AI-Driven Phytopathology Recognition and Remediation in Coffee Leaves: YOLOv8 Implementation with RAG-Enabled LLM Enhancement

PROJECT PHASE-1 REPORT

Submitted by

Imadh Ajaz Banday (1BM20CS059)
Vibha Venkatesh Shanbhag(1BM20CS184)
Manikantha Gada (1BM20CS194)
Afifah Khan (1BM20CS195)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

Dr. Selva Kumar S
Assistant Professor, BMSCE

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
2023-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project entitled "AI-Driven Phytopathology Recognition and Remediation: YOLOv8 Implementation with RAG-Enabled LLM Enhancement" is a bonafide work carried out by Imadh Ajaz Banday (1BM20CS059), Vibha Venkatesh Shanbhag (1BM20CS184), Manikantha Gada (1BM20CS194), Afifah Khan (1BM20CS195), in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the academic year 2023-24. The project report has been approved as it satisfies the academic requirements with respect to the project work prescribed for the said degree.

Guide	Head of Department
Dr. Selva Kumar S, Assistant Professor, Dept of CSE, B.M.S. College of Engineering	**Dr. JYOTHI S. NAYAK**, Professor and HOD, Dept of CSE, B.M.S. College of Engineering

Principal

Dr. S. Muralidhara

B.M.S. College of Engineering

External Viva

Name of the Examiners	Signature with Date
1.	
2.	

Table of Contents

Abstract

In the agrarian landscape of Karnataka, India, where agriculture is a cornerstone of the economy, the prevalence of diseases like Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust poses significant threats to crop yields. This research presents an innovative methodology that integrates YOLOv8 for precise disease identification and GPT-3.5 Turbo for comprehensive diagnosis, addressing challenges such as hallucination through Retrieval Augmented Generation (RAG).

The primary focus of this study is on remediation strategies, aiming to reshape traditional agricultural practices. By combining advanced AI and ML techniques, the RAG-centric approach facilitates automated disease detection, providing real-time, context-specific information crucial for effective remediation. The research extends its impact beyond the coffee sector, envisioning a transformation in the management of plant diseases across various agricultural domains.

Key objectives include revolutionizing disease identification precision, ensuring comprehensive diagnostic capabilities, and offering solutions to hallucination challenges. The proposed methodology not only contributes to securing food supplies and safeguarding livelihoods but also promotes eco-friendly farming practices by reducing reliance on pesticides.

This research, characterized by its succinct and purposeful nature, aspires to serve as a catalyst for positive transformations in agriculture. By advancing the integration of AI and ML technologies, the study seeks to usher in a new era of sustainable farming practices, aligning with the broader goals of agricultural sustainability, food security, and economic resilience in Karnataka and beyond.

Furthermore, the study emphasizes the potential systemic impact of the proposed methodology on the agricultural landscape. By leveraging YOLOv8 and GPT-3.5 Turbo in tandem, the research not only streamlines disease identification processes but also establishes a foundation for a dynamic and adaptive agricultural ecosystem. The synergy of these technologies ensures not only immediate remediation but also lays the groundwork for proactive measures, enabling farmers to anticipate and mitigate disease outbreaks. This proactive approach is crucial for the resilience of Karnataka's agriculture sector, offering a paradigm shift towards a more anticipatory and sustainable model. As the findings of this research permeate into broader agricultural sectors, the envisioned positive transformations have the potential to set new standards for agricultural practices, ultimately contributing to the global pursuit of sustainable and resilient food production systems.

DECLARATION

We, hereby declare that the Major Project Phase-1 work entitled "AI-Driven Phytopathology Recognition and Remediation: YOLOv8 Implementation with RAG-Enabled LLM Enhancement" is a bonafide work and has been carried out by us under the guidance of Dr. Selva Kumar S**,** Assistant Professor, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bengaluru**,** in partial fulfillment of the requirements of the degree of Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi.

We further declare that, to the best of our knowledge and belief, this project has not been submitted either in part or in full to any other university for the award of any degree.

Candidate details:

SL. NO.	Student Name	USN	Student's Signature
1	Imadh Ajaz Banday	1BM20CS059	
2	Vibha Venkatesh Shanbhag	1BM20CS184	
3	Manikantha Gada	1BM20CS194	
4	Afifah Khan	1BM20CS195	

Place: Bengaluru

Date: December 29, 2023

Certified that these candidates are students of the Computer Science and Engineering Department of B.M.S. College of Engineering. They have carried out the project work titled "AI-Driven Phytopathology Recognition and Remediation: YOLOv8 Implementation with RAG-Enabled LLM Enhancement" as Major Project Phase-1 work. It is in partial fulfillment for completing the requirement for the award of B.E. degree by VTU. The works are original and duly certify the same.

Guide Name   Signature

Dr. Selva Kumar S

Date:

Acknowledgment

We extend our heartfelt gratitude to all those who supported us on this journey. We are particularly grateful to the administration of BMSCE for providing essential resources and a supportive environment, and to Dr. S. Muralidhara, Principal BMSCE and Dr. Jyothi S. Nayak, Head of Department of Computer Science and Engineering, for generously providing both resources and the necessary framework to develop this project.

We are deeply indebted to our Project Guide Dr. Selva Kumar S and Dr. Umadevi V, our project coordinator, both of whose guidance, support, and invaluable insights have been instrumental in shaping this endeavor. Their expertise and encouragement have been pivotal throughout the entire project journey.

Moreover, we are grateful to our friends and family for their unwavering support, patience, and understanding during the challenging phases of this project. Their encouragement and belief in my abilities were fundamental in keeping me motivated.

Lastly, we extend our appreciation to all individuals, whether directly or indirectly involved, for their valuable insights, feedback, and support, which have significantly contributed to the completion of this literature survey.

List of Tables

List of Figures

Chapter 1: INTRODUCTION

1.1 Overview

This project delves into the innovative integration of YOLOv8 for precise phytopathology recognition and GPT-3.5 Turbo, enhanced with Retrieval Augmented Generation (RAG), for comprehensive diagnosis. Focused on addressing challenges in hallucination, the study explores an AI-driven approach to plant disease identification and remediation. With a primary emphasis on phytopathology in Karnataka, a major coffee-producing region facing issues such as Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust, the proposed methodology aims to revolutionize traditional agricultural practices. By harnessing the power of YOLOv8 and RAG-enabled language models, the research seeks to offer advanced, context-specific solutions for real-time remediation, contributing to sustainable agriculture and reduced reliance on pesticides.

The key components include:

- Real-Time Remediation Strategies: The study focuses on developing real-time remediation strategies based on the identified phytopathologies. This involves the integration of actionable insights generated by the AI models to provide timely and context-specific recommendations for farmers, enabling swift response to disease outbreaks.

- Context-Specific Language Model Outputs: GPT-3.5 Turbo, enhanced with RAG, plays a vital role in generating context-specific outputs related to phytopathology diagnosis. This ensures that the information provided is not only accurate but also tailored to the specific conditions and requirements of the agricultural context in Karnataka.

- Proactive Disease Mitigation: The research aims to establish a proactive approach to disease mitigation by leveraging the combined power of YOLOv8 and GPT-3.5 Turbo. The goal is to empower farmers with tools that enable them to anticipate and prevent disease outbreaks, contributing to a more resilient and sustainable agricultural ecosystem.

- Adaptability and Scalability: The proposed methodology is designed to be adaptable to various agricultural sectors beyond coffee production. Its scalability allows for widespread application, ensuring the relevance and impact of the research beyond the immediate context of Karnataka.

1.2 Motivation

The project "AI-Driven Phytopathology Recognition and Remediation: YOLOv8 Implementation with RAG-Enabled LLM Enhancement" is motivated by the imperative to overcome the limitations posed by LLMs' static nature and enhance their contextual relevance in the domain of precision agriculture. By addressing the challenge of hallucination through the introduction of Retrieval Augmented Generation (RAG), we aim to create a dynamic bridge that connects the user prompt, facilitated by YOLOv8 for disease identification, with external databases. RAG acts as a conduit for fetching the latest information, enriching the LLM's understanding and reducing the risk of generating outdated or inaccurate responses.

The intersection of Large Language Models (LLMs) and object detection technologies, exemplified by the deployment of YOLOv8 for disease identification in coffee leaves, holds immense potential for advancing precision agriculture. However, the inherent static nature of LLMs poses a significant challenge in providing accurate and up-to-date domain-specific information, often leading to the phenomenon of 'hallucination.' In the dynamic realm of disease diagnosis, particularly in the intricate context of the coffee industry, where factors such as climate, soil conditions, and evolving pathogens play pivotal roles, the need for real-time, context-aware responses is paramount.

The motivation behind this research extends beyond the confines of the coffee industry. The envisioned outcomes have the potential to revolutionize the synergy between object detection, language models, and precision agriculture, paving the way for optimal results not only in coffee cultivation but also in diverse agricultural sectors. The need for accurate, context-aware information is universal across agriculture, and the proposed approach strives to be a catalyst for transformative advancements that go beyond traditional static LLM limitations. By bridging the gap between dynamic disease identification and real-time contextual information, this research seeks to contribute significantly to the broader goal of sustainable and efficient agricultural practices.

1.3 Objective

The objective of our project is the following:

- Dynamic Disease Identification and Correction
- Real-Time Monitoring Implementation with RAG
- Dataset Expansion through RAG Collaboration
- Organizational Involvement in RAG Implementation
- Broader Range of Disease Classes Incorporation
- Future Developments in RAG-Integrated Object Detection Systems

1.3.1 Dynamic Disease Identification and Correction:

This objective centers on enhancing the agility and precision of disease identification processes in the context of precision agriculture, particularly focusing on diseases affecting coffee leaves. The term "dynamic" refers to the evolving nature of environmental conditions, pathogen strains, and other variables that influence disease manifestations. The goal is to implement a system that not only identifies diseases promptly but also adapts to real-time changes in the agricultural landscape.

