

INITIATIVE ON  
Digital Innovation

# AgroTutor

*Localised Generative AI Infrastructure for Agri Advisories*

Satish Nagaraji, Andrea Gardeazabal, Narasimhan KV, Swathi Vurakula, Sherin Maria Saji, Sandya NR, Fredrick Achar, Rosa Elena Bautista Ramirez and Sanjeev Mahto

**TECHNICAL PAPER**

December 2024

SUMMARY.....	3
1. INTRODUCTION .....	4
2. THE AGROTUTOR SYSTEM .....	4
<b>2.1. Core Components</b> .....	<b>4</b>
3. CONTEXT AND CHALLENGES.....	5
4. PILOT USE CASES.....	6
<b>4.1. Bihar, India: Enhancing Resilience in Rice and Wheat Farming in Bihar</b> .....	<b>6</b>
<b>4.2. Kenya: Climate-Smart Solutions for Dryland Farming</b> .....	<b>7</b>
<b>4.3. Mexico: Regenerative Farming in Bajío Region</b> .....	<b>7</b>
5. TECHNOLOGICAL ARCHITECTURE OF AGROTUTOR.....	7
<b>5.1. Indexing Pipeline</b> .....	<b>8</b>
5.1.1 Corpus Development.....	8
5.1.2 Corpus Data Translation .....	8
5.1.3 Tagging .....	9
5.1.4 Corpus Pre-Processing .....	9
5.1.5 Chunking Strategy .....	10
5.1.6 Embedding.....	11
5.1.7 Vector Database Creation.....	11
<b>Search Pipeline:</b> .....	<b>11</b>
<b>5.1.8 Phase 1: Query</b> .....	<b>11</b>
5.1.9 Phase 2: Memory .....	12
6. EVALUATION .....	13
6.6.1. Retriever Component Evaluation: This is .....	14
6.6.2. Generator Component Evaluation .....	14
7. STAKEHOLDER ENGAGEMENT AND FUTURE DIRECTIONS .....	14
8. CONCLUSION .....	15
9. REFERENCES.....	16

This publication has been prepared as an output of [CGIAR Initiative on Digital Innovation](#), which researches pathways to accelerate the transformation towards sustainable and inclusive agrifood systems by generating research-based evidence and innovative digital solutions. This publication has not been independently peer-reviewed. Any opinions expressed here belong to the author(s) and are not necessarily representative of or endorsed by [CGIAR](#). In line with principles defined in [CGIAR's Open and FAIR Data Assets Policy](#), this publication is available under a CC BY 4.0 license. © The copyright of this publication is held by [IFPRI](#), in which the Initiative lead resides. We thank all funders who supported this research through their contributions to the CGIAR Trust Fund.

## SUMMARY

Generative AI (GenAI) technologies offer transformative potential in agriculture, enabling precise and timely advisories tailored to regional needs. The GenAI module of the AgroTutor platform developed by CIMMYT pioneers a scalable approach to leveraging GenAI to deliver actionable and localised agronomic insights to smallholder farmers across the global south. By employing Retrieval-Augmented Generation (RAG), low-cost language models, and data integration, AgroTutor addresses critical challenges such as context localisation, resource optimisation, and quick deployment of RAG.

AgroTutor is positioned as GenAI infrastructure that enables local knowledge organisations to provide farmers with cutting-edge, AI-powered advisory services, particularly in the Global South. The system leverages a generative AI framework to mitigate critical knowledge gaps within the agricultural sector. Designed with localised contexts in mind, the system integrates RAG frameworks with Large Language Models (LLMs) to enhance advisory relevance and accuracy. Pilots conducted in India, Kenya and Mexico demonstrate its potential for scalability and ease of deployment, addressing challenges such as crop management, pest control, and climate adaptation.

This Technical Report paper outlines the process, methodology, pilot outcomes, and the way forward of AgroTutor for agriculture advisories.

**Keywords:** Generative AI, Digital Agriculture, GenAI for Farmer advisory, Large Language Model (LLM), Retrieval Augmented Generation (RAG), Global South

## 1. INTRODUCTION

The global agricultural sector faces unprecedented challenges due to climate variability, resource constraints, and knowledge gaps despite the emergence of cutting-edge technologies in AI and IoT (Fountas, S., et al, 2024). Generative AI (GenAI) models, such as OpenAI, GeminiAI, and Llama, have emerged as powerful tools (Ghosh, R.C. et al, 2024) for transforming agricultural practices by providing expert advisories (Kamilaris, A., & Prenafeta-Boldú, F. X. 2018). However, limitations in technological skills, capacities, and infrastructure for building AI architectures exist, and adapting these technologies to local contexts—considering regional languages, farming practices, and socioeconomic factors—remains a significant challenge (Liakos, K. G., et al, 2018).

**AgroTutor** is a digital platform powered by AI, created to address knowledge gaps in adapting to newer technologies and agricultural practices, providing decision support systems. In the long-term, this will enhance smallholder farmers' livelihoods through farmer advisory provision in their local languages and contextualised to their agroclimatic zones. The scalable architecture allows local organisations to develop tailored farming recommendations using their own knowledge base, which is provided directly to farmers in their native languages. Additionally, it leverages GenAI's capabilities to address these gaps. By combining RAG frameworks with localised datasets, the system delivers actionable insights tailored to smallholder farmers. This initiative represents a critical step towards leveraging evolving digital technologies like GenAI in achieving climate-smart and sustainable agriculture, especially in resource-constrained regions of the global south.

