Journal of Engineering and Sustainable Development
Vol. 29, No. 04, July 2025
https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd
ISSN 2520-0917
Research Article
https://doi.org/10.31272/jeasd.2732

Retrieval Augmented Generation Data Query Technique for Pineapple Cultivation

Badril Abu Bakar[1]*, Siti Noor Aliah Baharom[2], Mohd Taufik Ahmad[3], Mohd Nizam Zubir[4], Adli Fikri Ahmad Sayuti[5], Mohd Aufa Mhd Bookeri[6]

[1,2,3,4,5,6] Engineering Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Headquarters, 43400 Serdang, Selangor, Malaysia

*Email: badril@mardi.gov.my

Article Info

Received	**29/05/2024**
Revised	**13/04/2025**
Accepted	**17/05/2025**

Abstract

Generative artificial intelligence is advancing at a blistering pace. Large Language Models, in particular, have sped up the development of machine learning applications. This work presents a large language model-based technique to query data collected during MD2 pineapple crop production. Retrieval Augmented Generation was used to feed structured and unstructured data to two large language models (GPT-4 and LLAMA2) to train and fine-tune the models. The performance of the models was then measured using actual and predicted question-answer pairs. Results showed that the models had a 78% - 87% correct answer rate for structured and 75% - 79% correct answer rate for unstructured data. However, results showed that the models had a 61%-68 % correct answer rate when an answer to a question needed to refer to structured and unstructured data. These results showed that large language models can be further investigated to give farmers useful insights when making crop management decisions.

Keywords: Pineapple cultivation; Agriculture; Large language model; Retrieval augmented generation; Generative artificial intelligence

1. Introduction

Pineapple (Ananas comosus) is a tropical fruit highly sought after for its sweet taste [1]. Pineapple waste could be used for things such as fibres in composite materials and biogas [2], [3]. The market value of world pineapple production stood at USD 27.08 billion in 2022. The largest producers in the world are Indonesia (3.20 Mt), followed by the Philippines (2.91 Mt) and Costa Rica (2.90 Mt) [4]. Malaysia is the 25th largest producer with a total production of 287,799 tonnes. Although some mechanization and automation systems exist, pineapples are primarily planted manually worldwide [5]. This is also true in Malaysia [6].

The Malaysian government's National Agrofood Policy 2.0 aims to modernize the agricultural sector, encouraging the adoption of smart farming techniques and data-driven decision-making [7]. To align with these goals, the Malaysian Agricultural Research and Development Institute (MARDI) is developing a smart MD2 pineapple production system.

Central to developing smart agricultural practices is the ability to make informed management decisions based on data [8], [9]. Traditional agricultural data analysis mainly utilizes structured data, such as sensor outputs or manually recorded metrics, presented in tabular form [10]. However, much valuable information is often found in unstructured formats, like cultivation manuals, field notes, and reports. These data types are challenging to process using conventional data analysis techniques, usually falling short when dealing with natural language inputs [11].

Recent advances in artificial intelligence, particularly in generative models such as large language models (LLMs), have shown promise in processing unstructured data [12]. LLMs such as GPT-4 by OpenAI and LLAMA2 by Meta use the deep learning neural network algorithm to predict word sequences based on given input contexts. These words form coherent sentences and paragraphs [13], [14]. These capabilities make LLMs suitable for various applications, including natural language processing, question-answering tasks, and complex information synthesis.

1.1 Literature Review

LLMs have shown exceptional performance in answering questions about subjects that they have been trained in,

