

Sudan Agricultural Advising Consultant

Jessica Cannon

April 30, 2024

GitHub repo for this project:

https://github.com/jrcann11/DL4DS_FinalProject

Abstract

This project report describes the development of a question-and-answer (Q&A) chatbot, empowered by deep learning-based Language Model technology, to address the challenges faced by Sudanese farmers. I adapted and fine-tuned the existing GPT-4 LLM to the Sudanese agricultural context via RAG implementation, leveraging four large publicly available datasets from reputable sources. Evaluation of the chatbot's performance was conducted using the TruLens evaluation functions for Groundedness, Context Relevance and Answer Relevance [3], as well as a qualitative comparison of the RAG model's responses against the original model's responses to relevant queries. The results showed that the RAG model did well in both Context and Answer Relevance, as well as being domain specific and versatile within the domain. It also had a higher accuracy compared to the baseline when compared to ground truth data. However, the RAG model struggled with Groundedness and was less detailed with advice than the baseline model, which is most likely attributed to the source datasets not including enough specific contextual detail. Future work will ideally focus on further refining the LLM's capabilities and evaluating its long-term impact on agricultural productivity and livelihoods in Sudan.

Introduction

The elimination of poverty is a global imperative, and yet it remains an elusive goal, particularly in African regions like Sudan where agricultural sustainability plays a pivotal role in economic development. The challenges facing Sudanese farmers are multifaceted, ranging from environmental constraints to limited access to essential information and resources. In this context, leveraging advancements in deep learning technology presents a promising avenue to address these issues.

Sudan's agricultural sector serves as the backbone of its economy, employing a significant

portion of the population and contributing substantially to the nation's GDP. However, persistent poverty persists among farmers due to various factors, including inefficient farming practices, inadequate access to market information, and the adverse effects of climate change. These challenges not only hinder individual livelihoods but also impede the overall socio-economic development of this country.

In response to these pressing issues, my research aimed to combat poverty in Sudan by harnessing the power of deep learning-based Language Model (LLM) technology. Specifically, this report outlines the development of a question-and-answer (Q&A) chatbot tailored to address agriculture-related queries commonly encountered by Sudanese farmers. This chatbot serves as a virtual agricultural advisor, providing timely and accurate information on topics such as crop cultivation techniques, pest management strategies, legal policies, and market trends.

The innovative aspect of this approach lies in the integration of deep learning algorithms with natural language processing (NLP) techniques to create an intelligent and user-friendly interface for spreading agricultural knowledge. By leveraging large-scale datasets and a state-of-the-art language model, this chatbot is capable of understanding and responding to user queries in real-time, thereby empowering farmers with actionable insights to enhance their farming practices and decision-making processes.

Related Work

A team study at Microsoft that was conducted in 2024 [1] aimed to accomplish a comparable goal in a similar fashion. It presents a thorough investigation into the effectiveness of fine-tuned and Retrieval-Augmented Generation (RAG) language models in addressing context-related questions and answers in the agricultural domain. The study focuses on datasets from three major crop producer countries: the USA, Brazil, and India, utilizing diverse sources such as regulatory documents, scientific reports, agronomic exams, and knowledge databases. One of the key contributions of the paper lies in addressing the challenge of incorporating proprietary and domain-specific data into Large Language Models (LLMs). By leveraging popular LLMs including Llama2-13B, GPT-3.5, and GPT-4, the paper explores the tradeoffs between fine-tuning and RAG approaches, providing insights into their respective pros and cons.

The research methodology involves multiple stages, from data extraction and Q&A generation to model evaluation using metrics specifically designed for assessing the performance of the fine-tuning and RAG pipeline. The study focuses on the agricultural industry, aiming to provide location-specific insights to farmers, a domain with limited AI penetration. While this Q&A generation chatbot for agricultural advice directly aligns with what I set out to produce, the fact that these codes/models are so new and have yet to be released, alongside the fact that they are only specific to certain regions, provides much evidence

and value in recreating this model. I aspired to take what I learned from this study and apply it to an African region (Sudan) in hopes of providing similar access to resources and techniques.