## 2. THE AGROTUTOR SYSTEM

### 2.1. Core Components

The development of AgroTutor involved several critical components, each tailored to enhance the platform's functionality and accessibility for smallholder farmers. **Corpus collection and pre-processing** began with curating multilingual datasets sourced from credible entities such as agricultural research institutions and public repositories including NARS/NARES and through web scraping. The collected knowledge base was translated into English to ensure standardised processing and scalability, enabling seamless integration and application across diverse linguistic regions.

The **AI architecture** utilised OpenAI's GPT-4o model, integrated with a Retrieval-Augmented Generation (RAG) framework to deliver contextually relevant advisories. Localised corpora were embedded into '**vector databases**', allowing for advanced search capabilities. This approach ensured that advisories were tailored to specific regional conditions and agricultural practices, enhancing their practical applicability. **The RAG framework** serves as a bridge between general-purpose large language models and domain-specific agricultural knowledge. The system provided highly relevant and actionable insights by incorporating both structured and unstructured data, such as crop guides, extension manuals, research papers, field reports etc. The retrieval pipeline was optimised using similarity to retrieve information most pertinent to the user's query.

**RAG Pipeline** consists of different components:

1. **Retriever Component** that retrieves textual chunks which from a vector database for the query to be answered by the LLM.
2. **Generator Component** that generates an answer based on a prompt augmented with the retrieved information.

Other terminological aspects related to the RAG Pipeline are:

1. **Query:** This question (from farmers/extension agents) forms the input for the RAG pipeline.
2. **Response:** is the answer generated from the RAG pipeline, i.e., the output.
3. **Ground-truth** is simply the ground truth answer to the question. It is human-annotated information required for Evaluation Metrics. (Liu, Y., et al, 2023).

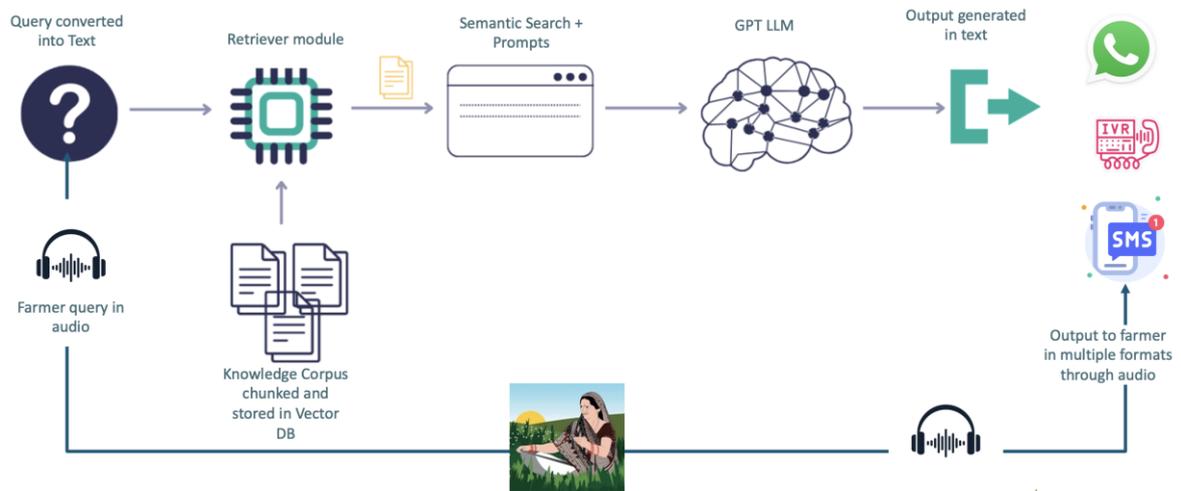


Figure 1: High-level architecture of the AgroTutor Framework

**Localisation Techniques** were crucial in ensuring the platform’s effectiveness across diverse regions. Corpus was developed using region-specific documents, enabling the generation of precise and relevant completions. Standardised terminologies for crops, pests, and farming practices were also established through ‘compendiums’ which mapped local dialect to English terms which might have otherwise eluded the GPT.

To maximise accessibility, AgroTutor employed a multi-channel **Delivery Mechanism** focusing on inclusivity. Farmers could access advisories via WhatsApp, where voice-to-text translation was utilised for input, followed by generating responses in local languages. These responses were delivered as audio messages, addressing literacy barriers and ensuring the platform’s usability among a broad demographic of farmers.

### 3. CONTEXT AND CHALLENGES

Localised agriculture needs: Agriculture is inherently local, with challenges that vary significantly depending on geographic, climatic, and socio-economic factors (Kuska, M. T., et al, 2024). In many regions, smallholder farmers face limited access to timely and expert agronomic advice, hindering their ability to adopt best practices, increased vulnerability to the impacts of climate change (Nguyen, V., et al, 2024), including unpredictable weather patterns and pest outbreaks, and inefficiencies in resource use, particularly concerning water, fertilisers, and pest management (Sapokta, R., et al, 2024). GenAI can provide valuable support by delivering context-specific, timely insights that enable farmers to make informed decisions and adopt climate-smart practices (Singh, N., et al, 2024). However, these solutions must be tailored to each region's unique agricultural landscape.

