Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AgriMemSynth: a synthetic benchmark for memory and multi-hop reasoning in agricultural question answering

Nur Arifin Akbar

nurarifin.akbar@unipa.it

University of Palermo

Rahool Dembani

SingularLogic (Greece)

Biagio Lenzitti

University of Palermo

Domenico Tegolo

University of Palermo

Article

Keywords:

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10090330/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AgriMemSynth: a synthetic benchmark for memory and multi-hop reasoning in agricultural question answering

Nur Arifin Akbar[1,*], **Rahool Dembani**[2], **Biagio Lenzitti**[1], **and Domenico Tegolo**[1]

[1]Department of Computer Science, University of Palermo, Palermo, Italy
[2]Singular Logic S.A., Athens, Greece
[*]Corresponding author: Nur Arifin Akbar, nurarifin.akbar@unipa.it

ABSTRACT

Agricultural advisory systems often need to combine current observations, earlier visits, and crop-disease knowledge, but existing agricultural benchmarks mainly evaluate image classification or short-form question answering. We present AgriMemSynth, a synthetic benchmark framework for evaluating agricultural question answering under two complementary requirements: multi-session conversational memory and structured multi-hop reasoning. AgriMemSynth provides AgriConvMem, a set of 500 generated advisory conversations, and AgriMultiHop, a set of 2,000 generated reasoning questions. Using held-out test splits with 146 conversational-memory questions and 200 multi-hop questions, we compare five retrieval organizations under a shared answer-generation pipeline: retrieval-augmented generation, a neuroscientific memory system, a hybrid retriever, BM25 with semantic reranking, and a graph-based memory system. Semantic evaluation reveals three stable benchmark patterns. First, lexical metrics substantially underestimate answer correctness. Second, conversational memory is harder than generated multi-hop reasoning. Third, retrieval organization is task-dependent: RAG has the highest conversational-memory score (85.6%, 95% CI: 79.0–90.4), whereas MemoryGraph has the highest multi-hop score (97.0%, 95% CI: 93.6–98.6). Temporal questions remain the weakest conversational category. These results support AgriMemSynth as a controlled benchmark for studying how agricultural AI systems organize, retrieve, and apply knowledge across longitudinal and relational tasks.

1 Introduction

AI-powered agricultural advisory systems need benchmarks that reflect the complexity of crop disease question answering. This paper introduces AgriMemSynth, a synthetic benchmark framework for conversational memory[1,2] and multi-hop reasoning[3,4] in agricultural[5] QA (Question & Answer). The contribution is primarily methodological: AgriMemSynth pairs longitudinal conversational recall with structured multi-hop reasoning, then compares memory architectures under a fixed backbone language model. We ask how memory architectures differ across conversational and multi-hop agricultural QA, which question categories are most difficult for retrieval systems, and how large the gap is between lexical and semantic evaluation in this domain. The next sections describe why agricultural question answering is difficult and why memory-augmented systems are a useful setting for studying it.

Knowledge Discovery Challenges in Agricultural Advisory Agriculture generates heterogeneous data spanning images, sensor readings, textual observations, and expert knowledge. In crop disease management, a single advisory interaction may involve visual symptoms captured under varying field conditions, temporal records of disease progression, severity assessments tied to growth stages, and treatment recommendations that depend on local regulations, climate, and crop variety[6,7]. For knowledge discovery systems, this setting creates a structural challenge: useful information is distributed across modalities, time points, entities, and relations. Effective agricultural QA therefore requires representations that preserve cross-modal relationships, temporal sequences, multi-hop reasoning chains, and contextual variability.

The problem has practical importance. Crop diseases account for an estimated 20–40% of global yield losses each year, and smallholder farmers in developing countries often have limited access to expert pathologists[8]. AI-powered advisory systems may help widen access to plant disease expertise, but benchmark coverage remains thin for the retrieval, memory, and reasoning requirements of agricultural question answering[9].

Current agricultural image repositories such as PlantVillage[10] and PlantDoc[8] have supported much of the work on disease classification models. PlantVillage provides 54,306 laboratory-controlled images across 38 crop-disease categories, while PlantDoc offers 2,598 images captured under realistic field conditions with natural backgrounds, lighting variation, and multiple diseases per leaf. However, these repositories store images in isolation and fail to capture the conversational context in which

farmers seek advice, the temporal progression of symptoms across multiple visits, or the complex reasoning required to connect treatments to specific disease, crop, and environmental conditions. This disconnect between available data and real-world advisory needs motivates the present study.

Complexity of Agricultural Question Answering Agricultural question answering differs from general-domain QA in several ways:

Multi-modal integration. A farmer's query about a diseased fruit plant involves visual evidence (leaf spots, discoloration patterns), textual descriptions ("the leaves started curling week ago"), and structured metadata (crop variety, location, season). Knowledge discovery methods must preserve these cross-modal relationships.

Temporal dependencies. Disease management is longitudinal. A farmer may consult an advisory system multiple times as symptoms evolve, treatments are applied, and outcomes are observed. This temporal context requires data structures beyond single-turn question-answer pairs.

Multi-hop reasoning. Questions like "Which fungicide works for both early blight on tomatoes and leaf spot on peppers?" require chaining knowledge across disease profiles, treatment databases, and crop compatibility tables. Answering this requires: (1) identifying that early blight is caused by *Alternaria solani* and leaf spot by *Xanthomonas campestris*, (2) finding treatments effective against both pathogens, (3) verifying compatibility with both tomato and pepper crops, and (4) checking application constraints. Each step depends on the previous one, forming a reasoning chain that standard flat datasets cannot support.

Contextual variability. The same disease presents differently depending on crop variety, growth stage, climate zone, and soil conditions. Advisory responses must account for this variability.

Evaluation challenges. Agricultural answers are often semantically correct but lexically diverse. A response recommending "copper-based fungicide spray" is equivalent to "apply Bordeaux mixture" in practice, but traditional lexical metrics such as F1 score and exact match would score this as incorrect.

Memory Systems as a Lens on Data Complexity We use memory-augmented AI systems as the evaluation setting for these knowledge discovery challenges. Such systems must handle heterogeneous information storage, temporal context, multi-hop retrieval, and knowledge integration across modalities.

We evaluate five memory architectures, each with a different way of organizing and retrieving agricultural knowledge. Retrieval-Augmented Generation (RAG) stores information as dense vector embeddings and retrieves relevant passages based on semantic similarity, treating agricultural knowledge as a flat collection of text chunks[11]. The Neuroscientific Memory System (NMS) organizes information into cognitively inspired memory types (working, episodic, semantic, and procedural), mirroring how human experts structure agricultural knowledge[12]. The Hybrid approach (NMS+RAG) combines neuroscientific memory organization with vector-based retrieval, attempting to get the benefits of both structured and unstructured approaches. BM25 with semantic reranking uses traditional term-frequency retrieval followed by neural reranking, representing a lightweight alternative to fully neural approaches. Finally, MemoryGraph extracts entities and relationships to build a knowledge graph, enabling explicit multi-hop traversal across agricultural concepts.

Our evaluation on 346 held-out test questions gives three main findings. First, no universal retrieval strategy exists: RAG achieves the highest conversational memory performance (85.6%) while MemoryGraph leads on multi-hop reasoning (97.0%), indicating that agricultural QA benchmarks should evaluate multiple retrieval paradigms rather than a single architecture. Second, graph-based knowledge organization substantially outperforms vector retrieval for multi-hop reasoning. Third, temporal information proves consistently challenging across all architectures (51.1–66.0% semantic correctness), demonstrating that longitudinal advisory datasets require explicit temporal information rather than relying only on unstructured dialogue text.

Performance differs by question category (disease identification, temporal reasoning, severity assessment, treatment planning) and question type (counting, negation, implicit, derived, and comparative). These differences identify which parts of agricultural QA are hardest to retrieve and reason over, and provide baseline evidence for memory-augmented agricultural advisory systems.

The AgriMemSynth Framework To address the challenges outlined above, we introduce AgriMemSynth, a framework that integrates two complementary modules. The AgriSynth module generates synthetic benchmark datasets using agricultural image repositories and structured crop-disease knowledge as grounding sources. AgriConvMem is generated from VLM-grounded descriptions of plant disease images, while AgriMultiHop is generated from a structured agricultural knowledge base using template-based reasoning patterns. The AgriMemory module then uses these datasets to evaluate memory-augmented systems, benchmarking five distinct architectures to reveal which aspects of agricultural knowledge are most challenging for AI systems. This two-stage approach (data generation followed by evaluation) is designed as a benchmark methodology rather than a single-model performance study: it separates dataset construction, memory organization, retrieval, and semantic evaluation so that memory-augmented systems can be compared under the same agricultural knowledge-discovery protocol.

