

# A Novel Explainable LLM-based Why-QA Framework for Climate Resilient and Sustainable Smart Agriculture

**Manvi Breja**

Department of CSE, The Northcap University, Gurugram, India

E-mail: [manvibreja@ncuindia.edu](mailto:manvibreja@ncuindia.edu)

ORCID iD: <https://orcid.org/0000-0003-0607-3094>

Received: April 29, 2026; Revised: June 8, 2026; Accepted: July 10, 2026; Published: August 8, 2026

**Abstract:** Sustainable smart agriculture and climate resilience are significant factors for maintaining food security with environmental management. Existing agricultural systems target predicting and generating reports but lack detailed explanations as to why the phenomenon occurs. The examples of such why are like “Why there is a significant decline in the yield?”, “Why the soil is degrading?”, “Why the level of water is declining?” and so on. To address this challenge, the paper presents a prototype for Explainable LLM based Why-QA framework for sustainable smart agriculture. The implemented prototype integrates domain-enriched LLM with knowledge graph, causal inference engine with explainability layer to provide detailed explanations to complex “why-questions”. Domain-specific LLMs are used to support the domain knowledge with knowledge graph analyzing the sustainable relationships in the answer, causal reasoning to produce the causes of events and explainability module to provide the detailed reasoning supported with benchmark sustainable metrics incorporating the soil, climate and crop data. The prototype implementation of framework is evaluated on a dataset of 500 annotated agricultural why-type questions constructed from FAOSTAT, USDA, and NOAA sources. Results clearly demonstrate promising improvements over five baseline systems developed across causal reasoning and explanation quality metrics, which help validating the architectural feasibility of the proposed framework.

**Index Terms:** Explainable AI, Question Answering System, Sustainability, Knowledge Graph, Smart Agriculture, Causality, LLM

## 1. Introduction

This section introduces the need of explainable why-QAS that can address the questions related to sustainability and agriculture. Further different research gaps are discussed which are required to be addressed through the proposed framework.

### 1.1. Background and Motivation

Agriculture plays a crucial role in ensuring the food security, environmental sustainability, and economic development. However, modern agricultural systems are increasingly affected by climate variability, soil degradation, water scarcity, pest outbreaks, and inefficient resource utilization. Existing agricultural QA systems have utilized AI and ML techniques to predict the crop yield and make decisions to forecast outcomes [27] but there is hardly any system which can provide explanations to the cause of agricultural phenomenon.

In real scenarios, the need for farmers and decision makers is to understand the real cause and effect of the agricultural event to predict the futuristic outcome [3]. For example, the answers to the questions like “Why did crop yield decrease?”, “Why is soil fertility declining?”, or “Why is irrigation demand increasing?” require detailed causal explanations to be supported by domain knowledge and based evidence. There are existing large language model (LLM)-based systems which can generate responses to these questions, but they usually suffer from hallucinations and provide limited causal reasoning. Such limitations to the existing systems reduce the user trust and thus prevent the actual deployment for decision-support systems.

Therefore, there is a growing need for explainable and causality-aware question-answering framework that can generate transparent, evidence-supported, and sustainability-oriented explanations for agricultural why-type questions. Thus, this paper proposes a novel Explainable LLM-based Why-Question Answering Framework (Why-QAS) for climate-resilient and sustainable smart agriculture. The framework integrates a domain-specific LLM, an agricultural

knowledge graph, a causal inference engine, a sustainability evaluation module, and an explainability layer within a unified multi-stage reasoning pipeline. The primary contribution lies in the novel integration of KG-validated hypothesis generation, quantitative causal reasoning, sustainability-aware assessment, and SHAP-based explainability capabilities that no existing agricultural QA system jointly addresses.

Further, different research gaps that are generally faced in the domain of agriculture why-QAs are discussed which are tried to be overcome through research contributions discussed in Section 2.

### **1.2. Research Gaps and Objectives Addressed**

This section discusses significant research gaps that are faced in the field of agriculture domain detailed as follows:

#### **RG1: LLM Hallucination in Domain QA**

In domains like agriculture, where there is a lack of ground knowledge and highly dependability on memory, large language models generate hallucinated, fabricated and unsupported information. Such hallucinations result in misleading decisions and recommendations for the tasks like sustainability reports or advising crop growth etc. Existing conventional QAS lack mechanisms to verify and validate the output generated which are resolved by our framework through incorporating RAG and knowledge graphs.

#### **RG2: Lack of Causal Reasoning**

Conventional QAS depends on correlations patterns rather than generating cause-effect relationship graphs. They find difficult to answer factual questions like “What happens if water level goes down?”. Without modeling causality in the answers, responses appear descriptive rather than factual justified. Thus, it limits the usability in the answering of agricultural and sustainability analysis.

#### **RG3: Limited Explainability**

Existing LLM based QAS cannot provide structured detailed explanations as to how answers are generated. Users cannot trace the ground evidence as to how the reasoning is provided in the answers. This lacks trust and accountability in answer generation. Proposed framework employs attention mechanism to support the evidence involved in generating the appropriate answers to the question.

#### **RG4: Few Sustainability-Enhanced QAS**

Currently QAS models are optimized for accuracy and efficiency but rarely has any model considered environmental and sustainability metrics. These QAS could not model carbon footprint, usage of water, water scarcity level or efficient usage of resources. Thus, the relation between environmental impact and productivity left unaddressed. In recent era, sustainability has become an integrated component which is considered as a foundation in the proposed framework.

#### **RG5: Unavailability of Agriculture Benchmark Datasets**

There are no standardized benchmark datasets available for agriculture and sustainability-enhanced QAS. Existing datasets consider general knowledge but don't capture domain-specific knowledge which makes it difficult to systematically evaluate and compare. This hinders the growth of research in the agricultural domain. To address this gap, this paper tries to integrate and utilize multi-domain datasets to be evaluated accurately.

Further, after identifying the above research gaps, the paper tries to fulfil the following research objectives briefly discussed as follows:

**RO1: Designing an explainable LLM-enhanced QAS** for answering causal and reasoning why-type questions from agriculture domain.

**RO2: Integrating multi-domain agricultural datasets** capturing data related to soil, climate, crop, pest, and land features etc. to develop one consolidated dataset.

**RO3: Incorporating causal reasoning** to inculcate the cause-effect relations in the question-answer pairs.

**RO4: Transparent explanations** are generated in the answers which are grounded and justified by the ground evidences.

**RO5: Evaluating sustainability impacts and metrics** from the question-answer generation.

Table 1 maps each identified gap to its corresponding research objective. With each mapping, evaluation criteria is also highlighted used to validate each objective.

Table 1. Research Gap-Research Objective Mapping Table

| Research Gap                   | Research Objective                | Evaluation criteria used in framework                                                                  |
|--------------------------------|-----------------------------------|--------------------------------------------------------------------------------------------------------|
| RG1: LLM Hallucination         | RO1: Explainable LLM-enhanced QAS | Hypothesis validation through knowledge graphs                                                         |
| RG2: Lack of Causal Reasoning  | RO3: Inculcating causal reasoning | Implementing causal module through Bayesian Networks representations                                   |
| RG3: Limited Explainability    | RO4: Transparent explanations     | Evaluation metrics like Explanation Completeness Score (ECS), Evidence Grounding Score (EGS), and SHAP |
| RG4: No Sustainability QAS     | RO5: Sustainability metrics       | Composite Sustainability Index (CSI) scoring                                                           |
| RG5: No Agriculture Benchmarks | RO2: Multi-domain datasets        | Collecting data from FAOSTAT + USDA + NOAA datasets                                                    |

The remainder of this paper is organized as follows: Section 2 describes each research contribution with hypothesis. Section 3 presents the literature review. Section 4 describes the proposed framework and its modules. Section 5 details the implementation of each module. Section 6 presents the results and discussion. Section 7 finally concludes the paper with future research directions.

## 2. Research Contributions and Hypotheses

Section 2 discusses each of the research contributions corresponding to the research gaps discussed in section 1. Based on each of the identified research gap in section 1, the research hypotheses are formulated. Based on the identified research gaps and objectives, the following measurable research hypotheses are formulated. Further section 6 maps each hypothesis to the evaluation metric which validates the practicality of the proposed framework.

### 2.1. Research Contributions

This paper proposes a framework for Explainable LLM-based Why-QA systems which are designed to answer questions based on climate-resilient and sustainable smart agriculture. The prime focus of this paper is on the integration of causal reasoning supported by evidence, assess the sustainable contributions and generating explainable answers within the unified agricultural why-QA systems. Further, there are four major research contributions which are implemented to address the research gaps discussed in section 1 are detailed below:

**RC1: Agriculture-Specific Causal Why-QA Framework:** A novel framework is proposed that specifically targets answering non-factoid why-type questions in the agricultural domain — such as "Why did crop yield decline?" and "Why is soil degrading?" by introducing domain-specific LLM reasoning with structured causal inference. Unlike existing QAS that addresses factoid question retrieval, this framework is designed for complex causal explanation generation which is supported by foundational agricultural domain knowledge.