including philosophy, mathematics, computer coding, and agriculture [15].
An interesting use case would be if LLMs can be trained on proprietary datasets such as confidential company documents or scientific experiments. Then, one can prompt questions about the data. However, in order to do this, whole documents would have to be fed in as part of the prompt. For large documents and datasets, this could be a challenge.
An LLM's business model usually charges a certain amount of money for a prompt or query based on the number of characters or words being fed to the LLM. A group of words or characters is called a token. The more tokens are passed in a query or prompt, the more expensive it would be. Feeding a document such as the MD2 crop production protocol with over 20,000 words equals roughly 27,000 tokens. Prompting a query with this amount of tokens to the GPT-4 model, for example, would cost USD 0.90 each for each prompt. This would get very expensive as the number of prompts increases.
The use of LLMs in agriculture is relatively new, but there has been growing interest in their potential applications. For instance, [11] provides a comprehensive survey on LLMs, noting their adaptability across multiple fields, including agriculture. Despite prior attempts, difficulties arise when LLMs need to process structured datasets such as tables or databases when no substantial preprocessing occurs; this is demonstrated in [16]. Hybrid methods that combine structured and unstructured data to address this shortcoming are already being researched. One example is the HybridQA dataset, which incorporates multi-hop question answering of both these data types.[17].
Retrieval Augmented Generation (RAG) is another promising approach to enhancing overall LLM processing and accuracy when using mixed types of data [18]. RAG data processing improves the efficiency of LLMs processing long files by breaking the files down into smaller units/sections of readable data. Subsequently, it uses embeddings to find the most suitable information about a prompt for LLMs to process.
Once the file is segmented, the contents of that segment are coded numerically, in a group of vectors called embeddings, and stored in a vector store. During the query, the prompt itself is also coded numerically and compared against the stored embeddings. The model selects a group of embeddings most similar to the original query "embedding" and feeds both the original prompt and the Group of embeddings to the model(s). Then the model(s) only process the feed prompting.
The advantage of RAG comes primarily from the methodology's potential to reduce the number of tokens required to prompt LLMs, which, in turn, reduces the cost of usage, saves computation power, and still maintains the overall response accuracy. Additional support for the processing efficiency, illusion of costs, and accuracy of RAG-supported models comes from reducing Kolmogorov complexity. Kolmogorov complexity suggests that by simplifying the total amount of input type data requested, LLMs and human cognitive efficiencies occur during the response processing [19]–[21]. Fig. 1 shows the overall workflow of RAG.

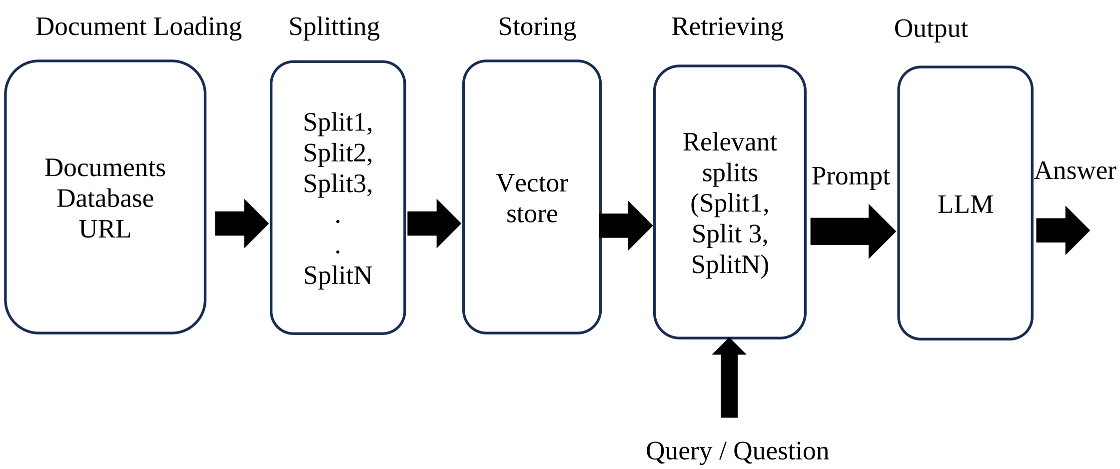

Figure 1. Overview of retrieval augmented generation.

The adoption of Retrieval-augmented generation (RAG) could assist the crop production of MD2 pineapples by querying data collected from the field and juxtaposing it against a crop production guide, a document, through the prompt generated by the RAG technique and populating that prompt into a large language model (LLM).

1.2 Objectives and Contributions

This paper expands upon the findings in the literature review while exploring the performance of LLMs when querying both structured and unstructured data about MD2 pineapple production. It provides a side-by-side comparison of GPT-4 and LLAMA2.
In summary, this study focuses on the feasibility and efficacy of the RAG approach when querying data considering the production of MD2 pineapples. The data sources included a 60-page crop production protocol (unstructured data) and field data collected over a 14-month, which included 5,500 data points across 40 variables (structured data). This approach evaluates the ability of LLMs to handle complex queries that span diverse data types, identifies the challenges of integrating structured and unstructured information, and explores potential solutions. The contributions of this study are:

1. Demonstrating the potential of RAG to simplify data inputs on a structured and/or unstructured data source(s) for the query.
2. Assessing the performance of GPT-4 and LLAMA2 with structured, unstructured and combination querying data.
3. Analyzing the computational and financial implications of the model, with a goal of practical application in agriculture.

