A
Report
On

Evaluating ARIA: A Multi-Model and Multi-Metric Assessment of an Agentic RAG-Based Research Assistant for Dryland Agriculture

Submitted by

Simritha Paruchuri

LSU No: 8636

Supervisor

Dr. Mukund Patil

Senior Scientist - Soil Physics, Digital Agriculture, Research Program - Resilient Farming Systems

Patancheru, Hyderabad, Telangana 502324
Website: www.icrisat.org

DECLARATION

I, (Simritha Paruchuri, LSU No.8636), hereby decl are that the internship report entitled "(**Evaluating ARIA: A Multi-Model and Multi-Metric Assessment of an Agentic RAG-Based Research Assistant for Dryland Agriculture**)" submitted to the **Learning Systems Unit (LSU), ICRISA T,** Hyderabad, is the result of original work carried out by me under the supervision of (**Dr.Mukund Patil**, Senior Scientist - Soil Physics, Digital Agriculture, Research Program - Resilient Farming Systems).

I also declare that no part of this report has been published or submitted elsewhere for any academic or professional purpose.

Place: ICRISAT

Date: Oct 22, 2026

Simritha

(Simritha Paruchuri)
(LSU No. 8636)

Dr. Mukund Patil,
Sr. Scientist, Digital
Agriculture, ICRISAT

Table Of Contents

1. Personal Reflections
2. Acknowledgments
3. Acronyms
4. Abstract
5. Introduction
6. Review of Literature
7. Methodology
8. Data Analysis - Results
9. Discussion
10. Conclusion
11. Recommendations
12. Limitations
13. Visual Gallery
14. References

Personal Reflections

Entering ICRISAT, a campus I've called my second home for the last 13 years, as an intern rather than a student at ISH was something I never expected as a child. Our school would always take us on trips to visit the labs or walk through the fields, so stepping into iHub on June 15th, 2026 felt like a foreign idea. I was the most nervous I had ever been, but also the most excited. I was scared I wouldn't know enough, yet thrilled to learn so much more. Although I was nervous, my first day on the Digital Agriculture team consisted of the start of an AI course that would broaden my horizons and a glimpse at all the amazing projects I could explore.

The project that spoke to me most was ARIA, which challenged me in so many ways and pushed me to think outside the box. Working on it has shown me just how powerful AI tools can be and how much they can drive the future of research. It has made me realise my passion for building, refining and evaluating AI models, as well as creating tools that can make a difference in our world. It has helped me develop skills I never knew I needed, and the AI course I took has changed the way I look at AI.

Looking back at the seven weeks I spent here, I couldn't be more grateful. Everyone treated me with kindness and guided me through the process even while I was still learning. I will use everything I learned from this internship in the future, and I hope to continue learning about AI tools like ARIA. I will miss the desk where I did all my work, the mouth-watering food and, most of all, the people who helped me through it all.

Acknowledgements

I want to ensure I thank and acknowledge every person who has helped make this internship so unforgettable and valuable.

I would like to thank **Dr. Mukund Patil**, my supervisor, for giving me the opportunity to work on such a valuable project, guiding me through every step, and helping me develop my research. I also want to acknowledge him for leading the AI course, which taught me to use AI in ways I didn't know were possible. I am grateful to **Dr. Padmaja Ravula** for meeting with me every week and mentoring me, not just on ARIA but on my future work. She helped me think outside the box and challenged me to consider the social and ethical issues of what I am working on and plan to work on. I appreciate the time and effort she gave to guide me and introduce me to others who helped push the boundaries of my thinking. I also want to thank **Ms. Sreevani Mitnala** for supporting me throughout this project. She always made time to work with me, even when she was busy. She was always there to discuss with me and see how we could move the project forward. I also thank her for helping to lead the AI course and for making the sessions so valuable.

I want to thank **Dr. Pranuthi Gogumalla** and **Dr. Cuba Perumal** for helping lead and teach the AI sessions. They provided me with important skills that will allow me to use AI effectively and efficiently. I appreciate the time they took to create these lessons and ensure that every one was meaningful. I would also like to thank **ICRISAT** and **LSU** as a whole for

providing such a valuable experience and allowing me to explore my interests and work on important tools of the future. Additionally, I thank **Mr. Prabash** at LSU for always keeping us interns informed about new sessions and forms we needed to complete. He never hesitated to help when I needed assistance.

In addition, I would also like to thank the people who took the time to offer guidance and mentorship. **Dr. Anthony Wenndt** at GAIN had a genuine conversation with me about my project and asked thought-provoking questions that helped push my research. He also gave me guidance on the college process and shared his own experiences. Likewise, **Dr. Niyati Singaraju** at CGIAR took the time to meet me in person and offered me ideas for future research. Our discussion was very helpful for my future plans, and the article she sent me has given me even more ideas for what I could further research in the future.