**RC2: KG-Gated Hypothesis Validation Pipeline:** Rather than directly generating answers from the LLM, the framework first generates multiple causal hypotheses which are then validated through knowledge graph traversal. Only hypotheses which are supported by valid causal paths in the agricultural knowledge graph are passed to next module for reasoning. This module specifically addresses the hallucination problem RG1 mentioned in section 1 by preventing unsupported causal claims from propagating through the pipeline.

**RC3: Sustainability-Aware Answer Generation:** The framework comprises a sustainable evaluation module that evaluates the Composite Sustainability Index (CSI) score by integrating sustainable scores like Water use efficiency (WUE), Nitrogen use efficiency (NUE), Carbon Emission Index (CEI), profitability score and farmer resilience index that helps in assessing the impact of environment, economy and social factors. This directly addresses RG4, which is not incorporated by any existing Why-QA systems.

**RC4: Multi-Layer Explainability with SHAP Attribution:** The framework generates interpretable answers by combining causal-path reasoning, evidence grounding, sustainability impact assessment, and SHAP-based feature attribution which enable farmers and decision-makers to trace exactly and retrieve evidence why the answer is generated and which factors have contributed most significantly.

### 2.2. Research Hypothesis

Based on each research contribution discussed above, the following research hypotheses are formulated. Each hypothesis is directly linked to a specific evaluation metric which is calculated in Section 6, enabling rigorous empirical validation of the proposed framework.

#### H1: Knowledge Graph Validation helps improving Causal Reasoning Accuracy

*"The integration of knowledge graph-based hypothesis validation into the LLM reasoning pipeline significantly*

*improves root cause identification accuracy compared to a baseline LLM without KG validation."*

- **Addresses:** RG1 (hallucination), RG2 (causal reasoning), RO1, RO3
- **Evaluation Metric:** Root Cause Accuracy, Precision, Recall, F1-score (Table 2)
- **Validation criterion:** Proposed framework achieves statistically significant improvement over Baseline LLM and LLM+KG in Root Cause Accuracy

## H2: Causal Reasoning Integration Improves Explanation Completeness

*"Incorporating explicit causal inference through Structural Causal Models and Bayesian Networks into the answer generation pipeline significantly improves the completeness and grounding of generated explanations compared to systems without causal reasoning."*

- **Addresses:** RG2 (causal reasoning), RG3 (explainability), RO3, RO4
- **Evaluation Metric:** Explanation Completeness Score (ECS), Evidence Grounding Score (EGS) (Table 2)
- **Validation criterion:** Proposed framework achieves  $ECS \geq 0.85$  and  $EGS \geq 0.85$  which shows the improvement over non-causal systems.

## H3: Sustainability-Aware Reasoning Improves Human Interpretability

*"The amalgamation of sustainable evaluation metrics like WUE, NUE, CEI, CSI for generating explainable answers help improving the human interpretability ratings compared to systems which don't incorporate sustainable reasoning."*

- **Addresses:** RG3 (explainability), RG4 (sustainability), RO4, RO5
- **Evaluation Metric:** Human Interpretability Rating (1–5 scale) (Table 2)
- **Validation criterion:** Proposed framework achieves Human Interpretability Rating  $\geq 4.5$ , indicating strong alignment between sustainability-grounded explanations and expert judgment

Table 2. Hypothesis–Metric Mapping Table

| Hypothesis                                             | Research Gaps Addressed | Evaluation Metric             | Target      | Expected Result |
|--------------------------------------------------------|-------------------------|-------------------------------|-------------|-----------------|
| H1: KG validation improves causal accuracy             | RG1, RG2                | Root Cause Accuracy           | $\geq 80\%$ | 85% (Table 2)   |
|                                                        |                         | F1-Score                      | $\geq 0.80$ | 0.84 (Table 2)  |
| H2: Causal reasoning improves explanation completeness | RG2, RG3                | ECS                           | $\geq 0.85$ | 0.89 (Table 3)  |
|                                                        |                         | EGS                           | $\geq 0.85$ | 0.91 (Table 3)  |
| H3: Sustainability reasoning improves interpretability | RG3, RG4                | Human Interpretability Rating | $\geq 4.5$  | 4.6 (Table 3)   |

Target thresholds were established based on domain expert consultation and review of comparable QA system benchmarks prior to experimentation. Reported values reflect prototype evaluation on the 100-question test set retrieved from 500 why-questions dataset.

## 3. Literature Review

This section discusses some of the significant research work carried out in why-question answer answering. The papers highlighting the role of explanations with different domains are briefed in Table 3 below:

Table 3. Literature Review

| Paper Title                                                                | Datasets Used                                          | Challenges Addressed                   | Gaps Identified                         | Techniques Used                                     | Results & Metrics                | Future Direction                   |
|----------------------------------------------------------------------------|--------------------------------------------------------|----------------------------------------|-----------------------------------------|-----------------------------------------------------|----------------------------------|------------------------------------|
| Analyzing Linguistic Features for Answer Re-Ranking of Why-Questions [1]   | Yahoo Answers, Quora, Answers.com (1,000 why-QA pairs) | Ranking answers for why-type questions | No deep causal reasoning                | Lexical, syntactic & semantic features + ExtraTrees | $MRR \approx 0.64$               | Add contextual & causal embeddings |
| Learning to Rank Answers to Non-Factoid Questions from web collections [2] | Web QA collections                                     | Long-answer ranking                    | No structured causality                 | Linguistic ranking features                         | 14–21% improvement (MRR, Prec@1) | Integrate semantic/causal models   |
| Creating Causal Embeddings for QA with minimal supervision [4]             | Yahoo Answers causal pairs                             | Improve causal relevance               | No KG or RAG integration                | Bootstrapped causal embeddings                      | ~7–8% rel. improvement (P@1)     | Combine with LLM + KG              |
| A Survey on Non-Factoid Question Answering Systems [5]                     | N/A (Survey)                                           | Overview of why-QA research            | Lack of benchmarks & causal integration | Comparative analysis                                | N/A                              | Unified causal QA frameworks       |

|                                                                                                                                                  |                                                                                                 |                                                                                        |                                                                                            |                                                                                                            |                                                                                                                              |                                                                                                               |
|--------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Causality for Question Answering [6]                                                                                                             | Conceptual                                                                                      | Role of causality in QA                                                                | No LLM-based implementation                                                                | Lexico-semantic causal modeling                                                                            | Conceptual                                                                                                                   | Neural causal QA models                                                                                       |
| CausalQA: A Benchmark for causal question answering[7]                                                                                           | 1.1M causal QA pairs                                                                            | Causal QA benchmarking                                                                 | Weak baseline models                                                                       | Dataset typology +                                                                                         | ROUGE-L $\approx 0.48$                                                                                                       | Advanced causal QA models                                                                                     |
| Developing a Why-How Question Answering system on community web boards with a causality graph including procedural knowledge [8]                 | Community boards (plant disease)                                                                | Domain why/how QA                                                                      | No LLM, limited explainability                                                             | Causality graph + ML classifier                                                                            | $\sim 91\%$ accuracy                                                                                                         | Integrate LLM & explanations                                                                                  |
| QED: A Framework and Dataset for Explanations in QA [9]                                                                                          | Natural Questions subset                                                                        | Explainable QA                                                                         | Not domain-specific                                                                        | Supervised explanation modeling                                                                            | Improved QA with explanations                                                                                                | Domain adaptation                                                                                             |
| MinosEval: Distinguishing Factoid and Non-Factoid for Tailored Open-Ended QA Evaluation with LLMs [10]                                           | Multi-domain QA datasets                                                                        | Factoid vs non-factoid evaluation                                                      | No reasoning integration                                                                   | Evaluation metric redesign                                                                                 | Better human alignment                                                                                                       | Causal evaluation metrics                                                                                     |
| Knowledge Graph-Extended RAG for QA[11]                                                                                                          | MetaQA                                                                                          | Reduce hallucination                                                                   | Limited why-type QA                                                                        | KG & RAG decomposition                                                                                     | Improved multi-hop accuracy                                                                                                  | Domain-specific QA                                                                                            |
| Document graphrag: knowledge graph enhanced retrieval augmented generation for document question answering within the manufacturing domain. [12] | SQuAD, HotpotQA                                                                                 | Structured retrieval                                                                   | No causal reasoning                                                                        | Graph-based RAG                                                                                            | Improved retrieval relevance                                                                                                 | Domain + explainable QA                                                                                       |
| Leveraging explainable AI for sustainable agriculture: a comprehensive review of recent advances [13]                                            | Agricultural ML datasets                                                                        | XAI in agriculture                                                                     | Not integrated with QA                                                                     | SHAP, LIME, Grad-CAM                                                                                       | Improved interpretability                                                                                                    | Apply XAI to LLM-based QA                                                                                     |
| Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [21]                                                                            | Natural Questions (NQ), WebQuestions, TriviaQA, Jeopardy                                        | Reducing hallucinations and improving factual grounding in QA systems                  | Limited support for causal reasoning and explainability                                    | Dense Passage Retrieval (DPR) + Retrieval-Augmented Generation (RAG)                                       | Improved Exact Match and F1-score across multiple QA benchmarks compared with parametric-only language models                | Integrate causal reasoning, domain knowledge graphs, and explainable answer generation                        |
| Finetuned Language Models Are Zero-Shot Learners (FLAN) [22]                                                                                     | More than 60 NLP benchmark tasks including QA, reasoning, classification, and dialogue datasets | Improving instruction following and zero-shot generalization across diverse NLP tasks  | Limited evidence grounding and lack of domain-specific causal validation                   | Multi-task Instruction Tuning (FLAN) using prompt-based training                                           | Significant improvement in zero-shot and few-shot performance over conventional pretrained language models                   | Combine instruction tuning with retrieval augmentation, domain adaptation, and explainable reasoning          |
| GraphRAG: Unlocking LLM Discovery on Narrative Private Data [23]                                                                                 | Narrative enterprise documents, private textual repositories, and graph-constructed corpora     | Improving multi-hop retrieval and contextual reasoning over large document collections | No explicit causal-effect modeling, sustainability reasoning, or explainability assessment | Knowledge Graph Construction + Community Detection + Graph-based Retrieval-Augmented Generation (GraphRAG) | Improved retrieval quality, contextual 306 completeness, and information discovery compared with conventional RAG approaches | Integration of causal inference, explainability mechanisms, and domain-specific decision-support applications |
| Self-Consistency Improves Chain of Thought Reasoning in Language Models [24]                                                                     | GSM8K, StrategyQA, MultiArith and reasoning benchmarks                                          | Improving reasoning reliability in LLMs                                                | Limited domain adaptation and evidence grounding                                           | Chain-of-Thought + Self-Consistency decoding                                                               | Significant improvement across reasoning benchmarks                                                                          | Integration with domain knowledge and explainable AI                                                          |
| AI for Crop Production – Where Can Large Language Models (LLMs) Provide Substantial Value? [25]                                                  | Literature and agricultural knowledge sources                                                   | Crop advisory, decision support, precision agriculture                                 | Hallucinations, limited domain knowledge, low explainability                               | LLMs, NLP, prompt engineering, literature review                                                           | Improved decision support; qualitative evaluation                                                                            | Domain-specific LLMs, IoT integration, improved accuracy and explainability                                   |
| Harnessing Agentic AI for Sustainable Innovation and Environmental Responsibility [26]                                                           | Sustainability case studies and conceptual examples                                             | Resource optimization, sustainability, autonomous decision-making                      | Ethical issues, governance, scalability, limited validation                                | Agentic multi-agent AI, systems, conceptual framework                                                      | Better resource efficiency; conceptual results                                                                               | Explainable Agentic AI, governance, real-world deployment, IoT integration                                    |

