multilingual Output and Audio Response**

The generated output is translated into the user’s preferred language. If audio output is enabled, the translated answer is passed to a Text-to-Speech (TTS) module, producing a spoken explanation. This increases accessibility for users who prefer auditory feedback or have limited literacy.

**Advantages of the Proposed System:**

**1. Reduced Hallucination and Improved Factuality**

By grounding outputs in retrieved, expert-verified documents, the system significantly reduces the risk of false or misleading recommendations.

**2. Continual Knowledge Updating Without Model Retraining**

The vector database can be refreshed with new research findings, updated pest lists, or revised regulations, enabling real-time system evolution without retraining the LMM.

**3. Enhanced Generalization Across Real-World Conditions**

The expanded and diverse dataset allows the visual encoder and LMM to perform robustly under variable field conditions such as uneven lighting, occlusion, and morphological variation.

**4. Multilingual and Voice-Based Accessibility**

Speech input/output and multilingual processing make the system usable for farmers across linguistic backgrounds, including those who may not read or write.

**5. Transparent and Trustworthy Recommendations**

The system provides citations and factual grounding, increasing user confidence and ensuring traceability of recommendations.

## Applications of the Proposed System:

### 1. Real-Time Farmer Assistance

A mobile application supporting image, text, and audio queries enables immediate, accessible diagnosis and treatment suggestions in the field.

### 2. Agricultural Extension and Advisory Services

Government agencies and extension officers can use the system to deliver standardized, fact-based guidance to farming communities.

### 3. Educational and Training Tool

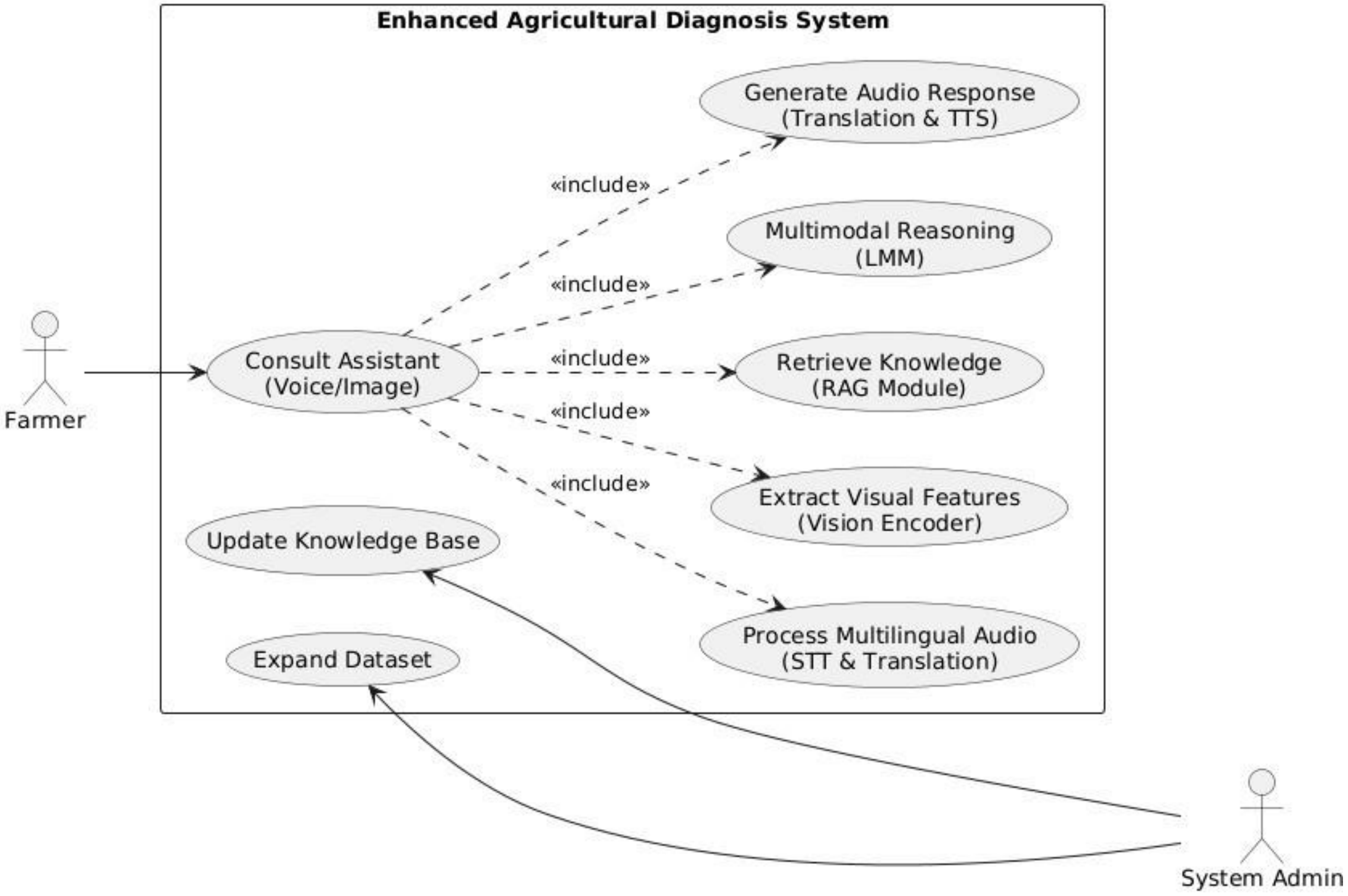
The system serves as an interactive learning platform for students and practitioners in agriculture, botany, and plant pathology.

