Are Re-Ranking in Retrieval-Augmented Generation Methods Impactful for Small Agriculture QA Datasets? A Small Experiment

Nur Arifin Akbar [1*]

[1]Dipartimento Matematica e Informatica, Università Degli Studi di Palermo, Palermo, Italy

Abstract. Agriculture requires accurate, location-specific information that would need the power of advanced Retrieval-Augmented Generation (RAG) models. To this end, we perform an experimental analysis on how integrating re-ranking strategies and in-memory computing into RAG models might affect performance on small agriculture question-answering (QA) datasets. This method envisages to enable real-time ground-truth kind of answers for agro-informatics sake, the proposed approach is to assist enhance document relevance and lower response latency. We trained the system on a large-scale agriculture QA dataset using high-level components like the Sentence Transformer for embedding generation, FAISS for fast vector search and a pre-trained language model for response generation. This is to keep the documents returned highly relevant, and zero-shot classification was used for re-ranking techniques. The efficacy of their algorithm across a range of QDMR transformation tasks was evaluated, and the experiment evaluation showed that rereading did not significantly increase performance over baselines. But the in-memory computing with FAISS greatly reduced retrieval latency which makes it appropriate for real-time applications in agriculture QA systems.

1 Introduction

1.1 Background

The agriculture sector, as the need for getting precise and timely information grows, to make an informed decision. Farmers and agricultural experts need the right information at the right time to deal with a range of issues, including crop management, pest control, and soil health, while policymakers need reliable data on which to base Agri-policy. In the area of data and information, the adaptation systems fail to provide contextually appropriate answers for complex agriculture-related queries, which makes traditional information retrieval systems ineffective [1].

The Retrieval-Augmented Generation (RAG) models that can mix retrieval and generative techniques have shown good results on this task [2]. The RAG models use a massive agricultural knowledge base to fetch relevant documents aligning with the user's input query and then generate an appropriate answer given the current situation. Such models can transform agriculture and farming QA systems, offering reliable and elaborate answers to a diverse set of questions.

The agriculture sector suffers from unique challenges such as information diversity, the need for real-time performance, etc. Hence, more improvements are needed to utilize the full potential of RAG models in this field.

1.2 Motivation

From the general QA point of view, RAG models have proven powerful results; however, its usefulness in the agriculture domain is not as simple as problems that must be considered to handle an agriculture framework. Questions relating to agriculture would be written in complex words, and jargon that are related to the domain of agriculture along with a wide range of sub-questions like types of crop varieties, methods for pest control in the field, and soil nutritional demands, etc. RAG models need to be architectural in a way that suitably handles these types of queries.

Also, the Agri sector generally requires a quick turnaround time. A farmer could require immediate answers to help them make business- or even life-changing decisions. As a result, the retrieval and generation processes in RAG models must be efficient to produce fast responses to user queries. Its scalability

* Corresponding author: nurarifin.akbar@unipa.it

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

is equally important with the amount of generation related to queries increasing as digital technologies in agriculture are getting adopted [3].

1.3 Research Objective and Contribution

This research investigates whether re-ranking strategies in RAG methods are impactful for small agriculture QA datasets. Specifically, we examine the effects of integrating re-ranking strategies and in-memory computing on the performance of RAG models in this domain. The key contributions of this research are as follows:

- Propose an approach to enhance RAG models in agriculture by incorporating re-ranking strategies and in-memory computing.
- Evaluate the performance of the enhanced RAG model on an agriculture QA dataset, analyzing accuracy and retrieval speed compared to baseline methods.
- Provide insights into the efficacy of re-ranking strategies and in-memory computing in enhancing RAG models for agriculture QA data.
- Contribute to advancing RAG architectures and their application in the agriculture sector, enabling efficient processing of agricultural data within generative AI systems.

Our study addresses this critical gap between the information needs of farmers and existing knowledge resources. It may also lead to advances in agriculture quality assurance systems and contribute to aid with data-driven decisions in the field of agriculture.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-augmented models have been widely explored in many domains. For example, they have been used in task-oriented dialog systems to generate responses from the chosen snippets with RAG [4]. Such models parametric and non-parametric memory with state-of-the-art deep learning methods for language generation [2]. Retrieval strategies for improving neural generative models have been addressed in the work of [5]. Their methods are designed to maximize the joint probability of the query with its retrieved contexts simultaneously [6]. Also, retrieval augmentation aids flexibility - non-parametric data- external to the model-level accuracy improvement [7][8].