#### 4. PILOT USE CASES

Considering the commonalities in the context of agriculture challenges, digital readiness and technology adoption patterns, India, Kenya, and Mexico were chosen. The lessons from these countries can help design solutions that work across the Global South and can be adapted to similar contexts globally. Considering the pilot duration and the scope, the crops covered in scope was aligned to the key crops cultivated in the pilot locations during the season.

Table 1: Scope of crops covered in each use case

Bihar, India	Kenya	Mexico
Maize Rice	Maize Irish Potato French Beans Green Gram Beans	Maize Wheat

##### 4.1. Bihar, India: Enhancing Resilience in Rice and Wheat Farming in Bihar

Bihar’s agriculture, dominated by rice and wheat, is heavily impacted by erratic rainfall and resource constraints (Hoda, A., et al, 2021). In regions like Bihar, where erratic rainfall and resource limitations severely affect agricultural practices, GenAI systems can deliver tailored advisories on climate-resilient crop choices. For example, recommending maize as an alternative to traditional rice crops during the kharif season could be a viable strategy for enhancing resilience. In a pilot use case deployed by our team in Bihar, AgroTutor was designed to deliver climate-smart advisories in a range of queries covering all aspects of crop production in rice and maize.



Figure 2: Mobile Interaction of AgroTutor

## 4.2. Kenya: Climate-Smart Solutions for Dryland Farming

In Kenya, where rainfed farming systems are increasingly vulnerable to drought (FAO), GenAI's advisory system is designed to provide drought-resilient crop recommendations, including promoting drought-tolerant varieties as per the best agriculture practices. These interventions helped improve farmer livelihoods and agricultural stability in the region. Considering the pilot, the scope of the crop that the advisory was provided was restricted to the pilot geographies.

## 4.3. Mexico: Regenerative Farming in Bajío Region

*Bajío Region* in Mexico is an agribusiness hub, where water scarcity and soil degradation pose significant challenges (Marañón, Boris. 2006). In this pilot under deployment in the *Bajío Region*, where soil degradation and water scarcity threaten agricultural productivity, AgroTutor was designed to provide actionable insights to support regenerative maize and wheat farming practices. By encouraging practices that restore soil health and improve water usage efficiency, GenAI is foreseen to contribute to the sustainability of farming systems, which aligns with our AgMission Initiative here.

## 5. TECHNOLOGICAL ARCHITECTURE OF AGROTUTOR

The technological foundation of AgroTutor is built upon several key components, each tailored to meet the unique challenges of delivering localised agricultural advisories. These components work together to ensure accurate, context-sensitive, and accessible solutions for smallholder farmers.

Below, we outline the step-by-step method for executing OpenAI's GPT (GPT3.5, GPT4, and GPT4o were tested) with RAG to deliver advisories to farmers and extension officers. Two pipelines have been developed—an '**indexing pipeline**' and a '**search pipeline**' for executing the advisories end-to-end, as outlined in Figure 3.