The contributions This paper contributes to knowledge discovery and agricultural QA in five ways. First, AgriMemSynth connects synthetic data generation with memory-system evaluation, allowing controlled comparison of how knowledge representations affect agricultural question answering. Second, the AgriSynth module generates two benchmark datasets: AgriConvMem, a VLM-grounded conversational memory dataset with 500 multi-session conversations, and AgriMultiHop, a structured knowledge-base dataset with 2,000 multi-hop reasoning questions. Third, we report baseline results on held-out test splits containing 146 AgriConvMem questions and 200 AgriMultiHop questions across five retrieval and memory architectures (RAG, NMS, Hybrid, BM25, MemoryGraph). Fourth, we report 95% Wilson confidence intervals for binary LLM-as-Judge outcomes, separating descriptive architectural trends from claims of statistical separation. Fifth, we show a large lexical-semantic evaluation gap in agricultural QA and interpret the architecture comparison as an ablation-style analysis of flat vector retrieval, structured cognitive memory, lexical retrieval with reranking, naive hybridization, and graph-based retrieval. Sixth, this study presents a gap analysis regarding the integration of conversational and multi-hop reasoning benchmarks. Agricultural image datasets, memory-augmented systems, and synthetic data generation have typically been examined independently. To the best of our knowledge, no prior research has combined these elements into conversational and multi-hop reasoning benchmarks for agricultural question answering. Table 1 illustrates how AgriMemSynth integrates the most relevant dataset models identified in the literature, positioning it as the most comprehensive framework for accurate analysis.

Dataset	Domain	Multi-session	Multi-hop	Source resources	Shareable
PlantVillage	Agriculture	No	No	54,306 images	Yes
PlantDoc	Agriculture	No	No	2,598 images	Yes
LoCoMo	General	Yes	No	No images	Yes
HotpotQA	General	No	Yes	Text corpus	Yes
AgriMemory	Agriculture	Yes	Yes	Images + KB	Yes

Table 1. Comparison of AgriMemory with existing datasets. AgriMemory uniquely combines agricultural domain specificity with multi-session conversational memory and multi-hop reasoning capabilities.

The paper is organized as follows. Section 2 reviews related work in agricultural datasets, memory-augmented systems, and synthetic data generation. Section 3 describes the materials and methods, including source data, the AgriSynth data generation module, the AgriMemory evaluation module, and evaluation metrics. Section 4 presents the benchmark results. Section 5 discusses implications for knowledge representation and retrieval in agricultural QA. Section 6 concludes the paper.

2 Related Work

AgriMemSynth draws on three areas: agricultural image and QA datasets, memory-augmented language models for knowledge-intensive tasks, and synthetic data generation for specialized domains. This section reviews these areas and identifies the gap addressed by the benchmark.

Agricultural Image and Question-Answering Datasets Agricultural datasets have mainly developed around image classification. Early efforts focused on curating standardized image repositories for disease classification. PlantVillage[10] established the first large-scale open-access collection of plant disease images, containing 54,306 images across 14 crop species and 26 diseases, all captured under controlled laboratory conditions. This dataset enabled foundational work in deep learning for plant pathology[7].

PlantDoc[8] complements PlantVillage with 2,598 images from field conditions. These images include natural backgrounds, variable lighting, multiple diseases per leaf, and the visual noise typical of actual farming environments. The contrast between the two datasets shows the difference between laboratory images and field imagery. Eldem and Eldem (2026) introduced an innovative approach that serves as a foundation for intelligent digital agricultural applications[13]. The authors introduced a large language model (LLM)-based question answering (QA) system for agricultural experts, utilizing Gemini and ChatGPT for advanced reasoning. Enhanced prompting techniques enable this system to outperform existing methods. The application of LLMs to knowledge-based questions in agriculture addresses a significant gap in the field. This approach was presented as a digital assistant infrastructure tailored for agricultural applications. Kabala et al. (2023) evaluated a federated learning approach, comparing convolutional neural network (CNN)-based models with attention-based models for disease detection. The study utilized four image datasets from the PlantVillage platform to provide a comprehensive assessment. The results highlight the advantages of federated learning in agricultural disease classification, particularly with respect to user data security and the protection of sensitive information[14].

More recent efforts have expanded beyond image classification to include textual agricultural knowledge. Question-answering datasets for agriculture have emerged, but they typically focus on factoid questions with short answers[15]. None of the

existing agricultural datasets capture the multi-session conversational dynamics, temporal reasoning, or multi-hop knowledge integration that characterize real advisory interactions.

Memory-Augmented Language Models External memory is widely used in language models for knowledge-intensive tasks. Retrieval-Augmented Generation (RAG) pioneered this approach by combining a neural retriever with a sequence-to-sequence generator, enabling models to access external knowledge at inference time[11]. Dense Passage Retrieval (DPR) further improved retrieval quality through learned dense representations[16]. A broader survey by Zhang et al.[17] organizes these external-memory mechanisms for language-model-based agents.

In addition to Retrieval-Augmented Generation (RAG), researchers have investigated cognitively inspired memory architectures. Sumers et al.[12] introduced a cognitive science-based framework for language agents that differentiates among working memory (active processing), episodic memory (event recall), semantic memory (factual knowledge), and procedural memory (learned skills). This taxonomy aligns with agricultural advisory scenarios, in which an expert maintains the current conversation in working memory, retrieves information about previous patient visits through episodic memory, accesses disease knowledge from semantic memory, and implements diagnostic procedures from procedural memory. It is described in the session 3.3.2 and shown in Figure 5 in which a coordinating NMS orchestrates the four memories.

The LoCoMo benchmark[18] evaluated long-term conversational memory in LLM agents, demonstrating that maintaining context across extended dialogues remains challenging. HotpotQA[19] established evaluation protocols for multi-hop reasoning, requiring systems to combine information from multiple sources. Recent work in Data Mining and Knowledge Discovery also shows the relevance of graph-based multi-hop reasoning beyond general-domain QA. Zhang and Cheng[20] embed entity–relation semantics into rule mining for medical knowledge graph reasoning, while Le et al.[21] use dual policy-guided multi-hop path reasoning for explainable knowledge graph recommendation. However, these settings do not address the specialized vocabulary, reasoning patterns, and evaluation challenges of agricultural applications.

Synthetic Data Generation for Agriculture Synthetic data generation is often used when specialized-domain data is scarce. Vision-language models (VLMs) can generate realistic textual descriptions from images, enabling the creation of paired image-text datasets at scale. In agriculture, this approach is useful because expert-annotated conversational data is expensive and time-consuming to collect.

Our work builds on this trend by using VLMs to transform existing image repositories into conversational advisory data and by using structured agricultural knowledge to generate multi-hop reasoning questions. Rather than generating synthetic images, we generate synthetic textual interactions and QA items grounded in real crop-disease resources, adding the conversational and reasoning layers needed for advisory system evaluation.

3 Materials and Methods

3.1 Study design

AgriMemSynth is presented as a descriptive benchmark, not a training method. In this study, we analyze how different retrieval organizations behave when the same agricultural question-answering pipeline is assessed using two types of evidence: longitudinal conversational records and structured multi-hop facts. Our approach clarifies three components that are often conflated in agricultural AI systems: benchmark example construction, organization of retrieved memory, and evaluation of answer correctness.

By systematically comparing these structures, we aim to illuminate how each influences the retrieval process, the interpretability of the pipeline's decisions, and the quality of answers generated. The use of longitudinal conversational records mimics real-world interactions, capturing temporal context and the evolution of information needs, while structured multi-hop facts offer a synthesized, stepwise reasoning path toward answers. Through rigorous evaluation across these scenarios, AgriMemSynth provides insights into the strengths and limitations of various retrieval strategies, highlighting the importance of memory organization in complex, knowledge-intensive agricultural tasks. This work also underscores the necessity of disentangling benchmark design, memory retrieval methodology, and answer evaluation: three factors that, when conflated, can obscure the true sources of performance gains in agricultural AI systems.

The framework has two modules. AgriSynth generates benchmark resources from agricultural source material. AgriMemory evaluates retrieval architectures on those generated resources under a shared answer-generation protocol. The shared protocol is important because the comparison is intended to reflect differences in retrieval organization, not changes in the language model backbone or answer-generation interface.

3.2 Resources and benchmark branches

AgriMemSynth uses open agricultural resources as grounding material. PlantVillage contains 54,306 healthy and diseased plant-leaf images captured under controlled conditions across multiple crop species and disease categories[10]. PlantDoc provides 2,598 images collected from internet sources under more natural field conditions, including variable lighting, natural

backgrounds, and occlusion[8]. These image resources support the construction of image-grounded advisory scenarios for conversational memory. Together, the two repositories contain 56,904 images across many crop-disease combinations and include both controlled and field conditions. This makes them suitable for grounding conversational advisory scenarios. Figure 1 shows representative samples from both datasets.



(a) Tomato, Healthy (PlantVillage). No symptoms, uniform gray background.