Approach/Methodology

The methodology for this project consists of several procedures. First, I acquired relevant textual datasets pertaining to agriculture in Sudan in the form of several substantial PDF files. These files were eventually converted to Markdown (.MD) for easier parsing, as they generally have a simpler format as well as consistency across the documents. Each file's data quality and validation has been assured, as they were acquired from reputable and trustworthy sources such as the World Bank and FAOLEX (Food and Agriculture Organization of the United Nations). From here, I selected a pre-trained language model suitable for fine-tuning, considering factors such as model size, performance, and compatibility with agricultural data. Specifically, I evaluated candidate models based on their ability to capture domain-specific nuances and generate coherent responses.

The LLMs that I experimented on were GPT-4, GPT-3.5 and Llama 2, similarly to the Microsoft study as these well-known models showcased immense promise. In order to do so, I first evaluated the quality of the responses by posing a list of questions that small farm farmers would likely ask to each LLM. This list was produced by ChatPDF, which was asked to generate a list of questions likely asked about the source documents (datasets). After receiving and analyzing initial responses to these questions, I performed multiple iterations of prompt-tuning, involving taking the queries that yielded unsatisfactory responses and revising them to be more clear, concise and specific. Once content with the responses given, I analyzed the overall quality of each LLM's responses in comparison with each other, and based on overall coherence, consistency and depth/insight, the model I selected was the GPT-4 model.

Leaderboard

| Model Name | Score | Date Tested | License Type |
|--------------------------------------|--------|-------------|--------------|
| claude-3-opus | 86.67% | 2024-03-11 | Proprietary |
| gpt-4 | 85.71% | 2024-01-15 | Proprietary |
| llama-3-70b-instruct | 84.76% | 2024-04-19 | Open Source |
| centeotl | 80.95% | 2024-02-21 | Proprietary |
| agronomYi-hermes-34b | 79.05% | 2024-01-15 | Open Source |
| mistral-medium | 77.14% | 2024-01-15 | Open Source |
| nous-hermes-yi-34b | 76.19% | 2024-01-15 | Open Source |
| mixtral-8x7b-instruct | 72.38% | 2024-01-15 | Open Source |
| claude-2 | 72.38% | 2024-01-15 | Proprietary |
| yi-34b-chat | 71.43% | 2024-01-15 | Open Source |
| norm | 69.52% | 2024-01-17 | Proprietary |
| openhermes-2.5-mistral-7b | 69.52% | 2024-01-15 | Open Source |
| gpt-3.5-turbo | 67.62% | 2024-01-15 | Proprietary |
| mistral-7b-instruct | 61.9% | 2024-01-15 | Open Source |

Figure 1. Results of the Agronomy LLM Benchmark, ranking all the LLMs tested based on accuracy.

This choice was further supported by a quantitative analysis. In order to get a gauge of these LLMs' overall performance and accuracy on agricultural data, I utilized an Agronomy LLM Benchmark GitHub repo. This test benchmarked the ability of various large scale LLMs to give correct answers to a multitude of agronomy-based questions. By looking at Figure 1, we can see that out of all the candidates, GPT-4 scored the second highest with a score of 85.71%, only below claude-3 (86.67%) and just beating out llama-3. Both this and my qualitative findings on Sudan-specific agricultural data influenced my decision to employ the GPT-4 model for this chatbot.

After the pretrained model selection, I began constructing my model architecture, which is built on a combination of the Retrieval-Augmented Generation approach integrated with LangChain and OpenAI's GPT-4 Turbo model. The data was divided into chunks with a size of 500 and an overlap of 100 to ensure continuity. These chunks were then indexed/stored in Chroma vector store and converted into vector embeddings, facilitating efficient similarity searches. The application then interfaces with the database to fetch relevant document pieces into response to user queries.