Retrieval-augmented models have also demonstrated success in increasing model robustness, suggesting their value to boost modularity [9]. Retrieval methods have been proposed in other tasks, such as commonsense generation [10] or semantic parsing [11], showing the versatility of retrieval approaches.

2.2 BM25 Algorithm and Its Applications

BM25 is a probabilistic retrieval algorithm widely used due to its effectiveness and simplicity [25]. It ranks documents based on the query terms appearing in each document, weighted by term frequency and inverse document frequency. While BM25 is a strong baseline in many IR tasks, its performance in conjunction with RAG models and re-ranking strategies for small datasets requires further examination.

2.3 Re-ranking Strategies

Re-ranking strategies are crucial for enhancing the performance of information retrieval systems. Various methods have been proposed to refine document rankings after the initial retrieval process, addressed document re-ranking to boost the ranking of pertinent documents post-initial retrieval [12], while others presented a re-ranking method that uses label propagation-based semi-supervised learning algorithms, exploiting the intrinsic structure within large document datasets [13].

Re-ranking strategies have been studied in different contexts. For instance, visual re-ranking for multi-aspect information retrieval [14[15] introduced a re-ranking method that uses information from relevance feedback to improve rankings. It emphasizes the significance of entropy-based clustering in enhancing document re-ranking, positioning it as an intermediary step between initial retrieval and query expansion in information retrieval systems [16].

Studies have also demonstrated that re-ranking algorithms are effective in various domains, including resequencing of web data based on knowledge domains [17] and passage re-ranking using cross-encoder architectures [18].

2.4 Existing Agriculture QA Systems

QA system for agricultural planting technology has been developed, using text matching algorithms and crop planting knowledge graphs to improve precision in answering farming-related questions. [1], while an agriculture-focused question-answering system designed to respond to farmers' queries [19]. Additionally, an Agriculture Chatbot project focused on utilizing machine learning techniques, specifically

recurrent neural network (RNN) algorithms, to provide accurate responses related to the agriculture domain.

3 Methodology

3.1 Overview of Proposed Approach

The proposed approach aims to enhance RAG models in the agriculture domain by incorporating re-ranking strategies and in-memory computing. The objective is to improve the relevance of retrieved documents, reduce response latency, and deliver precise real-time answers for agriculture-based queries. Fig. 1 illustrates the architecture of the proposed model.

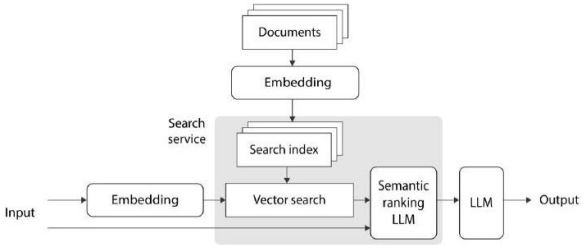

Figure 1. Proposed RAG Model Architecture

3.2 Dataset and Preprocessing

This study utilizes the Agriculture QA dataset (English version) from the KisanVaani dataset. The dataset is a comprehensive compilation of questions and corresponding answers on agriculture-related subjects, covering various topics such as crop management, pest control, and soil health.
To prepare the dataset for training, the following preprocessing steps were undertaken:
- Load the dataset using the HuggingFace Datasets library for seamless integration and preprocessing convenience.
- Generate embeddings for each question using the Sentence Transformer model "all-mpnet-basev2".
- Save the embeddings for use in retrieval processes.
- Chunk the documents into manageable pieces using Recursive Character Text Splitter to optimize retrieval performance.

For each question $q \in Q$, where Q is the set of questions in the dataset, the embedding E_q is generated as follows :

$$E_q = f_e(q) \quad (1)$$

where f_e is the SentenceTransformer model.

3.3 Dataset Characteristics

The Agriculture QA dataset from the KisanVaani project [21] consists of approximately 10,000 question-answer pairs covering:
- Crop Management: Planting schedules, crop rotation, fertilization techniques.
- Pest and Disease Control: Identification and treatment of common pests and diseases.
- Soil Health: Soil testing, nutrient management, organic amendments.
- Climate and Weather: Impacts of weather patterns, climate change adaptations.
- Market Information: Pricing trends, demand forecasting, supply chain logistics.