1.3 Structure of the Paper

The remainder of this paper is organized as follows: Section 2 outlines the methodology used, including data processing techniques and evaluation criteria. Section 3 presents the results and discusses the findings in detail. Section 4 provides conclusions and future work recommendations, focusing on enhancing the model's diagnostic capabilities in agricultural settings.

2. Materials and Methods

2.1 Software Platform

The Python 3.11.3 programming language was used to prompt two large language models (LLMs) through their application programming interface (API). The two LLMs chosen were GPT-4 and LLAMA2 due to their state-of-the-art capabilities in generative artificial intelligence [13], [14]. Both of these models were selected to compare performance with structured and unstructured recovery augmented generation (RAG) agricultural data.

2.2 MD2 Pineapple Cultivation Data Input Sources

Data for the study was sourced from two primary inputs:

1. Unstructured Data: The MD2 pineapple crop production protocol, a 60-page document, guided cultivation practices, including land preparation, planting procedures, crop maintenance, and harvesting techniques [22]. This document was treated as unstructured data due to its natural language format.
2. Structured Data: Field data was collected over 14 months during the 2023–2024 planting season. This data comprised 5,500 data points across 40 parameters, covering production metrics (e.g., plot size, yield), crop characteristics (e.g., leaf length, fruit size), environmental conditions (e.g., temperature, humidity), and work logs (e.g., task details, equipment usage). Table 1 shows the complete data fields that were collected.

These parameters were structured in a tabular format, representing the' conventional data collection approach of agricultural studies. The integration of these data sources aimed to evaluate LLMs' ability to answer queries requiring information from diverse data types, highlighting the challenges of processing structured versus unstructured data in a unified framework.

2.3 Retrieval Augmented Generation

The RAG technique was implemented to improve the querying process by reducing the data complexity and optimizing the amount of information fed into the LLMs. The methodology consisted of several key stages:

1. Document Loading: The crop production protocol was converted into a digital format and processed as a series of text segments. The structured field data was imported in a tabular format. Both data sources were split into smaller units, with the protocol segmented into 198 splits and the structured data divided into 5,500 splits.

Table 1. Data parameters were collected throughout the planting season. Four data types were collected. They are: production, crop, environmental, and work log data.

Data type	Parameter
Production	plot name, plot size, planting date, harvesting date, plant count, induction date, seedling size, seedling origin, yield, total input use
Crop	plant height, leaf length, leaf width, leaf colour, leaf count, fruit size, and crown size
Environment	temperature, relative humidity, pressure, solar radiation, wind speed, precipitation, soil PH, soil EC, soil salinity, soil nutrient level
Work log	task name, task date, task location, task time, task report, operator name, equipment uses, input use, task image

2. Data Splitting and Vector Embedding: The text segments and tabular data entries were numerically encoded as vectors, known as embeddings, using an embedding model compatible with the LLMs. These embeddings were stored in a vector database to facilitate efficient retrieval during the query process.
3. Retrieval Methods: Five retrieval strategies were evaluated: Similarity Search (SS), Maximum Marginal Relevance (MMR), Self-Query (SQ), Compression (Comp), and a combined MMR+Comp approach. Each method was designed to optimize the retrieval of relevant data segments for the LLMs.
4. Query Generation and Processing: Fifty query-answer pairs were manually generated to serve as a control for evaluating retrieval performance. For each query, the relevant splits were compared to the manually generated answer to assess their similarity in terms of context. A score of 1 was given if the split was relevant to the query context. A score of 0 was given if the split was irrelevant to the context. The score was aggregated and divided by the total number of queries. This indicated how well the retrieval method performed. The method with the highest score was chosen as the retriever.

The most relevant data splits, identified through the retrieval methods, were combined with the original query and submitted to the LLMs for processing. The LLM's response was then compared to the control answers.