From the above literature review, it is evident that although substantial progress has been made in the individual areas of retrieval-augmented generation, knowledge-graph-enhanced reasoning, explainable artificial intelligence, and causal question-answering systems, but still suffers from several limitations. RAG systems improve the factual answer generation with minimal hallucinations but lack casual reasoning. Although it is expensive to maintain the knowledge graphs, the approaches help provide semantic reasoning. Causal relationships in QAS model the cause effect relationships and improve the reasoning capabilities in the answer. Explainable AI helps provide transparency in answer generation. Graph-enhanced retrieval methods improve the multi-hop reasoning through structured knowledge representations. Further, agent-based AI systems have demonstrated the need to plan, monitor and recommend management tasks in agricultural domain. These individual advancements highlight the importance of integrating retrieval grounding, explainability, and intelligent decision-support capabilities within a unified domain-specific agricultural why-QA framework. Table 4 provides a comparison of our proposed framework with existing traditional systems, RAG based QA, KG-RAG, causal QA and explainable QA.

Table 4. Comparison of our proposed framework with existing traditional systems, RAG based QA, KG-RAG, causal QA and explainable QA.

| Type of QA         | Scalability | Explainability Quality | Computational Complexity | Deployment Limitation            |
|--------------------|-------------|------------------------|--------------------------|----------------------------------|
| Traditional Why-QA | High        | Low                    | Low                      | Limited causal reasoning         |
| RAG-based QA       | High        | Low–Moderate           | Moderate                 | Weak causal explanations         |
| KG-RAG QA          | Moderate    | Moderate               | High                     | Graph maintenance overhead       |
| Causal QA          | Moderate    | Moderate–High          | High                     | Limited sustainability reasoning |
| Explainable QA     | Moderate    | High                   | Moderate                 | Weak domain adaptation           |
| Proposed Framework | Moderate    | High                   | Moderate–High            | Requires domain KG construction  |

Comparative analysis reveals that existing approaches generally optimize one aspect of intelligent question answering, such as retrieval accuracy, causal inference, or explainability, while neglecting other important dimensions. Sustainability-aware reasoning remains largely unexplored in agricultural Why-QA systems. Furthermore, few studies provide mechanisms for simultaneously validating generated hypotheses, identifying causal factors, assessing sustainability impacts, and producing interpretable explanations. These limitations motivate the development of the proposed Explainable LLM-based Why-QA framework. Table 5 further compares each of the capabilities addressed in different types of QA.

Table 5. Comparison of QA based on different capabilities

| Capability                | RAG     | KG-RAG  | CausalQA | Explainable QA | Proposed |
|---------------------------|---------|---------|----------|----------------|----------|
| Agricultural Adaptation   | Partial | Partial | Partial  | Partial        | ✓        |
| Causal Inference          | ×       | ×       | ✓        | ×              | ✓        |
| Sustainability Assessment | ×       | ×       | ×        | ×              | ✓        |
| Evidence Grounding        | ✓       | ✓       | Partial  | Partial        | ✓        |
| Explainability            | Partial | Partial | Moderate | High           | High     |

#### 4. Proposed Work

RAG and knowledge graphs approaches help in reducing hallucinations but they have shown large contribution in factoid QAS rather than non-factoid QAS. However, agricultural AI addresses predictions but rare studies incorporated explainable, reasoning and justified sustainable-enhanced QA systems. This motivated the need to implement and develop an integrated explainable AI-LLM-based why-QA framework that integrates RAG, knowledge graph, causal inference, explainability layer illustrated in Figure 1.

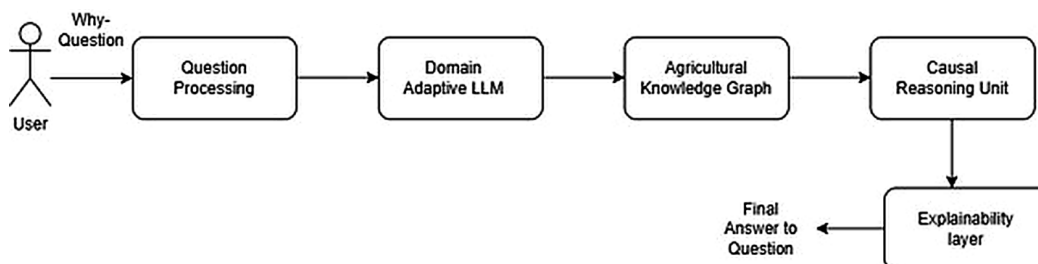


Fig. 1. Architecture of Explainable LLM-based why-QAS for Sustainable Agriculture

The framework depicts the explainable and causality-enhanced QAS pipeline. It transforms the why-questions asked by the user into grounded, reasoned and interpretable answers by employing structured knowledge with sustainability reasoning and explanations. In this, the framework is adapted to utilize domain-specific LLM to answer agricultural why-questions, knowledge graph to represent sustainable parameters, causal reasoning module to infer the causes of anomalies and explainability layer to deduce the detailed reasoning with benchmark sustainability evaluation metrics. The framework is tested on the dataset comprising 500 why-type agriculture questions deduced from multiple public datasets comprising soil, climate, and crop yield data. The detailed flow of information in each module is illustrated as below in Figure 2:

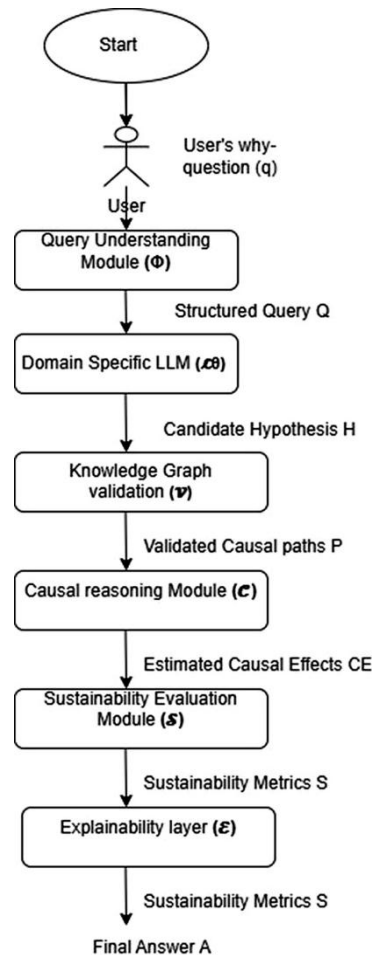


Fig. 2. Detailed Flow of Information in each module of the proposed architecture

#### 4.1. Query Understanding Module:

Query Understanding Module represents the users' input natural language why-type question into a machine-interpretable form whose components are illustrated in Figure 3. Why-questions related to agricultural domain target various factors as crop type, soil properties, climate conditions with geographical regions etc. This module requires pre-processing of questions which is performed by semantic parsing with entity extraction. To ensure the semantic consistencies, identified entities are aligned with the agricultural ontologies to represent the structured query from the question.