1.3.2 Real-Time Monitoring Implementation with RAG:

Real-time monitoring with RAG aims to transform traditional agricultural monitoring into a proactive and adaptive process. By integrating dynamic data acquisition, processing, and contextual enhancement, this objective facilitates a deeper understanding of the agricultural environment. The use of RAG ensures that the system is not only aware of current conditions but can also retrieve external knowledge to enrich its responses, fostering a more intelligent and responsive approach to agricultural management.

1.3.3 Dataset Expansion through RAG Collaboration:

This objective focuses on using Retrieval Augmented Generation (RAG) to collaboratively expand agricultural datasets. By dynamically incorporating information from external sources, this approach ensures the dataset's ongoing relevance and comprehensiveness. The goal is to create a more adaptive and informed dataset, enabling machine learning models to better capture the evolving nuances of agricultural dynamics.

1.3.4 Organizational Involvement in RAG Implementation:

This objective emphasizes the active participation and collaboration of the organization in the implementation of RAG. By involving the organization in the integration process, the aim is to ensure seamless adoption and alignment with organizational goals. The objective seeks to leverage RAG as a strategic tool, aligning it with organizational objectives to maximize its impact on knowledge retrieval, decision-making, and overall operational efficiency.

1.3.5 Broader Range of Disease Classes Incorporation:

This objective focuses on expanding the scope of disease identification by incorporating a more diverse range of disease classes. By broadening the categories considered, the aim is to enhance the system's ability to detect and diagnose various diseases across different crops. This proactive approach ensures a more comprehensive and adaptable solution, contributing to the versatility and effectiveness of the disease identification system in diverse agricultural contexts.

1.3.6 Future Developments in RAG-Integrated Object Detection Systems:

This objective envisions advancing RAG-integrated object detection systems for enhanced functionality and broader applicability. The focus is on continual improvements, exploring innovations in both Retrieval Augmented Generation (RAG) and object detection methodologies. The goal is to propel these systems towards increased accuracy, adaptability, and efficiency, fostering a dynamic and intelligent framework capable of addressing evolving challenges in diverse domains.

1.4 Scope

The scope of this project is expansive, aiming to revolutionize precision agriculture through the integration of advanced technologies. At its core, the project employs YOLOv8 for highly accurate object detection, specifically focusing on diseases prevalent in the coffee industry in Karnataka, such as Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust. The use of YOLOv8 ensures the system's capability to promptly and reliably identify diseases affecting crops, forming a foundational element of the project's scope.

A critical aspect involves addressing the limitations of static Large Language Models (LLMs) in disease diagnosis, commonly leading to inaccuracies known as 'hallucination.' To combat this, the project introduces Retrieval Augmented Generation (RAG), acting as a dynamic bridge to fetch real-time, contextually relevant information. This ensures that the diagnostic capabilities of LLMs remain adaptive and informed, particularly crucial in the dynamic agricultural environment.

The project extends beyond disease identification to incorporate real-time monitoring systems, integrating sensor networks for continuous assessment of environmental conditions. This not only contributes to adaptive disease management but also aligns with the broader goal of proactive agricultural practices.

Dataset expansion forms another integral component of the project's scope, emphasizing collaboration through RAG. This collaborative approach aims to keep the dataset comprehensive and adaptable by dynamically incorporating external information. The involvement of the organization in the implementation process is pivotal, ensuring seamless integration and alignment with overarching goals, fostering a culture of innovation and adaptability.

Furthermore, the project envisages a broader scope by incorporating a diverse range of disease classes, making the solution applicable to various crops and agricultural domains. It anticipates ongoing advancements in RAG-integrated object detection systems, exploring innovations for increased accuracy, adaptability, and efficiency in addressing emerging challenges. In summary, the holistic nature of this project positions it as a transformative force in precision agriculture, with far-reaching implications for sustainable and efficient farming practices globally.

1.5 Existing System

The following applications exist in the agricultural industry to support the farmers :

- **Plantix:** An AI-powered mobile app for plant disease detection. Users can take pictures of plant symptoms, and the app provides information about the likely disease and potential solutions.

- **DeepAgro:** A deep learning-based platform that uses computer vision to identify and analyze crop diseases. It provides farmers with real-time insights for disease management.

- **AgroSight:** Combines satellite imagery and artificial intelligence to monitor crop health. It offers early detection of diseases and provides actionable insights for farmers.

- **eAgroSys:** An integrated system that employs IoT (Internet of Things) sensors, drones, and AI to monitor crops, soil conditions, and detect diseases. It provides farmers with real-time data for decision-making.

- **CropX:** Utilizes soil sensors and data analytics to optimize irrigation and fertilization. While not focused on disease detection, it contributes to overall crop health management.

- **IBM Watson Decision Platform for Agriculture:** Offers AI-powered insights for farmers, including weather forecasts, crop disease models, and recommendations for crop management.

- **PlantVillage Nuru:** An AI-driven platform for plant disease diagnosis. It uses a combination of machine learning and human expertise to identify diseases based on images of plant symptoms.

- **Gamaya:** Integrates hyperspectral imaging and AI for crop monitoring. It provides detailed information about crop health, nutrient levels, and potential diseases.

- **Digital Green:** Combines digital technology and community engagement to provide farmers with agricultural information, including disease management practices.

- **Taranis:** Uses high-resolution aerial imagery, machine learning, and computer vision to monitor crops for signs of diseases, pests, and nutrient deficiencies.

1.6 Proposed System

The proposed system aims to integrate cutting-edge technologies to address challenges in precision agriculture, with a specific focus on disease identification and remediation. The system combines several key components:

1. **YOLOv8 for Object Detection:** YOLOv8 is proposed for precise and real-time object detection, specifically targeting diseases affecting crops. This component ensures swift and accurate identification of diseases in crops, particularly focusing on varieties such as Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust in the context of the coffee industry in Karnataka.

2. **Retrieval Augmented Generation (RAG):** To overcome the limitations of static Large Language Models (LLMs) and mitigate hallucination challenges, the system incorporates RAG. Acting as a dynamic bridge, RAG retrieves up-to-date and contextually relevant information from external databases. This enhances the overall disease diagnosis process, providing accurate and timely insights to farmers.

3. **Real-Time Monitoring System:** The project includes the implementation of a real-time monitoring system equipped with sensor networks. These sensors continuously assess environmental conditions such as temperature, humidity, and soil moisture. The real-time data acquired contributes to adaptive disease management and provides valuable insights for farmers to make informed decisions.

4. **Collaborative Dataset Expansion through RAG:** The system emphasizes collaborative dataset expansion using RAG. This involves dynamically incorporating external information to keep the dataset comprehensive and adaptable. Collaborative efforts ensure that the dataset remains diverse, addressing a broader range of diseases and agricultural scenarios.

5. **Organizational Involvement:** The organization is actively involved in the implementation process, aligning the proposed system with overarching goals. This collaboration fosters a culture of innovation and adaptability, ensuring the system's seamless integration into existing agricultural practices.

6. **Versatility Across Crop Types:** While initially tailored for the coffee industry in Karnataka, the proposed system aims for versatility across various crop types. The objective is to create a scalable solution applicable to diverse agricultural domains, contributing to a broader impact beyond the coffee sector.

7. **Proactive Disease Management:** The system anticipates future developments, aiming for continual advancements in RAG-integrated object detection systems. This forward-looking approach ensures the system's adaptability to emerging challenges, contributing to proactive and intelligent disease management in agriculture.

In summary, the proposed system is a comprehensive and forward-thinking solution that leverages state-of-the-art technologies to address the complexities of disease identification and remediation in precision agriculture. The integration of YOLOv8, RAG, real-time monitoring, and collaborative dataset expansion reflects a holistic approach to revolutionize agricultural practices for sustainability and efficiency.

1.7 Work Plan

Phase 1: Project Initiation (Month 1-3)

1. Define Project Objectives and Scope:
   - Clarify project goals, objectives, and deliverables.
   - Specify the scope, including target crops and geographical focus.
2. Stakeholder Engagement:
   - Identify and engage key stakeholders, including farmers, agricultural experts, and organizations.
3. Literature Review:
   - Conduct an in-depth review of existing AI-driven precision agriculture systems.
   - Identify relevant technologies, challenges, and best practices.

Phase 2: System Design and Planning (Month 3-4)

1. Technology Selection:
   - Finalize the selection of technologies, including YOLOv8, RAG, and real-time monitoring components.
2. System Architecture Design:
   - Develop a comprehensive system architecture, outlining the integration of chosen technologies.
   - Detailed data flow, interactions, and dependencies.
3. Data Collection and Dataset Preparation:
   - Identify and collect relevant datasets for training and testing.
   - Begin the preparation and augmentation of datasets to cover diverse scenarios.

Phase 3: Development (Month 4-6)

1. YOLOv8 Implementation:
   - Implement and fine-tune YOLOv8 for disease detection, focusing on the identified diseases in coffee crops.
2. RAG Integration:
   - Integrate RAG as a dynamic bridge for information retrieval, linking user prompts with external databases.
3. Real-Time Monitoring System Development:
   - Develop the real-time monitoring system using sensor networks.
   - Implement algorithms for data acquisition and analysis.

4. Collaborative Dataset Expansion:

- Implement mechanisms for collaborative dataset expansion through RAG.
- Ensure dynamic and diverse dataset growth.

Phase 4: Testing and Optimization (Month 6-7)

1. System Integration:

- Integrate all components into a cohesive system.
- Address interoperability issues and ensure seamless interactions.