Finally, I would like to thank my fellow interns. Whether they were helping me find a building or working with me on a group project for the AI course, I never felt hesitant to ask them for anything. I would also like to thank my parents for driving me to the ICRISAT campus every day and supporting me in pursuing such a valuable internship.

Acronyms

RAG	Retrieval-Augmented Generation
LLM	Large Language Model
ICRISAT	International Crops Research Institute for the Semi-Arid Tropics
API	Application Programming Interface
RAGAS	Retrieval-Augmented Generation Assessment
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation(Longest Common Subsequence)
GAIN	Global Alliance for Improved Nutrition
CGIAR	Consultative Group on International Agricultural Research
TCI	Tata-Cornell Institute for Agriculture and Nutrition

Abstract

ARIA is an agentic retrieval-augmented chatbot developed by ICRISAT for answering queries related to dryland and semi-arid agriculture research. While ARIA has a built-in evaluation system, it is unclear if its scores matched its real performance. This study evaluated ARIA and examined what matters more for improving a RAG-based model: the choice of LLM or the design of prompts. To investigate this, we compared ARIA against five open-source models using 360 test questions, followed by an evaluation on 90 benchmark research queries. Performance metrics included BERTScore, Token F1, ROUGE-L, RAGAS frameworks, and an LLM judge. Results indicate that overlap-based metrics and BERTScore are poorly suited for ARIA's extended responses, with scores fluctuating based on the reference model. When evaluated by Claude, the Gemini-backed version of ARIA ranked first at 57.8%, outperforming Claude (37.8%) and OpenAI (4.4%), suggesting that LLM choice was not the main problem. An error log of 11 issues was compiled, and code changes were tested on 450 questions. These changes gave a small improvement in the overall score (3.86 to 3.96 out of 5), increased the proportion of strong answers (67.8% to 70.7%), but also increased the proportion of weak answers (16.4% to 18.4%). ARIA also performed better through the API than through the chat interface, and fabricating details on unanswerable questions remained the biggest problem. Overall, prompt design and code appear to matter more than LLM choice, and changes should be carefully tested before being implemented.

Introduction

Every day, countless researchers browse through countless papers and spend hours trying to find what they need. In the age of AI, researchers shouldn't have to spend so much time finding, verifying and validating information. In agricultural research, ICRISAT has tried to bridge this gap through ARIA, an agentic retrieval-augmented chatbot for dryland and semi-arid agriculture research, with a built-in evaluation system. ARIA searches ICRISAT's open-access repository of 12,268 publications. It has nine workflow-aware response models: Factual Query, Overview, How-to Guide, Literature Review, Policy Brief, Experiment Design, Comparative Analysis, Practical Guidance and Research Co-Scientist, all of which help researchers work more efficiently (ICRISAT, n.d.).

Even AI tools with built-in evaluation systems benefit from independent external evaluation, since self-reported scores aren't necessarily accurate. External evaluation allows us to refine and improve the chatbot. ARIA includes its own evaluation pipeline, which scores responses on Confidence, Faithfulness, Relevance, Completeness, Hallucination and more. However, early comparisons raised questions about whether these self-reported scores reflected ARIA's actual performance, especially compared to open-source models.

Understanding whether this gap is real, and what might be causing it, is essential to improving ARIA's reliability and accuracy as a tool. At the moment, it is not clear why ARIA produces worse responses than other models, or what is preventing it from reaching their level.

This study therefore asks: *What matters more in improving a RAG-based model, LLM choice or prompt design?*

Review of Literature

6.1 AI Tools and Large Language Models in Agricultural Research

Agriculture has been a bedrock of human society for years. It has always changed to meet the needs of the world and today those needs are growing rapidly. For example, by 2050, agriculture will need to produce about 50 percent more food, feed and fibre than in 2012, while also using water and land more carefully (FAO, 2025). To meet these growing demands, agricultural research needs tools that can help researchers find and use knowledge faster. With the rise of AI, large language models (LLMs) are now being heavily explored in agriculture, with applications including agricultural question answering(Zhu et al., 2025). AI tools that answer agricultural questions can help effectively guide and support researchers.

However, while agriculture is an emerging field for LLM research, it is still highly underrepresented. This is because LLMs rely heavily on large amounts of data and because there are no standard evaluation frameworks. Using specific data is especially difficult in a field where answers to agricultural questions can depend on region, climate and culture. This is why retrieval-augmented generation from a specific database is generally preferred for research questions, as it allows answers to be tailored to particular criteria. For example, tools like

AgriGPT address this by using specialized databases (Agri-342K) and advanced search tools (Tri-RAG) to give accurate, evidence-backed answers (Yang et al., 2025).