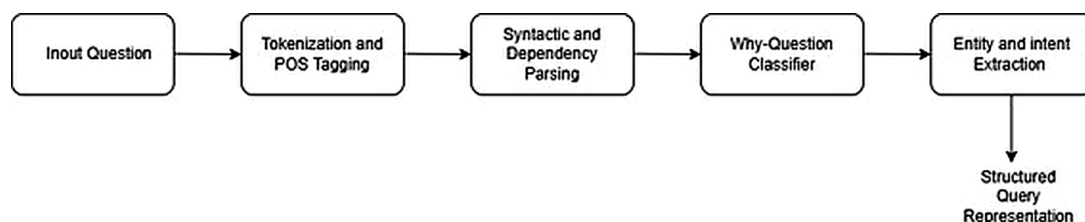


Fig. 3. Detailed Flow to transform question into structured query

## Mathematical Formulation

Given a natural language query:  $q$   
The structured representation is:

$$Q = \Phi(q) = (A, C, R_g, T, \mathcal{F}) \quad (1)$$

Where:

- $q$  is the natural language input why-question
- $\Phi$  is the transformer-based parsing function fine-tuned on agricultural why-type questions
- $A \in A\_set$  is the observed anomaly from the set of anomalies such as yield decline, soil degradation etc.
- $C \in C\_set$  is the crop type chosen from the sets of crops such as rice, wheat etc.
- $R_g \in R\_set$  is the geographical region
- $T$  is the temporal scope which denotes season and year of the crop irrigation
- $F = \{f_1, f_2, \dots, f_n\}$  represents the contextual feature vector comprising soil properties, climate indicators

### 4.2. Domain-Specific LLM Module:

Domain-Specific LLM Module helps generate the explanations of the questions. To align the reasoning with the agricultural domain knowledge, domain-specific LLM is fine-tuned with the agricultural corpora including information related to crop yields, soil properties, irrigation analysis and sustainability reports etc. This LLM doesn't produce the answer to a question, rather it helps in generating hypothesis based on cause-effect relationships which are identified from question and its respective answers.

Let  $L_\theta$  denote the domain-specific LLM output with parameters  $\theta$ , based on soil properties, climate indicators, and management practices

The model generates candidate hypotheses as  $H = L_\theta(Q)$  where  $Q$  is the structured query representation of question  $q$ .

Where:

$$H = L_\theta(Q) = \{h_1, h_2, \dots, h_k\} \quad (2)$$

Where  $k$  is the maximum number of candidate hypotheses generated per query ( $k=5$ ).

Each hypothesis  $h_i$  has an associated probability  $P(h_i | Q)$  calculated using softmax scoring function.

Hypotheses are ranked in the descending order and the top- $k$  ranked hypotheses are passed to the Knowledge Graph Validation module.

#### Flow Diagram

Structured Query  $Q$



Domain-Specific LLM  $L_\theta(Q)$



Candidate Hypotheses  $H = \{h_1, h_2, \dots, h_k\}$



Top-Ranked Hypotheses

### 4.3. Knowledge Graph Validation Module:

This module verifies and validates the generated hypothesis from previous module. The hypothesis generated should be supported by agricultural knowledge sources. This module helps in generating the knowledge graph [15] from the hypothesis comprising entities such as crop, soil types, nutrients and fertilizers, climatic factors and weather conditions, geographical regions. Each hypothesis is verified by the generated knowledge graph to deduce whether the cause path is valid or invalid. If a direct or indirect causal path is found, the hypothesis is considered valid or otherwise invalid. The invalid hypothesis is removed and filtered out which ensures that the causal explanation of the question is consistent with the domain agricultural knowledge.

## Mathematical Formulation

Let the agricultural knowledge graph represent as

$$G = (E, R_g) \quad (3)$$

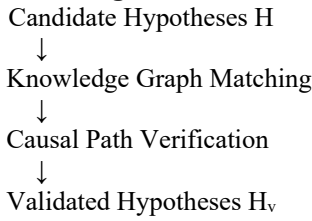
where  $E = \{e_1, e, \dots, e_n\}$  is the set of domain entities like crops, soil factors, climate, nutrients etc. and  $R_g$  is the set of causal relationships represented as  $(e_i, r, e_j)$  where  $r$  belongs to  $\{\text{causes, leads\_to, results\_in and so on}\}$ .

For each hypothesis  $h_i$ , a causal path validation function  $V$  is defined as:

$$\begin{aligned} V(h_i) &= 1 \text{ if a valid causal path exists in } G \\ V(h_i) &= 0 \text{ otherwise} \end{aligned}$$

The final set of validated hypotheses are represented finally as  $H_v = \{h_i \in H \mid V(h_i) = 1\}$  which are passed to the next causal reasoning module.

#### Flow Diagram



For example the question is: “Why did rice yield decrease in Region X during 2023?”

The knowledge graph consists of:

- **Nodes (Entities):** Crop, Region, Time, Soil Factors, Climate Factors, Management Factors, Anomaly
- **Edges (Relations):** Causes, Leads\_To, Results\_In, Observed\_In, Cultivated\_In

**Context Layer generates labeling of question as:**

- **Crop (C):** Rice
- **Region (R):** Region X
- **Temporal Scope (T):** 2023
- **Observed Anomaly (A):** Yield Decrease

#### Directed Causal Relationships

1. Low Rainfall  $\rightarrow$  Soil Moisture Reduction
2. Soil Moisture Reduction  $\rightarrow$  Crop Stress
3. High Temperature  $\rightarrow$  Crop Stress
4. Reduced Fertilizer Application  $\rightarrow$  Soil Nitrogen Deficiency
5. Soil Nitrogen Deficiency  $\rightarrow$  Poor Growth
6. Crop Stress  $\rightarrow$  Yield Decrease
7. Poor Growth  $\rightarrow$  Yield Decrease
8. Pest Infestation  $\rightarrow$  Yield Decrease
9. Rice  $\rightarrow$  Cultivated In  $\rightarrow$  Region X
10. Yield Decrease  $\rightarrow$  Observed In  $\rightarrow$  2023

#### 4.4. Causal Reasoning Module:

This module helps in estimating the extent of associations between the validated cause and effect. For example, this module helps in evaluating how the crop yield will be modified if the rainfall is decreased or fertilizers are introduced. In this module, statistical models are applied on the datasets with controlled factors which helps in estimating the degree to which each factor contributes to the anomaly.

#### Mathematical Formulation

For a cause variable  $X$  affecting anomaly  $A$ :  
Causal Effect (CE) is estimated as:

$$CE = E(A \mid do(X = x_1)) - E(A \mid do(X = x_0)) \quad (4)$$

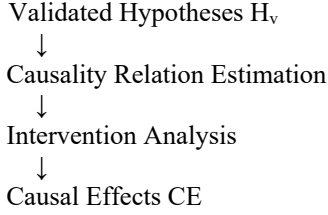
Where:

- $do(X = x_1)$  represents the intervention setting to the active state
- $do(X = x_0)$  represents the intervention setting to the baseline state
- E represents expected value of anomaly outcome

The module produces:

$CE = \{ce_1, ce_2, \dots, ce_m\}$  where each  $ce_i$  represents quantified causal impact. The positive value corresponds to the positive causal contribution to anomaly whereas negative value presents the negative contribution and zero represents the neutral contribution to the anomaly.

#### Flow Diagram



#### 4.5. Sustainability Evaluation Module (S):

This module utilizes the sustainability features useful in evaluating the long-term and short-term implications of the causal effects. In this module, in case any sustainability factor declines, its impact will be accessed. For example, if the fertilizer amount is increased, yield is increased but soil health is damaged. Thus, this integrates the score of causal effects into the sustainability score. Thus, the output of this module helps the decision makers to find out the impact of causes on the long-term resilience of agricultural economy.

#### Mathematical Formulation

Let sustainability dimensions be:

$$S = (S_{env}, S_{econ}, S_{soc}) \quad (5)$$

Where  $S_{env}$  represents the environment,  $S_{econ}$  represents the economic and  $S_{soc}$  represents the social sustainability scores respectively.