Relevant document pieces are determined through similarity searches, where if a satisfactory match is found, it is used to aid the model in generating a contextually aware answer. A user interface was also built with streamlit, allowing users to input questions and view both the chatbot’s responses and the document context used for those responses more efficiently. Once the model was complete, I implemented retrieval augmented generation (RAG) of the source documents and repeated a cycle of prompt tuning and response evaluation. Finally, I was able to evaluate the RAG results and compare them to the baseline model to see how well the responses improved. Further detail of the evaluation can be found below.

Datasets

The input for my LLM consists of four large Markdown (.MD) file datasets (approximately 100 pages of text each) that include a multitude of textual Sudan agricultural information, such as crop data (types, growth patterns), soil data (composition, fertility), market data (prices, demand), policy data, farm management practices, and historical data. I am using what is publicly available and provided by the World Bank [6] , Ministry of Agricultural and Food Security [4] , the Food and Agricultural Organization of The United Nations [5] , and the Agricultural Research Corporation (ARC) for Sudan [2] .

Evaluation/Results

The final evaluation consisted of both a quantitative and qualitative analysis of the chatbot’s responses. For the former, I utilized evaluation metrics from the TruLens website [3], implementing specialized functions from their website into my code. These functions tested the overall groundedness, context relevance, and answer relevance of the responses produced from queries, each category earning a score between 0 and 1 (where 0 is the worst, and 1 is the best).

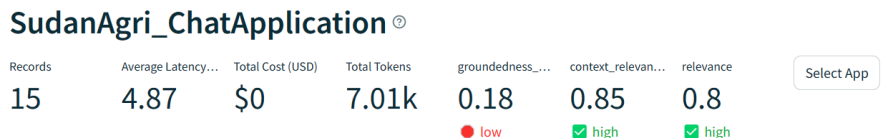


Figure 2. Average scores of the TruLens evaluation functions for Groundedness, Context Relevance, and Answer Relevance across 15 unique user queries.

Here we see the outcome of the TruLens evaluation on the RAG model. Looking at the

average scores from Figure 2, we can see that the model did quite well in terms of context and answer relevance, earning scores of 0.85 and 0.8 respectively. However, the model struggled a bit with its groundedness, averaging a score of only 0.18 across queries. Figure 3 below, which shows the distribution of the scores across the number of queries for the three evaluation functions, demonstrates how this metric also proved to be the least consistent. Its score fluctuated quite a bit (earning anywhere between 0.0 and 0.6 per response) compared to that of context and answer relevance, which remained very consistent throughout (earning between 0.8 and 1 for each response).

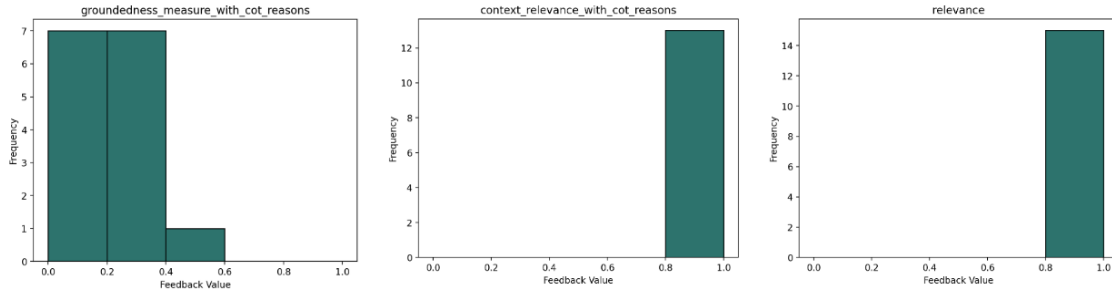


Figure 3. A series of bar graphs detailing the feedback score against the frequency of responses for the Groundedness, Context Relevance, and Answer Relevance metrics.