### 4. Early Warning and Surveillance Systems

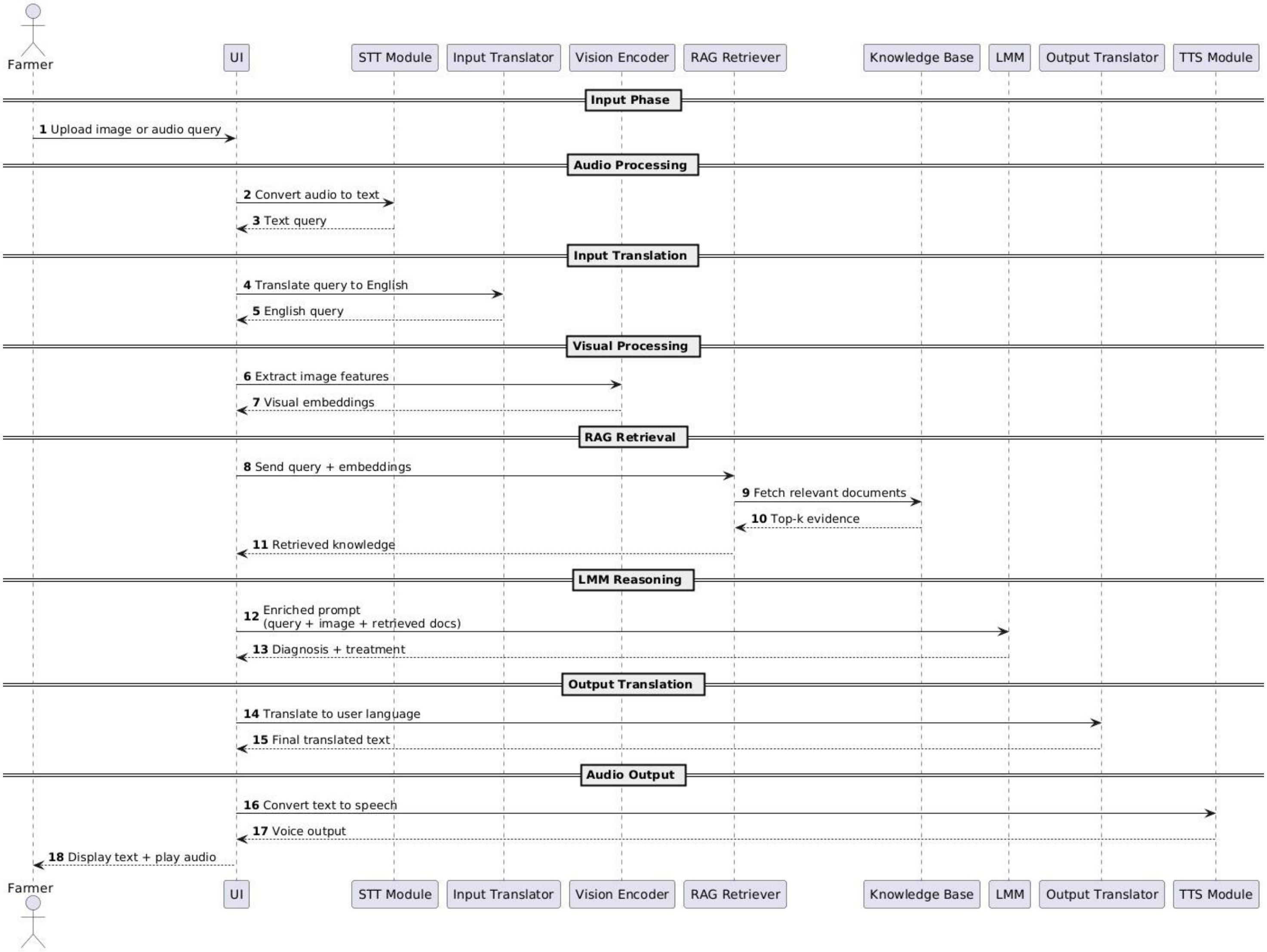
Aggregated user data can be analyzed to identify geographical patterns of disease outbreaks, supporting proactive decision-making and regional intervention planning.