(b) Potato, Early Blight (PlantVillage). Dark brown *A. solani* lesions on yellowing leaf, plain background.



(c) Potato, Late Blight (PlantVillage). Water-soaked *P. infestans* lesions, controlled conditions.



(d) Tomato, Late Blight (PlantDoc). Late blight in field setting with fruit and natural lighting.



(e) Potato, Early Blight (PlantDoc). Wilted and necrotic leaflets alongside healthy foliage in the field.



(f) Pepper, Bacterial Spot (PlantDoc). Subtle *Xanthomonas* symptoms under natural sunlight with overlapping foliage.

Figure 1. Representative disease images from PlantVillage (top row, laboratory conditions) and PlantDoc (bottom row, field conditions). The domain shift between controlled imagery with uniform backgrounds and field imagery with natural lighting, surrounding vegetation, and variable quality is a core challenge for agricultural knowledge extraction. The AgriSynth module draws from both sources to generate advisory scenarios spanning this spectrum of visual complexity.

For structured reasoning, AgriMemSynth uses crop-disease-treatment facts to create linked reasoning contexts. These contexts are used to generate questions that require more than one relation, for example connecting a crop to a disease, symptoms, and candidate management actions. The benchmark is therefore synthetic but grounded: it is not a collection of real farmer-expert conversations, and it is not an official evaluation on LoCoMo or HotpotQA, but it provides controlled agricultural QA tasks that can be shared and re-evaluated. AgriMemSynth generates two complementary text benchmark resources. AgriConvMem contains 500 multi-session agricultural advisory conversations split into 350 training, 75 validation, and 75 test conversations. The test split contains 146 evaluation questions. AgriMultiHop contains 2,000 structured reasoning questions split into 1,600 training, 200 validation, and 200 test questions. AgriConvMem evaluates disease identification, temporal tracking, severity assessment, and treatment planning. AgriMultiHop evaluates counting, negation, implicit, derived, and comparative reasoning.

The AgriSynth module generates structured conversational and reasoning datasets from two grounding sources: plant disease images and structured crop-disease knowledge. Figure 2 shows the workflow. The AgriMemory module then uses these datasets for memory-system evaluation.

Stage 1: Scenario Selection. For each data sample, we define an agricultural scenario by selecting a crop species, disease type, geographic location, and severity level. Scenarios are designed to cover realistic combinations encountered in agricultural practice, with severity levels ranging from early/mild to advanced/severe.

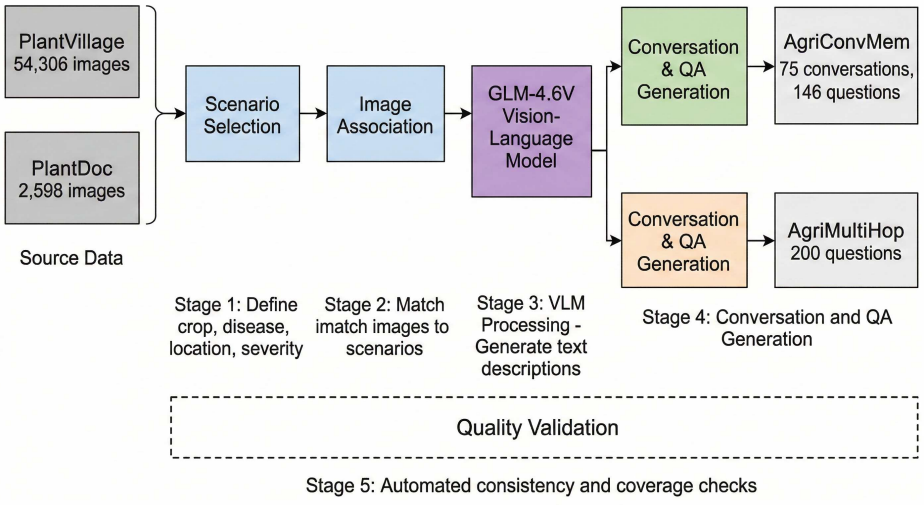


Figure 2. AgriSynth data-generation pipeline. Agricultural source resources are converted into benchmark examples through scenario selection, grounding, vision-language or structured-knowledge processing, question generation, and validation.

Stage 2: Grounding Source Association. Each scenario is paired with an appropriate grounding source. For AgriConvMem, scenarios are paired with relevant images from PlantVillage and/or PlantDoc to simulate a farmer sharing photos across multiple advisory sessions. For AgriMultiHop, scenarios are paired with structured crop-disease-treatment facts from the agricultural knowledge base to support cross-referencing questions.

Stage 3: VLM and Knowledge-Base Processing. For AgriConvMem, we use GLM-4.6V[22], a vision-language model, to generate detailed textual descriptions from the associated images. The VLM produces structured observations including visible symptoms, estimated disease stage, affected plant parts, and environmental context visible in the image. For AgriMultiHop, the generator uses structured crop-disease-treatment facts to construct reasoning contexts that require computation or relation chaining.

Stage 4: Conversation and QA Generation. We generate two types of outputs. For AgriConvMem, VLM-generated descriptions ground multi-session conversations that simulate 3–5 sessions between a farmer and an agricultural expert, covering disease identification, symptom progression, severity assessment, and treatment recommendations, with questions generated to test recall across sessions. For AgriMultiHop, template-based generation constructs multi-hop question-answer pairs that require information from multiple disease profiles, treatment databases, or crop management guidelines, with each question annotated by its reasoning type (counting, negation, implicit, derived, or comparative). The detailed structure and characteristics of these two generated datasets are described in the next paragraph.

Stage 5: Quality Validation. Generated samples undergo automated validation to ensure consistency with their grounding sources, appropriate difficulty distribution, valid answer formats, and coverage across agricultural categories.

The following paragraphs analyze the types of data generated, offers a clear justification for their inclusion, presents their percentage distributions, and evaluates their relevance to the identified challenges in agricultural question answering. This analysis facilitates an assessment of the database's overall validity by demonstrating how each generated subset corresponds to a specific retrieval or reasoning challenge.

Figure 3 presents the overall structure of the datasets generated by the AgriSynth module and their relationship to the AgriMemory evaluation module. The figure separates the benchmark database into two complementary branches. The first branch, AgriConvMem, represents multi-session agricultural advisory conversations grounded in plant disease images; it is linked to four question categories: disease identification, temporal tracking, severity assessment, and treatment planning. The second branch, AgriMultiHop, represents structured crop-disease knowledge and is linked to five reasoning types: counting, negation, implicit, derived, and comparative questions. These two branches are connected to the evaluation stage, where memory-augmented systems are tested on the held-out questions. The relationship among the modules is therefore sequential and functional: source data are transformed by AgriSynth into task-specific benchmark datasets, the datasets define the question categories and reasoning types, and AgriMemory uses these outputs to evaluate how different memory architectures retrieve and combine agricultural knowledge. Table 2 provides an overview of the generated datasets. Accordingly, this dataset will support the evaluation of enhanced memory system in AgriMemory framework for answering agricultural questions.

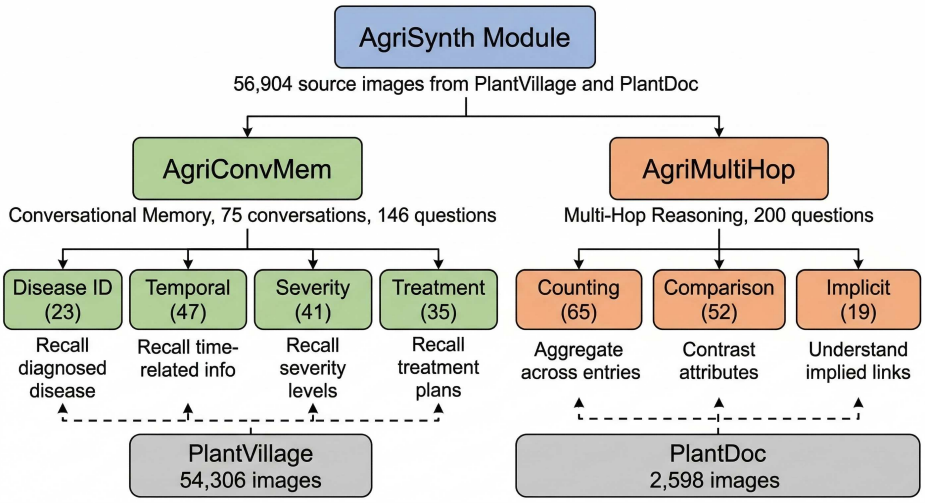


Figure 3. Overview of the AgriMemSynth benchmark resources. AgriConvMem evaluates multi-session conversational memory, while AgriMultiHop evaluates structured multi-hop reasoning.