2. Testing and Validation:

- Conduct rigorous testing, including unit testing, system testing, and validation against diverse datasets.
- Gather feedback from stakeholders for refinement.

3. Optimization and Performance Tuning:

- Optimize algorithms for speed, accuracy, and resource efficiency.
- Fine-tune the system based on testing results.

Phase 5: Deployment (Month 7-9)

1. Deployment:

- Deploy the system with a user friendly website/application.
- Monitor system performance in real-world conditions.

2. User Training:

- Provide training sessions for end-users, including farmers and agricultural professionals.
- Ensure users are familiar with system functionalities.

3. Feedback Collection:

- Collect feedback from users and stakeholders.
- Identify areas for improvement and address any issues that arise.

Phase 6: Documentation and Knowledge Transfer (Month 9-10)

1. Documentation:

- Create comprehensive documentation fora complete user guide.

2. Knowledge Transfer:

- Transfer knowledge to relevant personnel for ongoing maintenance and support.

3. Project Conclusion:

- Conclude the project with a final report summarizing achievements, challenges, and recommendations for future enhancements.

Gantt Chart

The following Gantt chart gives a timeline for the project.

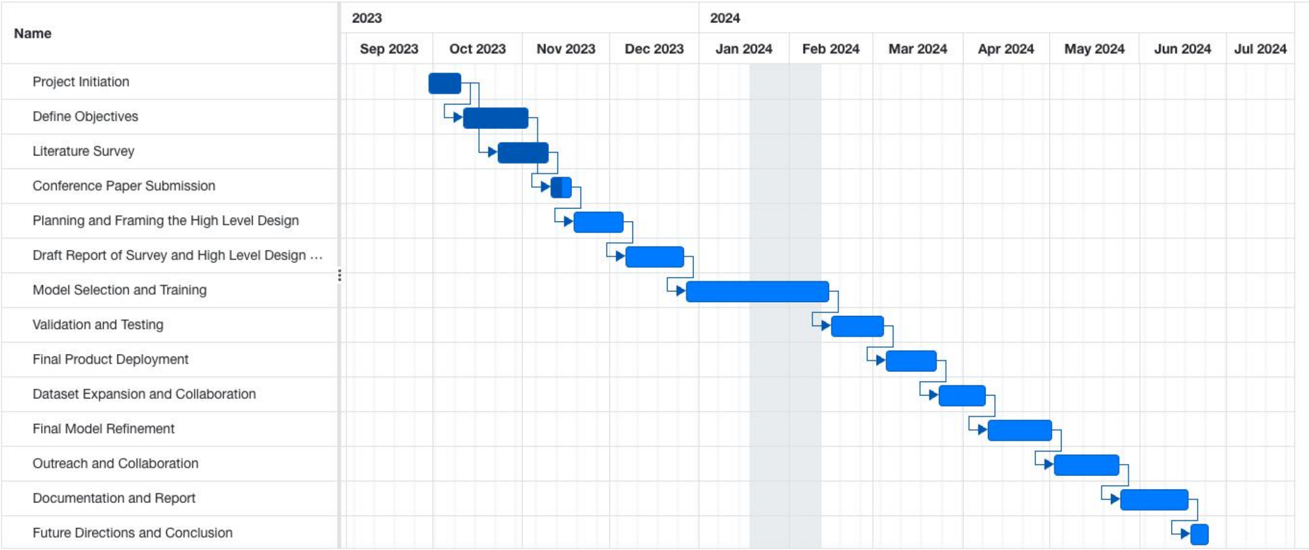


Figure 1.1: Work Plan of the project

Chapter 2: LITERATURE SURVEY

This section describes the main proposals found in the literature review.

Sajitha P et al. [1] introduced a system that integrates YOLO v7 for leaf disease detection and GPT-3 for providing corrective measures. And Jiajun Qing et al. in their paper [2], combine the logical reasoning capabilities of GPT-4 with YOLOPC, a lightweight YOLO variant, to achieve real-time agricultural disease diagnosis. While YOLOPC attains 94.5% accuracy with 75% fewer parameters, GPT-4 demonstrates 90% reasoning accuracy in generating diagnostic reports. Both [1] and [2] face the limitation that the language model may not consistently yield the most accurate answers.

Jean Kaddour et al. [3] exlpore the challenges with LLMs through their paper “Challenges and Applications of Large Language Models”. The paper explores large language models (LLMs), highlighting challenges like complex datasets and high costs. It focuses on improving LLM behavior and knowledge, discusses fine-tuning methods, and emphasizes the need for comprehensive evaluation. In [4], the authors explore the issue of hallucination in LLMs, they propose the LVLM Hallucination Revisor (LURE), an algorithm designed to address object hallucination in Large Vision-Language Models (LVLMs). Evaluation on six open-source LVLMs demonstrates a substantial 23% improvement in general object hallucination metrics compared to prior approaches. Further Junyi Li et al. through their paper [5] introduce the Hallucination Evaluation benchmark for Large Language Models (HaluEval), a comprehensive collection of 35,000 hallucinated and normal samples for analyzing and evaluating LLMs. The study presents a two-stage framework for generating and annotating hallucinated samples, revealing that existing LLMs often fail to recognize hallucinations in text and tend to generate such content. [6] provides an overview of the challenges posed by hallucination in LLMs. It delves into the impact of noisy data during pre-training on LLMs' parametric knowledge and the subsequent occurrence of hallucinations. The survey explores various mitigation strategies, including data curation, filtering, and supervised fine-tuning, as well as the use of high-quality reference corpora. In summary [4], [5], and [6] collectively reveal that hallucination is a pervasive challenge in Large Language Models, prompting the exploration of algorithms, benchmarks, and mitigation strategies to enhance their performance and reliability.

In [7] the authors address hallucination in Large Language Models (LLMs) through the creation of the HILT dataset, utilizing 75,000 text passages generated by 15 LLMs. They emphasize the need for continuous updates due to the evolving nature of the field and present detailed statistics on factual mirage (FM) and silver lining (SL) categories.

Nitin Liladhar Rane et al. [8] explore the multifaceted contributions of large language models, such as ChatGPT, in scientific and research progress across diverse domains. It underlines the potential for these models to revolutionize knowledge dissemination while acknowledging the ethical and societal implications, emphasizing the importance of responsible development, deployment, and regulation.
In [9] Abdullahi Saka et al. shift the focus to the application of GPT models in the construction industry. The authors delve into opportunities, limitations, and a use case validation involving NLP techniques for processing construction project documents and Building Information Modeling (BIM) data. Abdullahi Saka et al. suggest formulating ethical use policies, exploring novel applications, and researching solutions for GPT model limitations in construction.

The paper [10] "Assessing the Strengths and Weaknesses of Large Language Models" presents a balanced assessment of large language models (LLMs). It highlights the sophisticated inductive learning and inference capabilities of LLMs, including their ability to identify hierarchical syntactic structure and complex semantic relations. Additionally, LLMs have shown promise in tasks such as medical image analysis and diagnostics, as well as predicting properties of proteins and new molecular structures. However, Shalom Lappin also acknowledges limitations, such as the potential for LLMs to hallucinate plausible-sounding narratives with no factual basis, their susceptibility to adversarial testing, and the need for further research into improving their performance on specific tasks, addressing biases, and developing smaller, more lightweight models.

[11] introduces the Retrieval-Augmented Generation (RAG) model, showcasing its state-of-the-art performance in open-domain question answering tasks, emphasizing its ability to generate diverse and factual content. The next paper [12], Self-RAG, presents a framework that enhances large language models (LLMs) through retrieval and self-reflection without compromising creativity. It employs instruction and demonstration pairs, achieving significant improvements in model performance, factuality, and citation accuracy, with future work aiming to address factual errors and enhance self-reflection mechanisms. The authors of [13] discuss the integration of large language models into information retrieval systems, highlighting potential benefits and challenges, proposing research directions for improvement, and addressing limitations such as bias and data requirements. Jiawei Chen et al. [14] introduce a benchmark for evaluating the performance of LLMs in retrieval-augmented generation tasks, identifying limitations in noise robustness and information integration abilities, and suggesting directions for improvement. [15] proposes ARM-RAG, a system leveraging RAG to enhance problem-solving capabilities of LLMs, demonstrating superior performance by utilizing Neural Information Retrieval for reasoning chains derived from solving math problems and suggesting avenues for further enhancements.

In [16] "InPars," introduces a method using large language models (LLMs) for few-shot labeled data generation in information retrieval (IR) tasks, demonstrating superior performance and emphasizing potential with domain-specific training data. Limitations include the lack of pretraining and limited suitability for non-neural retrieval algorithms. The next paper [17], "Retrieval-based Evaluation for LLMs," proposes Eval-RAG, a method for evaluating LLM-generated texts in the legal domain, outperforming existing methods in correlation with human evaluation and factual error identification. Future work includes refining Eval-RAG, exploring applicability to other domains, and addressing potential limitations. The authors in [18], "Retrieval Meets Long Context LLMs," compare retrieval-augmented language models (RAG) and long context LLMs, demonstrating RAG's significant performance improvement in question answering and summarization tasks.
Future work aims to explore combined retrieval and long context LLMs for enhanced accuracy, with limitations not explicitly mentioned.