2.4 Evaluation of Model Performance

Model performance was evaluated based on the accuracy of the responses generated by GPT-4 and LLAMA2. In evaluating the

LLMs (Language Learning Models) with RAG (Retrieval Augmented Generation), the predicted answers produced by the model were compared directly against a control set of answers produced manually. Points were accumulated for the number of correct answers to each query and divided by the total number of answers generated by queries. This was the LLM score. The evaluation procedure examined the percentage of correct answers for three categories of queries:

1. Unstructured Data Queries: Questions requiring information from the Crop Production Protocol.
2. Structured Data Queries: Questions that were based on the collection of field data.
3. Combined Data Queries: Questions requiring information synthesis of unstructured and structured data.

This categorization was functional to assess the model's capabilities relative to the specific data types and to gain insights into the challenges associated with retrieving structured and unstructured data. The retrieval component of the RAG framework is essential for completing the task.

3. Results and Discussion

3.1 Vector Store Retrieval

The retrieval component of the RAG framework is essential for completing the task. Table 2 presents performance differences of the various retrieval strategies: Similarity Search (SS), Maximum Marginal Relevance (MMR), Self-Query (SQ), Compression (Comp), and MMR+Comp.

Table 2. Performance of different retriever variations. They are similarity search (SS), maximum marginal relevance (MMR), self-query (SQ), compression (Comp), and a combination of MMR and Comp (MMR+Comp).

Variation	GPT-4	LLAMA 2
SS	0.72	0.63
MMR	0.85	0.79
SQ	0.83	0.80
Comp	0.73	0.66
MMR+Comp	0.91	0.85

The combination of MMR+Comp demonstrated the best retrieval performance across GPT-4 and LLAMA2. This combination leverages MMR, prioritizing the most relevant responses and removing unwanted information to achieve balance with respect to relevance/diversity. The SS method yielded the lowest performance, as this method may seek responses that are clustered together as similar but not necessarily diverse or contextually relevant. Using a hybrid retrieval method, MMR+Comp worked exceptionally well for queries that involved synthesizing more complex information.

From the 40 assessed parameters, crop physical characteristics (fruit size, leaf length, etc.) and environmental characteristics (soil EC, temperature, etc.) demonstrated a high relevance score in the retrieval process. Inquiring about attributes like leaf length, fruit size, temperature, and precipitation yielded a score of no less than 0.85 for GPT-4 and LLAMA2, thus indicating these parameters are significant to answering crop questions. This result implies that prioritizing particular features could improve the relevance of the results returned, potentially improving the RAG system's overall accuracy.

3.2. Model Performance

The models' accuracy was measured by comparing the predicted answers to a set of manually generated correct answers across three query categories: unstructured data, structured data, and a combination of both. The results are presented in Table 3.

Table 3. Model performance for queries on structured data (St), unstructured data (Un), and structured+unstructured data (St+Un)

Data type	GPT-4	LLAMA 2
Un	0.87	0.78
St	0.79	0.75
St+Un	0.68	0.61

GPT-4 outperformed LLAMA2 in all categories, with the most notable difference observed in the combined data queries. The higher accuracy for unstructured data (87% for GPT-4 and 78% for LLAMA2) can be attributed to the model's training on diverse natural language datasets, which aligns well with the nature of unstructured input.

Structured data presented a greater challenge, reflected in the lower accuracy rates (79% for GPT-4 and 75% for LLAMA2), due to LLMs' limitations in interpreting tabular data and inferring relationships from isolated data points. The combined data type showed the lowest accuracy (68% for GPT-4 and 61% for LLAMA2), emphasizing the difficulty of synthesizing structured and unstructured information.

This challenge arises from differences in data representation: structured data is typically organized in tabular form with explicit relationships, while unstructured data consists of free-text descriptions with implicit context. LLMs' limitations in correctly synthesizing these data types would necessitate an advanced hybrid approach to eventually convert structured data into a more flexible format for processing by an LLM (e.g., embedding-based representations, semi-structured prompts, etc.).

Results indicated that GPT-4 considerably outperforms LLAMA2 in all queries and that the most significant difference is observed in the combined data query. This is likely due to GPT-4 having a larger pre-training history on unfocused datasets; this experience helps it interpret/make sense of complex information. Moreover, the benefits of LLAMA2 being an open-source system come with limitations with respect to processing mixed data dimensions with similar levels of sophistication as GPT-4.