AgriConvMem: Conversational Memory Dataset. AgriConvMem contains 500 generated multi-session conversations split into 350 training, 75 validation, and 75 test conversations. The benchmark evaluation in this paper uses the held-out test split, which contains 75 conversations and 146 evaluation questions. Each conversation simulates a realistic farmer-expert interaction spanning multiple visits. Questions are categorized into four agricultural dimensions: disease Identification questions (23 test questions) test whether the system correctly recalls the diagnosed disease from previous sessions, for example: "What disease was identified in my tomato plants during our first consultation?" Temporal Tracking questions (47 test questions) evaluate recall of time-related information such as when symptoms first appeared, how long treatments have been applied, or the progression timeline. Severity Assessment questions (41 test questions) test recall of disease severity levels, affected area percentages, and progression assessments. Treatment Planning questions (35 test questions) evaluate recall of recommended treatments, application schedules, and follow-up instructions.

AgriMultiHop: Multi-Hop Reasoning Dataset. AgriMultiHop contains 2,000 generated questions split into 1,600 training, 200 validation, and 200 test questions. The benchmark evaluation in this paper uses the 200-question held-out test split. Test questions require information synthesis across multiple knowledge sources and are distributed across five reasoning types: (i) Counting questions (65, 32.5%) require aggregating information across multiple entries to produce numerical answers, such as "How many diseases affect both tomato and potato crops?"; (ii) negation questions (19, 9.5%) require identifying exclusions or conditions that do not hold; (iii) nmplicit questions (19, 9.5%) require understanding implied relationships that are not explicitly stated in the knowledge base, such as inferring that if a fungicide controls a particular disease and that disease affects a specific crop; (iiii) then the fungicide protects that crop. Derived questions (45, 22.5%) require deriving an answer from intermediate crop-disease or symptom-disease relationships; (iiiii) Comparative questions (52, 26.0%) require contrasting attributes of different diseases, treatments, or crops to identify similarities or differences.

Table 2. Summary of the datasets generated by the AgriSynth module. Evaluation results in this paper use the held-out test splits.

Attribute	AgriConvMem	AgriMultiHop
Total samples	500 conversations	2,000 questions
Train/validation/test split	350/75/75	1,600/200/200
Test evaluation questions	146	200
Grounding source	Plant disease images	Structured knowledge base
Source repositories	PlantVillage, PlantDoc	Crop-disease-treatment facts
Question categories/types	4 categories	5 reasoning types
Average conversation length	3–5 sessions	N/A
Data format	JSON	JSON

3.3 AgriMemory Module: Evaluating Memory Systems on Generated Data

The AgriMemory module uses the datasets generated by AgriSynth to benchmark memory-augmented retrieval architectures on agricultural question answering. This subsection describes the five memory system architectures evaluated, their organizational principles, and how they differ in handling agricultural knowledge.

Memory System Architectures (MSA) Memory-augmented systems organize knowledge by structure, function, and temporal context rather than relying only on keyword matching or semantic similarity. Rather than treating all information uniformly, these systems can distinguish between factual knowledge (disease characteristics), procedural knowledge (diagnostic steps), episodic events (farmer visits), and active reasoning state (current conversation context). This structured organization enables more precise retrieval: temporal queries can target timestamped episodes, procedural questions can access step-by-step routines, and multi-hop reasoning can traverse explicit entity relationships. For agricultural QA, this means that information can be stored and retrieved in ways that preserve structure, sequence, and relationships.

MSA orchestrates five memory systems with different approaches to organizing and retrieving agricultural knowledge. Figure 4 illustrates parallel architectural differences, and a backbone language module in which all five memory systems use DeepSeek-V3[23]. DeepSeek-V3 is a Mixture-of-Experts model with 671 billion total parameters (37 billion activated per token), providing strong reasoning capabilities while maintaining computational efficiency. Using a single backbone model across all systems ensures that performance differences reflect the memory architecture rather than the language model.

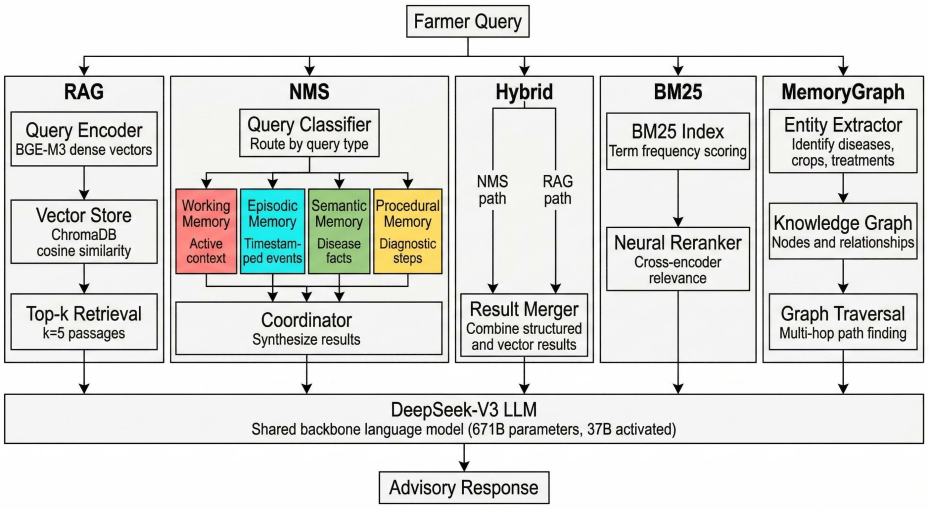


Figure 4. Architectural comparison of the five memory systems evaluated in the AgriMemory module. RAG uses flat vector storage with cosine similarity retrieval. NMS organizes knowledge into four cognitively inspired stores (working, episodic, semantic, procedural). Hybrid combines both NMS and RAG paths. BM25 uses two-stage lexical retrieval with neural reranking. MemoryGraph builds a knowledge graph of entities and relationships for explicit multi-hop traversal. All five systems share DeepSeek-V3 (671B parameters, 37B activated per token) as the backbone language model, ensuring that performance differences reflect memory architecture rather than language model capabilities. The shared LLM receives retrieved context from each memory system and generates the final advisory response.

3.3.1 Retrieval-Augmented Generation (RAG)

RAG implements a two-stage retrieve-then-generate pipeline[11]. Agricultural conversations and knowledge documents are chunked into passages and encoded into dense vector representations using BGE-M3 embeddings[24]. At query time, the system retrieves the top-k most semantically similar passages and provides them as context to the language model. RAG treats all agricultural knowledge uniformly as text chunks, without distinguishing between disease facts, treatment protocols, or temporal observations.

3.3.2 Neuroscientific Memory System (NMS)

NMS organizes information into four cognitively inspired memory stores based on the framework of Sumers et al.[12]. Working Memory maintains the current conversation context and active reasoning state, holding only the most recent and immediately relevant information, analogous to a farmer's active attention during a consultation. Episodic Memory stores specific interaction events with temporal markers, recording each farmer visit, observation, and recommendation as a distinct episode to enable recall of "what happened when," which is important for tracking disease progression across sessions. Semantic Memory

contains general agricultural knowledge including disease characteristics, treatment protocols, and crop management guidelines, representing the expert's accumulated factual knowledge. Procedural Memory encodes diagnostic and treatment procedures as structured routines, such as the step-by-step process for identifying early blight (check leaf spots, examine concentric rings, assess stem involvement).

A coordinator module routes incoming queries to the appropriate memory store(s) based on query type classification, then synthesizes retrieved information before passing it to the language model. Figure 5 illustrates this taxonomy and how each memory type maps to agricultural advisory functions.

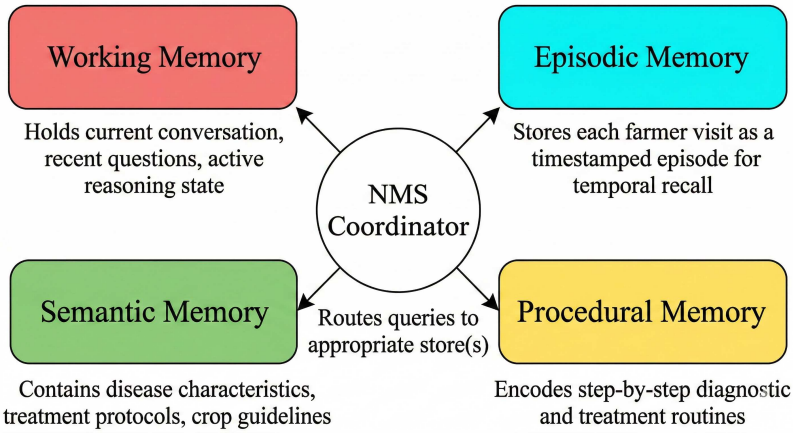


Figure 5. Taxonomy of memory types in the Neuroscientific Memory System (NMS) and their mapping to agricultural advisory functions. Working memory holds the active conversation state. Episodic memory stores timestamped farmer visits and observations. Semantic memory contains general disease and treatment knowledge. Procedural memory encodes diagnostic and treatment routines.