Zhangyin Feng et al. [19] introduces Retrieval-Generation Synergy Augmented Large Language Models, showcasing an iterative framework that significantly enhances the reasoning ability of large language models (LLMs) for knowledge-intensive tasks, particularly open-domain question answering. Through experiments on four datasets, the proposed method outperforms previous baselines, demonstrating improved LLM reasoning. [20] focuses on Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. It presents an end-to-end methodology that leverages a "retrieve-then-read" pipeline, employing a retrieval-augmented generator (RAG) approach with LLMs. The authors fine-tune on a task-specific dataset and introduce the Long-form Legal Question Answering (LLeQA) dataset. The authors highlight the positive aspects of this approach, emphasizing its potential for generating syntactically correct answers relevant to legal questions. The last paper [21], Establishes Performance Baselines in Fine-Tuning, Retrieval-Augmented Generation, and Soft-Prompting for Non-Specialist LLM Users, explores methods to enhance LLM performance for non-specialist users. It compares unmodified GPT 3.5, fine-tuned GPT 3.5, and RAG using a limited dataset and technical skill. RAG stands out as an effective strategy, outperforming fine-tuning, showcasing positive results in the context of the LayerZero cryptocurrency bridging project. The paper discusses the accessibility of these techniques to non-technical users and emphasizes the positive impact of RAG on LLM performance.

Summary of Literature Survey:

SI. No.	Title	Year	Conference Name	Journal Name	Datasets Used	Data Preprocessing	Technique	Limitations	Results	Future Work
1.	Leaf Disease Detection & Correction using YOLO V7 with GPT3 integrated	June 2023		IJERT	Dataset: PlantLeaf, 2,750 images, diseases (early leaf blight, rust), 1,900 training images, 850 testing/validation images.	The uploaded data underwent pre-processing steps and augmentation, which were selected in Roboflow. Image resizing, Data augmentation, Normalisation, Data splitting	Deep Learning, CNN, Computer Vision, Object detection, YOLO v7 (155 frames per second. This significantly surpasses other state-of-the-art object detection algorithms), GPT-3.	Limited dataset, Limited classes of diseases, Limited evaluation metrics, Limited discussion on ethical considerations	YOLO v7, achieved 96% accuracy in leaf disease detection and classification.	Expand dataset, enable real-time monitoring via mobile apps, establish collaborative learning for continuous model improvement.
2.	GPT-aided diagnosis on agricultural image based on a new light YOLOPC	Oct 2023		Science Direct-Elsevier - Computers and Electronics in Agriculture	Open agricultural dataset with images of crops, diseases, and deficiencies.	Standardization, augmentation, and multimodal fusion of image and text data.	Combined GPT-4 for logical reasoning with YOLOPC, a lightweight YOLO variant, for real-time agricultural disease diagnosis.	GPT-4's potential language bias, YOLOPC's reliance on specific features, and the need for continuous model improvement.	YOLOPC achieves 94.5% accuracy with 75% fewer parameters; GPT-4 demonstrates 90% reasoning accuracy in generating diagnostic reports.	Explore language model improvements, address bias in agricultural language, enhance YOLOPC's feature extraction, and expand the dataset.

3.	Challenges and Applications of Large Language Models	July 2023	arXiv preprints		Dataset details not explicitly mentioned, but size and diversity of pre-training datasets have rapidly increased.	-	Briefly mentions instruction fine-tuning and references specific datasets for diverse tasks and multilingual settings..	Unfathomable datasets, high pre-training costs, inference latency, limited context length, prompt brittleness, misaligned behavior, outdated knowledge.	Discusses fine-tuning techniques, emphasizes comprehensive LLM research evaluation, and outlines applications in chatbots, computational biology, law, and medicine.	Tackling dataset challenges, cutting pre-training costs, refining model behavior and knowledge, improving evaluation methods.
4.	Analyzing and Mitigating Object Hallucination in Large Vision-Language Models	Oct 2023	Arxiv preprints		COCO and Flickr30k	Curating a training dataset with images and their corresponding descriptions, generating hallucinatory descriptions using GPT-3.5	LVLM Hallucination Revisor (LURE)	The reliance on GPT-3.5 for generating hallucinatory descriptions may introduce biases, and the evaluation is limited to object hallucination.	LURE achieved a 23% improvement in general object hallucination on six open-source LVLMs.	Investigating the impact of LURE on downstream tasks and exploring the use of other language models.
5.	HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large	Dec 2023	Proceedings of the 2023 Conference on Empirical Methods in Natur		HaluEval, a large-scale collection of generated and human-annotated hallucinated samples.	Two-step automatic process to generate hallucinated samples for evaluation, which is completely based	Two-stage framework for generating and annotating hallucinated samples, and empirical evaluation of large	Quality of hallucinated samples is limited by the capacity of ChatGPT in following the	Existing LLMs mostly fail to recognize the hallucinations in text and tend to generate hallucin	Investigating the underlying reasons behind the appearance of hallucinations in

	Language Models		al Language Processing			on LLMs.	language models in recognizing and generating hallucinations.	complex instruction of hallucination sampling.	ated content.	LLMs.
6.	Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models	Sep 2023	Arxiv Preprints		The paper discusses the challenges of curating training data during pre-training due to the vast scale of pre-training corpora	Cleaning and pre-training data using similarity to high-quality reference corpora, carefully extracting high-quality data from the web via heuristic rules, and pre-training on filtered "textbook-like" synthetic data	Collecting pre-training data from credible text sources, and up-sample data from factual sources, and prepend topic prefixes to sentences in factual documents during pretraining.	Potential biases introduced by the selection of pre-training data sources and the scalability of these techniques to larger LLMs and more diverse domains	These approaches aim to mitigate hallucinations by ensuring the quality and factuality of the pre-training data, leading to powerful LLMs with reduced hallucination tendencies	Developing novel algorithms for data curation, and investigating the impact of different pre-training data sources on hallucination mitigation.
7.	The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and	Oct 2023	Arxiv Preprints		HILT dataset, which consists of 75,000 text passages generated by 15 LLMs using prompts from NYTimes tweets and the Politifact dataset	Data preprocessing involved obtaining sentence-level annotations for hallucination orientations and categories using the Amazon Mechanical Turk (Amazon) and the	The techniques/algorithms used included utilizing 15 LLMs to generate text passages and obtaining annotations for hallucination orientations and categories.	Limitations include the ever-evolving nature of the field and the need for continuous updates and contributions to the research community.	The result was the creation of a publicly available hallucination dataset, HILT, with detailed statistics on factual mirage (FM) and	Future work involves fostering an environment of continuous updates and contributions to the HVI benchmark leaderboards.

	Prescriptive Remediations					MACE tool			silver lining (SL) categories	
8.	Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: a double-edged sword	Oct 2023		IJRETS	-	-	The paper surveyed and researched ChatGPT's contributions across disciplines to understand its potential impact on scientific and research progress.	Need for continuous improvement due to varying performance based on specific tasks and training data quality. Additionally, bias-related challenges pose significant concerns in the ethical utilization of ChatGPT.	Highlights ChatGPT's potential to revolutionize fields like public health, climate change, programming, and education, with an emphasis on responsible and ethical development, deployment, and regulation.	Refining AI models, establishing clear guidelines, fostering interdisciplinary collaboration, and addressing ethical concerns to ensure equitable and effective educational paradigms
9.	GPT models in construction industry: Opportunities, limitations, and a use case validation	Dec 2023		Science Direc, Elsevier - Developments in the Built Environment	Construction project documents including architectural drawings,engineering plans, material databases, and historical project data for cost and schedule analysis.	Clean and standardize text data from construction documents, extract relevant information using NLP techniques, and format data for GPT models.	NLP for text analysis, fine-tuning GPT models on construction data, and integrating GPT with BIM for decision support and optimization.	Confidentiality, intellectual property, trust in AI, and ownership of data for training GPT models in construction.	Improved materials selection and optimization efficiency through GPT models integrated with BIM, enhancing decision support for	Ethical use policies, novel applications in project management, and solutions for GPT model limitations in construction research.

									construction stakeholders.	
10.	Assessing the Strengths and Weaknesses of Large Language Models	Oct 2023		Journal of Logic, Language and Information-Springer			Transformers' inductive capacity, reinforcement learning with LLMs, and incorporation of syntactic tree structure into BERT.	The use of truncation of matrices and the need for additional learning biases for LLMs.	The achievement of subtle inductive inferences, the ability to identify hierarchical syntactic structure and complex semantic relations.	Experimenting with additional learning biases for LLMs and developing smaller, more lightweight models, among other research concerns.
11.	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	2020	Advances in Neural Information Processing Systems 33 (NeurIPS 2020)		Natural Questions (NQ) and MS-MARCO.	Passage retrieval and tokenization.	Retrieval-augmented generation (RAG) models.	Limited to tasks where relevant documents can be retrieved from a large corp	Improved performance on various knowledge-intensive NLP tasks	Exploring more complex retrieval strategies and incorporating external knowledge sources.
12.	Self-RAG: Learning To Retrieve, Generate, And Critique Through Self-Reflection	Oct 2023	arXiv preprints		SELF-RAG is assessed across diverse datasets, including PopQA, PubHealth, and ASQA, to gauge its performance in different downstream	SELF-RAG prompts GPT-4 with "##"-separated instruction and demonstration pairs (max 200 tokens). Non-compliance results	SELF-RAG elevates LLM performance with seed dataset training, reflection tokens for task adaptation, and a self-reflective mechanism for adaptive	SELF-RAG enhances model performance, factuality, and citation accuracy; however, a recognized limitation	SELF-RAG experiments show substantial improvements in model performance, factuality, and citation	Future work on SELF-RAG aims to rectify factual errors, improve performance with expanded training data, and

					tasks.	in discarding, and manual analysis checks prediction accuracy.	retrieval and generation aligned with specific task needs.	is the potential for factual errors in outputs, signaling a need for continued refinement in addressing this aspect.	accuracy, especially with larger training datasets in PopQA and ASQA. Human evaluations affirm the outputs' plausibility and evidence support.	refine self-reflection mechanisms. Its versatility extends to other domains such as summarization and dialogue generation.
13.	[Information Retrieval meets Large Language Models: A strategic report from Chinese IR community]()	Aug 2023		Science Direct - Elsevier			The authors propose a framework for integrating LLMs into IR systems and discuss the potential benefits of using LLMs for personalized and conversational IR, as well as knowledge refinement.	The need for large amounts of training data, potential for bias, and credibility concerns.	Benefits of integrating large language models (LLMs) into information retrieval (IR), highlighting the need for improved evaluation metrics, personalized and conversational IR, and knowledge refinement.	Better evaluation metrics, more effective integration of LLMs into IR systems, and improved personalized and conversational IR.