6.2 Retrieval-Augmented Generation (RAG) Systems and Evaluation Methods

Retrieval-Augmented Generation (RAG) is a framework that helps make AI responses more reliable. RAG connects an LLM to an outside knowledge base so its answers do not depend solely on the training data (Lewis et al., 2020; Belcic, 2024). When a user asks a chatbot powered by RAG a question, it will search for relevant information from a knowledge base instead of giving a general answer. RAG also allows the chatbot to give domain-specific responses. This is especially useful for fields like agriculture where the data and information highly vary depending on location.

Lewis et al. (2020) found that RAG models produced more specific and factual text than a model without retrieval. However, evaluating these systems presents a myriad of challenges. A proper evaluation includes ensuring that relevant passages are extracted, checking if the model uses those retrieved passages properly, and measuring the overall quality and relevance of the generated response (Es et al., 2024). Commonly used string-matching metrics like Token F1 and ROUGE-L are not highly recommended for RAG evaluation since they rely on lexical similarity with a ground truth. This means they can struggle with comparing long-form system responses to concise reference responses. This limitation appeared in this evaluation of ARIA, where metrics such as BERTScore, Token F1 and ROUGE-L proved unsuitable for ARIA's long responses (see Section 7). To overcome these limitations, modern RAG evaluation frameworks like RAGAS

(Retrieval-Augmented Generation Assessment) include metrics such as faithfulness and answer relevancy, which do not need a reference answer, as well as metrics like semantic similarity, which compare a response with one (Es et al., 2024; Ragas, 2025).

6.3 ARIA and ICRISAT's Digital Agriculture Initiatives

To help farmers, researchers and policymakers in dry climate regions, the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) integrates digital tools directly into its agricultural research. ICRISAT's digital agriculture team develops AI systems, satellite mapping tools and online dashboards to make scientific knowledge easier to access and use (ICRISAT, n.d.). Examples of these technologies include yield prediction systems, spatial modeling for crop stress, mobile applications and open-access knowledge platforms (ICRISAT, n.d.).

One of ICRISAT's digital agriculture initiatives is ARIA, which is an agentic retrieval‑augmented assistant for dryland and semi‑arid agriculture research. ARIA directly searches ICRISAT's Open Access Repository (OAR) corpus of 12,268 indexed publications, using a hybrid retrieval strategy that combines dense semantic search, BM25 keyword search and Reciprocal Rank Fusion (RRF). Unlike normal conversational models, ARIA has nine workflow-aware response modes designed for specific scientific tasks. These include overview queries, literature reviews, experiment design, policy briefs, comparative analysis, practical guidance and a research co-scientist. Since ARIA is still in active development, ensuring that the tool can reach maximum reliability requires rigorous evaluation.

Methodology

To start the evaluation process, we compared ARIA's responses with those of open-source models such as DeepSeek and LLaMA by generating responses to 360 questions. The questions were generated by Ms. Sreevani Mitnala using the Gemini API and included research questions as well as irrelevant questions to test the chatbot's ability to handle off-topic queries. Responses from the open-source models were generated automatically through their API keys. They were evaluated using Token F1, ROUGE-L, retrieval recall and precision, MRR, latency and error rate, and each model was compared against a gold-standard response generated by Gemini. ARIA, however, had no API endpoint, so its responses could not be generated automatically, and typing out 360 questions manually would have been neither efficient nor feasible. To overcome this, a script was developed. It used an active ARIA session ID to simulate user interaction, with timing controls to account for delays and latency in ARIA's answers. This allowed us to collect ARIA's responses to all 360 questions and compare it with the other models.

Before comparing ARIA with the open-source models directly, we first wanted to evaluate ARIA's responses on their own. After some research, we compared ARIA's responses against the Gemini gold responses using BERTScore F1, a metric that measures semantic similarity through the harmonic mean of precision and recall. However, the scores were unexpectedly low. We concluded that BERTScore was not suitable here because of the length mismatch between ARIA's long responses and the shorter gold responses, which makes a direct mathematical comparison unreliable.

After BERTScore proved unsuitable, we explored other evaluation metrics and chose RAGAS-based metrics, which are designed for RAG systems and better suited to comparing responses of different lengths. We then evaluated ARIA and LLaMA, the best-performing open-source model, against the gold responses on relevancy and semantic similarity. We combined these into an accuracy score calculated as accuracy = 0.5 × relevancy + 0.5 × semantic similarity. The two chatbots were then compared using these scores.

After reviewing these evaluations, we suspected ARIA was scoring low because of both the question set and its own design. Many of the 360 questions were not research-related or designed for one of ARIA's workflow-aware response modes. To test this, we switched to ARIA's 90 benchmark queries, which are the type of questions ARIA is built to answer. Because this required a new set of gold responses, and because we wanted to check whether an unfair gold standard was also lowering ARIA's scores, we generated gold responses using both the Gemini API and the OpenAI API. ARIA's responses were then compared against each.