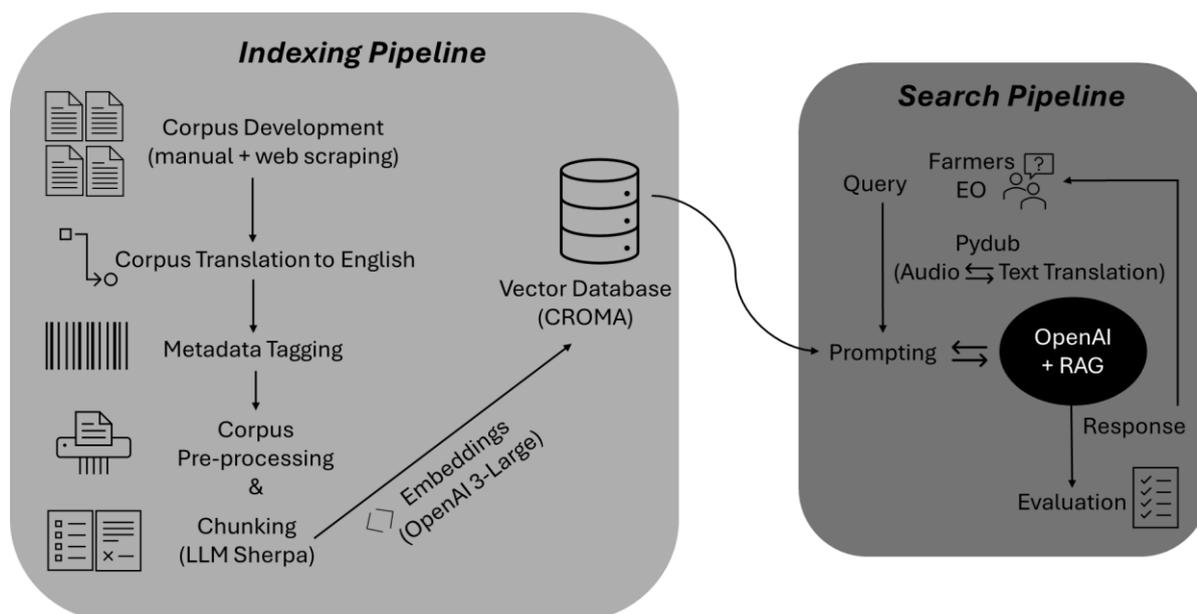


Figure 3: Overview of the Technological Architecture of AgroTutor

## 5.1. Indexing Pipeline

### 5.1.1 Corpus Development

For India, 310 documents were collected between 29 July 2023 to 23 Oct 2024. While the first 118 documents were collected via manual search, a combination of manual searching and website scraping was done to collect the following 192 documents. The SCOPUS database was used for manual searching with the keywords ‘Climate Smart Agriculture AND Bihar OR Indo-Gangetic Plains’ and ‘Climate Smart Agriculture AND Rice’. The search scope focused mainly on rice and maize (two significant crops grown in Bihar) and climate-smart agriculture in Bihar, India. Additionally, research papers published in the last 10-12 years were selected for addition to the corpus for the Bihar use-case. Web scraping was done for the latter mainly from State and National Government sources such as agriculture extension websites, Farmers’ Corner (a repository that has localised recommendations for Bihar and similar cropping systems), NRRI (National Rice Research Institute, Cuttack, Orissa) website owing to the similar cropping patterns making its advisories and recommendations applicable across the States of Orissa, Bihar, UP etc. A first revision of the corpus was done to remove about forty documents that were not useful in answering the queries, and a second revision of the corpus was done to remove documents on crops other than rice and maize.

For Mexico, 148 documents were collected between 29 July 2023 and 5 Sep 2024. While the first 147 documents were collected via manual search, manual searching and website scraping were done to collect the last document. The scope of the search is mainly on maize and wheat, with other crops such as beans, barley, and oats included. The location is in Mexico, primarily the Bajío region (Guanajuato, Queretaro, Michoacan and some parts of the nearest States). However, we have documents about other important Agricola regions, such as the north of Mexico and the Peninsula.

Additionally, documents from the last 10 years were selected to add to the corpus of this knowledge base. Additionally, for research articles, only indexed journals and technical words were used as selection criteria for the sources. Important and reputable agricultural institutions in Mexico, such as INIFAP, CIMMYT, and SADER, have been chosen for their information reliability for bulletins and manuals. For Kenya, 21 documents were collected from the Kenya Agricultural and Livestock Research Organization (KALRO) website <https://laikipiagaps.kalro.org/> and were in Swahili. The KALRO knowledge base is a public-access repository of researched step-by-step production advisory for most crops and livestock commonly grown and kept in Kenya. These resources are well curated and provided on web pages and downloadable PDFs for users to access and utilise. KALRO, a national agriculture research institute in Kenya with a high reputation for validated agricultural advisory, has been chosen for its information reliability. The scope of KALRO focused on five main crops: Maize, Irish potatoes, French beans, Green grams, and Common beans.

The documents added to the corpus included primarily PDF and TXT file formats.

### 5.1.2 Corpus Data Translation

OpenAI’s GPT4o was used to convert the corpus documents to local languages like Hindi, Bhojpuri, Swahili, etc., to English. All the documents were kept in English, and the queries from farmers and/or extension officers were also translated into English (as detailed in step 7: Prompting of Methodology) for following reasons. First is that the models of OpenAI, while robust enough to generate good results for multiple languages, have been optimised for use in English. The second is that the popular open-source datasets, such as Common Crawl, The Pile, Refined Web, C4, Wikipedia, etc., are in English for LLMs for OpenAI training. The third is that multilingual embedding makes it challenging to search the vector database. Hence, all the documents in the latter are also

chosen to be in English to allow for better embedding in further steps of the methodology. Finally, this also allows us to scale our advisories to other locations with a different language. Therefore, no matter the language used in any of the locations where the advisory is issued, translation of the language to English has been performed to ensure simpler workflows and scalability of advisories.

The system inherently answers in English but provides a translated answer in local languages. After this, Naraakeet converts the text in the local language text to audio to make it accessible to local farmers and extension officers (as described further in the prompting step of the method).

### 5.1.3 Tagging

The corpus was then tagged (auto-tagging using OpenAI followed by manual verification of the tagging) for filtering into the following three categories:

- a. Category (For example, pest management).
- b. Sub-Category (For instance, pest management method, IPM etc).
- c. Crop (For example, if the document collected is concerning rice or maize or any other crop in question).

This tagging allows us to do post-query filtering using a waterfall approach, wherein the full match is first searched for. If not found, categories, subcategories, and crops are used to search the corpus for relevant texts for a query. Finally, if none of the tags match, the entire corpus without metadata tagging is used to search for relevancy. Additionally, based on the tags, we can understand the distribution of topics, file types, size, etc.