# DESIGN DIAGRAMS

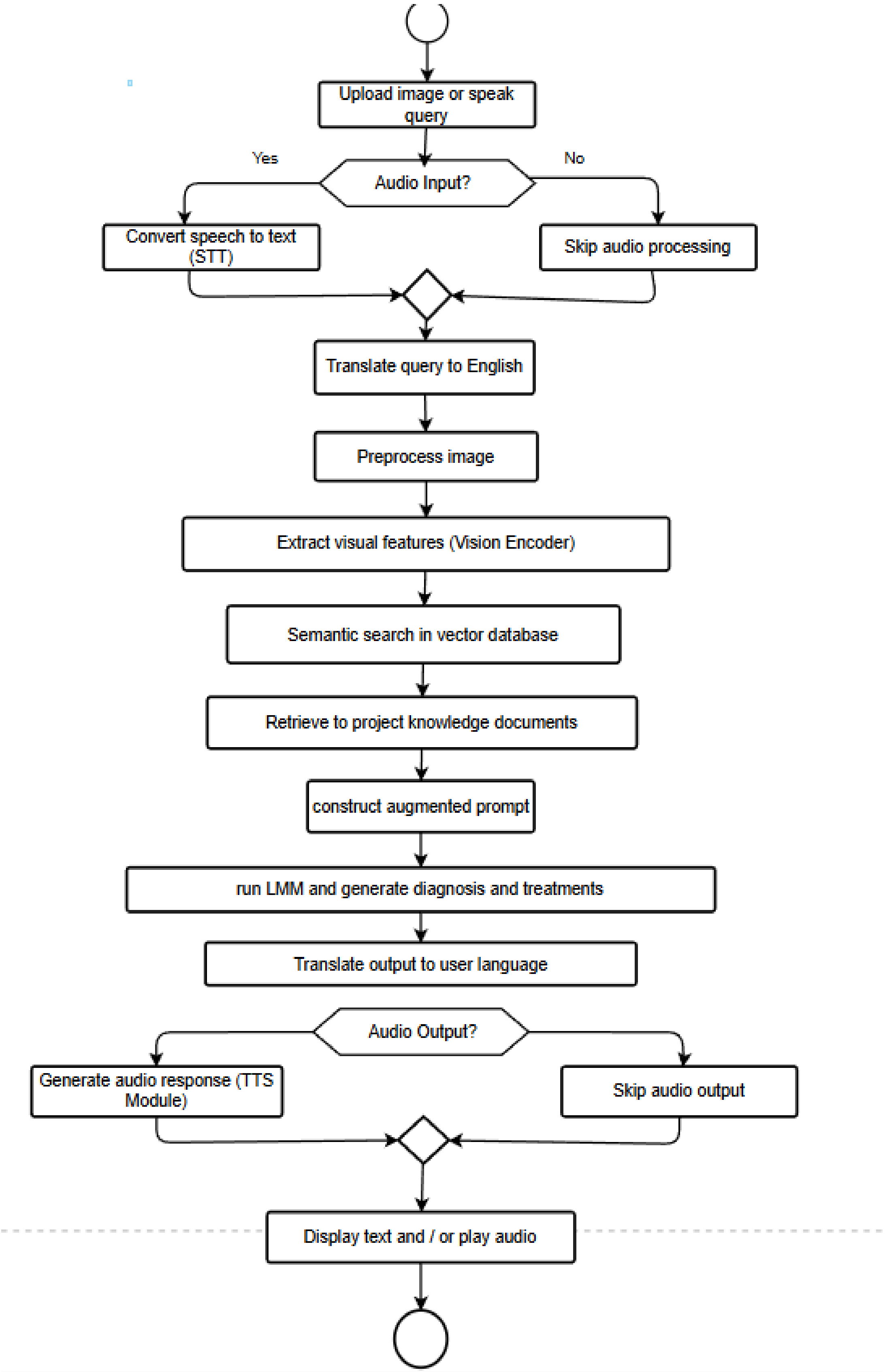
**Use Case Diagram:**


Enhanced Agricultural Diagnosis System
Generate Audio Response (Translation & TTS)
Multimodal Reasoning (LMM)
Retrieve Knowledge (RAG Module)
Extract Visual Features (Vision Encoder)
Process Multilingual Audio (STT & Translation)
«include»
«include»
«include»
«include»
«include»
Consult Assistant (Voice/Image)
Update Knowledge Base
Expand Dataset
Farmer
System Admin

### Sequence Diagram:



**Activity Diagram:**


Upload image or speak query
Yes
No
Audio Input?
Convert speech to text (STT)
Skip audio processing
Translate query to English
Preprocess image
Extract visual features (Vision Encoder)
Semantic search in vector database
Retrieve to project knowledge documents
construct augmented prompt
run LMM and generate diagnosis and treatments
Translate output to user language
Audio Output?
Generate audio response (TTS Module)
Skip audio output
Display text and / or play audio

# CONCLUSION

The rapid rise of crop pests and diseases continues to pose significant challenges to global agricultural productivity, particularly in regions where farmers lack timely access to expert diagnostic support. Although recent advances in Large Multimodal Models such as Agri-LLaVA have demonstrated promising capabilities in automated plant disease recognition and image-based agricultural reasoning, their limitations—including hallucinations, static internal knowledge, insufficient dataset diversity, monolingual constraints, and limited real-world generalization—restrict their reliability for field deployment. These constraints are especially impactful for rural farming communities where language diversity, variable environmental conditions, and limited digital literacy pose additional barriers.

To address these shortcomings, this work proposes an enhanced agricultural diagnostic framework that integrates Retrieval-Augmented Generation (RAG), a comprehensive Data Expansion Module, and extended support for both audio-based interaction and multilingual processing. The incorporation of RAG ensures that the system grounds its outputs in verified, up-to-date agricultural knowledge, thereby significantly