Each sustainability dimension  $S_k$  is computed as:

$$S_k = \sum_{i=1}^m w_{k,i} \cdot CE_i, \quad k \in \{env, econ, soc\} \quad (6)$$

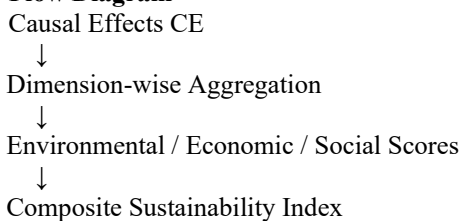
Where

- $w_{\{k,i\}}$  depicts the importance weight of causal factor  $i$  for each dimension  $k$  which is determined from the domain knowledge of experts
- $ce_i$  depicts the causal effect score with factor  $i$
- $m$  is the total number of causal factors

Composite Sustainability Index integrates all three dimensions and is calculated as

$$CSI = \alpha S_{env} + \beta S_{econ} + \gamma S_{soc} \quad (7)$$

#### Flow Diagram



#### 4.6. Explainability Layer (E):

Explainability Layer generates explanation based on quantitative causal and sustainable hypothesis. It makes use of causal paths, determining the contributing and influencing factors to summarize the sustainability applications. Rather than returning raw facts and numbers, this module helps structuring the explanations in a way like: “Decline of yield in region  $R_g$  is due to the reduced rainfall and depletion of soil nitrogen during the time period T”. To estimate the correct prediction and deduce perfect answer, confidence score is calculated for each generated explanation.

#### Mathematical Formulation

$\hat{A}$  is represented as the final structured answer generated by the Explainability layer. This module helps in determining the human-readable explanations of answers corresponding to the why-question.

$$\hat{A} = \varepsilon(Q, H_v, CE, S, \phi) \quad (8)$$

Where:

- $Q$  is the structured query
- $H_v$  is the validated hypotheses
- $CE$  is the causal effects
- $S$  are the sustainability scores
- $\phi$  is the SHAP attribution values which are computed for each causal factor
- $\varepsilon$  is the explainability generation function

#### Flow Diagram

Structured Data ( $Q, H_v, CE, S$ )

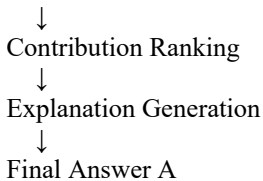


Table 6 provides the details of the notations used in each of the above modules.

Table 6. Details of the Notations

| Symbol    | Definition                     |
|-----------|--------------------------------|
| $q$       | Natural language why-question  |
| $\Phi$    | Semantic parsing function      |
| $A$       | Agricultural anomaly           |
| $C$       | Crop type                      |
| $R_g$     | Geographic region              |
| $T$       | Temporal scope                 |
| $F$       | Context feature vector         |
| $H$       | Candidate hypotheses           |
| $H_v$     | Validated hypotheses           |
| $CE$      | Causal effect                  |
| $CSI$     | Composite Sustainability Index |
| $\hat{A}$ | Final answer                   |

#### 4.7. Novel Contributions of the Proposed Framework

The novelty of the proposed framework does not comprise any single component individually rather address the integration of multiple reasoning components within a unified agricultural Why-QA architecture. Unlike conventional retrieval-augmented or knowledge-graph-enhanced question answering systems, the proposed framework **presents** a multi-stage validated reasoning pipeline that **uniquely combines** hypothesis generation, causal validation, sustainability assessment, and explainability generation within a **single unified architecture for agricultural Why-QA**.

Thus, the major novel contributions are summarized as follows:

##### 1. Validated Hypothesis Generation Framework

Instead of directly generating answers, the domain-specific LLM first produces multiple causal hypotheses. These hypotheses are subsequently verified through knowledge graph traversal and causal-path validation before being incorporated into the final explanation.

## 2. Integration of Sustainability-Aware Reasoning

Existing Why-QA systems primarily focus on factual correctness or retrieval accuracy. The proposed framework introduces sustainability evaluation through environmental, economic, and social indicators.

## 3. Multi-Layer Explainability Architecture

The framework combines evidence grounding, causal-path reasoning, sustainability impact assessment, and SHAP-based feature attribution to generate interpretable answers rather than simple textual responses.

## 4. Agriculture-Specific Causal Why-QA Framework

Unlike traditional open-domain QA systems, the proposed architecture is specifically designed for climate-resilient agriculture and supports explanation of complex why-type non-factoid questions involving crop yield decline, soil degradation, water scarcity, climate variability, and sustainability impacts. Table 7 compares the proposed framework with existing approaches.

Table 7. Comparison with existing Approaches

| Feature                          | RAG-QA [11] | KG-RAG [11-12] | CausalQA [7] | Explainable QA [13] | Proposed Framework |
|----------------------------------|-------------|----------------|--------------|---------------------|--------------------|
| Why-type QA                      | ✓           | ✓              | ✓            | ✓                   | ✓                  |
| Knowledge Graph Validation       | X           | ✓              | X            | X                   | ✓                  |
| Explicit Causal Inference        | X           | X              | ✓            | X                   | ✓                  |
| Sustainability Assessment        | X           | X              | X            | X                   | ✓                  |
| Hypothesis Validation Pipeline   | X           | X              | X            | X                   | ✓                  |
| SHAP-based Attribution           | X           | X              | X            | Partial             | ✓                  |
| Agricultural Domain Adaptation   | Partial     | Partial        | Partial      | Partial             | ✓                  |
| End-to-End Explainable Reasoning | Limited     | Moderate       | Moderate     | Moderate            | Comprehensive      |

## 5. Implementation Details

This section discusses in detail the techniques and approaches utilized by each of the module. The framework integrates the agricultural datasets either in structured or unstructured format to enable explainable reasoning.

### 5.1. Query Understanding Module Implementation:

The Query Understanding Module transforms the natural language why-question into structured query representation. It is implemented by utilizing a transformer-enhanced semantic parsing model which is fine-tuned on annotated why-type questions and extract the named entities. Named entities are identified by applying Named Entity Recognition techniques which help in identifying the crop types, soil features, rainfall and temperature parameters and so on. The extracted entities are mapped through similarity measures to ensure consistency [14]. The final output resulted in the form of structured query format as  $Q = (A, C, R, T, F)$  whose representation is stored in the JSON format.

#### Output Format:

```
{
  "crop": "Rice",
  "location": "District X",
  "time_period": "Kharif 2023",
  "anomaly": "Yield decrease"
}
```

### 5.2. Domain-Specific LLM Module:

This module fine-tunes the transformer-based model on data from agricultural corpora which includes reports from crop yields, soil-health, climate and weather conditions etc. Fine-tuning helps in ensuring the domain related information and reasoning. Each candidate hypothesis is generated in structured template form which helps to determine the effects corresponding to the cause and helps in validation of knowledge graphs. The top-ranked hypotheses are passed to the next module.

### 5.3. Knowledge Graph Validation Implementation:

The agricultural knowledge graph is constructed by utilizing agricultural domain ontologies which help in identifying causal relationships between nodes. It is implemented using Neo4j Graph Database. Each generated hypothesis is converted into RDF triples for semantic representation. A confidence score for each hypothesis is assigned based on the causal path length and relationships accuracy.

### 5.4. Causal Reasoning Module Implementation:

Causal reasoning module helps in identifying the root causes behind the anomalies which are quantified using historical datasets. The module is implemented using Bayesian Networks, Structural Causal Models (SCM) where variables are identified from the knowledge graph, after that the causal graph is constructed which is finally ranked on the basis of cause-effect relationships size as CE score is calculated as  $CE = \{ce_1, ce_2, \dots, ce_m\}$ .

### 5.5. Sustainability Evaluation Module Implementation:

To rate each cause-effect relation with the sustainability metrics, multiple parameters such as Nitrogen Use Efficiency (NUE), Water Use Efficiency (WUE), Soil Organic Carbon Index, and Carbon Emission Index are used [16]. These metrics are calculated by applying mathematical modeling, which is finally compared with a threshold value to determine the sustainability score and categorizing risk levels as low, moderate and high.

### 5.6. Explainability Layer Implementation:

The Explainability Layer module integrates the output from query parser, causal engine, sustainability score and knowledge graph modules to deduce the grounded explanation supported by hypothesis. This module helps in generating final answer with cause and effect supported by its sustainability impact as evidence. Explainability is implemented by using SHAP API [17].

### 5.7. Experimental Configuration and System Implementation

The proposed Explainable LLM-based Why-QA framework was implemented and evaluated using a dataset including 500 agricultural why-type questions, which were collected from numerous agricultural knowledge sources. The framework integrates the various components including query understanding, domain-specific language modelling, knowledge graph validation, causal reasoning, sustainability assessment and explainability generation. The complete framework was evaluated using benchmark agricultural datasets containing crop yield, soil health, climate, irrigation, and sustainability-related question-answer pairs.

The implementation described in this section constitutes a prototype system designed to validate the architectural feasibility of the proposed framework. The objective of this paper is not to deploy the large-scale agricultural model but to demonstrate the impact and further improvements by integrating KG-validated hypothesis, causal reasoning, sustainable assessment and enhancing explainability. This section further provides the detailed experimental setup for each module of the framework.

#### 5.7.1. Base Language Model and Fine-Tuning

The framework uses **Llama-3-8B-Instruct** as the baseline large language model which is known for its capability and adapting the domain knowledge. The model is aligned with domain knowledge by fine-tuning the parameters on agricultural corpora which comprises crop yields, soil yields, climate impact and sustainable orientation by utilizing **Low-Rank Adaptation (LoRA)** method.

#### 5.7.2. Training Configuration

Table 8 highlights the settings of the parameters used while prototype implementation of the framework.