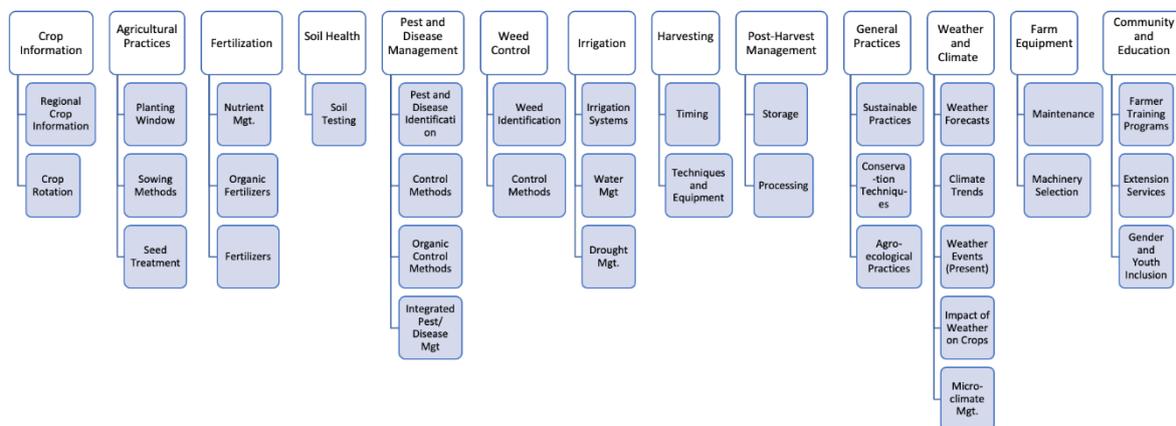


Figure 2: Meta-tagging Process

### 5.1.4 Corpus Pre-Processing

We tested the pre-processing of the corpus documents using recursive splitting, Llama Parse and LLM Sherpa to remove noise, redundancy, and irrelevance and improve the LLM performance. Recursive splitting was ineffective. Llama Parse is a paid alternative to LLM Sherpa which automatically removes bibliography, headers, and footers, including numerical information such as page numbers and other irrelevant text information. It also pre-processed the corpus to remove tables in the corpus documents. Hence, LLM Sherpa was finally chosen as the apt package for corpus pre-processing and the next chunking step, as it has the added advantage of keeping the document structure coherent when creating chunks.

### 5.1.5 Chunking Strategy

The next step is chunking – dividing large pieces of text into smaller, digestible chunks that can effectively feed information into the LLM within its optimal context window. The LLM context window is the amount of text the LLM can analyse at a time – often a critical limitation in LLM applications that hinders its performance in tasks requiring contextual understanding, such as searches, essential for localised agriculture advisory.

```
Chunk 6:
Management of Insect Pests
The newly hatched larvae crawl over the leaf for about 15-30 minutes and then feed on the rolled leaves.
Leaf eaten by young larva when unfurled, displays pin holes in horizontal row, The grown up larvae There are only three serious and regular pests in maize ecosystem that too are divided in three different seasons.
Chilo partellus is infesting in kharif and also in late spring maize.
Sesamia inferens is more prevalent in rabi and spring maize.
Atherigona spp.
is a regular pest of spring in northern part of India.
Besides, there are nearly one dozen pests which occur sporadically and cause considerable crop loss at times.
The loss caused by insect pests in maize crop ranges from 5- 15%.
The colossal production of pollen grains attract large number of natural enemies, further, the plant architecture offers a suitable niche for them to hide during harsh conditions, thus naturally protecting the crop from the ravage of pests considerably.

Chunk 7:
Management of Insect Pests
MAIZE is one of the most important cereal crops in the world; provides nutrients for humans, animals and serves as a basic raw material for the production of starch, oil, alcoholic beverages, food sweeteners and more recently fuel, besides myriads of other industrial products.
It is an important source of carbohydrate, protein, iron, vitamin B, and minerals.
In Indian agriculture, maize occupies a prominent position and each part of the maize plant is put to one or the other use.
It is a versatile crop, grown across a wide range of agro ecological zones.
Insect pests and diseases severely limit the production of maize.
Over 130 insect pests have been reported causing varying degree of damage attacking from seedling to maturity stage and some pests destroy stored products in godowns, bins, storage structures and packages, causing huge amount of loss to the stored food and also deterioration of food quality.

Chunk 8:
Management of Insect Pests > Spotted Stem borer Chilo partellus (Swinhoe) (Lepidoptera : Crambidae)
| Distribution: distributed throughout maize | This pest is | Chilo partellus Swinhoe - the most serious pest of maize (a) Adult (b) Larva
| --- | --- | ---
| Cocoons of Cotesia sp.; T. chilonis (inset) | | S. inferens, (a) adult, (b) larva damaging stem, (c) dead-heart

Chunk 9:
Management of Insect Pests > Spotted Stem borer Chilo partellus (Swinhoe) (Lepidoptera : Crambidae)
feed downward thus the holes made are large and vertically oblong.
As the larvae move downward feeding upon a 10-20 day-old-plant it reaches the meristem which is also fed upon.
The central leaf of such plant dries up making dead heart and the plant usually dies or gives rise to tillers.
```

Figure 3: Chunking using LLM Sherpa

We deployed LLM Sherpa (<https://github.com/nlomatics/llmsherpa>) for corpus pre-processing and chunking for further search. While recursive text splitters do not work for contextualised deployment of OpenAI due to splitting chunks between concepts, LLM Sherpa has been shown to better hold contexts due to its adherence to document structure and hierarchy.