Research was done to find the best metrics to evaluate ARIA. I used a script to automate this comparison. It calculated Token F1 and ROUGE-L, and it used the SentenceTransformers model (all-MiniLM-L6-v2) to generate sentence embeddings and measure answer relevancy and semantic similarity through cosine similarity. Claude was then used to generate a report comparing ARIA's results against the Gemini gold standard with those against the OpenAI gold standard. This raised questions about whether the metrics captured genuine response quality or only mathematical similarity to another model's response.

After discussion, we hypothesised that ARIA's weaknesses came either from the choice of LLM or from prompt design and code errors. To test the first possibility, we changed the underlying model and evaluated responses from Gemini (ARIA's current model), OpenAI and Claude. Since gold-standard comparison had proved unreliable, we evaluated each model using another LLM as a judge instead. We chose Claude as the judge because of its strength at analysis and evaluation tasks. A script prepared for this used Claude to automatically evaluate and compare the responses from ARIA (running on Gemini), OpenAI and Claude on the 90 benchmark research questions. Each response was scored from 0 to 100 on directness, completeness, relevance and quality, and the judge selected the best response for each question. This identified which model worked best as the "brain" of ARIA. To investigate the second possibility, we created an error log of the main errors and prompt issues, based on front-end testing and a review of ARIA's scripts.

The error log was then presented, with suggestions and code changes that would help improve ARIA responses. Some of the code changes recommended in the error log were made and the re-evaluation process started. We decided to evaluate both the 90 research style benchmark queries as well as the 360 questions that were generated using the Gemini API from the start and compare them to the old ARIA responses from before the code changes using Claude to evaluate.

After making some of the code changes suggested in the error log, the evaluation process was repeated to check whether ARIA's responses had improved. Two evaluations were carried

out. The first compared responses generated through the API using the original code with responses generated through the API using the modified code. The second compared responses generated through the API using the modified code with responses collected manually through ARIA's chat interface. Both evaluations used the 90 benchmark queries and the original 360 questions, giving a total of 450 questions. An LLM-based evaluation script was used to assess the responses based on correctness, completeness, and whether they contained unsupported information. The results were compared to identify differences in performance and any remaining errors. The findings were used to identify further improvements needed in ARIA's code, retrieval process, and response generation.

Data Analysis - Results

1. Model Performance Evaluation

In the first evaluation, ARIA and five open-source models (LLaMA, Mistral, DeepSeek, Gemma and Qwen) were tested on 360 agricultural research questions generated using the Gemini API. ARIA returned an error on every question (100% error rate) because it could not connect to the ARIA chatbot API endpoint, so no scores could be calculated for it in this run. This failure led to the session-based collection method described in the Methodology. Among the open-source models, LLaMA scored highest on Token F1 and ROUGE-L, with perfect retrieval recall and a mean latency of 20.2 seconds (Table 1).

Table 1. Initial 360-question benchmark of ARIA and open-source models

Model	Token F1	ROUGE-L	Retrieval Recall	Mean Latency (s)	Error Rate
ARIA	n/a	n/a	n/a	n/a	100%
LLaMA	0.594	0.508	1.000	20.2	0%
Mistral	0.540	0.456	1.000	26.7	0%
DeepSeek	0.416	0.326	0.962	229.3	4%
Gemma	0.261	0.189	0.000	21.0	0%
Qwen	0.283	0.202	0.000	13.4	0%

Before comparing ARIA to other models, we wanted to evaluate ARIA's responses on its own. Once ARIA's 360 responses had been collected through an active session, they were evaluated. ARIA's responses were compared with the Gemini gold responses using BERTScore F1. The scores were unexpectedly low and didn't accurately reflect ARIA's performance. This was due to the length mismatch between ARIA's long, structured responses and the shorter gold answers, so BERTScore was not used further.

Table 2. ARIA Evaluation using BERTScore

Model	Token F1	ROUGE-L	Mean Latency (s)
ARIA	0.0707	0.0516	12.5

We then switched to using RAGAS-based evaluations. ARIA and LLaMA, the strongest open-source model, were compared using relevancy, semantic similarity and an accuracy score (0.5 × relevancy + 0.5 × semantic similarity). ARIA scored 0.620 for accuracy compared with 0.814 for LLaMA (Table 3).

Table 3. RAGAS-based comparison of ARIA and LLaMA (360 questions)

Model	Relevancy	Semantic Similarity	Accuracy
ARIA	0.653	0.587	0.620
LLaMA	0.804	0.825	0.814

Overall, it was concluded that overlap-based metrics scored ARIA much lower than embedding-based metrics did, which is consistent with ARIA's longer and more structured responses. The metrics also cannot say whether ARIA's answers were correct, only how closely they resembled a reference. We therefore used RAGAS-based metrics for the next stage of testing.