Table 8. Experimental Configuration Parameters

| Parameter               | Value                                        |
|-------------------------|----------------------------------------------|
| Base LLM                | Llama-3-8B-Instruct                          |
| Fine-tuning Method      | LoRA (Low-Rank Adaptation)                   |
| LoRA Rank (r)           | 16                                           |
| Learning Rate           | $2 \times 10^{-5}$                           |
| Batch Size              | 16                                           |
| Number of Epochs        | 5                                            |
| Optimizer               | AdamW                                        |
| Hardware                | NVIDIA A100 GPU (40 GB VRAM)                 |
| Deep Learning Framework | PyTorch 2.2                                  |
| Knowledge Graph         | Neo4j Community Edition                      |
| Explainability Tool     | SHAP (SHapley Additive exPlanations)         |
| Causal Modeling         | Bayesian Networks + Structural Causal Models |

### 5.7.3. Prompt Engineering Strategy

This strategy first prepares a domain-specific template for prompt that helps retrieving correct output from LLM. The prompt is described as:

*"Given the following agricultural factors as Crop = [C], Region = [R], Time Period = [T], and Observed Anomaly = [A], generate (1) Root causes supported by causal reasoning, (2) Supporting domain-knowledge evidences, (3) Assessment of impact of sustainability factors and (4) Confidence score corresponding to each generated hypothesis."*

This prompt structure is directly aligned with the structured query format  $Q = (A, C, R, T, F)$  which is discussed in Section 4.1. The generated response to the prompt ensures the proper alignment between query understanding module with LLM generation module.

Example Query:

Crop = Rice  
Region = Haryana  
Time = Kharif 2023  
Anomaly = Yield Decrease

Generated Hypothesis:

H1: The low intensity of rainfall results in depletion of soil moisture which reduces the yield of crop.

Evidence:

Historical climate records indicate rainfall deficit of 18% during the reproductive growth stage.

The generated hypothesis is represented in JSON format as:

```
{
  "hypothesis_id": "H1",
  "cause": "Low Rainfall",
  "mechanism": "Soil moisture depletion",
  "effect": "Yield reduction",
  "confidence": 0.87
}
```

### 5.7.4. Knowledge Graph Construction

The agricultural knowledge graph was constructed using **Neo4j Graph Database** with nodes representing domain entities (crops, soil types, climate factors, regions, nutrients) and directed edges representing causal relationships. Some example nodes with corresponding entities are discussed below in Table 9.

Table 9. Example Node Types

| Node Category             | Example Entities                                        |
|---------------------------|---------------------------------------------------------|
| Crop                      | Rice, Wheat, Maize                                      |
| Soil Factors              | Soil Moisture, Nitrogen Deficiency, Soil Organic Carbon |
| Climate Factors           | Rainfall, Temperature, Humidity                         |
| Management Factors        | Fertilizer Application, Irrigation Practice             |
| Pest Factors              | Brown Plant Hopper, Stem Borer                          |
| Region                    | Haryana, Punjab, District X                             |
| Sustainability Indicators | WUE, NUE, CEI                                           |

Some of the examples of causal relationships are discussed in Table 10.

Table 10. Some causal relationships

| Subject                 | Relationship  | Object                  |
|-------------------------|---------------|-------------------------|
| Low Rainfall            | CAUSES        | Soil Moisture Reduction |
| Soil Moisture Reduction | LEADS_TO      | Crop Stress             |
| High Temperature        | CAUSES        | Crop Stress             |
| Nitrogen Deficiency     | CAUSES        | Poor Plant Growth       |
| Crop Stress             | LEADS_TO      | Yield Reduction         |
| Yield Reduction         | OBSERVED_IN   | Region X                |
| Rice                    | CULTIVATED_IN | Region X                |

### 5.7.5. Explainability Implementation

The feature-oriented explanations were generated using the **SHAP (SHapley Additive exPlanations)** library, which assigns contribution scores to each input feature in the causal reasoning and sustainability evaluation modules. SHAP values were computed accordingly, which helped identify which factors (e.g., rainfall deficit, soil nitrogen level, temperature anomaly) contributed most significantly to the predicted anomaly. These features were passed to the causal reasoning model to support the generated answer.

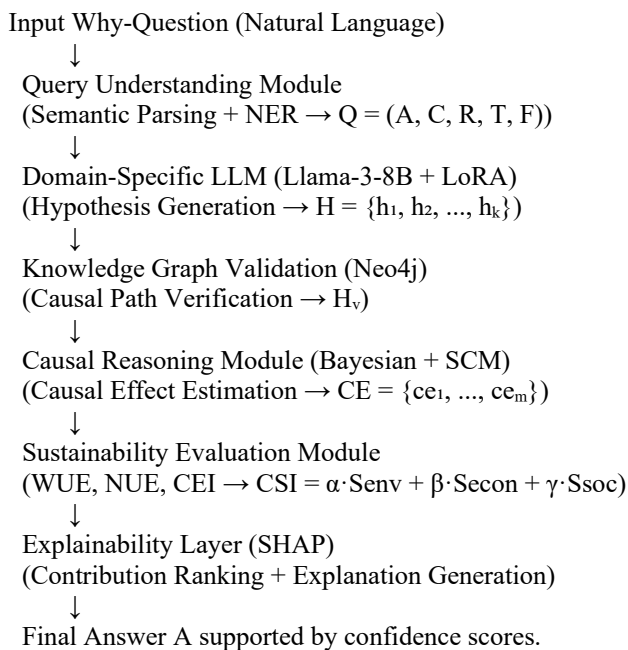
For example, for the question “**Why did rice yield decrease in Region X during 2023?**”, the SHAP-based explanation identified low rainfall, high temperature, and nitrogen deficiency as the most influential factors contributing to yield decline. Table 11 provides an example of the output generated by SHAP.

Table 11. Example SHAP output table

| Feature                 | SHAP Contribution | Interpretation                                |
|-------------------------|-------------------|-----------------------------------------------|
| Low rainfall            | 0.31              | Strong positive contribution to yield decline |
| Soil moisture reduction | 0.24              | Increased crop stress                         |
| Nitrogen deficiency     | 0.19              | Reduced plant growth                          |
| High temperature        | 0.15              | Increased evapotranspiration stress           |
| Pest incidence          | 0.07              | Minor contribution                            |
| Irrigation frequency    | -0.04             | Partially reduced yield loss                  |

### 5.7.6. System Pipeline

Further, the complete end-to-end implementation pipeline is illustrated below:



The objective of the above implementation is to validate the feasible integration of modules in framework rather than developing the deployment-oriented model.

## 5.8. Dataset Construction and Annotation

### 5.8.1. Process of Dataset Preparation

A dataset of 500 agricultural why-type questions is prepared by integrating data from multiple agricultural repositories which are publicly available. The prime source of agricultural data includes FAOSTAT (Food and Agriculture Organization), USDA National Agricultural Statistics Service (NASS) and NOAA climate databases. FAOSTAT provides data related to crop yield, fertilizer usage, irrigation, and production statistics, NASS provides data related to soil health, crop condition, and agricultural management, and NOAA climate databases comprise data records for rainfall, temperature, drought, and weather anomaly.

The process of preparing the dataset involves four stages detailed below:

#### Stage 1: Data Collection

Data records for different instances are integrated from FAOSTAT, USDA, and NOAA data repositories.

#### Stage 2: Event Identification

Different agricultural events are identified which are associated with anomalies such as decline in crop yield, degradation of soil, decrease in water level, deficiency of nutrients, pest manifestation and variability in climate and weather conditions.

#### Stage 3: Why-Question Generation

This step involves the role of domain experts who transform the identified events in step 2 into the why-type questions.

For example, the event identified is: “Yield of crop Rice is declined by 18% in Region X during Kharif 2023”

This event is transformed to the question as: “Why there was a decline in yield of crop Rice in Region X during Kharif 2023?”

#### Stage 4: Ground-Truth Explanation Creation

Domain experts help identify and prepare the causal explanations supported by ground-truth agricultural evidence.

### 5.8.2. Distribution of Questions corresponding to each category

| Category                      | Questions |
|-------------------------------|-----------|
| Crop Yield Decline            | 140       |
| Soil Health                   | 90        |
| Climate Impact                | 95        |
| Irrigation & Water Management | 70        |
| Nutrient Deficiency           | 55        |
| Pest & Disease Incidence      | 50        |
| Total                         | 500       |

The dataset is intentionally balanced across different domains of agriculture so that the proposed framework is evaluated on diverse scenarios rather than focusing on a single task.

### 5.8.3. Annotation Guidelines

Each question-answer pair is annotated by three agricultural domain experts supported by following guidelines:

1. Identify the prime root cause of the agricultural event.
2. Identify the other secondary factors contributing the agricultural event.
3. Verify the causal relationships supported by agricultural domain knowledge.
4. Validate the supporting evidence of each causal event.
5. Assess the impact of sustainable factors to each causal event.

The disagreement by experts is resolved by brainstorming sessions, discussions and majority decisions. The final annotation forms the basis for ground-truth explanations to each causal event.