14.	Benchmarking Large Language Models in Retrieval-Augmented Generation	Sep 2023	Arxiv Preprints		RGB dataset assesses LLMs' retrieval-augmented generation abilities using aggregated news information across 4 testbeds.	Latest news prompts generate events, questions, and answers. Google's API fetches 10 relevant web pages, with dense retrieval models selecting the top-30 matching text chunks.	Evaluation uses Hugging Face's retrieval, paired with ChatGPT, for generating content from news articles.	Current LLM limitations in managing noisy documents and integrating information underscore the need for advancements in these areas.	Research reveals current LLM shortcomings, especially in noise robustness and information integration for retrieval-augmented generation tasks.	Improving LLMs by addressing noise, integrating information, and enhancing accuracy through better external knowledge integration.
15.	Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation	Nov 2023	arXiv preprints		Datasets: GSM8K (grade-school math), CommonsenseQA (common-sense scenarios), STaR (math problems with reasoning).	GSM8K: 5,000 training and 2,473 testing examples. Pyserini, a Python library, used dense representations with Faiss for retrieving reasoning chains.	ARM-RAG extends RAG, blending models with search engine knowledge using Neural Information Retrieval for grade-school math, powered by gpt-3.5-turbo via OpenAI API.	ARM-RAG improves, but untapped potential noted. Authors propose refining retrieval for enhancements and exploring abstract/classify methods in future research.	ARM-RAG outperforms baseline relying solely on LLMs, demonstrating positive influence from storing and retrieving reasoning chains in grade-school math problems.	Future research involves refining retrieval for enhanced performance and proposing methods to fully abstract or classify questions within a specific taxonomy.
16.	InPars: Data Augmentation for Information Retrieval using	Feb 2022	Arxiv Preprints		TREC-COVID dataset from BEIR: 50 queries, 1,326 judged documents per query,	Triple generation: randomly sample 100,000 documents, generate	Using large pretrained language models for IR tasks with few-shot capabilities achieves	Method skips pre-training for target corpus adaptation, avoids loss function	Models solely fine-tuned on unsupervised data outperform	Exploring the potential of domain-specific training data to further

	Large Language Models				171k articles from COVID-19 literature.	one question per document using GPT-3's Curie, and discard documents with less than 300 characters.	better zero-shot transfer than models fine-tuned only on supervised data.	modifications, less suited for non-neural retrieval algorithms.	BM25 and self-supervised methods. Retrievers fine-tuned on both supervised and synthetic data excel in zero-shot transfer.	improve the performance of neural models in information retrieval tasks
17.	Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA	Dec 2023	Association for Computational Linguistics - Proceedings of the Natural Legal Language Processing Workshop 2023		Korean Legal QA tasks use Korea Legal Aid Corp. and Korea Legislation Research Institute for QA pairs, with Law&Good's client cases as the query set.	GPT-4 generates questions for provisions and cases, with three answer types per query: legal professional, ChatGPT as a lawyer, and including relevant documents.	Main technique: Eval-RAG uses retrieval-based evaluation for LLM-generated texts, assessing based on the relevant retrieved document.	Document notes challenges with LLM-generated texts, including potential factual errors, motivating the proposal of Eval-RAG.	Eval-RAG excels in Korean Legal QA tasks, surpassing existing evaluation methods in correlating with human evaluation and identifying factual errors in LLM-generated texts.	Future work involves refining and validating Eval-RAG, exploring its applicability to other domains, and enhancing it to address identified limitations.
18.	Retrieval Meets Long Context Large Language Models	Oct 2023	Arxiv Preprints		Study includes seven datasets: Scroll, ELI5, FLAN, Open Assistant, Dolly, and a proprietary conversatio	The context documents are chunked into 300-word segments, and both questions and chunks are	The study employs retrieval-augmented language models and instruction tuning to train pretrained LLMs for question answering	-	Retrieval-augmented LLMs perform as well or better than fine tuned LLMs with	Authors propose combining retrieval and long context LLMs for higher accuracies and explorin

					nal dataset.	independently encoded with corresponding encoders	and text summarization		extended context windows, requiring less computation.	g their applications in natural language processing.
19.	Retrieval-Generation Synergy Augmented Large Language Models	Oct 2023	Arxiv Preprints		QA experiments on four datasets, including single and multi-hop tasks, compare the proposed method to GPT-3.5 and baselines.	Data preprocessed by retrieving top-5 paragraphs per query, setting max iterations (T) to 5. Answers evaluated using standard exact match metric (EM score).	Proposed framework comprises two crucial steps: generation-augmented retrieval and retrieval-augmented generation, forming a closed loop with mutual improvement through multiple iterations.	The proposed framework still has limitations in terms of scalability and efficiency	Empirical results show that the proposed method significantly improves the reasoning ability of large language models and outperforms previous baselines on all four datasets	Explore effective document retrieval and generation methods, investigating the impact of various strategies on the proposed framework's performance.
20.	Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models	Sep 2023	arXiv preprints		Introduced LLeQA dataset: 1,868 French legal questions annotated by professionals with detailed answers, referencing 27,942 statutory articles.	Used dense retriever to extract legislative articles; LLM processed them for detailed answers. Fine-tuning on LLeQA employed in-context learning and parameter-efficient fine-tuning strategies.	Utilized "retrieve-then-read" pipeline: dense retriever extracted legislative articles, LLM processed for detailed answers. Explored retrieval-augmented generator (RAG) approach using LLMs for legal question answering.	Results noted limitations: inaccuracies and potential hallucinations in answers and rationales. Cautioned conventional metrics may not precisely measure answer quality, possibly leading to misinterpretations.	"Retrieve-then-read" generally accurate but occasional fabricated facts. Retrieval-augmented LLMs perform well on metrics, yet qualitative analysis reveals inaccuracies and	Future work: enhance answer accuracy, interpretability, address hallucinations, and devise robust metrics. Emphasize responsible development and deployment of legal aid technologies.

									errors in responses.	
21.	Establishing Performance Baselines in Fine-Tuning, Retrieval-Augmented Generation and Soft-Prompting for Non-Specialist LLM Users	Nov 2023	arXiv preprints		Created corpus from publicly available info on LayerZero cryptocurrency bridging project. Used for fine-tuning model and creating vectorized RAG database.	Collected relevant info on LayerZero cryptocurrency bridging project via web search, split into paragraphs, converted to text, and vectorized in pkl format for RAG.	Paper tested LLM performance with unmodified GPT-3.5, fine-tuned GPT-3.5, and RAG using KIPLEY.AI. Also assessed the impact of soft prompts.	Study limited by dataset, time, and technical constraints. Focused on cryptocurrency, specific questions, potentially limiting generalizability.	RAG excelled over fine-tuning, superior to unmodified OpenAI model. Soft prompts enhanced all. Unmodified model often guessed LayerZero answers, errors providing insightful insights	Future work: explore larger datasets, advanced fine-tuning/RAG techniques, investigate soft prompt impact on LLM performance, and generalize results.

Table 2.1: Literature Survey Summary

CHAPTER 3: REQUIREMENT ANALYSIS AND SPECIFICATIONS

3. Requirement Analysis and Specification

3.1 Functional Requirements:

- **Disease Identification:**

  The system should use YOLOv8 for precise identification of phytopathologies in crops, focusing on diseases such as Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust.

- **Contextual Diagnosis:**

  GPT-3.5 Turbo with Retrieval Augmented Generation (RAG) should provide context-specific and comprehensive diagnostic information based on identified diseases.

- **Real-Time Monitoring:**

  Implement a real-time monitoring system that collects and analyzes data from sensor networks, including environmental factors like temperature, humidity, and soil moisture.

- **Dynamic Disease Mitigation:**

  The system should enable proactive disease mitigation by leveraging YOLOv8 and GPT-3.5 Turbo, allowing farmers to anticipate and prevent disease outbreaks.

- **Adaptability and Scalability:**

  Design the prototype to be adaptable to various crops beyond coffee production and scalable to accommodate future enhancements and features.

3.2 Non-Functional Requirements

- Performance:

  Response Time: The system should identify phytopathologies in crop images within 2 seconds.
- Reliability:

  Uptime: The system should be available 99% of the time for continuous access.
- Security:

  Data Security: Encrypt data at rest and in transit.

  User Access: Implement secure user authentication and authorization.
- Scalability:

  User Growth: Design the system to handle a growing user base.

  Crop and Disease Expansion: Easily accommodate additional crops and diseases.
- Usability:

  Intuitive UI: Design a user-friendly interface.

  Accessibility: Ensure the system is accessible to users with varying technical expertise.
- Maintainability:

  Documentation: Well-document code for future maintenance.

  Version Control: Use Git for code version control.
- Availability:

  24/7 Operation: The system should be available round the clock.

  Maintenance Communication: Communicate planned maintenance windows in advance.
- Auditability: Logging:

  Implement logs for user interactions, system events, and diagnostic outcomes.
- Interoperability:

  Data Compatibility: Ensure compatibility with common agricultural data formats and standards.
- Compliance:

  Regulatory Adherence: Adhere to data protection and privacy regulations.
- Adaptability:

  Environmental Changes: Design the system to adapt to changes in environmental conditions, crop types, and disease characteristics.