The five memory systems differ in their retrieval mechanisms and in how they organize information into memory types. Table 3 summarizes the memory organization across all evaluated systems.

3.3.3 Hybrid (NMS+RAG)

The Hybrid system combines NMS's structured memory organization with RAG's dense retrieval capability. Information is stored in both the four-store NMS architecture and a parallel vector index. At query time, the system retrieves from both sources and merges the results, aiming to capture both the structured recall of NMS and the semantic flexibility of RAG.

The Hybrid architecture tests whether combining complementary memory paradigms yields additive benefits or introduces interference. Dual embedding (storing the same information in both structured and vector formats) increases storage overhead and retrieval complexity. The architectural question is whether the combined retrieval paths provide better context to the language model. In practice, the Hybrid system's performance shows that naive combination of memory architectures does not guarantee improved results (the merging strategy, deduplication logic, and context window management become important factors). This finding is relevant to knowledge discovery systems that combine heterogeneous retrieval backends: systems should select architectures based on dominant query patterns (conversational recall vs. multi-hop reasoning) or implement intelligent routing that directs queries to the most appropriate memory type. The Hybrid system is therefore a cautionary baseline: architectural complexity must be justified by measured performance gains rather than assumed synergy.

3.3.4 BM25 with Semantic Reranking

This system uses BM25, a classical term-frequency-based retrieval algorithm, as the first-stage retriever. Retrieved passages are then reranked using a neural cross-encoder that scores query-passage relevance. This two-stage approach is computationally lighter than fully neural retrieval and provides a baseline that relies more on lexical matching than semantic understanding.

3.3.5 MemoryGraph

MemoryGraph extracts entities (crops, diseases, treatments, symptoms) and their relationships from agricultural text to construct a knowledge graph. At query time, the system identifies relevant entities in the question, traverses the graph to find connected information, and provides the subgraph context to the language model. This architecture is particularly suited for multi-hop reasoning, where answering a question requires following chains of relationships (e.g., disease $\rightarrow$ treatment $\rightarrow$ application method $\rightarrow$ crop compatibility).

Table 3. Memory organization across the five evaluated systems. NMS and Hybrid use structured memory types, while RAG, BM25, and MemoryGraph treat information uniformly.

Memory Type	RAG	NMS	Hybrid	BM25	MemoryGraph
Working Memory	–	✓	✓	–	–
Episodic Memory	–	✓	✓	–	–
Semantic Memory	–	✓	✓	–	–
Procedural Memory	–	✓	✓	–	–
Vector Retrieval	–	–	✓	✓	–
Term-Frequency Index	✓	–	–	–	–
Knowledge Graph	–	–	–	–	✓

3.4 Retrieval systems

AgriMemory evaluates five retrieval architectures under the same benchmark protocol. RAG uses dense vector retrieval over flat text chunks[11]. NMS uses working, episodic, semantic, and procedural memory stores inspired by cognitive memory architectures[12]. Hybrid combines NMS and RAG retrieval paths. BM25 with semantic reranking provides a lexical-first baseline. MemoryGraph constructs an entity-relation graph and retrieves through graph traversal. Figure 4 summarizes the retrieval architectural system.

All systems are evaluated with the same answer-generation pipeline. This setup makes the comparison architectural: it describes how flat dense retrieval, structured memory, lexical retrieval, hybrid retrieval, and graph traversal behave on the same agricultural tasks. It should not be read as a causal ablation of a single system.

3.4.1 Evaluation metrics and uncertainty

We employ two categories of evaluation metrics to assess system performance. This dual-metric approach captures both surface-form answer fidelity (lexical metrics) and meaning-level correctness (semantic metric). Agricultural QA requires validation at both levels: systems must retrieve relevant information accurately while also generating responses that are semantically correct even when expressed differently from reference answers.

Lexical metrics: measure surface-level text overlap between generated and reference answers, providing a baseline for retrieval precision and answer fidelity. F1 is a general precision–recall measure with a long history in information retrieval and classification evaluation[25,26]. In question answering, token-level F1[27] computes the harmonic mean of precision and recall between predicted and reference answers, balancing the retrieval of relevant information against the inclusion of irrelevant content. Exact Match (EM)[27] is a binary score indicating whether the predicted answer exactly matches the reference, serving as a strict measure of retrieval accuracy particularly important for factual agricultural data like disease names and treatment protocols. BLEU-1[28] measures unigram precision of the predicted answer against the reference and is applied to AgriConvMem only. BLEU-1 is not applied to AgriMultiHop because AgriMultiHop consists of short-form factual answers (e.g., disease names, counts, treatment protocols) where exact token matching is more appropriate, whereas AgriConvMem contains longer conversational responses where n-gram overlap better captures answer quality.

Semantic metric: captures meaning-level correctness, addressing the fundamental challenge that agricultural knowledge can be expressed in many lexically different but semantically equivalent ways. LLM-as-Judge[29] uses an independent language model to evaluate whether the generated answer is semantically correct given the question and reference answer. This metric matters for agricultural QA because correct answers can be expressed in many lexically different ways (e.g., “Alternaria solani” vs. “early blight fungus”, or “apply copper-based fungicide” vs. “use copper hydroxide spray”). Semantic evaluation therefore helps determine whether retrieved and generated agricultural knowledge retains its meaning even when surface-level wording differs.

Combining lexical and semantic evaluation is important for agricultural data validation. The gap between the two metric categories shows why agricultural knowledge is difficult to evaluate with surface-form matching alone.

Statistical Uncertainty for Judge-Based Evaluation: Because LLM-as-Judge correctness is recorded as a binary outcome for each held-out test question, we estimate uncertainty using 95% Wilson confidence intervals for binomial proportions. For a system with k judged-correct answers out of n questions, the interval is computed as:

$$\hat{p}_W = \frac{\hat{p} + z^2/(2n)}{1 + z^2/n}, \qquad \text{halfwidth} = \frac{z}{1 + z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}, \tag{1}$$

where $\hat{p} = k/n$ and $z = 1.96$ for a 95% interval.

Table 4. Overall benchmark evaluation results across five memory-augmented systems on AgriConvMem (146 questions) and AgriMultiHop (200 questions). LLM-as-Judge scores consistently exceed lexical metrics, revealing the evaluation gap inherent in agricultural QA. Note: Lexical metrics (F1, EM, BLEU-1) are reported as decimals (0-1 range) following standard NLP conventions, while Judge scores are shown as decimals here for direct comparison.

	AgriConvMem				AgriMultiHop		
System	**F1**	**EM**	**BLEU-1**	**Judge**	**F1**	**EM**	**Judge**
RAG	0.287	0.027	0.691	0.856	0.362	0.100	0.960
NMS	0.269	0.014	0.657	0.842	0.373	0.105	0.955
Hybrid	0.268	0.027	0.655	0.815	0.360	0.095	0.950
BM25	0.263	0.014	0.625	0.822	0.356	0.100	0.955
MemoryGraph	0.267	0.021	0.643	0.808	0.362	0.095	0.970

Accurate interpretation of confidence intervals requires distinguishing between different sources of variability. The Wilson confidence interval (CI) specifically addresses sampling uncertainty: for a fixed set of n test questions, it quantifies how much the observed proportion $\hat{p}$ would vary if another random sample of the same size were drawn from the same question population. A wide interval (for example, ± 6 to 8 percentage points) suggests that the test split is small, so the true underlying accuracy could reasonably be several points higher or lower than the observed value. In contrast, a narrow interval suggests that the observed proportion is a reliable estimate of system performance for that question type.

The intervals do *not* constitute new experiments and should not be interpreted as validation on held-out data beyond what was already evaluated. Specifically, they do not account for (i) variability in the LLM judge itself across repeated evaluations of the same answer, (ii) sensitivity to the particular questions selected for the test split, or (iii) generalization to agricultural domains or language models not represented in the benchmark. Readers should therefore treat a system whose 95% CI is entirely above another system's CI as *descriptively* superior on this benchmark, without claiming statistical significance in the hypothesis-testing sense.

We use the intervals for two practical purposes. First, they identify *stable benchmark patterns*: differences that persist even after accounting for sampling variability, such as the consistently large gap between lexical metrics and LLM-as-Judge scores (a gap of 2–3$\times$ that is far wider than any confidence interval). Second, they flag *ambiguous rankings*: when the confidence intervals of two architectures overlap substantially, the observed ordering should be read as a descriptive trend rather than evidence that one system is definitively superior. In the results that follow, both stable patterns and overlapping rankings are identified explicitly so that readers can draw appropriately calibrated conclusions.

4 Results

Aggregate performance across the five memory systems follows a consistent pattern: semantic evaluation is much higher than lexical metrics, which supports the need for domain-aware validation in agricultural QA. Table 4 presents the aggregate performance of all five memory systems across both dataset subsets.