2. Benchmark Evaluation

To test whether ARIA performed better on the kind of questions it was designed for, ARIA's responses to its 90 benchmark research queries were scored against a new set of reference responses. One set was generated by the Gemini API as before and one by the OpenAI API. This was to test if the reference answers had an effect on the evaluation. Scores combined answer relevancy and semantic similarity (overall = 0.5 × relevancy + 0.5 × semantic similarity).

Table 4. ARIA scored against two different reference models

Reference model	Relevancy	Semantic Similarity	Overall Score
Gemini API	0.795	0.638	0.717
OpenAI API	0.845	0.806	0.826

ARIA scored 0.717 against the Gemini references and 0.826 against the OpenAI references (Table 4), showing that the choice of reference response has a large effect on evaluation scores.

3. LLM-Based Evaluation

To test whether the underlying language model was responsible for ARIA's weaknesses, Claude was used as an independent judge to compare answers from ARIA (running on Gemini), OpenAI and Claude on the 90 benchmark questions. Each answer was scored from 0 to 100 on directness, completeness, relevance and quality, and the judge selected the best response for each question. We tested using an LLM as a judge instead of a reference answer.

Table 5. LLM-judged comparison of ARIA, Claude and OpenAI

Model	Directness	Completeness	Relevance	Quality	Overall Avg	Times Best	Win %
ARIA (Gemini)	64	78	70	63	69	52	57.8%
Claude	61	72	72	61	66	34	37.8%
OpenAI	49	67	56	42	53	4	4.4%

ARIA led on directness, completeness and quality, and Claude narrowly led on relevance. OpenAI scored the worst on every metric, often citing sources without synthesizing the underlying findings. One Claude response (bq083) failed with a processing error and was scored 0 on all metrics, which slightly lowers Claude's averages. ARIA's own model using the Gemini

API ranked first, so replacing it with Claude or OpenAI will not help improve the responses. This means prompt design and code issues are likely what is holding back ARIA's responses.

4. Error Analysis

Front-end testing and thorough review of ARIA's scripts helped us produce an error log with 11 key issues. Table 6 shows the error log.

Table 6. ARIA Error log

File	Lines	Issue
agent.py	247–307	- Prompt classification is too vague (around 10 words) - Comparative analysis is triggered incorrectly sometimes
prompts.py	643–644	- Strict format structure forced on every factual answer
prompts.py	597	- No ability to summarize; can add here?
agent.py	198–207	- max_output_tokens too low — table cuts off in comparative analysis
prompts.py	66	- "Comprehensive" instruction forcing unnecessarily long answers
n/a	n/a	- Frequently unable to properly display tables - Could be a token limit?
agent.py	1885-1910	- User has to explicitly say 'give me a short answer' or 'give me a detailed answer' to get an appropriately sized response - Otherwise, it forces a 600- 800-word response
evaluator.py	155-167	- If LLM evaluation fails, it estimates a placeholder mathematically - Could lead to some errors in evaluation
evaluator.py	n/a	- Evaluation prompt measures faithfulness, relevance, and completeness only - In future, could also assess readability, formatting quality, or whether the response length matches user intent. (suggestion)

File	Lines	Issue
prompts.py	n/a	- Responses are not automatically adjusted to the complexity of the user's question - Simple questions often receive unnecessarily long answers.
prompts.py	n/a	- Prompts encourage paragraph-style responses instead of other formats like bullet points, even if other formats are more appropriate

The main patterns were that ARIA forced long answers (600–800 words) unless the user explicitly asked for something shorter, forced a fixed structure on every factual answer, cut tables off because of a low output-token limit and sometimes mistakenly triggered comparative analysis. In the evaluator, a failed LLM evaluation was replaced by a mathematical estimate, which could inaccurately create scores. Table 7 below contains all the changes made to ARIA as a result of the error log.