reducing factual inconsistencies and hallucinations. Simultaneously, the expansion of the training dataset with diverse real-world agricultural images improves robustness under challenging field conditions, including poor lighting, occlusions, and mixed infections. The addition of Speech-to-Text and Text-to-Speech modules enables voice-based interaction, while multilingual translation modules allow users to communicate in their preferred local languages, enhancing accessibility for farmers with limited literacy or those living in linguistically diverse regions.

Through multimodal fusion of visual embeddings, transcribed or translated user queries, and retrieved domain-specific documents, the proposed architecture delivers more accurate, transparent, and trustworthy diagnostic and treatment recommendations. The refined workflow illustrates how next-generation AI systems for agriculture can evolve from static, model-only reasoning toward dynamic, evidence-backed decision support systems that continually adapt to new research findings, regional disease patterns, and updated agricultural policies.

The advantages demonstrated—such as continuous knowledge updating, improved generalization across environmental conditions, reduced hallucination, and inclusive multilingual and audio-enabled communication—highlight the potential of this architecture to empower farmers, extension workers, educators, and policymakers. The system’s accessible interface positions it as a valuable tool for both individual users and institutional agricultural advisory networks.

Future work may further extend this system by integrating additional data modalities such as soil chemistry, weather parameters, satellite imagery, and temporal crop monitoring. Incorporating region-specific disease forecasting models, on-device inference for offline use, and cross-lingual knowledge retrieval can further enhance its reliability and scalability. With continued development, retrieval-grounded, multimodal, multilingual, and audio-enabled diagnostic systems can play a transformative role in advancing precision agriculture and supporting global food security through accessible, reliable plant health diagnostics.