3.3. Cost Implications and Strategies for Mitigation

The research compared the computational duration and monetary expenditure needed for GPT-4 and LLAMA2 processing a standard query, about 27,000 tokens. For the typical query, GPT-4 had an average processing time of 8.2 seconds and cost of USD 0.90, while the cost-free and open-source LLAMA2 required 11.4 seconds. The increased processing time for LLAMA2 can be attributed to its architecture being designed for affordability rather than speed. Regarding practical usage in agriculture, LLAMA2 can be a more economical alternative if the desired application, such as a real-time decision support tool for large-scale deployment, does not require immediate results.

The cost of processing queries is proportional to the cost of several prompts and tokens. We recommend working in practical settings to reduce cost with methods such as pre-filtering the data using RAG, batching the prompts, and using weight models for data-filtering methods before using LLMs such as GPT-4. These methods and others can reduce the cost to consumers of implementing LLMs while maintaining high accuracy in actual agricultural fields.

The results of our work also indicated that there is potential to use LLMs to query public proprietary data. This could be proprietary data from scientific experiments and confidential organizational documents. Furthermore, there is potential to use LLMs to diagnose crop conditions during production and understand why crops are in the condition they are in by comparing field data to known protocols.

4. Conclusion

In this work, we established the principles of a Retrieval Augmented Generation (RAG)-based method to query data that has been collected from MD2 pineapple crop production. The method married structured and unstructured data sources consisting of a crop production protocol document and empirically collected field data, for over 14 months, to test the performance of LLMs, GPT-4 and LLAMA2 at synthesizing both data types. Our data showed that both LLMs performed accurately and independently using a data type, but both LLMs decreased in performance when synthesizing both types of information. This illustrates the challenge of combining disparate data formats and the need for methodological development.

The significant issues with integrating structured and unstructured data stem from their individual properties. While structured data has natural relationships between data points due to its tabular nature, unstructured data has context primarily coded implicitly in natural language form. The problem of merging these two types of formats, with the subsequent reduced accuracy of the whole model when making combined queries, should lead to future experiments with preprocessing models like converting structured data to a semi-structured format to improve accuracy.

Regarding computational efficiency, GPT-4 shows that it has overall better accuracy than LLAMA2 in terms of workloads that were constructed with the robustness of handling data queries with datasets of structured and unstructured forms. That said, this comes with a higher financial cost - approximately USD 0.90 per typical query for all processing that includes 27,000 tokens worth of data. While LLAMA2 has a lower operational cost to run an average query, it also performs with a little less accuracy, particularly with mixed datasets, it presents the structure of cost efficiency versus accuracy in operational labor. Approaches to control most of these costs to be more efficient in operational situations with agricultural datasets may be by batching prompts or utilizing real-time agriculture specialty/knowledge context and static modeling to pre-filter using RAG or RAG-like approaches, or using lightweight models to consider as initial contextual operation models to improve accuracy once real data process model may have time constraints in query processing time.

Future work with the model as an assessment instrument will emphasize developing agricultural datasets that will include more dynamic input data to facilitate and support better diagnostics in supporting evidence in crop issues, e.g., dynamic real-time disease reporting systems and/or real-time high-resolution imagery.

Acknowledgements

This work was supported financially by the 12th Malaysian Plan development fund.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Author Contribution Statement

Authors Badril Abu Bakar and Siti Noor Aliah Baharom proposed the research problem.

Authors Mohd. Taufik Ahmad, Adli Fikri Ahmad Sayuti and Mohd Nizam Zubir: developed field experimental design.

All authors discussed the results and contributed to the final manuscript.

References

[1] R. M. Q. De Ramos and E. B. Taboada, "Cradle-to-gate life cycle assessment of fresh and processed pineapple in the Philippines," *Nat. Environ. Pollut. Technol.*, vol. 17, no. 3, pp. 783–790, 2018. https://neptjournal.com/upload-images/NL-65-15-(13)D-740.pdf

[2] R. D. Giamasrow, A. N. Azman, N. Zainol, M. S. A. Karim, and N. A. T. Yusof, "Fabrication of Cellulose Powder Dielectric Composite Material using Pineapple Leaves Fiber," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 38, no. 2, pp. 1–15, 2024. doi: https://doi.org/10.37934/araset.38.2.115