3.4 Incorporation of Local Wisdom

While primarily focused on Indian agriculture, the dataset includes region-specific questions and answers that reflect local practices, which are analogous to those in Indonesia due to similar agro-climatic conditions. This enables the exploration of integrating local wisdom into the QA system to enhance its relevance in different regional contexts [20].

3.5 RAG Model Components

3.5.1 Retriever

The retriever component is responsible for fetching relevant documents from a large corpus based on the input query. FAISS (Facebook AI Similarity Search), an efficient vector search library, was utilized for fast and scalable retrieval tasks capable of handling large volumes of agricultural data.

3.5.2 Generator

The generator component produces a contextually appropriate answer using the retrieved documents. The pretrained language model Meta-Llama-3-8B was employed to generate high-quality responses by leveraging the context from the retrieved documents. This ensures that the answers are pertinent, linguistically fluent, and informative.

3.5.3 Re-ranking Strategy

Re-ranking strategies were employed to enhance the relevance of the retrieved documents. The re-ranking process involves scoring the retrieved documents based on their relevance to the query and reordering them to prioritize the most pertinent content. Zero-shot classification was used for relevance scoring by applying the pre-trained model "facebook/bart-large-

mnli". This model can assess the relevance of each document to the query without specific training for the task, allowing for flexible and accurate relevance assessment across diverse queries.

The re-ranking algorithm was implemented in three steps:

- Retrieve a set of candidate documents from FAISS.
- Score each document using the zero-shot classification model.
- Re-rank the documents based on the relevance scores, prioritizing higher-scoring documents.

This re-ranking process ensures that the most relevant documents are utilized for generating the final response, improving the accuracy and appropriateness of the answers.

3.5.4 In-Memory Computing with FAISS

To reduce retrieval latency and improve real-time performance, in-memory computing with FAISS was employed. By indexing the question embeddings with FAISS and storing them in memory, significant reductions in retrieval time were achieved. This approach enables swift vector searches and significantly reduces the time needed to find relevant documents.

4 Results and Discussion

4.1 Accuracy Evaluation

The accuracy of the RAG model was evaluated with and without re-ranking strategies. The model was tested using a subset of the Agriculture QA dataset from KisanVaani. The accuracy was calculated by comparing the generated responses to the ground truth answers.

Table 1. Accuracy Comparison Between Baseline and Re-ranked RAG Models

Model	Accuracy (%)
Baseline RAG Model	33.33
Re-ranked RAG Model	33.33
BM25 without Re-ranking	33.33

The experimental results showed that both the baseline RAG model, and BM-25 and the re-ranked RAG model achieved an accuracy of 33.33%. The inclusion of reranking strategies did not deliver substantial benefits in terms of precision compared to the baseline. This indicates that the re-ranking process utilized in this study had a limited impact on enhancing the model's ability to generate accurate responses to agriculture-related queries.

4.2 Retrieval Time Evaluation

To assess retrieval time performance, two methods were compared: disk-based retrieval and in-memory retrieval using FAISS. In the disk-based approach, a file containing question embeddings was read to simulate data retrieval from storage and perform a similarity search. This was compared with in-memory retrieval, where FAISS was employed to index the question embeddings and perform fast similarity searches.

Table 2. Retrieval Time Comparison Between Disk-based Retrieval and In-memory Retrieval with FAISS

Method	Average Retrieval Time (s)
Disk-based Retrieval	0.180653
In-memory Retrieval (FAISS)	0.000456

The retrieval time improvement factor F was calculated as:

$$F = \frac{T_{\text{disk}}}{T_{\text{memory}}} \quad (2)$$

where T_{disk} is the average retrieval time for disk-based retrieval, and T_{memory} is the average retrieval time for in memory retrieval.

4.3 Discussion

Low gain in accuracy with re-ranking strategies indicates that the re-ranking methodology used is not suitable for small agriculture QA datasets. This may be because the dataset has a narrower distribution or because of the need to adapt the re-ranking model to field specifics.

Nevertheless, in-memory computing offers a noticeable gain of lowering the retrieval latency; therefore, we can conclude it is beneficial to make use of it for improving the real-time response times like QA systems. It allows users to get an immediate response of what they want from the database, and retrieval times are very critical in the agriculture sector, where time is valuable because you need to make decisions so quickly.

In agriculture, these rapid answers are not possible without coupling in-memory computing with RAG models. These systems can access the abundance of agricultural data available and deliver decision-

relevant information as soon as the farmer can use it, expert recommendations, personalized advice, and alerts to enable informed decisions.