Table 7. ARIA Change log

#	Feedback Item (as reported)	Change Made	File(s)
1	evaluator.py 155-167 — If LLM evaluation fails, it estimates a placeholder mathematically; could lead to errors in evaluation.	Added evaluation_degraded flag to EvalResult, set True only on the fallback path, included in to_dict() so downstream consumers can distinguish a real LLM score from a heuristic guess.	part2/evaluator.py
2	(Surfaced during investigation of #1) — the fallback's placeholder feedback text was being fed back into the retry loop as	expand_queries now skips injecting feedback/missing_aspects into the retry-query prompt when	part2/agent.py

	if it were a real critique.	evaluation_degraded is True.	
3	prompts.py 643-644 — Strict format structure forced on every factual answer.	Format directive (## Summary / ## Key Evidence / ## Context and Implications) is now applied only when spec.enforce_sections is True. Factual/procedural default-length answers get plain prose instead of forced headings.	part2/agent.py
4	agent.py 198-207 — max_output_tokens too low; table cuts off in comparative analysis.	Added _WIDE_DEFAULT_WORKFLOWS (comparative, experiment_design); default max_tokens raised from 4096 to 6144 for these workflows, since their prompts mandate a fixed multi-section structure including a table.	part2/response_spec.py
5	n/a — "Frequently unable to properly display tables / could be a token limit?"	Same fix as #4 — confirmed the token-limit hypothesis (comparative's 5 mandatory sections + table routinely exceeded the old 4096-token default) and resolved via the budget increase.	part2/response_spec.py

5. Post-Modification Evaluation

To evaluate whether the code and prompt changes improved ARIA, a total of 450 questions (combining the 90 benchmark research queries and the 360 general agricultural questions) were tested after code modifications. Because ARIA lacked a direct external API

endpoint, an automated Python script using an active session ID was developed to simulate user interactions and fetch responses across all 450 questions.

The overall score rose from 3.86 to 3.96 out of 5, the share of strong answers (4-5) rose from 67.8% to 70.7%, and the share of fully answered questions rose from 53.8% to 60.4%. The share of weak answers (1-2) also rose slightly, from 16.4% to 18.4%. Strong and weak answers both became more common in the Fixed version, which suggests its responses were more polarised. The increase in weak answers indicates that the changes introduced some new problems that need to be addressed.

Table 8. Automated Session ID Performance Before and After Code/Prompt Changes

Metric (450 questions, 1-5 judge)	**Original**	**Fixed (Code/API)**
Overall score	3.86	3.96
Strong answers (4-5)	67.8%	70.7%
Weak answers (1-2)	16.4%	18.4%
Fully "answered"	53.8%	60.4%

Next, we wanted to see if ARIA's responses would change when accessed manually through the chat UI vs. programmatically through the API for the same 450 questions.

Table 9. ARIA Performance Comparison: Automated Session ID vs. Manual Chat UI (450 Questions)

Metric	Automated Session ID (Code/API)	Manual Chat Interface
Average Correctness	59 / 100	46 / 100
Average Completeness	52 / 100	37 / 100
Hallucination-Free Score	63 / 100	55 / 100
Average Response Time	~11.0 sec	~11.5 sec

The automated session ID responses scored higher than the manual chat responses on all three quality measures: correctness by 13 points, completeness by 15 points and hallucination-free score by 8 points. Response time was almost the same (about 11.0 vs. 11.5 seconds), so the quality gap was not explained by speed. We therefore concluded that the Code/API answers consistently outperformed the Chat Interface answers.

Based on these final results, the following bug changes must be fixed (Table 10).

Table 10. Bug Fixes

ID / Item	Bug / Issue Description	Severity / Status	Found In / Confirmed In

BUG-01	Over-eager "not my domain" refusal firing on valid in-domain agricultural queries.	High / Open	Fixed (Code/API); Chat Interface; Code/API
BUG-02	Backend tool failures surface raw technical errors or canned messages instead of falling back.	High / Open	Chat Interface; Code/API
BUG-03	Comparison-mode routing misfires on non-comparison questions, returning generic crop templates.	High / Open	Chat Interface; Code/API
BUG-04	Evaluation scoring pipeline drops 6% to 7% of questions due to unparseable JSON/glitches.	Medium / Open	Chat Interface; Code/API
BUG-05	Fabrication of fake study names, numbers, and citations on unanswerable questions.	High / Open	Code/API
BUG-06	Retrieval/ranking gap causing the chat interface to miss or add information.	Medium / Open	Chat Interface

6. Summary of Results

Overlap-based metrics (Token F1, ROUGE-L) and BERTScore scored ARIA low, and RAGAS metrics scored it lower than LLaMA (0.620 vs. 0.814). The metrics were sensitive to response length, and ARIA's score changed with the reference model chosen (0.717 against Gemini, 0.826 against OpenAI). Similarity to a reference answer was therefore treated as an unreliable measure of quality.

In the LLM-judged comparison, ARIA running on Gemini ranked first (57.8% of questions), ahead of Claude (37.8%) and OpenAI (4.4%). This suggests the choice of underlying model was not the main source of ARIA's weaknesses.

The error log identified 11 issues, mainly forced answer length, rigid formatting, token limits and prompt routing.

After the code and prompt changes, the overall score increased slightly from 3.86 to 3.96 out of 5. Strong answers increased from 67.8% to 70.7%, and fully answered questions increased from 53.8% to 60.4%. However, weak answers also increased from 16.4% to 18.4%, suggesting that the changes improved some responses but introduced or exposed problems in others.