The LLM Sherpa performs intelligent chunking, wherein the document reader/loader knows the document structure and hierarchies. This results in,

- The text identifies sections and subsections, merges lines into coherent paragraphs and establishes connections/links between sections and paragraphs.
- For tables, the table layout is preserved along with table headers and sub headers.
- For lists, all list items are in a single chunk, list items from the following page are merged into this chunk, and the lead-in sentence with the context for the list stays with this chunk.

The following parameters were evaluated for optimising the chunking strategy:

- Chunk size
- Chunk volume
- Number of chunks

Table2: Use case-wise corpus volume, number of chunks and word length

Use case	Number of words	Approx Number of Pages in A4	Number of chunks	Chunk Volume
Bihar, India	2,233,070	4,962	34960	1022 Mb
Kenya	744,519	1,654	8768	280 Mb
Mexico	1,125,789	2,501	7627	164 Mb

### 5.1.6 Embedding

Keyword matching is inefficient for LLM applications needing contextualisation, like that of agricultural advisory. We, therefore, employed embedding in arrays for richer comprehension and, thereby, for a better understanding of the context for optimal advisory generation. We tested OpenAI 3-Large and ADA as our embeddings and found OpenAI 3-Large to give better embeddings and generated arrays for further steps.

### 5.1.7 Vector Database Creation

The arrays generated here are stored in a vector database. We used ‘Croma’ as our vector database, in which our corpus is embedded as distinct arrays.

To develop our application powered by LLMs, we used LangChain as the framework to tie everything, i.e., the indexing pipeline, with the search pipeline.

<https://python.langchain.com/docs/introduction/>

Name	Size	Changed	Rights	Owner
⋮		1/18/2025 4:38:01 PM	rwxrwxr-x	ubuntu
d55781fa-b7c5-4eba-867e-618bd9e2c4e6		1/15/2025 8:36:17 AM	rwxrwxr-x	ubuntu
chroma.sqlite3	631,428 KB	1/15/2025 5:38:35 PM	rw-rw-r--	ubuntu

Figure 4: Vector Database of Bihar use case

### Search Pipeline:

Prompting and response generation is done in two phases:

#### 5.1.8 Phase 1: Query

In the first phase, queries are sent by farmers and/or extension officers in audio format (voice messages on WhatsApp in the local languages such as Hindi, Bihari and Bhojpuri in the Bihar region of India, Swahili, etc.). Four libraries, namely Pydub, Librosa (Python package for audio and music analysis), Google Cloud and OpenAI Whisper, were tested for the audio-to-text translation. The audio-to-text translation was evaluated, and it was found that Pydub and Open AI Wisper performed better among the packages tested. Since PyDub was an open source, it was used in the model. The text generated is then translated into English by OpenAI.

Additionally, in Bihar State of India, local dialects and differential pronunciations of agriculture-related terminologies in Bihari and Bhojpuri languages led to translation issues of farmer queries. Agronomists manually prepared a ‘Key Compendium’ to negate these primary translation issues and include colloquial names of seasons, pests, diseases, and weeds. Similarly, agronomists manually

prepared a 'Key Compendium' for our pilot in Kenya, linking terminologies in Swahili to English for crop names, weeds, pests, and disease names.

The English queries are then used to search for similarity in the vector database, i.e., a similarity retriever in langchain, which most likely implements cosine similarity, a similarity measure between two non-zero vectors defined in an inner product space. This fetching of similar chunks happens in two ways – firstly, we can currently choose the five closest chunks, and secondly, the search can be expanded around these five closest chunks. Once the closest/relevant chunks are retrieved, they are sent to OpenAI again with both the query in English and the relevant chunks. Based on this, OpenAI answers the queries in English, which are then translated back to the local languages such as Hindi, Swahili, etc., and sent to the final users, i.e., farmers and/or extension officers. At the end of phase one, all interactions from queries are saved with phone numbers as object names.

```

15m ✓ ▶ from langchain.embeddings import OpenAIEmbeddings
    from langchain.vectorstores import Chroma
    embeddings = OpenAIEmbeddings()
    docsearch = Chroma.from_documents(documents, embeddings)

WARNING:chromadb:Using embedded DuckDB without persistence: data will be transient

15m ✓ ▶ from langchain.embeddings import OpenAIEmbeddings
    from langchain.vectorstores import Chroma
    embeddings = OpenAIEmbeddings()
    docsearch = Chroma.from_documents(documents, embeddings)

WARNING:chromadb:Using embedded DuckDB without persistence: data will be transient

```

Figure 5: Embedding of text into vector database

### 5.1.9 Phase 2: Memory

In the second phase, when the single user (farmer/EO) sends another or multiple sequential queries, the queries get tagged with a unique ID linked to their WhatsApp phone number. When there are numerous queries from the same ID, the last five interactions are stored in the memory of OpenAI, which generates a 'standalone question' that looks at the history of queries from a unique ID and the relevant response. Once these steps happen, the original query follows the same steps in phase one of prompting. Thus, with every query, memory instances are created using 'conversational memory retriever' in Langchain to result in learning and better contextualisation, before OpenAI answers a query.