# REFERENCES

1. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al., “Flamingo: a visual language model for few-shot learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
2. J. Bai, S. Bai, S. Yang, et al., “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, 2023.
3. Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al., “Sparks of artificial general intelligence: Early experiments with gpt-4,” *arXiv preprint arXiv:2303.12712*, 2023.
4. Feilong Chen, Minglun Han, Haozhi Zhao, et al., “X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages,” *arXiv preprint arXiv:2305.04160*, 2023.
5. Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, et al., “Allava: Harnessing gpt4v-synthesized data for a lite vision-language model,” *arXiv preprint arXiv:2402.11684*, 2024.
6. Lin Chen, Jinsong Li, Xiaoyi Dong, et al., “Sharegpt4v: Improving large multi-modal models with better captions,” *arXiv preprint arXiv:2311.12793*, 2023.
7. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al., “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
8. Wenliang Dai, Junnan Li, Dongxu Li, et al., “Instructblip: Towards general-purpose vision-language models with instruction tuning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
9. Peng Gao, Jiaming Han, Renrui Zhang, et al., “Llama-adapter v2: Parameter-efficient visual instruction model,” *arXiv preprint arXiv:2304.15010*, 2023.
10. Tao Gong, Chengqi Lyu, Shilong Zhang, et al., “Multimodal-gpt: A vision and language model for dialogue with humans,” *arXiv preprint arXiv:2305.04790*, 2023.
11. Cyril Goutte and Eric Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *European Conference on Information Retrieval*, pp. 345–359, Springer, 2005.
12. Lakshay Goyal, Chandra Mani Sharma, Anupam Singh, and Pradeep Kumar Singh, “Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture,” *Informatics in Medicine Unlocked*, vol. 25, p. 100642, 2021.
13. David Hughes, Marcel Salathé, et al., “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” *arXiv preprint arXiv:1511.08060*, 2015.

14. Xin Lai, Zhuotao Tian, Yukang Chen, et al., “Lisa: Reasoning segmentation via large language model,” *arXiv preprint arXiv:2308.00692*, 2023.

15. Teven Le Scao, Angela Fan, Christopher Akiki, et al., “Bloom: A 176b-parameter open-access multilingual language model,” 2023.

16. Chunyuan Li, Cliff Wong, Sheng Zhang, et al., “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

17. Yanwei Li, Yuechen Zhang, Chengyao Wang, et al., “Mini-gemini: Mining the potential of multi-modality vision language models,” *arXiv preprint arXiv:2403.18814*, 2024.

18. Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, “Improved baselines with visual instruction tuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

19. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, “Visual instruction tuning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

20. Gen Luo, Yiyi Zhou, Tianhe Ren, et al., “Cheap and quick: Efficient vision-language instruction tuning for large language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

21. Chenyang Lyu, Minghao Wu, Longyue Wang, et al., “Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration,” *arXiv preprint arXiv:2306.09093*, 2023.

22. Muhammad Maaz, Hanoona Rasheed, Salman Khan, et al., “Video-chatgpt: Towards detailed video understanding via large vision and language models,” *arXiv preprint arXiv:2306.05424*, 2023.

23. Ben Mann, N. Ryder, M. Subbiah, et al., “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.

24. OpenAI, “Gpt-4v(ision) system card,” 2023. [Online].

25. Renjie Pi, Jiahui Gao, Shizhe Diao, et al., “Detgpt: Detect what you need via reasoning,” *arXiv preprint arXiv:2305.14167*, 2023.

26. Yixuan Su, Tian Lan, Huayang Li, et al., “Pandagpt: One model to instruction-follow them all,” *arXiv preprint arXiv:2305.16355*, 2023.

27. Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al., “Stanford alpaca: An instruction-following llama model,” 2023.

28. Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

29. Junke Wang, Lingchen Meng, Zejia Weng, et al., “To see is to believe: Prompting gpt-4v for better visual instruction tuning,” *arXiv preprint arXiv:2311.07574,* 2023.

30. Weihan Wang, Qingsong Lv, Wenmeng Yu, et al., “Cogvlm: Visual expert for pretrained language models,” *arXiv preprint arXiv:2311.03079*, 2023.

31. Wenhai Wang, Zhe Chen, Xiaokang Chen, et al., “Visionllm: Large language model is also an open-ended decoder for vision-centric tasks,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

32. Jason Wei, Maarten Bosma, Vincent Y. Zhao, et al., “Finetuned language models are zero-shot learners,” *arXiv preprint arXiv:2109.01652,* 2021.

33. Xiaoping Wu, Chi Zhan, Yu-Kun Lai, et al., “Ip102: A large-scale benchmark dataset for insect pest recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 8787–8796, 2019.

34. Zhiyang Xu, Ying Shen, and Lifu Huang, “Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning,” *arXiv preprint arXiv:2212.10773,* 2022.

35. Qinghao Ye, Haiyang Xu, Guohai Xu, et al., “mplug-owl: Modularization empowers large language models with multimodality,” *arXiv preprint arXiv:2304.14178*, 2023.