### 5.8.4. Dataset Partitioning

The dataset is partitioned into training, validation, and testing subsets whose distribution is given as:

| Dataset    | Questions |
|------------|-----------|
| Training   | 350 (70%) |
| Validation | 50 (10%)  |
| Testing    | 100 (20%) |
| Total      | 500       |

The training dataset is used for fine-tuning and configuring the system parameters, validation dataset is used for hyperparameter selection, and testing dataset is used for final evaluation.

#### 5.8.5. Expert Annotation and their Evaluation

The evaluation is performed by three experts having respective experience in crop management, soil health and implementing agricultural decision-support systems. There are two roles of experts, (1) Providing the ground-truth annotation of causal explanations and (2) Evaluating the interpretability of explanations. Each QA-pair is reviewed by each individual expert, then after collective discussions, evaluation is conducted. Each explanation to QA pair is rated on 5-point Likert scale based on Clarity in explanations, causal reasoning, ground truth of evidence, relevance to sustainable factors, and ultimate impact for agricultural decision-support. The final score of each QA pair is calculated by taking the average of all expert ratings.

#### 5.8.6. Evaluation Reliability Analysis

The consistency of annotation by all experts having different expertise is assessed by measuring inter-annotator agreement using Fleiss' Kappa coefficient. The Fleiss' Kappa value to the existing implementation is calculated as 0.78 which indicates good substantial agreement among evaluators. This value ensures that the annotations generated are highly consistent and reliable for further comparative evaluations.

## 6. Results and Discussion

The system is evaluated using multiple datasets from Food and Agriculture Organization (FAOSTAT) [18] comprising information of crop yield, fertilizer usage, irrigation statistics, United States Department of Agriculture (USDA) [19] comprising data of soil parameters and National Oceanic and Atmospheric Administration (NOAA) [20] comprising climate records. For small study, scope is limited to rice and wheat crops, multi-district agricultural zones are considered. The evaluation was conducted on a dataset consisting of 500 why-type questions collected from the above multiple data sources. The dataset is divided into 70% training and 20% testing and 10% validation datasets. Each why-question is annotated by three experts to support the evidence of the ground-truth.

The proposed framework is evaluated against five representative individual baseline systems covering retrieval-augmented generation, knowledge-graph-enhanced reasoning, and instruction-tuned language models. The selected baselines include:

- Baseline LLM: A standard instruction-following LLM without knowledge graph validation or causal reasoning.
- LLM + Knowledge Graph (KG): A language model integrated with agricultural knowledge graph retrieval.
- RAG-QA: A retrieval-augmented generation model that retrieves relevant agricultural documents before answer generation.
- KG-RAG: A knowledge-graph-enhanced retrieval-augmented generation system combining document retrieval with graph-based reasoning.
- Instruction-Tuned LLM: A domain-adapted instruction-following model fine-tuned for explainable question answering.

The proposed Explainable Why-QAS is compared against these baselines using causal reasoning accuracy, explanation quality, evidence grounding, and interpretability metrics as described in Table 12.

Table 12. Comparison of proposed framework with other baseline QA

| Baseline                          | Description                                                    |
|-----------------------------------|----------------------------------------------------------------|
| Baseline LLM (FLAN inspired) [22] | Direct answer generation without retrieval or graph validation |
| LLM + KG (KG-QA inspired) [11]    | Knowledge graph supported answer generation                    |
| RAG-QA [21]                       | Retrieval-Augmented Generation using agricultural documents    |
| KG-RAG [11-12]                    | Retrieval + Knowledge Graph reasoning                          |
| Instruction-Tuned LLM [22]        | Domain-adapted instruction-following QA model                  |
| Proposed Framework                | LLM + KG validation + causal reasoning + sustainability + SHAP |

### 6.1. Causal Reasoning Accuracy:

The accuracy of causal reasoning module is measured by the ability of the system to identify the root causes accurately against the annotated ground truth by experts. The results are depicted in Table 13 below.

Table 13. Comparative Results of Causal Reasoning Module

| Model                 | Root Cause Accuracy | Precision   | Recall      | F1-score    |
|-----------------------|---------------------|-------------|-------------|-------------|
| Baseline LLM          | 58%                 | 0.60        | 0.55        | 0.57        |
| LLM + KG              | 71%                 | 0.72        | 0.69        | 0.70        |
| RAG-QA                | 74%                 | 0.75        | 0.72        | 0.73        |
| Instruction-Tuned LLM | 76%                 | 0.77        | 0.74        | 0.75        |
| KG-RAG                | 79%                 | 0.80        | 0.77        | 0.78        |
| Proposed Why-QAS      | 85.2 ± 1.8%         | 0.86 ± 0.02 | 0.83 ± 0.02 | 0.84 ± 0.02 |

Where,

$$Accuracy = \frac{Correct\ Root\ Causes}{Total\ Test\ Queries} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

To assess whether performance improvements are statistically significant, McNemar's test was applied to compare the proposed framework against each baseline on the 100-question test set. Results indicate significant improvement over Baseline LLM ( $p < 0.01$ ), LLM+KG ( $p < 0.01$ ), RAG-QA ( $p < 0.05$ ), Instruction-Tuned LLM ( $p < 0.05$ ), and KG-RAG ( $p < 0.05$ ), confirming that the observed gains are not attributable to chance.

### 6.2. Explanation Quality Evaluation:

The explainability of the framework is evaluated by calculating Explanation Completeness Score (ECS), Evidence Grounding Score (EGS) and Human Interpretability Rating (1–5 scale). The results of each score corresponds to different models are depicted in Table 14.

Table 14. Comparative Results of Explanation Quality Module

| Model                 | ECS  | EGS  | Interpretability |
|-----------------------|------|------|------------------|
| Baseline LLM          | 0.52 | 0.41 | 3.1              |
| LLM + KG              | 0.71 | 0.69 | 3.8              |
| RAG-QA                | 0.74 | 0.76 | 3.9              |
| Instruction-Tuned LLM | 0.78 | 0.79 | 4.1              |
| KG-RAG                | 0.83 | 0.85 | 4.3              |
| Proposed Why-QAS      | 0.89 | 0.91 | 4.6              |

Where the Explanation Completeness Score (ECS) depicts if all the reasoning appears in the explanation of the answer or not and is calculated as:

$$ECS = \frac{\text{Number of required explanation components present}}{\text{Total required components}} \quad (13)$$

Evidence Grounding Score (EGS) measures how each explanation is supported by the ground-truth evidences and is calculated as:

$$EGS = \frac{\text{Correctly grounded evidence}}{\text{Total evidence statements}} \quad (14)$$

Further, human interpretability rating is provided by averaging the scaling provided individually by three experts based on clarity, logical flow and relevancy of the explanations.

### 6.3. Sustainability Assessment Analysis

In addition to causal reasoning and explanation quality, the proposed framework incorporates sustainability-aware reasoning through the evaluation of environmental, economic, and social sustainability dimensions. The Sustainability Evaluation Module utilizes indicators such as Water Use Efficiency (WUE), Nitrogen Use Efficiency (NUE), Carbon Emission Index (CEI), profitability-related factors, and resilience-oriented indicators to estimate the long-term

implications of identified causal factors. The purpose of the sustainability module is not to replace specialized agricultural sustainability assessment tools but rather to provide decision-support information that complements causal explanation generation. Table 15 provides analysis of sustainable indicators used with the examples of observation in Table 16.

Table 15. Sustainability Indicators

| Sustainability Dimension | Indicator                     | Purpose                              |
|--------------------------|-------------------------------|--------------------------------------|
| Environmental            | Water Use Efficiency (WUE)    | Assess irrigation efficiency         |
| Environmental            | Nitrogen Use Efficiency (NUE) | Evaluate fertilizer utilization      |
| Environmental            | Carbon Emission Index (CEI)   | Estimate environmental impact        |
| Economic                 | Profitability Score           | Assess economic viability            |
| Economic                 | Yield Stability               | Measure productivity resilience      |
| Social                   | Farmer Resilience Index       | Assess adaptation capability         |
| Social                   | Resource Accessibility        | Evaluate sustainability of practices |

Table 16. Sustainability Example

| Question                                                   | Root Cause                        | Sustainability Observation                  |
|------------------------------------------------------------|-----------------------------------|---------------------------------------------|
| Why did rice yield decrease in Haryana during Kharif 2023? | Rainfall deficit                  | Reduced WUE and increased irrigation stress |
| Why is soil fertility declining?                           | Excessive fertilizer use          | Reduced NUE and long-term soil degradation  |
| Why did irrigation demand increase?                        | High temperature and low rainfall | Increased water consumption pressure        |

#### 6.4. Discussion of Evaluation Limitations

Although the proposed framework demonstrates consistent improvements across causal reasoning and explanation quality metrics, the present study is intended as a proof-of-concept validation of the proposed architecture rather than a large-scale benchmarking study. The evaluation was conducted on a benchmark dataset of agricultural why-type questions annotated by domain experts to assess the feasibility of integrating knowledge graph validation, causal reasoning, sustainability assessment, and explainability within a unified framework.

The authors acknowledge that additional large-scale validation involving broader datasets, larger expert panels, statistical significance testing, confidence interval estimation, and robustness analysis across diverse agricultural regions would further strengthen the empirical findings. These aspects constitute an important direction for future work and will be considered in subsequent versions of the framework.