3.3 Hardware Requirements

Server Infrastructure:

- Processor: Minimum quad-core processor (e.g., Intel Core i5 or equivalent) to ensure efficient execution of YOLOv8 for phytopathology identification and GPT-3.5 Turbo with RAG for language model enhancement.

- RAM: 16GB RAM to support simultaneous processing tasks, especially during real-time monitoring and language model operations.

- Storage: A minimum of 500GB SSD storage for fast data retrieval and efficient model performance.

- Graphics Processing Unit (GPU): A dedicated GPU (e.g., NVIDIA GeForce GTX 1660 Ti or higher) to accelerate image processing tasks, particularly beneficial for YOLOv8's object detection.

Networking:

- High-Speed Internet Connection: A reliable and high-speed internet connection for seamless data retrieval in real-time monitoring and language model enhancement through RAG.

Backup and Redundancy:

- Backup Systems: Implement a backup solution to ensure data integrity and system stability.

- Redundancy Measures: Introduce redundancy for critical components to minimize downtime in case of hardware failures.

3.4 Software Requirements

- **Operating System:**

  Server OS: Ubuntu Server 20.04 LTS for a stable and well supported server environment.

- **Development Environment:**

  Python: Version 3.8 or higher for coding the AI algorithms and integration with models.

  Integrated Development Environment (IDE): Use a]\IDE such as Visual Studio Code or PyCharm for efficient development.

- **AI Model Frameworks:**

  YOLOv8 Framework: Install the YOLOv8 framework for object detection in images.

  GPT3.5 Turbo API: Access the OpenAI GPT3.5 Turbo API for language model integration and enhancement.

- **Web Framework:**

  Flask: Use Flask as the web framework for deploying and managing the AIdriven application.

- **Database Management:**

  Database System: MongoDB for storing relevant data, such as crop images and diagnostic information.

- **Version Control:**

  Git: Implement Git for version control to track changes, collaborate, and manage codebase versions effectively.

- **Networking:**

  RESTful API: Develop and implement a RESTful API for communication between the frontend and backend components.

  Socket.io: Use Socket.io for real time communication in monitoring systems.

- **Dependency Management:**

  pip: Utilize pip as the package installer for Python dependencies.

- **Security:**

  SSL/TLS: Implement SSL/TLS for secure data transmission.

  Firewall: Configure a firewall to control and monitor network traffic.

3.5 Cost Estimate Prototype:

The table below gives an estimated cost of each component of the proposed prototype.

Sl. no:	Item	Cost
1	Graphic Processing Unit (GPU)	On premises allotment would be needed
2	Cloud Backup Services	Rs. 2000-Rs15000

Table 3.1: Cost Estimation

- The GPU is a critical component for machine learning tasks, especially for training deep neural networks like GPT-2. It accelerates the computation required during the training process.
- Cloud backup services are crucial for ensuring data integrity and security. In the context of the project, they can be used for backing up datasets, models, and other critical project-related data.

CHAPTER 4: DESIGN

4.1 High-level design

Subsystem 1:

This subsystem focuses on the identification of plant diseases using the YOLOv8 (You Only Look One-level) object detection algorithm.

- **Components:**
- YOLOv8 Model: This component includes the pre-trained YOLOv8 model for detecting phytopathological symptoms in plant images.
- Image Preprocessing Module: Responsible for preprocessing input images to feed into the YOLOv8 model.
- Disease Classification Module: Extracts and classifies identified diseases based on YOLOv8 predictions.

Subsystem 2:
Remediation Assistance (LLM Enhancement with RAG Framework)

This subsystem aims to provide remediation suggestions using a Language Model (LLM) enhanced with a RAG (Retrieval-Augmented Generation) framework, similar to GPT-3.5T but with a focus on avoiding hallucinations.

- **Components:**
- Language Model (LLM): Utilizes a sophisticated language model (GPT-3.5T or similar) to generate contextually relevant suggestions for remediation.
- RAG Framework: Enhances the LLM by integrating a Retrieval-Augmented Generation framework, allowing for better context understanding and more accurate remediation suggestions.
- Database of Remediation Data: Stores a repository of information related to effective plant disease remedies, which can be retrieved and utilized by the RAG framework.

Subsystem 3:

Integration and Communication

This subsystem facilitates communication and integration between the Phytopathology Recognition and Remediation Assistance subsystems.

- **Components:**
- Communication Interface: Establishes a communication channel between the YOLOv8 subsystem and the LLM + RAG subsystem to exchange information.
- Decision Fusion Module: Integrates the disease identification results from YOLOv8 with the remediation suggestions from the LLM + RAG, ensuring a cohesive and effective
- recommendation for plant disease management.
- User Interface: Provides an interface for users to interact with the system, input images, view disease identification results, and receive remediation suggestions.

Workflow:

- The Phytopathology Recognition subsystem processes input images using YOLOv8, identifying and classifying plant diseases.
- The identified diseases are communicated to the Remediation Assistance subsystem.
- The LLM + RAG subsystem generates context-aware remediation suggestions based on the recognized diseases and relevant information from the remediation database.
- The Integration and Communication subsystem facilitate the exchange of information between the two subsystems.
- The Decision Fusion Module integrates disease identification and remediation suggestions to provide a comprehensive recommendation.
- The User Interface allows users to interact with the system, providing input images and receiving detailed reports on identified diseases and suggested remediation measures.

This high-level design ensures a seamless integration of YOLOv8 for accurate phytopathology recognition and an enhanced LLM with RAG for effective and contextually relevant remediation suggestions, all while avoiding hallucinations in the remediation process.

4.1.1 System Architecture

The detailed description of the system architecture has been given below along with the explanation of the subsystems.

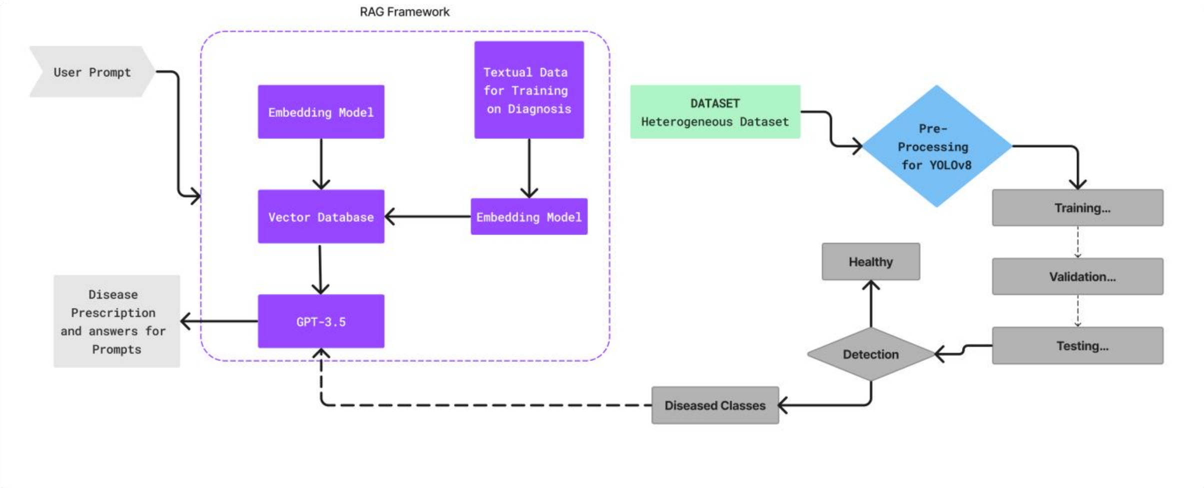


Figure 4.1: System Architecture of the proposed system

4.1.2 Abstract Specification of Subsystems

Subsystem 1:

Phytopathology Recognition (YOLOv8 Implementation)

1. **YOLOv8 Model:**

   Functionality: Performs object detection on input images to identify phytopathological symptoms.

   Specifications:

   - Utilizes YOLOv8 architecture with trained weights for plant disease detection.
   - Handles input images of various resolutions and formats.

2. **Image Preprocessing Module:**

   Functionality: Prepares input images for the YOLOv8 model.

   Specifications:

   - Resizes and normalizes input images to ensure consistency.
   - Applies any required augmentation techniques for improved model generalization.

3. **Disease Classification Module:**

   Functionality: Classifies identified diseases based on YOLOv8 predictions.

   Specifications:

   - Extracts region-of-interest (ROI) information from YOLOv8 output.
   - Maps ROIs to known plant diseases for accurate classification.
   - Provides confidence scores for disease classifications.

Subsystem 2:

Remediation Assistance (LLM Enhancement with RAG Framework)

1. **Language Model (LLM):**

   Functionality: Generates contextually relevant suggestions for remediation.

   Specifications:

   - Incorporates a sophisticated language model (e.g., GPT-3.5T) for natural language understanding and generation.
   - Supports contextual analysis to understand the specific context of the identified plant diseases.

2. **RAG Framework:**

   Functionality: Enhances the LLM with a Retrieval-Augmented Generation framework.

   Specifications:

   - Integrates a retrieval mechanism to fetch relevant information from a remediation database.
   - Augments the generation process with retrieved information to ensure contextually accurate suggestions.
   - Mitigates hallucinations by prioritizing retrieved information in the generation process.

3. **Database of Remediation Data:**

   Functionality: Stores information related to effective plant disease remedies.

   Specifications:

   - Maintains a structured database with information on plant diseases, symptoms, and corresponding remediation measures.
   - Supports efficient retrieval of relevant information based on disease classifications.

Subsystem 3:

Integration and Communication

1. **Communication Interface:**

   Functionality: Establishes communication channels between subsystems.

   Specifications:

   - Defines communication protocols for data exchange between the Phytopathology Recognition and Remediation Assistance subsystems.
   - Ensures secure and reliable transmission of data.