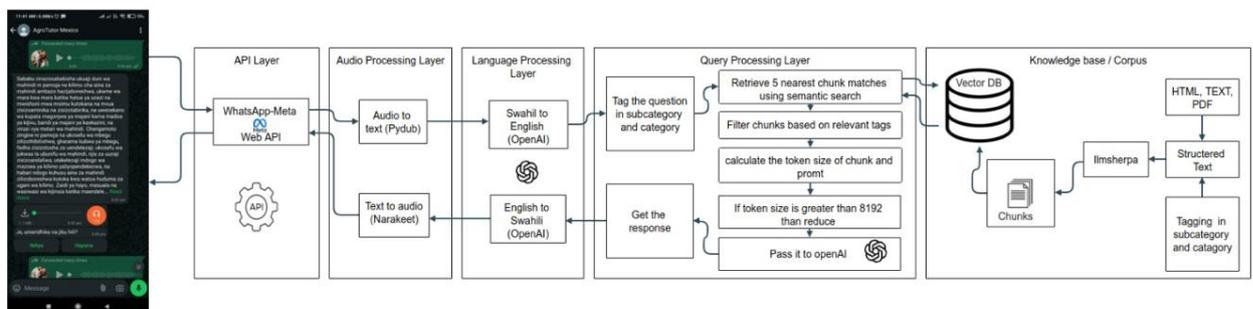


Figure 6: Process of the Technological Architecture of AgroTutor

## 6. EVALUATION

Building the satisfactory GenAI infrastructure for effective agriculture advisories using LLMs and RAG requires the careful consideration of various factors, such as:

### 6.1. Corpus Collection: Manual vs Web Scraping

While manual collection ensures high accuracy, reliability, and domain specificity of data, it is simultaneously labour-intensive, time-consuming, and expensive as it requires expertise to curate and validate. On the other side, web scraping is faster and cost-effective for gathering large datasets from agricultural blogs, government resources, research papers etc. While web scraping provides diverse perspectives, it requires data cleansing to remove noise and irrelevant information, compliance with legal and ethical requirements, and often can result in non-localized data (Zhou, L., et al, 2017) which is needed for agricultural advisory systems such as AgroTutor. Considering the need for localization as well as the resource intensiveness, we deployed a hybrid approach, wherein both manual collection and web scraping was done to build a representative and localised corpus.

### 6.2. Chunking Strategy: Fixed-size Chunking vs Semantic Chunking vs Smart Chunking

While fixed-size chunking i.e., dividing the corpus documents into fixed-size tokens or sentences results in simpler implementation and consistent memory usage during RAG retrieval, its key limitation is in splitting of contextually coherent information that reduces retrieval relevance. A better alternative is ‘semantic chunking’ that divides documents based on semantic coherence using NLP models. While the latter preserves contextual integrity of the information, it can be computationally expensive, especially for large-scale corpora. As AgroTutor is designed to provide localised agriculture advisories, smart chunking i.e., chunking within sections following PDF hierarchy, has been employed to improve answer relevance and user satisfaction.

### 6.3. Compendium Building

The limited availability of agricultural data in non-English languages and variability in dialects and regional languages usage for agricultural terminologies are the biggest obstacles in ensuring multilingual understanding and accuracy of the generated advisory (Conneau, A, et al., 2020). To solve this challenge, agriculture experts collected and listed agricultural terminologies in local dialects and built it into a ‘Key Compendium’ addendum of the AgroTutor platform.

### 6.4. Speech-to-Text translation for Voice-based Advisory

The key challenges in speech-to-text translation include the generation of accurate translations that requires robust models to handle agricultural terminologies and regional accents as well as real-time processing that needs low-latency translations for conversational systems (Baevski, A., et al, 2020). AgroTutor therefore using OpenAI Whisper technology which is a pre-trained speech model.

### 6.5. Cost-effectiveness per Advisory

Cost considerations in generating localised and accurate agricultural advisories for application in price-sensitive economies of the Global South is firstly the ‘LLM Size’ wherein smaller models on domain-specific data are more cost-effective for inference. Other considerations are leveraging of open-source models wherever possible to reduce licensing costs, retrieval optimization through efficient indexing or meta-tagging. AgroTutor leverages all the above considerations to generate advisories at \$0.018 per advisory, making it a practically deployable platform for farmers and other stakeholders in the Global South.

## 6.6. Evaluation Metrics

The system's performance was continuously refined through an iterative evaluation process for both the Retriever and Generator Components.

6.6.1. Retriever Component Evaluation: This is done through metrics such as precision@k and recall@k. 'Precision@k' measures the proportion of relevant items among the top k retrieved items. 'Recall@k' measures the proportion of relevant items retrieved within the top k results.

Setting 'k': We compute precision@k and recall@k we set k at two different values i.e., k=5 and k=10 to find the right setting at which the best retrieval takes place.

6.6.2. Generator Component Evaluation: is done through a customised human evaluation by agronomy experts, categorising the generated response into one of the following:

- a. Not relevant - When the response is wrong or is not in any way related to the question asked.
- b. Generic - When the response is related to the question asked, but the response is vague and impractical, i.e., not actionable without further search.
- c. Satisfactory - When the response is related to the question asked and reasonable, i.e., conclusive based on the chatbot's information, the user can take action. However, the request still needed further action, which was provided as the next steps in the feedback. (Examples include 'call this number for more information', 'I have no information about that; please try other sources', etc.)
- d. Excellent - When the response is related to the question asked and adequately justifies the question asked, i.e., the feedback provided directly addressed the question and was comprehensively actionable.