5 Conclusion

5.1 Summary of Experimental Results

The experimental results revealed that incorporating reranking did not significantly increase accuracy compared to the baseline RAG model without re-ranking. Both models achieved an accuracy of 33.33% on a subset of the Agriculture QA dataset from KisanVaani. This suggests that the re-ranking approach used in this study had a limited impact on enhancing the model's ability to generate accurate responses for agriculture-related queries.

However, the benchmarking experiments demonstrated the superior performance of in-memory retrieval using FAISS in reducing retrieval latency. The average retrieval time required by in-memory retrieval was substantially less than that of disk-based retrieval, indicating significant efficiency improvements achieved by incorporating in-memory computing with FAISS for fast similarity search.

5.2 Implications for Agriculture QA Systems

The findings of this study provide valuable insights for the development of agriculture QA systems. While the re-ranking strategies did not yield significant accuracy improvements with the current model, the notable reduction in retrieval latency achieved by utilizing in-memory computing with FAISS has significant implications for real-time performance.

Implementing in-memory retrieval can dramatically reduce response times, which is a critical factor affecting the user experience of agriculture QA systems. By providing rapid responses to user queries, these systems become more effective in supporting timely decision-making and day-to-day operations for farmers. The enhanced scalability also makes them suitable for large-scale deployment.

5.3 Future Research Directions

The research offers insightful knowledge and information on the re-ranking strategies and in-memory computing that might be used in agriculture QA systems. Yet, there are also many opportunities that we believe the research community can pursue to increase the strength and usability of such systems.

Another experiment with other reranking algorithms or tuning the model may improve accuracy depending on domain-specific knowledge and associated features used.

To evaluate across all query types and throughout larger populations of the agriculture QA dataset, it is necessary to have general understanding of the limitations and areas for improvement as well

The feedback system may be implemented at scalable way so it will be a key learning for the continual improvement of the agriculture QA apart from human intervention.

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. G. Blanchy, G. Bragato, C. Di Bene, N. Jarvis, M. Larsbo, K. Meurer, S. Garré, Soil and crop management practices and the water regulation functions of soils: a qualitative synthesis of meta-analyses relevant to European agriculture. SOIL 9, 1–20 (2023). https://doi.org/10.5194/soil-9-1-2023.
2. L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, J. Li, A survey of knowledge enhanced pre-trained language models. IEEE Trans. Knowl. Data Eng. 36, 1413–1430 (2024). https://doi.org/10.1109/TKDE.2023.3310002.
3. M. Lewenstein, A. Gratsea, A. Riera-Campeny, A. Aloy, V. Kasper, A. Sanpera, Storage capacity and learning capability of quantum neural networks. Quantum Sci. Technol. 6, 045002 (2021). https://doi.org/10.1088/2058-9565/ac070f.
4. A. Fan, C. Gardent, C. Braud, A. Bordes, Augmenting transformers with KNN-based composite memory for dialog. Trans. Assoc. Comput. Linguist. 9, 82–99 (2021). https://doi.org/10.1162/tacl_a_00356.
5. A. Scalercio, A. Paes, Masked transformer through knowledge distillation for unsupervised text style transfer. Nat. Lang. Eng. 30, 973–1008 (2024). https://doi.org/10.1017/S1351324923000323.
6. S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, S. Nanayakkara, Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Trans. Assoc. Comput. Linguist. 11, 1–17 (2023). https://doi.org/10.1162/tacl_a_00530.
7. Z. Ahmad, A. Ekbal, S. Sengupta, P. Bhattacharyya, Neural response generation for task completion using conversational knowledge graph. PLoS ONE 18, e0269856 (2023). https://doi.org/10.1371/journal.pone.0269856.
8. U. Kamal, M. Zunaed, N. B. Nizam, T. Hasan, Anatomy-XNet: An anatomy aware convolutional neural network for thoracic disease classification in chest X-rays. IEEE J. Biomed. Health Inform. 26, 5518–5528 (2022). https://doi.org/10.1109/JBHI.2022.3199594.