ARIA's answers were noticeably better through the API than through the chat interface (59 vs. 46 correctness), and fabrication on unanswerable questions remained the most important unresolved problem.

Discussion

The main question of this study was whether LLM choice or prompt design matters more for a RAG-based model. Our results suggest prompt design and code matter more. When Claude judged the answers, ARIA running on Gemini won the most often, so switching to Claude or OpenAI would likely not have helped. The error log supports this, since most of the issues were about forced long answers, rigid formats and token limits, not the LLM itself.

The results also showed how hard it is to evaluate a RAG system. Token F1, ROUGE-L and BERTScore gave ARIA very low scores because ARIA's answers are long and structured while the gold answers were short. RAGAS gave higher scores, but ARIA's score still changed from 0.717 to 0.826 just by changing the reference model from Gemini to OpenAI. This means these metrics mostly show how similar an answer is to another model's answer, not how good it is. That is why we moved to LLM-based judging.

Another important finding was that changing a prompt does not always make things better. The code and prompt changes led to a small improvement in the overall score, from 3.86 to 3.96, while the proportion of strong answers increased from 67.8% to 70.7%. The share of fully answered questions also increased from 53.8% to 60.4%. However, weak answers increased from 16.4% to 18.4%, suggesting that the changes had mixed effects. This shows that every change should be tested carefully before it is kept.

Finally, ARIA performed better through the API (59 correctness) than through the chat interface (46 correctness). The two routes also failed differently, so the chat interface may retrieve or rank sources differently. The biggest unsolved problem was that ARIA sometimes said

there was no evidence and then added made-up figures and citations. For a research tool this is a serious issue because researchers need to be able to trust the sources.

Conclusion

This study evaluated ARIA using multiple models and metrics. We found that similarity-based metrics were not reliable for ARIA, and that the LLM was not the main reason for ARIA's weaker responses. Prompt design and code issues were more likely the cause. The changes we tested produced a small improvement in overall scores and increased the proportion of strong and fully answered responses. However, weak answers also increased, showing that some problems remain. ARIA is a promising tool for dryland agriculture research, but it needs to fix its remaining bugs before researchers can fully rely on it.

Recommendations

The first priority should be fixing the main bugs found during testing. These include ARIA wrongly refusing agricultural questions, showing technical errors instead of searching the documents, starting comparison mode when it is not needed, and making up information when no answer is available. The most important issue is fabrication. When ARIA cannot find an answer in its documents, it should clearly say that there is not enough evidence instead of making up numbers, studies or citations. This is important because researchers need to trust the information it provides.

Next, every prompt or code change should be tested before being used. The changes made to ARIA slightly improved the overall scores, but the number of weak answers also increased. This shows that changes can improve some answers while causing problems with others. Future changes should be tested on the same set of questions using an LLM judge before they are accepted. It would also help to investigate why the API gives better results than the chat interface, especially how information is found and selected. ARIA should also give shorter answers to simple questions and longer answers when more detail is needed.

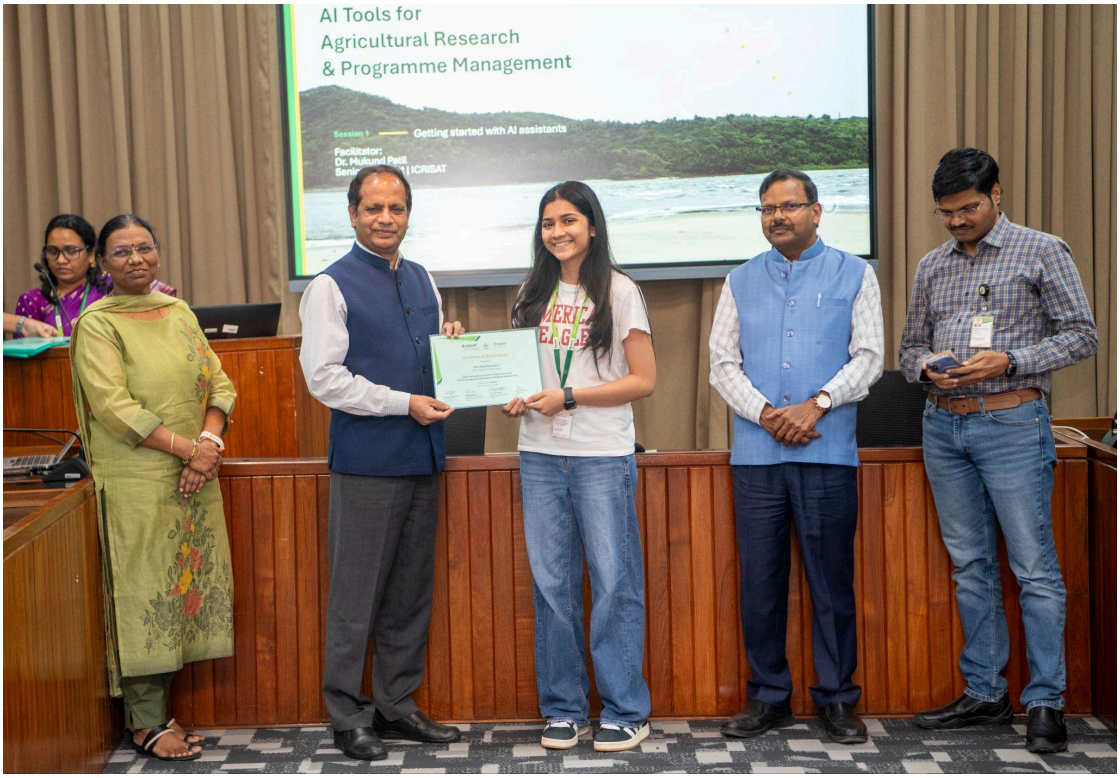
For future testing, human experts should check ARIA's answers alongside LLM judges. Using more than one judge model could also help reduce bias. The scoring system should be fixed so that all questions receive a score. Finally, ARIA should be tested with more real questions from researchers and with different LLMs. This would help show whether the improvements work across a wider range of questions and situations.