| User Input | Response | relevance | context_relevance_with_co... | groundedness_measure_wi... |
|-----------------------------------|------------------------------------|-----------|------------------------------|----------------------------|
| "Are there any specific polic... | "Yes, there are specific polic... | 0.8 | 0.8 | 0 |
| "How can I adapt to potentia... | "To adapt to potential climat... | 0.8 | 0.9 | 0 |
| "Are there specific recomme... | "To increase crop yields and ... | 0.8 | 0.9 | 0.4333333333 |
| "How can the findings and i... | "The document suggests tha... | 0.8 | 0.8 | 0 |
| "What support will be availa... | "Support for small farmers i... | 0.8 | 0.9 | 0.3333333333 |
| "Will there be training or cap... | "Yes, there will be training or... | 0.8 | 0.8 | 0.2666666667 |
| "How can small farmers leve... | "Small farmers can leverage ... | 0.8 | 0.9 | 0.2333333333 |
| "What steps can small farme... | "Small farmers can improve ... | 0.8 | 0.8 | 0 |
| "Are there examples of succe... | "Yes, there are successful ini... | 0.8 | 0.8 | 0.3333333333 |
| "How can small farmers eng... | "Small farmers can engage ... | 0.8 | 0.9 | 0.1 |
| "Are there any specific mark... | "Yes, the report highlights fiv... | 0.8 | 0.8 | 0.3333333333 |
| "How can small farmers acc... | "Small farmers can access fi... | 0.8 | 0.9 | 0.3333333333 |

Figure 4. A chart that gives a further breakdown of the different queries posed to the RAG model, as well as the unique scores earned for each query.

Figure 4 further highlights the diversity in the groundedness evaluation responses across the different queries. While the exact reason these scores remain so low is unknown, my hypothesis is that the datasets the RAG model incorporated into formulating its responses did not include enough specific contextual detail. Finding reliable and sufficient data sources on Sudanese agriculture proved to be a challenge, given the scarcity of publicly

available data in general, especially in English and on smaller African countries whose agricultural data sectors are underdeveloped compared to that of North American regions like the United States.

However, I was still able to obtain two main takeaways from this experimentation. The first is that increasing the chunk size and overlap aided in producing better values for the groundedness metric, which was even lower to begin with. Also, utilizing 3 context chunks with the highest similarity search scores to formulate responses allowed for both convenience as well as overall balanced scores.

ChatGPT

Adapting to climate change threats to safeguard your farming livelihood involves several strategic and practical measures. Here are some key approaches you might consider:

- 1. Diversify Crops and Livestock:** Growing a variety of crops can reduce the risk of total loss from any single climate event. Crop diversity can also improve soil health and reduce pest and disease pressures. Similarly, diversifying livestock can help stabilize income as different animals may be more resilient to varying conditions.
- 2. Implement Water Conservation Techniques:** Water scarcity is a major concern under climate change scenarios. Techniques like drip irrigation, rainwater harvesting, and the construction of water storage facilities can help ensure water availability during dry spells.
- 3. Improve Soil Health:** Healthy soils retain water more efficiently and are more resilient to erosion and degradation. Practices such as cover cropping, reduced tillage, and organic amendments can enhance soil structure and fertility.
- 4. Adopt Climate-Resilient Varieties:** Some crops have been bred or genetically modified to withstand stresses like drought, heat, or salinity. Switching to these varieties can help maintain productivity even under altered climatic conditions.



To adapt to potential climate change threats and safeguard your farming livelihood, consider the following strategies based on the provided context:

- 1. Access to Information:** Ensure you have access to the right information at the right time. This includes updates on weather patterns, pest attacks, and disease outbreaks. Utilize resources that provide emergency advisory services and improved weather forecasting systems.
- 2. Education and Awareness:** Increase your knowledge and understanding of climate change and its impacts on agriculture. Participate in educational programs at schools and universities, and seek information about agriculture best practices, including those related to climate adaptation.
- 3. Linkages with Experts:** Enhance connections with researchers, extension services, and other farmers. This network can provide vital support and share innovative farming techniques and solutions that are resilient to climate change.
- 4. Financial and Insurance Schemes:** Gain awareness and make use of available financial and insurance products designed to support farmers in times of uncertainty and loss due to climatic changes.
- 5. Risk Management:** Develop and implement a proper risk management system tailored to your specific farming conditions and potential climate threats.
- 6. Diversification:** Consider diversifying your crop types or farming activities to reduce dependence on a single crop or method susceptible to climate threats.