The lexical-semantic gap. As shown in Table 4, across all systems, LLM-as-Judge scores are 2–3 times higher than F1 scores. On AgriConvMem, systems achieve F1 scores of 0.263–0.287 but semantic correctness of 0.808–0.856. On AgriMultiHop, the gap is even more pronounced: F1 of 0.356–0.373 versus semantic correctness of 0.950–0.970. This gap reflects agricultural language itself: the same concept can be expressed in different words while remaining semantically equivalent. Table 5 reports 95% Wilson confidence intervals for the binary LLM-as-Judge outcomes. The intervals support two points. First, the conversational-memory task has lower semantic correctness than the multi-hop task across all memory systems. Second, the best-performing architecture differs by task, but the small differences among top systems should be read as descriptive benchmark trends rather than definitive statistical separation.

Conversational memory is harder than multi-hop reasoning. As shown in Table 5, all systems perform better on AgriMultiHop (95.0–97.0% semantic correctness) than on AgriConvMem (80.8–85.6%). This suggests that maintaining context across multi-session agricultural conversations is more challenging than synthesizing information from structured knowledge sources, an important consideration for knowledge discovery architectures. Figure 6 shows the gap between lexical(blue bars) and semantic(orange bars) evaluation metrics across all five memory systems. The gap indicates that agricultural QA needs validation methods that capture semantic equivalence beyond surface-level text matching.

Other two observations from these results are relevant to agricultural data complexity:

No single architecture dominates. As shown in Tables 4 and 5, RAG achieves the highest conversational memory score (85.6%, 95% CI: 79.0–90.4), while MemoryGraph leads on multi-hop reasoning (97.0%, 95% CI: 93.6–98.6). Because

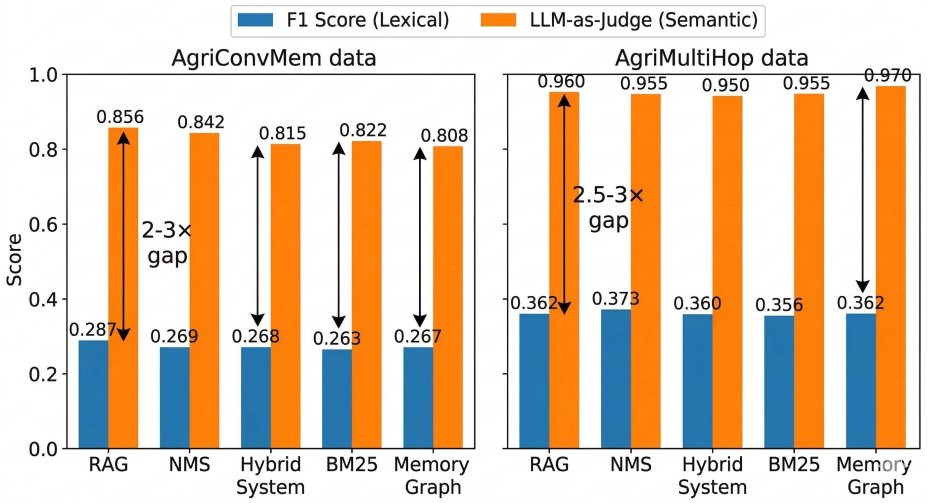


Figure 6. Lexical-semantic evaluation gap across five memory systems on AgriConvMem and AgriMultiHop datasets. The figure shows grouped bar charts comparing F1 scores (lexical metric, blue bars) against LLM-as-Judge scores (semantic metric, orange bars) for each system. Left panel: AgriConvMem results showing F1 scores of 0.263–0.287 versus semantic correctness of 0.808–0.856 (2–3× gap). Right panel: AgriMultiHop results showing F1 scores of 0.356–0.373 versus semantic correctness of 0.950–0.970 (2.5–3× gap). The consistent gap across all systems indicates that agricultural knowledge can be expressed in many lexically different but semantically equivalent ways. Traditional NLP metrics therefore substantially underestimate system performance in agricultural QA.

confidence intervals overlap among several top systems, these rankings should be interpreted as architecture-sensitive benchmark patterns rather than proof that one system is universally superior. The task-dependent pattern still indicates that agricultural data has multiple dimensions of complexity that no single retrieval strategy fully addresses.

Hybrid architecture shows combination challenges. As shown in Table 4, the Hybrid system (NMS+RAG), despite combining two complementary approaches, achieves lower performance (81.5% on AgriConvMem, 95.0% on AgriMultiHop) than either NMS or RAG alone. This result suggests that dual embedding can introduce retrieval interference: the system must merge results from structurally different memory stores (cognitively organized vs. flat vectors), leading to context dilution when both retrieval paths return partially overlapping information. The finding suggests that knowledge discovery systems should prioritize architectural specialization over naive combination or should route queries intelligently to the most appropriate retrieval backend.

4.1 Category-Level Analysis: AgriConvMem

The conversational memory dataset enables analysis of how different question categories challenge memory systems. Disease identification and severity assessment represent relatively straightforward retrieval tasks, while temporal reasoning and treatment planning require more complex information integration. Table 6 breaks down performance by agricultural question category, revealing which aspects of crop disease data are most challenging to share and retrieve.

Lexical metrics offer one perspective on system performance, whereas semantic evaluation reveals which question categories present substantive challenges. Category-stratified lexical F1 outputs are unavailable for BM25; therefore, Table 6 presents only the four systems with available category-level lexical results, avoiding unsupported estimations. BM25 is included in the

Table 5. LLM-as-Judge correctness with 95% Wilson confidence intervals on held-out test splits. Counts are reconstructed from binary judge outcomes: 146 AgriConvMem questions and 200 AgriMultiHop questions.

System	AgriConvMem correct (95% CI)	AgriMultiHop correct (95% CI)
RAG	125/146, 85.6% (79.0–90.4)	192/200, 96.0% (92.3–98.0)
NMS	123/146, 84.2% (77.5–89.3)	191/200, 95.5% (91.7–97.6)
Hybrid	119/146, 81.5% (74.4–87.0)	190/200, 95.0% (91.0–97.3)
BM25	120/146, 82.2% (75.2–87.5)	191/200, 95.5% (91.7–97.6)
MemoryGraph	118/146, 80.8% (73.7–86.4)	194/200, 97.0% (93.6–98.6)

Table 6. AgriConvMem token-level F1 scores. Rows denote question categories and columns denote retrieval systems. BM25 is omitted because category-stratified lexical outputs were not available for this analysis; its aggregate F1 score is reported in Table 4, and its category-level semantic scores are reported in Table 7.

Question category	#	Token-level F1 score by retrieval system			
		RAG	NMS	Hybrid	MGraph
Disease ID	23	0.658	0.645	0.635	0.670
Severity	41	0.226	0.228	0.220	0.206
Treatment	35	0.202	0.193	0.187	0.185
Temporal	47	0.221	0.179	0.191	0.184

aggregate comparison in Table 4 and in the category-level semantic comparison in Table 7.

Table 7. AgriConvMem LLM-as-Judge scores (%) by question category. Disease identification and severity consistently achieve high scores across all systems and are omitted. Treatment and temporal questions show meaningful variation. Note: Judge scores are presented as percentages in category-level and type-level tables for improved readability.

Category	#	RAG	NMS	Hybrid	BM25	MGraph
Treatment	35	85.7	82.9	80.0	82.9	85.7
Temporal	47	66.0	63.8	59.6	57.4	51.1

The category-level analysis shows a hierarchy of data complexity in agricultural advisory:

Disease identification and severity are well-handled by all systems, achieving high semantic correctness (Table 7). These categories involve relatively stable, factual information that maps well to standard retrieval approaches.

Treatment recommendations present moderate difficulty (80.0–85.7% semantic correctness, Table 7). Treatment information is more variable because it depends on disease stage, crop variety, and local conditions, requiring systems to retrieve and contextualize multiple pieces of information.

Temporal information is consistently the most challenging category (51.1–66.0% semantic correctness, Table 7). Questions about when symptoms appeared, how long treatments have been applied, or the timeline of disease progression require precise recall of time-anchored information from specific conversation sessions. This result indicates that temporal context matters for disease management and remains difficult for current memory architectures to preserve.

4.2 Question Type Analysis: AgriMultiHop

The multi-hop reasoning dataset tests systems' ability to chain information across multiple retrieval steps. Different question types (counting, negation, and comparative reasoning) require different traversal strategies through the knowledge base. Table 8 presents the multi-hop reasoning results by question type.

Table 8. AgriMultiHop LLM-as-Judge scores (%) by question type on the 200-question test split.