Limitations

This study has a few limitations. First, it was done in 7 weeks, so not every change in the error log could be tested. Second, Claude was used as the judge while Claude was also one of the models being compared, so there could be some bias. Also, no human expert checked the answers. Third, the 360 questions and the gold answers were generated by LLMs, so they may not be perfect. Fourth, the chat interface answers were pasted in manually, which could add small

errors. Finally, the grading pipeline failed to score about 6–7% of questions, and only one ARIA system was tested, so the results may not apply to other RAG tools.

Visual Gallery


AI Tools for
Agricultural Research
& Programme Management
Session 1
Getting started with AI assistants
Facilitator:
Dr. Mukund Patil
ICRISAT

AI Use Statement

AI tools were used to help plan and organise this report and to assist with identifying and formatting citations and references. All content and references were reviewed by the author.

References

BasicAI. (n.d.). *7 applications of AI in agriculture*. https://www.basic.ai/blog-post/7-applications-of-ai-in-agriculture

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

GeeksforGeeks. (n.d.). *What is retrieval-augmented generation (RAG)?* https://www.geeksforgeeks.org/nlp/what-is-retrieval-augmented-generation-rag/

International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). (n.d.). *ARIA*. https://aria.icrisat.org/

International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). (n.d.). *Digital agriculture and geospatial sciences*. https://icrisat.org/research/digital-agriculture/about

International Food Policy Research Institute (IFPRI). (2025, November 14). *How accurately do the large language models (LLMs) powering chatbots respond to questions on gender equality and women's empowerment in Indian agriculture?* https://www.ifpri.org/blog/how-accurately-do-the-large-language-models-llms-powering-

chatbots-respond-to-questions-on-gender-equality-and-womens-empowerment-in-indian-agriculture/

IBM. (2024, October 21). *What is retrieval augmented generation (RAG)?* https://www.ibm.com/think/topics/retrieval-augmented-generation

Khanifar, J. (2025). *Large language models in AI-driven agriculture*. ResearchGate. https://www.researchgate.net/publication/393280360_Large_Language_Models_in_AI-Driven_Agriculture

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*. https://www.sciencedirect.com/science/article/pii/S0168169924003156

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://arxiv.org/abs/2005.11401

Magalhaes, M., Singaraju, N., & Koo, J. (2025, November 14). How accurately do the large language models (LLMs) powering chatbots respond to questions on gender equality and women's empowerment in Indian agriculture? *International Food Policy Research Institute*. https://www.ifpri.org/blog/how-accurately-do-the-large-language-models-llms-powering-chatbots-respond-to-questions-on-gender-equality-and-womens-empowerment-in-indian-agriculture/

Shaikh, T. A., Rasool, T., Koildurai, V., & Yaseen, S. M. (2025). The role of large language models in agriculture: Harvesting the future with LLM intelligence. *Progress in Artificial Intelligence, 14*(2), 117–164. https://www.researchgate.net/publication/393280360_Large_Language_Models_in_AI-Driven_Agriculture

United Nations. (2013, December). *United Nations News article*. https://news.un.org/en/story/2013/12/456912

Frontiers in Plant Science article. (2025). Frontiers. https://doi.org/10.3389/fpls.2025.1579355

arXiv preprint arXiv:2508.08632. (2025). arXiv. https://arxiv.org/html/2508.08632v1