Figure 5. The GPT-4 and RAG model's responses to the prompt "How can I adapt to potential climate change threats to safeguard my farming livelihood?" The baseline response is on the left while the chatbot's response is on the right.

Finally, I performed prompt/response evaluations against the baseline model. The responses produced were compared to each other and measured on accuracy (compared against ground truth data obtained from the source documents), clarity, and overall relevance. Within Figure 5, we can see example responses from both GPT-4 and the RAG model given the same prompt.

The responses in general were pretty similar, but a few differences stood out. One is that the chatbot response was more domain specific, containing practices that were more specific and imperative to the region and social climate. It was also more diverse within this domain overall, and had a greater accuracy rate than the baseline when I compared them to the ground truth data. However, my implementation's responses had a few downsides, one being that its responses were less detailed with the advice it suggested compared to that of GPT-4. Also, for some responses, the clarity of actions was not always as satisfactory

as those of the baseline model. Similarly to the subpar TruLens scores for Groundedness, I believe this lack of clear and specific detail stems from the source documents not including enough specific contextual detail. With more time, this is something I would consider for the future, perhaps even reaching out to trusted Sudanese organizations myself to obtain relevant data in greater quantity.

Conclusion

This project presents a novel approach to addressing poverty in Sudan by leveraging Language Model-based chatbots for agricultural support. Through fine-tuning the existing GPT-4 architecture with domain-specific data, this LLM has the potential to offer a scalable and effective solution to empower Sudanese farmers with localized knowledge and assistance. As of now, while the retrieval-augmented generation chatbot saw improvements in domain specificity, the advice for actions was often general. As stated above, I believe this could be greatly improved by incorporating data with more/better contextual detail, and taking the initiative to reach out and obtain private domain data instead of limiting myself to what was publicly available may serve as a huge aid in accomplishing this. Also, larger chunk size and chunk overlap generally produced better results for this experiment, however this may vary for other LLMs depending on the specific goals of the model and the nature of its data. Future work for this project will ideally focus on further refining the LLM's capabilities, expanding its reach, and evaluating its long-term impact on agricultural productivity and livelihoods in Sudan.

Repo Survey

- <https://github.com/avrabyt/RAG-Chatbot>
- https://github.com/gbstox/agronomy_llm_benchmarking
- <https://github.com/pixegami/langchain-rag-tutorial>

References

- [1] Balaguer, Angels., Benara, Vinamra., Cunha, Renato. *Rag vs Fine-Tuning: Pipelines, Tradeoffs, And A Case Study On Agriculture*. Microsoft, 2024.
- [2] El-Siddig, Prof. Kamal, and Prof. Imad-eldin Ahmed Ali Babiker. *News from APGRC*. Agricultural Research Corporation (ARC), arc.gov.sd/. Accessed 16 Feb. 2024.
- [3] *Langchain Quickstart*. LangChain Quickstart - TruLens, www.trulens.org/trulens_eval/getting_started/quickstarts/langchain_quickstart/#import-from-langchain-and-trulens. Accessed 30 Apr. 2024.

- [4] *Republic of South Sudan*. Ministry of Agriculture and Food Security South Sudan, 17 Aug. 2023, mafs.gov.ss/.
- [5] *Sudan*. Food and Agriculture Organization of the United Nations, www.fao.org/countryprofiles/index/en/?iso3=SDN. Accessed 16 Feb. 2024.
- [6] “*Sudan - Agriculture Value Chain Analysis*.”. World Bank, 1 July 2020, documents.worldbank.org/en/publication/documents-reports/documentdetail/731741593616746051/sudan-agriculture-value-chain-analysis. Accessed 30 Mar. 2024.