[3] A. F. A. Hamzah, M. H., Hamzah, H. C., Man, N. S., Jamali, and S. I. Siajam, "Influence of Subcritical Water Pretreatment Temperature on Pineapple Waste Biogas Efficiency: Experimental and Kinetic Study," *J. Eng. Sustain. Dev.*, vol. 28, no. 2, pp. 143–159, 2024, doi: https://doi.org/10.31272/jeasd.28.2.1

[4] FAO, "Agriculture production data, License: CC BY-NC-SA 3.0 IGO," Mar. 2022, Accessed: Mar. 25, 2024. [Online]. Available: https://www.fao.org/faostat/en/#data/QCL

[5] S. Cotabato, "A Study on the Production Methods of Conventionally-grown Pineapples in the Philippines," M*agsasaka at Siyetipiko para sa Pag-unlad ng Agrikultura*, February, pp. 1–25, 2015. https://pdfcoffee.com/conventional-pineaple-production-philippines-pdf-free.html

[6] B. Abu Bakar *et al.*, "A Review of Mechanization and Automation in Malaysia's Pineapple Production," *Adv. Agric. Food Res. J.*, vol. 2, no. 1, pp. 1–13, 2021, doi: https://doi.org/10.36877/aafrj.a0000206

[7] MAFI, "Dasar Agromakanan Negara 2.0 2021 - 2030 (National Agrofood Polizy 2.0 2021 - 2030)" Putrajaya, 2021.

[8] A. S. Bujang and B. H. Abu Bakar, "Agriculture 4.0: Data-Driven Approach to Galvanize Malaysia's Agro-Food Sector Development.", In Proceedings of the FFTC-RDA International Symposium on "Developing Innovation Strategies in the Era of Data-driven Agriculture". Jeonju, Republic of Korea, p.1631, 2019, [Online]. Available:

https://ap.fftc.org.tw/article/1631

[9] A. Knierim, M. Kernecker, K. Erdle, T. Kraus, F. Borges, and A. Wurbs, “Smart farming technology innovations – Insights and reflections from the German Smart-AKIS hub,” *NJAS - Wageningen J. Life Sci.*, vol. 90–91, Dec. 2019, doi: https://doi.org/10.1016/j.njas.2019.100314

[10] C. Y. N. Norasma, A. R. M. Shariff, E. Jahanshiri, M. Amin, S. Khairunniza-Bejo, and A. R. Mahmud, “SCIENCE & TECHNOLOGY Web-Based Decision Support System for Paddy Planting Management,” *Pertanika J. Sci. Technol*, vol. 21, no. 2, pp. 343–364, 2013, [Online]. Available: http://www.pertanika.upm.edu.my/

[11] W. X. Zhao *et al.*, “A Survey of Large Language Models,” pp. 1–97, 2023.

[12] A. Q. Jiang *et al.*, “Mixtral of Experts,” 2024, [Online]. Available: https://arxiv.org/abs/2401.04088

[13] H. Touvron *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” 2023, [Online]. https://arxiv.org/abs/2307.09288

[14] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 2023, [Online]. Available: https://arxiv.org/abs/2303.08774

[15] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science? Under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license," *Comput. Linguist.*, vol. 50, no. 1, 2024, doi: https://doi.org/0.1162/coli.

[16] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, “TabLLM: Few-shot Classification of Tabular Data with Large Language Models,” *Proc. Mach. Learn. Res.*, vol. 206, pp. 5549–5581, 2023, [Online]. Available: https://arxiv.org/abs/2210.10723

[17] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Wang, “HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data,” *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020*, pp 1026–1036, 2020, doi: https://doi.org/10.18653/v1/2020.findings-emnlp.91

[18] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models : A Survey,” Dec. 2023, [Online]. Available: https://arxiv.org/abs/2312.10997

[19] Y. Lee, S. Kim, T. Yu, R. A. Rossi, and X. Chen, “Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models,” Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.14195

[20] H. Kabir and N. Garg, “Machine learning enabled orthogonal camera goniometry for accurate and robust contact angle measurements,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, 2023, https://doi.org/10.1038/s41598-023-28763-1.

[21] V. Bolón-Canedo and B. Remeseiro, “Feature selection in image analysis: a survey,” *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2905–2931, 2020, https://doi.org/10.1007/s10462-019-09750-3.

[22] PIP, “Crop production protocol; Production Guide for the Production of the Pineapple Variety MD2 (a handbook for farm managers and technicians),” vol. 2, no. December, p. 60, 2011, [Online]. Available: www.coleacp.org/pip