2. **Decision Fusion Module:**

   Functionality: Integrates disease identification results with remediation suggestions.

   Specifications:

   - Receives disease identification results from the Phytopathology Recognition subsystem.
   - Combines disease information with remediation suggestions from the Remediation Assistance subsystem.
   - Generates a comprehensive recommendation for plant disease management.

3. **User Interface:**

   Functionality: Provides an interface for user interaction.

   Specifications:

   - Supports user input of plant images for disease identification.
   - Displays detailed reports on identified diseases and suggested remediation measures.
   - Facilitates user feedback and input for system improvement.

4.1.3 Interface Design

The interface specifications, requirements and use cases are all mentioned below:

4.1.3.1. Home Page:

- Description: The Home Page serves as the landing page of the mobile app, providing users with access to different functionalities.
- Components:
  - Logo and App Name: The branding elements, including the app logo and name.
  - Navigation Buttons: Buttons that allow users to navigate to different sections of the app (Phytopathology Recognition, Remediation Assistance).
  - Brief Overview or Instructions: A concise description or instructions to guide users.

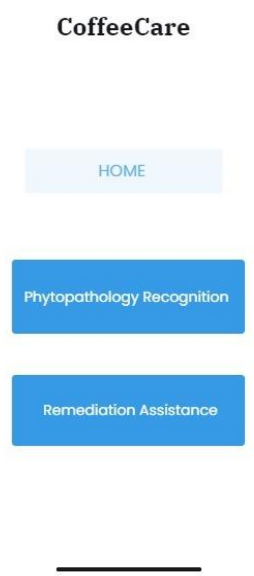


Figure 4.2: Home Page

4.1.3.2. Phytopathology Recognition Page:

- Description: This page allows users to input plant images for disease identification and view the results.
- Components:
  - Back Button: Allows users to return to the Home Page.
  - Image Upload Section: Enables users to upload images for disease identification.
  - Detection Results: Displays identified diseases and their confidence scores.
  - View Detailed Report Button: Allows users to access more detailed information about the identification results.

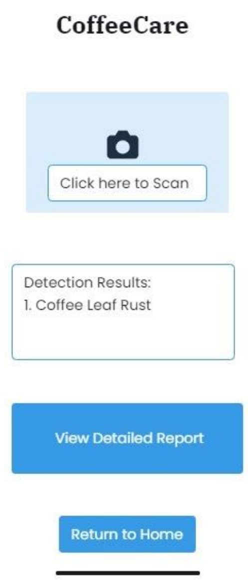


Figure 4.3: Phytopathology Recognition Page

4.1.3.3. Detailed Report Page:

- Description: Provides detailed information about the identified diseases, confidence scores, and potential remedies.
- Components:
  - Back Button: Allows users to return to the Phytopathology Recognition Page.
  - Disease List: Lists the identified diseases with their confidence scores.
  - Remediation Suggestions: Displays potential remedies for the identified diseases.
  - Remediation Button: Allows users to access the Remediation Assistance Page for further suggestions.

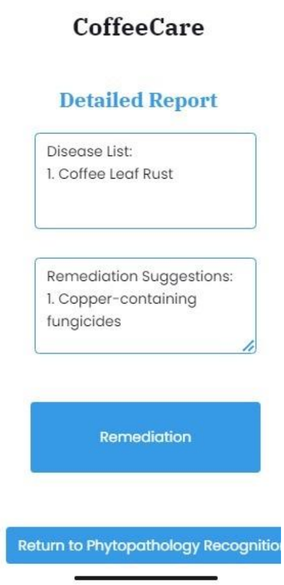


Figure 4.4: Detailed Report Page

4.1.3.4 Remediation Assistance Page:

- Description: This page provides context-aware remediation suggestions based on user input or identified diseases.

- Components:
  - Back Button: Allows users to return to the Home Page or the Detailed Report Page.
  - Disease Input: Text input field where users can describe the disease context or select from the identified diseases.
  - Remediation Suggestions: Displays context-aware suggestions for remediation.
  - View Detailed Button: Allows users to access more detailed information about the remediation suggestions.

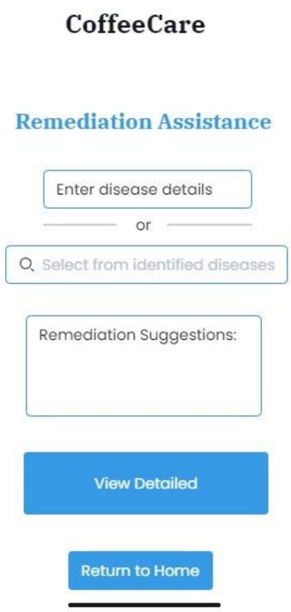


Figure 4.5: Remediation Assistance Page

4.1.3.5 QnA Page:

- Description: This page allows users to provide feedback on the system.
- Components:
  - Back Button: Allows users to return to the previous page.
  - Prompt Input: A text input field where users can prompt for further questions.

Figure 4.6: Q&A

4.1.3.6. Detailed Assistance Page:

- Description: Offers detailed information about the remediation suggestions, including context and sources.
- Components:
  - Back Button: Allows users to return to the Remediation Assistance Page.
  - Remediation Context: Provides context information for the suggested remediation.
  - Suggestion List: Lists the remediation suggestions along with their sources.

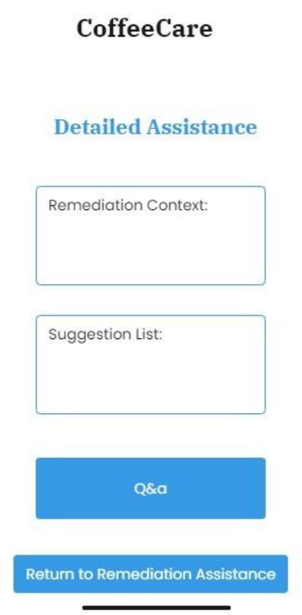


Figure 4.7: Detailed Assistance Page

4.2 Methodology

4.2.1. Phytopathology Recognition (YOLOv8 Implementation):

Algorithm Workflow:

- Input Image Preprocessing:
    - Normalize and resize input images.
    - Apply data augmentation techniques for improved model generalization.
- YOLOv8 Object Detection:
    - Use the pre-trained YOLOv8 model for plant disease detection.
    - Obtain bounding boxes and confidence scores for identified diseases.
- Disease Classification:
    - Extract region-of-interest (ROI) information from YOLOv8 output.
    - Map ROIs to known plant diseases for classification.
    - Assign confidence scores to disease classifications.

Strengths:

- YOLOv8 provides real-time object detection capabilities, making it suitable for applications where timely detection is crucial.
- The algorithm can handle multiple object classes simultaneously, allowing for the identification of various plant diseases in a single image.

Limitations:

- YOLOv8 may struggle with detecting small or heavily overlapped objects.
- The model's accuracy is highly dependent on the quality and diversity of the training dataset.

4.2.2. Remediation Assistance (LLM Enhancement with RAG Framework and Vector Database):

Algorithm Workflow:

- Language Model (LLM) for Natural Language Understanding:
    - Use a sophisticated language model (e.g., GPT-3.5T) to understand natural language input.
    - Analyze the context of identified plant diseases and generate initial remediation suggestions.

- RAG Framework Integration with Vector Database:
    - Employ a Retrieval-Augmented Generation (RAG) framework.
    - Utilize a vector database for efficient and context-aware information retrieval based on identified diseases.
    - Integrate retrieved vector representations into the generation process to enhance contextual accuracy.
- Context-Aware Remediation Suggestions:
    - Generate contextually relevant remediation suggestions, combining information from the LLM and the RAG framework.

Strengths:

- The vector database allows for efficient and vector-based information retrieval, enhancing the speed and relevance of suggestions.
- The language model's understanding of context allows for more accurate and contextually relevant suggestions.
- RAG integration ensures that generated suggestions are grounded in real-world remediation data.

Limitations:

- The quality of suggestions heavily relies on the comprehensiveness and accuracy of the vector database.
- The system may struggle with generating suggestions for rare or less-documented plant diseases.

4.2.3. Integration and Communication:

Workflow:

- Communication Interface:
    - Establish communication channels between the Phytopathology Recognition and Remediation Assistance subsystems.
    - Define protocols for secure and reliable data exchange.
- Decision Fusion Module:
    - Receive disease identification results from the Phytopathology Recognition subsystem.
    - Combine disease information with remediation suggestions from the Remediation Assistance subsystem.
    - Generate a comprehensive recommendation for plant disease management.

- User Interface:
    - Provide an interface for users to input plant images.
    - Display detailed reports on identified diseases and suggested remediation measures.
    - Facilitate user feedback and input for system improvement.

Technical Components:

- Communication Interface: RESTful API for communication between subsystems.
- Decision Fusion Module: Python script for data integration and recommendation generation.
- User Interface: Web-based interface using HTML, CSS, and JavaScript.

Technical Considerations:

- Frameworks/APIs:
    - YOLOv8: Leveraging the Darknet framework for YOLO implementation.
    - GPT-3.5T: Utilizing the OpenAI GPT-3.5T API for natural language understanding and generation.
    - RAG Framework with Vector Database: Utilizing a vector database (e.g., Faiss, Annoy) for storing and retrieving vector representations of remediation data.
- Storage:
    - Vector Database: Utilizing a vector database for storing and retrieving vector representations of remediation data.
    - Image Storage: File system or cloud storage for storing input images.
- Security:
    - Implement secure communication protocols (HTTPS) between subsystems.
    - Ensure data privacy and user confidentiality.
- Scalability:
    - Design subsystems to scale horizontally to accommodate increased load.
    - Consider cloud-based solutions for scalability and resource management.