Type	#	RAG	NMS	Hybrid	BM25	MGraph
Counting	65	93.8	92.3	92.3	95.4	98.5
Negation	19	100.0	100.0	100.0	100.0	100.0
Comparative	52	94.2	92.3	90.4	92.3	90.4

The multi-hop analysis shows that:

Counting questions show the most inter-system variation (92.3–98.5%, Table 8), with MemoryGraph's explicit entity-relationship structure providing an advantage for aggregation tasks. This is consistent with graph-based representations supporting counting operations over agricultural entities.

Negation questions are solved consistently by all systems in the test split (100.0%, Table 8), suggesting that these generated items are less discriminative for the evaluated memory architectures than counting or comparative questions.

Comparative questions show moderate inter-system variation (Table 8), reflecting the challenge of aligning information across different disease profiles when the comparison criteria are not explicitly stated.

The category-level and question-type analyses show that agricultural data complexity differs across conversational memory and multi-hop reasoning tasks. Conversational memory struggles most with temporal information, while multi-hop reasoning

shows the advantages of graph-based representations for aggregation and relationship inference. These results inform how agricultural data should be structured and shared.

5 Discussion

The benchmark results clarify several design requirements for memory-augmented agricultural advisory systems. This section discusses implications for knowledge representation and retrieval, applications beyond agriculture, synthetic data, and the scope of the study. Several scope boundaries should be considered when interpreting the findings. The evaluation employs DeepSeek-V3 as the backbone language model across all memory systems; performance patterns may vary with alternative language models, potentially affecting the relative ranking of architectures. The current benchmark does not include an LLM-only no-memory baseline: that is, a condition in which the language model answers questions solely from its pre-trained knowledge, without retrieving any external context. This omission matters because, without such a baseline, it is impossible to determine how much of each system's accuracy comes from the retrieval mechanism itself versus what the language model would have produced anyway. A reader might reasonably ask: does adding a memory system actually help, or does DeepSeek-V3 already "know" most of the answers from training? The present results cannot answer this question directly. They should therefore be read as a comparison *among* five retrieval and memory architectures, showing which organizational strategy is most effective for each question type, rather than as evidence that any of the five systems outperforms a closed-book model. Establishing the retrieval gain over a no-memory baseline is an important direction for follow-up work. The datasets are generated in English, which constrains immediate applicability to multilingual agricultural advisory contexts prevalent in developing regions where language diversity is substantial. The vision-language model used for data generation (GLM-4.6V) processes disease images through its learned representations, which may systematically emphasize certain symptom presentations while underrepresenting others. The benchmark focuses on single-turn retrieval scenarios and does not capture the iterative, interactive dynamics of real-world advisory sessions where farmers progressively refine their queries based on system responses. The source image repositories cover 14+ crop species and span both controlled laboratory conditions (PlantVillage) and realistic field conditions (PlantDoc), providing a broad foundation for grounding agricultural advisory scenarios. Together, these scope considerations define the setting in which AgriMemSynth operates and point toward natural extensions, such as additional backbone models, multilingual datasets, and a wider crop inventory, that future work can build upon.

5.1 Relation to Recent Agricultural AI Systems

Recent agricultural AI research clarifies why a benchmark must evaluate more than image classification accuracy, real-time CNN-based diagnosis has emphasized rapid and scalable inference from plant images[30], while drone imagery combined with environmental sensing has extended disease monitoring to large and heterogeneous agricultural areas[31]. Other work has addressed deployment and trust through lightweight explainable CNN architectures suitable for mobile or edge settings[32] and through deep diagnostic models equipped with visual explanation methods[33]. AgriMemSynth builds on these foundations, and on others reported in section 2, by introducing an extended evaluation layer that rigorously tests the system's ability to maintain robust connectivity between retrieved pieces of agricultural evidence. Specifically, it assesses whether information supported by multi-hop relationships can be accurately preserved, integrated, and referenced throughout a conversation. This approach enables a more nuanced analysis of how retrieval strategies impact the continuity and coherence of evidence, ensuring that critical contextual information is not lost as the dialogue progresses. AgriMemSynth facilitates a deeper understanding of both the strengths and limitations of current agricultural AI systems, ultimately guiding the design of more effective, reliable, and transparent advisory solutions.

5.2 The Role of Synthetic Data in Agricultural AI

AgriMemSynth shows that synthetic data generation using vision-language models can produce evaluation datasets that capture meaningful complexity in agricultural question answering. The generated conversations and reasoning questions expose real architectural trade-offs between memory systems, suggesting that the synthetic data captures meaningful properties of agricultural advisory interactions.

However, synthetic data generation involves certain trade-offs. The conversations are generated by a language model and may not fully capture the linguistic patterns of actual farmers, particularly those communicating in non-native languages or using regional agricultural terminology. The disease descriptions, while grounded in real images, are filtered through the VLM's training distribution and may miss rare or atypical presentations. Validation against real farmer-expert interactions would help quantify the synthetic-to-real gap.

5.3 Implications Beyond Agriculture

Although AgriMemSynth is grounded in crop disease advisory, the benchmark methodology addresses a broader knowledge discovery problem: how to evaluate AI systems that must retrieve, preserve, and combine domain knowledge distributed

across observations, time, and structured relations. Similar requirements arise in clinical decision support, environmental monitoring, industrial maintenance, disaster response, recommendation, and scientific literature assistance. Recent Data-Mining-Knowledge-Discovery studies in medical graph reasoning and explainable knowledge graph recommendation show that entity–relation semantics, contextual subgraphs, and multi-hop path reasoning are now central concerns across applied knowledge discovery[20,21]. In each setting, useful answers may depend on longitudinal records, domain-specific terminology, implicit relationships, and evidence scattered across heterogeneous sources.

The results suggest lessons for other domains. First, benchmarks for domain-specific QA should evaluate both conversational memory and structured reasoning because these capabilities stress different retrieval assumptions. Second, semantic evaluation is necessary when expert terminology admits multiple correct paraphrases. Third, graph-based and vector-based representations should be treated as complementary design choices rather than interchangeable retrieval backends. AgriMemSynth contributes a controlled protocol for comparing these choices in a concrete application setting where errors have practical consequences.

5.4 Implications for Knowledge Representation and Retrieval

The benchmark results have several implications for how agricultural data should be represented, retrieved, shared, and evaluated across AI systems. The consistent difficulty of temporal questions across all memory architectures, with semantic correctness ranging from 51.1% to 66.0% in Table 7, indicates that time-related agricultural information is difficult to recover reliably when it remains embedded only in unstructured conversation. Agricultural knowledge representations should therefore encode timestamps for observations, durations of symptom progression, treatment start and end dates, ordering relations between visits, and explicit links between interventions and subsequent outcomes. Such temporal structure would allow retrieval systems to distinguish the current state from historical observations and to answer questions about change, recurrence, and treatment response without relying solely on implicit linguistic cues. The large gap between lexical F1 scores of 0.263–0.287 and semantic scores of 0.808–0.856 on AgriConvMem (Table 4) further shows that representation quality cannot be judged adequately through surface-form overlap alone. Agricultural terminology permits abbreviations, synonyms, paraphrases, and answers expressed at different levels of specificity, so validation protocols should combine reproducible lexical measures with semantic assessment and, where practical, expert review. The multi-hop results also suggest that representation should match the relational structure of the task. MemoryGraph achieves the highest AgriMultiHop score at 97.0% (Table 4), indicating that explicit entities, relations, and traversable paths can support questions that require connecting crops, diseases, symptoms, and treatments across several facts. However, RAG performs best on conversational memory, which shows that graph structure is not a universal replacement for dense retrieval over narrative records. A practical agricultural QA system should therefore support interoperable representations, using vector stores for semantically rich dialogue and document retrieval, graph structures for relational inference, and structured temporal records for longitudinal reasoning. The Hybrid results also caution that simply querying and merging multiple backends is insufficient: its lower performance than NMS or RAG alone indicates that overlapping evidence, inconsistent ranking scales, duplicated context, and limited context windows can offset the expected benefits of combining representations. Multi-backend systems should instead use query-aware routing, calibrated relevance scoring, provenance-preserving deduplication, and explicit context allocation so that each question is handled by the representation most suited to its information needs. Overall, the findings support a retrieval architecture in which representation choices are task-dependent, semantic evaluation is treated as a necessary complement to lexical scoring, and added architectural complexity is justified by measured gains rather than assumed to improve performance automatically.

6 Conclusion

This paper presents AgriMemSynth, a framework that integrates synthetic data generation with memory system evaluation for agricultural question answering. The framework operates through two complementary modules: AgriSynth generates benchmark datasets from plant disease images and structured agricultural knowledge, while AgriMemory evaluates memory-augmented systems using these generated datasets. The AgriSynth module produces two structured benchmarks: AgriConvMem, a conversational memory dataset with 500 conversations, and AgriMultiHop, a multi-hop reasoning dataset with 2,000 questions. The benchmark evaluation uses held-out test splits containing 146 AgriConvMem questions and 200 AgriMultiHop questions. The AgriMemory module uses these test splits to benchmark five memory architectures (RAG, NMS, Hybrid, BM25, and MemoryGraph), revealing which aspects of agricultural knowledge pose the greatest challenges for AI systems.