### 7. Conclusion and Future Research Directions

The paper presents an explainable LLM-based Why-QA framework for sustainable smart agriculture which integrates the different layers of domain knowledge, causal reasoning, interpretability and explainable layers. As clearly seen from the results, the proposed framework outperforms the traditional LLM approaches in identifying and determining the causal answers to the sustainable agricultural queries. The paper demonstrates the inculcation of advanced AI to make critical decisions in the agricultural sector which ultimately supports the sustainability dimensions. The framework can be further enhanced by inculcating multilingual QA support with federated learning to ensure privacy and secure data and further improving the explainability performance by taking user feedback in iterations.

Unlike conventional explainable question-answering systems that primarily focus on answer correctness and interpretability, the proposed framework additionally incorporates sustainability-oriented reasoning by considering environmental, economic, and social indicators. This enables generated explanations to provide decision-support information relevant to climate resilience, resource utilization, and long-term agricultural sustainability.

### All the Declarations and Statements

#### Author Contributions Statement

Paper has single author. That author has individually made the contribution towards conceptualization, data curation, implementation, model training, validation, performance evaluation, statistical analysis, and writing.

#### Conflict of Interest Statement

The authors declare no conflicts of interest.

#### Funding Declaration

This research is not funded by any organization/institution.

## Data Availability Statement

None

## Ethical Declarations

Ethical approval was not required for this study because it did not involve human participants, animals or identifiable personal data.

## Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

## Declaration of Generative AI in Scholarly Writing

The author used ChatGPT (OpenAI) only for language editing, grammar correction and readability enhancement during manuscript preparation. All the contents were reviewed and verified by the author, who take full responsibility of the manuscript.

## Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence  
LLM - Large Language Model  
QA - Question Answer  
QAS - Question Answering System  
FAOSTAT - Food and Agriculture Organization Statistics  
USDA - United States Department of Agriculture  
NOAA - National Oceanic and Atmospheric Administration  
SHAP - SHapley Additive exPlanations  
ECS - Explanation Completeness Score  
EGS - Evidence Grounding Score  
KG - Knowledge Graphs  
CSI - Composite Sustainability Index  
WUE - Water use efficiency  
NUE - Nitrogen use efficiency  
CEI - Carbon Emission Index  
RC - Research Contribution  
RO - Research Objective  
RG - Research Gap  
MRR - Mean Reciprocal Rank  
RAG - Retrieval-Augmented Generation  
XAI - Explainable AI  
CE - Causal Effect  
SCM - Structural Causal Models

## Appendix A/B/C..., with appendix title

None

## References

- [1] Breja, M., & Jain, S. K. (2022). Analyzing linguistic features for answer re-ranking of why-questions. *Journal of Cases on Information Technology (JCIT)*, 24(3), 1-16. <https://doi.org/10.4018/JCIT.20220701.0a10>
- [2] Surdeanu, M., Ciaramita, M., & Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Computational linguistics*, 37(2), 351-383. [https://doi.org/10.1162/COLI\\_a\\_00051](https://doi.org/10.1162/COLI_a_00051)
- [3] Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3), 345-361. <https://doi.org/10.1016/j.jksuci.2014.10.007>
- [4] Sharp, R., Surdeanu, M., Jansen, P., Clark, P., & Hammond, M. (2016). Creating causal embeddings for question answering with minimal supervision. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (pp. 138-148). Association for Computational Linguistics (ACL). <https://doi.org/10.18653/v1/d16-1014>
- [5] Breja, M., & Jain, S. K. (2022). A survey on non-factoid question answering systems. *International Journal of Computers and Applications*, 44(9), 830-837. <https://doi.org/10.1080/1206212X.2021.1949117>

- [6] Breja, M., & Jain, S. K. (2020). Causality for Question Answering. In COLINS (pp. 884-893). <https://ceur-ws.org/Vol-2604/>
- [7] Bondarenko, A., Wolska, M., Heindorf, S., Blübaum, L., Ngomo, A. C. N., Stein, B., ... & Potthast, M. (2022, October). CausalQA: A benchmark for causal question answering. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3296-3308). <https://aclanthology.org/2022.coling-1.291/>
- [8] Pechsiri, C., & Piriyaikul, R. (2016). Developing a Why-How Question Answering system on community web boards with a causality graph including procedural knowledge. *Information Processing in Agriculture*, 3(1), 36-53. <https://doi.org/10.1016/j.inpa.2016.01.002>
- [9] Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E., Soares, L. B., & Collins, M. (2021). QED: A framework and dataset for explanations in question answering. *Transactions of the Association for computational Linguistics*, 9, 790-806. [https://doi.org/10.1162/tacl\\_a\\_00398](https://doi.org/10.1162/tacl_a_00398)
- [10] Fan, Y., Wang, Y., Wang, G., Jie, Z., Liu, J., Ye, Q., & Ruan, T. (2025, July). MinosEval: Distinguishing Factoid and Non-Factoid for Tailored Open-Ended QA Evaluation with LLMs. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 10517-10548). <https://doi.org/10.48550/arXiv.2506.15215>
- [11] Linders, J., Tomczak, J.M. Knowledge graph-extended retrieval augmented generation for question answering. *Appl Intell* 55, 1102 (2025). <https://doi.org/10.1007/s10489-025-06885-5>
- [12] Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025). Document graphrag: knowledge graph enhanced retrieval augmented generation for document question answering within the manufacturing domain. *Electronics*, 14(11), 2102. <https://doi.org/10.3390/electronics14112102>
- [13] Rajbongshi, A., Johora, F.T., Hossain, A. et al. Leveraging explainable AI for sustainable agriculture: a comprehensive review of recent advances. *Artif Intell Rev* 59, 105 (2026). <https://doi.org/10.1007/s10462-025-11459-5>
- [14] Breja, M., & Jain, S. K. (2022). Why-Type Question to Query Reformulation for Efficient Document Retrieval. *International Journal of Information Retrieval Research (IJIRR)*, 12(1), 1-18. <https://doi.org/10.4018/IJIRR.289948>
- [15] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021, June). QA-GNN: Reasoning with language models and knowledge graphs for question answering. In Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 535-546). <https://doi.org/10.18653/v1/2021.naacl-main.45>
- [16] Li, Z.W., Wang, G.Y., Khan, K. et al. Irrigation combines with nitrogen application to optimize soil carbon and nitrogen, increase maize yield, and nitrogen use efficiency. *Plant Soil* 499, 605–620 (2024). <https://doi.org/10.1007/s11104-024-06480-6>
- [17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
- [18] Food and Agriculture Organization of the United Nations (FAO). (2024). FAOSTAT Statistical Database. Rome, Italy: FAO. Available at: <https://www.fao.org/faostat/>
- [19] United States Department of Agriculture (USDA). (2023). USDA National Agricultural Statistics Service (NASS) Quick Stats Database. Washington, DC: USDA. Available at: <https://www.nass.usda.gov/datasets>
- [20] National Oceanic and Atmospheric Administration (NOAA). (2023). NOAA National Centers for Environmental Information (NCEI) Climate Data Online. Silver Spring, MD, USA. Available at: <https://www.ncdc.noaa.gov/cdo-web/>
- [21] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, pp. 9459-9474. <https://dl.acm.org/doi/abs/10.5555/3495724.3496517>
- [22] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024, May). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In International conference on learning representations (Vol. 2024, pp. 9112-9141). <https://doi.org/10.48550/arXiv.2310.11511>
- [23] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. <https://doi.org/10.48550/arXiv.2404.16130>
- [25] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations (ICLR)* <https://doi.org/10.48550/arXiv.2203.11171>
- [26] Kuska, M. T., Wahabzada, M., & Paulus, S. (2024). AI for crop production—Where can large language models (LLMs) provide substantial value?. *Computers and electronics in agriculture*, 221, 108924. <https://doi.org/10.1016/j.compag.2024.108924>
- [27] Christadoss, J., & Panda, M. R. (2025). Harnessing Agentic AI for Sustainable Innovation and Environmental Responsibility. *Futurity Proceedings*, (5), 269-280. <https://doi.org/10.5281/zenodo.17662706>

## Authors' Profiles



**Dr. Manvi Breja** is currently working as an Assistant Professor (Selection Grade) in the department of Computer Science & Engineering, Northcap University Gurugram. She has received her PhD degree from National Institute of Technology, Kurukshetra, completed her M.Tech with first position from YMCA University, Faridabad and B.Tech from MDU Rohtak. She has over 9 years of experience in teaching. In addition to this, she has qualified CSIR UGC NET with JRF with rank 72, UGC-NET with Assistant Professor and GATE 2012, 2013, 2015. Her area of interests are in the field of Data Mining, Information Retrieval and Natural Language Processing and Cyber Security. She has in her records around 30 research papers in reputed international journals and conferences, and 3 patents. She is active reviewer of several research papers belonging to SCI and Scopus journals.

### **How to cite this paper**

Manvi Breja, "A Novel Explainable LLM-based Why-QA Framework for Climate Resilient and Sustainable Smart Agriculture", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.302-323, 2026. DOI:10.5815/ijem.2026.04.21