Strengths and Limitations

- Strengths:
    - Real-time disease identification with YOLOv8.
    - Contextually relevant remediation suggestions with LLM, RAG, and vector database integration.
    - User-friendly interface for seamless interaction.

- Limitations:
  - YOLOv8 limitations in small or heavily overlapped object detection.
  - Quality of remediation suggestions depends on the comprehensiveness and accuracy of the vector database.
  - Potential challenges in generating suggestions for rare diseases.

This methodology incorporates the use of a vector database in the Remediation Assistance subsystem, providing more efficient and context-aware information retrieval for generating remediation suggestions. The strengths and limitations remain focused on providing a comprehensive AI-driven solution for phytopathology recognition and remediation, acknowledging potential challenges and areas for improvement.

CHAPTER 5: CONCLUSION

5.1 Conclusion

In conclusion, the proposed AI-driven precision agriculture system, leveraging YOLOv8 for disease identification and Retrieval Augmented Generation (RAG) for context-aware diagnosis, represents a transformative approach to revolutionize agricultural practices. The fusion of advanced object detection with language models offers a comprehensive solution to the challenges posed by diseases such as Brown Eye Spot, Coffee Wilt Disease, and Coffee Leaf Rust in Karnataka's crucial coffee production sector.

The innovative methodology not only addresses the inherent hallucination challenges associated with Large Language Models (LLMs) but also introduces dynamic disease identification and remediation strategies. By incorporating real-time monitoring, collaborative dataset expansion, and organizational involvement, the system ensures adaptability and relevance in diverse agricultural settings beyond coffee.

The envisaged impact extends beyond mere automation, aiming to secure food supplies, protect livelihoods, and promote eco-friendly farming practices. The reduction in pesticide use, facilitated by precise disease identification, contributes to sustainable and environmentally conscious agriculture.

As we look to the future, the project's scope expands into broader disease classes and envisions continuous development in RAG-integrated object detection systems. The emphasis on scalability, reliability, and usability positions this research as a catalyst for positive transformations in agriculture, aligning with global efforts toward sustainable and technologically enhanced food production.

In summary, the AI-driven precision agriculture system outlined in this research strives to be a beacon for positive change, ensuring the well-being of agricultural economies, fostering innovation, and providing a model for sustainable practices in the ever-evolving landscape of global agriculture.

References

1. Malone, B.S., 1963. *Some Solids Residence Time Studies in Tapered Gas Fluidised Beds*. University of Surrey (United Kingdom).
2. Qing, J., Deng, X., Lan, Y. and Li, Z., 2023. GPT-aided diagnosis on agricultural image based on a new light YOLOPC. *Computers and Electronics in Agriculture*, *213*, p.108168.
3. Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R. and McHardy, R., 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
4. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M. and Yao, H., 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.
5. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y. and Wen, J.R., 2023, December. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6449-6464).
6. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y. and Wang, L., 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*.
7. Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S.M., Chadha, A., Sheth, A.P. and Das, A., 2023. The Troubling Emergence of Hallucination in Large Language Models--An Extensive Definition, Quantification, and Prescriptive Remediations. *arXiv preprint arXiv:2310.04988*.
8. Rane, N.L., Tawde, A., Choudhary, S.P. and Rane, J., 2023. Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, *5*(10), pp.875-899.
9. Saka, A., Taiwo, R., Saka, N., Salami, B.A., Ajayi, S., Akande, K. and Kazemi, H., 2023. GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, p.100300.
10. Lappin, S., 2023. Assessing the Strengths and Weaknesses of Large Language Models. *Journal of Logic, Language and Information*, pp.1-12.
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, pp.9459-9474.
12. Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H., 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.
13. Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F.

and Gao, S., 2023. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, *4*, pp.80-90.

14. Chen, J., Lin, H., Han, X. and Sun, L., 2023. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.
15. Melz, E., 2023. Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation. *arXiv preprint arXiv:2311.04177*.
16. Bonifacio, L., Abonizio, H., Fadaee, M. and Nogueira, R., 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.
17. Ryu, C., Lee, S., Pang, S., Choi, C., Choi, H., Min, M. and Sohn, J.Y., 2023, December. Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023* (pp. 132-137).
18. Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M. and Catanzaro, B., 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.
19. Feng, Z., Feng, X., Zhao, D., Yang, M. and Qin, B., 2023. Retrieval-Generation Synergy Augmented Large Language Models. *arXiv preprint arXiv:2310.05149*.
20. Louis, A., van Dijck, G. and Spanakis, G., 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2309.17050*.
21. Dodgson, J., Nanzheng, L., Peh, J., Pattirane, A.R.J., Alhajir, A.D., Dinarto, E.R., Lim, J. and Ahmad, S.D., 2023. Establishing Performance Baselines in Fine-Tuning, Retrieval-Augmented Generation and Soft-Prompting for Non-Specialist LLM Users. *arXiv preprint arXiv:2311.05903*.

APPENDIX A: Details of publications

Author Names: Dr. Selva Kumar S, Afifah Khan Mohammed Ajmal Khan, Imadh Ajaz Banday, Manikantha Gada, Vibha Venkatesh Shanbhag

Paper Title: A comprehensive survey on LLM Challenges and RAG-Driven Precision in Coffee Leaf Disease Remediation.

Name of the Conference or Journal: ICETCS 2024 - BMSCE, IEEE Bangalore Section.

Place of the conference: B. M. S. College of Engineering, Bull Temple Rd, Basavanagudi, Bengaluru, Karnataka 560019

Conference Date: 25/04/2024

APPENDIX B : PROGRAM OUTCOMES

PO	Level (3/2/1) 3-High 2-Medium 1-Low	Justification if addressed
PO1	3	Engineering Knowledge: Applying computational algorithms and sensor integration to design an innovative, non-invasive glucose monitoring system based on robust engineering principles and computer science expertise.
PO2	2	Problem Analysis: Identifying and assessing the challenges associated with continuous glucose monitoring in the elderly population, leveraging data-driven approaches and technological advancements in computer science.
PO3	3	Design/Development of Solutions: Creating a user-centric, non-intrusive glucose monitoring solution tailored to the specific needs of the elderly, considering user interface design and seamless integration of sensor technology.
PO4	2	Conducting Investigations: Employing data analysis methodologies and experimental validation to ensure accuracy and reliability in glucose measurements, aligning with the project's scientific and engineering objectives.
PO5	3	Modern Tool Usage: Utilizing cutting-edge software tools and machine learning models for predictive analytics, optimizing the monitoring device's efficiency and performance.
PO6	2	Engineering and Society: Addressing healthcare accessibility and ethical considerations in developing technology for elderly care, ensuring the device's usability and adherence to safety standards.
PO7	2	Environment and Sustainability: Emphasizing sustainable design principles in the device's manufacturing and disposal processes to minimize environmental impact.
PO8	2	Ethics: Upholding ethical standards in data privacy, ensuring informed consent, and maintaining transparency in data handling and usage.
PO9	3	Individual and Teamwork: Collaborating within a multidisciplinary team, fostering diverse perspectives to innovate and refine the glucose monitoring system effectively.
PO10	2	Communication: Effectively communicating technical concepts to stakeholders, including potential users, healthcare providers, and the scientific community, ensuring clarity in instructions and documentation.
PO11	2	Project Management and Finance: Efficiently managing project resources, considering budgetary constraints, and implementing effective project timelines to achieve successful device development.
PO12	3	Lifelong Learning: Embracing continuous learning to adapt to technological advancements, exploring emerging research, and staying updated with advancements in glucose monitoring technology for ongoing improvements.

PROGRAM SPECIFIC OUTCOMES

PSO	Level (3/2/1) 3-High 2-Medium 1-Low	Justification if addressed
PSO1	3	Applying Software Engineering Principles: Implementing rigorous software engineering practices to develop a robust software system that interfaces seamlessly with the glucose monitoring hardware, ensuring reliability and scalability.
PSO2	2	Designing and Developing Network, Mobile, and Web-based Systems: Creating an integrated system that allows for remote monitoring of glucose levels via mobile or web interfaces, adhering to real-world constraints such as data security, accessibility, and user-friendliness for the elderly population.
PSO3	3	Designing Efficient Algorithms and Developing Effective Code: Crafting algorithms for data processing and analysis that optimize resource usage while maintaining accuracy in glucose level predictions. Developing clean and efficient code to ensure the device operates effectively within the computational constraints of the hardware.

APPENDIX C: Plagiarism Report

DrillBit

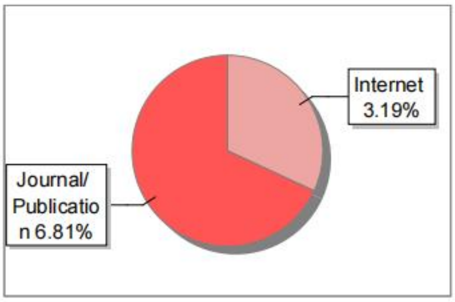
The Report is Generated by DrillBit Plagiarism Detection Software

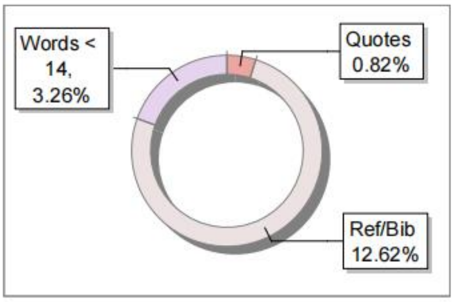
Submission Information

Author Name	Imadh
Title	PP1-Report
Paper/Submission ID	1277709
Submitted by	hod.cse@bmsce.ac.in
Submission Date	2023-12-26 14:07:56
Total Pages	47
Document type	Thesis

Result Information

Similarity **10 %**





Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words %	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File