The AgriMemory evaluation on AgriSynth datasets shows that agricultural data has complexity that standard NLP metrics do not capture well. The large gap between lexical and semantic evaluation metrics confirms that agricultural question answering requires domain-aware evaluation methods. The difficulty hierarchy across question categories (temporal $>$ treatment $>$ severity $>$ disease identification) provides a roadmap for prioritizing improvements in agricultural data standards.

The key contributions are AgriMemSynth, a benchmark framework that integrates data generation and memory evaluation; two benchmark datasets, AgriConvMem and AgriMultiHop, covering multi-session conversational memory and multi-hop reasoning in agricultural advisory scenarios; the AgriMemory module for evaluating memory-augmented architectures on the

generated datasets; empirical evidence supporting semantic evaluation, graph-based representation for multi-hop reasoning, and explicit temporal information for longitudinal advisory scenarios; and a comparative baseline across five memory architectures.

The complete AgriMemSynth framework (including the data generation module AgriSynth and the evaluation module AgriMemory), the generated datasets, and all evaluation tools will be made publicly available upon publication to support reproducible research on agricultural knowledge discovery and memory-augmented question answering.

Data availability

The AgriMemSynth benchmark code, released text datasets, result artifacts, and audit reports are available in the GitHub supplementary repository at https://github.com/syaikhipin/scirep-agrimem.

Code availability

The released benchmark code and configuration files are available in the same GitHub repository. The manuscript source is not included in the supplementary repository.

Author contribution statement

N.A.A. conceptualized the study, designed the AgriMemSynth framework, developed the benchmark generation and evaluation workflow, implemented the experiments, analyzed the results, prepared the figures and tables, and wrote the main manuscript text. R.D. contributed to the methodological design, reviewed the benchmark evaluation approach, and supported the interpretation of memory-augmented system results. B.L. supervised the research, contributed to the study design, reviewed the experimental methodology, and revised the manuscript critically for intellectual content. D.T. supervised the research, contributed to the knowledge discovery framing, reviewed the results and discussion, and revised the manuscript critically for intellectual content. All authors reviewed and approved the final manuscript.

Declaration of competing interest

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

Acknowledgments

The authors gratefully acknowledge the support of the European Union's Horizon research and innovation programme.

Funding

This project has received funding from the European Union's Horizon research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101073381.

References

1. L. Stafford, J.A. Daly, CONVERSATIONAL MEMORY: The Effects of Recall Mode and Memory Expectancies on Remembrances of Natural Conversations (1984) Human Communication Research, 10 (3), pp. 379 - 402. doi:10.1111/j.1468-2958.1984.tb00024.x
2. V. Turgut, Q.H. Mahmoud, Agentic Retrieval for Conversational Memory in LLM-Based Chatbots (2026) 2026 IEEE 16th Annual Computing and Communication Workshop and Conference, CCWC 2026, pp. 356 - 362 DOI:10.1109/CCWC67433.2026.11393694
3. S. Zheng, Y. Zhu, Y. Zhang, X. Nie, J. Hou, Multi-hop reasoning with fine-grained entity representations and LLM-augmented actions over few-shot knowledge graphs (2026) Knowledge and Information Systems, 68 (1), art. no. 86, DOI:10.1007/s10115-026-02711-7
4. X. Yang, P. Fan, Convolutional end-to-end memory networks for multi-hop reasoning (2019) IEEE Access, 7, art. no. 8835083, pp. 135268 - 135276. DOI:10.1109/ACCESS.2019.2940707
5. J. Lelong, A. Errazine, A. Blangero, Agentic RAG with Knowledge Graphs for Complex Multi-Hop Reasoning in Real-World Applications (2025) Frontiers in Artificial Intelligence and Applications, 413, pp. 5163 - 5166. DOI:10.3233/FAIA251445

6. N. A. Akbar, B. Lenzitti, D. Tegolo, Proof of federated learning contribution (PoFLC) for secure agri-data sharing and collaborative model training, in: 2024 International Conference on Information and Communications Technology (ICOIACT), IEEE, 2024, pp. 258–263. doi:10.1109/ICOIACT64819.2024.10913369.
7. S. P. Mohanty, D. P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection, Frontiers in Plant Science 7 (2016) 1419. doi:10.3389/fpls.2016.01419.
8. D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, N. Batra, PlantDoc: A dataset for visual plant disease detection, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Association for Computing Machinery, New York, NY, USA, 2020, pp. 249–253. doi:10.1145/3371158.3371196.
9. N. A. Akbar, B. Lenzitti, D. Tegolo, A novel approach for leveraging agent-based experts on large language models to enable data sharing among heterogeneous IoT devices in agriculture, in: Advances in Artificial Intelligence and Machine Learning, Springer Nature Switzerland, Cham, 2025, pp. 12–22. doi:10.1007/978-3-031-80607-0_2.
10. D. P. Hughes, M. Salathé, An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing, (2015) arXiv preprint, arXiv:1511.08060.
11. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 9459–9474. ISBN:9781713829546
12. T. R. Sumers, S. Yao, K. Narasimhan, T. L. Griffiths, Cognitive architectures for language agents, (2024) arXiv preprint arXiv:2309.02427. arXiv:2309.02427.
13. A. Eldem, H. Eldem, The development and evaluation of agricultural question-answering systems based on large language models, Scientific Reports 16 (2026) 5357. doi:10.1038/s41598-026-35003-9.
14. D. M. Kabala, A. Hafiane, L. Bobelin, et al., Image-based crop disease detection with federated learning, Scientific Reports 13 (2023) 19220. doi:10.1038/s41598-023-46218-5.
15. N. A. Akbar, Are re-ranking in retrieval-augmented generation methods impactful for small agriculture QA datasets? a small experiment, BIO Web of Conferences 167 (2025) 01001. doi:10.1051/bioconf/202516701001.
16. V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.T. Yih, Dense passage retrieval for open-domain question answering (2020) EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 6769 - 6781 doi:10.18653/v1/2020.emnlp-main.550
17. Z. Zhang, Q. Dai, X. Bo, C. Ma, R. Li, X. Chen, J. Zhu, Z. Dong, J.-R. Wen, A survey on the memory mechanism of large language model-based agents, ACM Transactions on Information Systems (2024). doi:10.1145/3748302.
18. A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, Y. Fang, Evaluating very long-term conversational memory of LLM agents, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 13851–13870. doi:10.18653/v1/2024.acl-long.747.
19. Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 2369–2380. doi:10.18653/v1/d18-1259.
20. S. Zhang, Q. Cheng, Research on entity relationship semantic embedded rule mining model for medical graph reasoning, Data Mining and Knowledge Discovery 40 (2026) 52. doi:10.1007/s10618-026-01217-8.
21. T. Le, H. A. Nguyen, B. Le, Dual policy-guided multi-hop path reasoning for explainable knowledge graph recommendation, Data Mining and Knowledge Discovery 40 (2026) 53. doi:10.1007/s10618-026-01225-8.
22. GLM Team, GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, arXiv preprint (2025). arXiv:2508.06471.
23. DeepSeek-AI, DeepSeek-V3 technical report, (2024). arXiv:2412.19437.
24. J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 2318–2330. doi:10.18653/v1/2024.findings-acl.137.
25. P. Christen, D.J. Hand, N. Kirielle, A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives (2023) ACM Comput. Surv. 56, 3, Article 73 (March 2024). doi:10.1145/3606367

26. C.J.V. Rijsbergen, (1975) Information Retrieval. Butterworth-Heinemann. ISBN:10.1145/3606367

27. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 2383–2392. doi:10.18653/v1/d16-1264.

28. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

29. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, et al., Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Vol. 36, 2024, pp. 46595–46623.

30. D. Devarajan, R. Allafi, M. Obayya, et al., AI based real time disease diagnosis in plants using deep learning driven CNNs, Scientific Reports 16 (2026) 4587. doi:10.1038/s41598-025-34681-1.

31. H. M. Manoj, D. L. Shanthi, B. N. Lakshmi, et al., AI-driven drone technology and computer vision for early detection of crop disease in large agricultural areas, Scientific Reports 16 (2026) 2479. doi:10.1038/s41598-025-32384-1.

32. C. Pal, S. Karmakar, I. Mukherjee, et al., A lightweight and explainable CNN model for empowering plant disease diagnosis, Scientific Reports 15 (2025) 30720. doi:10.1038/s41598-025-94083-1.

33. S. Natarajan, P. Chakrabarti, M. Margala, Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI, Scientific Reports 14 (2024) 13695. doi:10.1038/s41598-024-64601-8.