9. L. Zhang, Q. Zhou, CRISPR/Cas technology: A revolutionary approach for genome engineering. Sci. China Life Sci. 57, 639–640 (2014). https://doi.org/10.1007/s11427-014-4670-x.
10. J. Guan, F. Huang, Z. Zhao, X. Zhu, M. Huang, A knowledge-enhanced pretraining model for commonsense story generation. Trans. Assoc. Comput. Linguist. 8, 93–108 (2020). https://doi.org/10.1162/tacl_a_00302.
11. O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, Y. Shoham, In-context retrieval-augmented language models. Trans. Assoc. Comput. Linguist. 11, 1316–1331 (2023). https://doi.org/10.1162/tacl_a_00605.
12. D. Ji, S. Zhao, G. Xiao, Chinese document re-ranking based on automatically acquired term resource. Lang. Resour. Eval. 43, 385–406 (2009). https://doi.org/10.1007/s10579-009-9106-z.
13. S. A. Seyedi, A. Lotfi, P. Moradi, N. N. Qader, Dynamic graph-based label propagation for density peaks clustering. Expert Syst. Appl. 115, 314–328 (2019). https://doi.org/10.1016/j.eswa.2018.07.075.
14. Y. Zhang, Q. Qian, H. Wang, C. Liu, W. Chen, F. Wang, Graph convolution based efficient re-ranking for visual retrieval. IEEE Trans. Multimedia 26, 1089–1101 (2024). https://doi.org/10.1109/TMM.2023.3276167.
15. Z. Li, K. C. K. Lee, B. Zheng, W.-C. Lee, D. Lee, X. Wang, IR-Tree: An efficient index for geographic document search. IEEE Trans. Knowl. Data Eng. 23, 585–599 (2011). https://doi.org/10.1109/TKDE.2010.149.
16. H. Yu, X. Wang, G. Wang, X. Zeng, An active three-way clustering method via low-rank matrices for multi-view data. Inf. Sci. 507, 823–839 (2020). https://doi.org/10.1016/j.ins.2018.03.009.
17. G. Zhao, X. Zhang, Re-ranking web data per knowledge domain. Int. J. Serv. Knowl. Manag. 3, 66–84 (2019). https://doi.org/10.52731/ijskm.v3.i1.274.
18. D. S. Sachan, M. Lewis, D. Yogatama, L. Zettlemoyer, J. Pineau, M. Zaheer, Questions are all you need to train a dense passage retriever. Trans. Assoc. Comput. Linguist. 11, 600–616 (2023). https://doi.org/10.1162/tacl_a_00564.
19. S. Godara, J. Bedi, R. Parsad, D. Singh, R. S. Bana, S. Marwaha, AgriResponse: A real-time agricultural query-response generation system for assisting nationwide farmers. IEEE Access 12, 294–311 (2024). https://doi.org/10.1109/ACCESS.2023.3339253.
20. I. Annamoradnejad, G. Zoghi, ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor. Expert Syst. Appl. 249, 123685 (2024). https://doi.org/10.1016/j.eswa.2024.123685.
21. KisanVaani/agriculture-qa-english-only · Datasets at Hugging Face [Online]. 2024. https://huggingface.co/datasets/KisanVaani/agriculture-qa-english-only [11 Mar. 2024].
22. Johnson, M. Douze, H. Jegou, Billion-scale similarity search with GPUs. IEEE Trans. Big Data 7, 535–547 (2021). https://doi.org/10.1109/TBDATA.2019.2921572.
23. M. G. Sohrab, M. Asada, M. Rikters, M. Miwa, BERT-NAR-BERT: A non-autoregressive pre-trained sequence-to-sequence model leveraging BERT checkpoints. IEEE Access 12, 23–33 (2024). https://doi.org/10.1109/ACCESS.2023.3346952.
24. Z. Sun, G. Pedretti, E. Ambrosi, A. Bricalli, D. Ielmini, In-memory eigenvector computation in time O (1). Adv. Intell. Syst. 2, 2000042 (2020). https://doi.org/10.1002/aisy.202000042.
25. I. Yang, X. Huang, Y. Li, H. Zhou, Y. Yu, H. Bao, J. Li, S. Ren, F. Wang, L. Ye, Y. He, J. Chen, G. Pu, X. Li, X. Miao, Self-selective memristor-enabled in-memory search for highly efficient data mining. InfoMat 5, e12416 (2023). https://doi.org/10.1002/inf2.12416.
26. S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. 3, 333–389 (2009). https://doi.org/10.1561/1500000019.
27. Q. H. Ngo, T. Kechadi, N.-A. Le-Khac, OAK: Ontology-based knowledge map model for digital agriculture, in Future Data and Security Engineering, edited by T. K. Dang, J. Küng, M. Takizawa, T. M. Chung (Springer International Publishing, 2023), pp. 245–259.