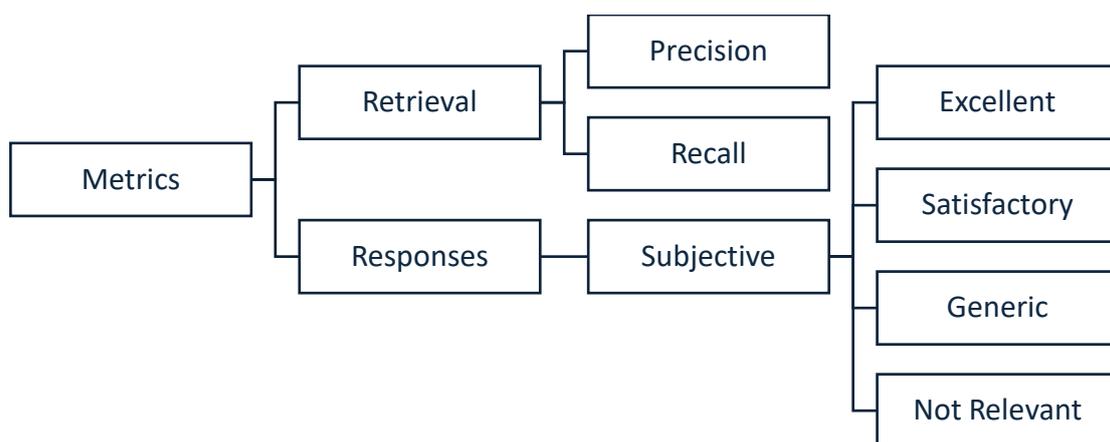


Figure 9: Evaluation Metrics

## 7. STAKEHOLDER ENGAGEMENT AND FUTURE DIRECTIONS

Looking ahead, AgroTutor aims to deepen its engagement with these stakeholders by fostering co-creation opportunities. Collaborative workshops, stakeholder-specific dashboards, and shared decision-making processes are envisioned to ensure the platform remains adaptive, inclusive, and

impactful. By aligning the needs and expertise of its stakeholders, AgroTutor sets a precedent for AI-driven solutions in agriculture that are both scalable and sustainable.

Most importantly, AgroTutor is built to be a low-cost platform for regional organisations to plug in their knowledge base through API integration for localised advisory provision to farmers in the global South. Our initial studies have shown that the cost of generating advisory using our customised platform (LLM and RAG integrated) is \$0.018 per advisory, which is very attractive in terms of cost to our stakeholders in the global South.

## 8. CONCLUSION

Localised generative AI systems, such as the AgroTutor platform outlined in this paper, offer a promising solution to addressing the specific challenges faced by smallholder farmers in the global south. These systems deliver actionable, timely, and context-specific advice by leveraging RAG frameworks, localised datasets, AI translation and voice integration. GenAI can enhance sustainability, improve climate resilience, and increase agricultural productivity in vulnerable regions.

Through our pilots in three different locations in the global South with limited and cost-efficient access to advanced AI systems with built-in capabilities to leverage LLMs and RAG, AgroTutor offers a plug-in platform wherein local organisations can directly deploy their APIs and make actual use of their localised knowledge base – thereby showcasing the transformative potential of generative AI in addressing localised agricultural challenges. By enabling precise, scalable, and accessible advisories, it empowers smallholder farmers to adopt sustainable practices, mitigating climate risks and enhancing productivity. As the initiative evolves, its contributions will be pivotal in shaping resilient and inclusive agrifood systems globally.

## REFERENCES

- Fountas, S., Espejo-Garcia, B., Kasimati, A., Gemtou, M., Panoutsopoulos, H., & Anastasiou, E. (2024). Agriculture 5.0: Cutting-Edge Technologies, Trends, and Challenges. *IT Professional*, 26(1), 40–47. <https://doi.org/10.1109/MITP.2024.3358972>
- Ghosh, R. C., Shailendra, P., & Singh, G. B. (2024). Large Language Model in Various Fields: Opportunities, Challenges and Risks. *Lecture Notes in Networks and Systems*, 1023 LNNS, 587–596. [https://doi.org/10.1007/978-981-97-3604-1\\_39](https://doi.org/10.1007/978-981-97-3604-1_39)
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70-90.
- Liakos, K. G., et al. (2018). AI and Agriculture: A Comprehensive Review. *Sensors*, 18(8), 2674.
- Liu, Y., Iyer, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment, May 2023. *arXiv preprint arXiv:2303.16634*, 6.
- 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*, 221. <https://doi.org/10.1016/j.compag.2024.108924>
- Nguyen, V., Karimi, S., Hallgren, W., Harkin, A., & Prakash, M. (2024). My Climate Advisor: An Application of NLP in Climate Adaptation for Agriculture. <https://www.elsevier.com/en-au/about>
- Sapkota, R., Qureshi, R., Hassan, S. Z., Shutske, J., Shoman, M., Sajjad, M., Dharejo, F. A., Paudel, A., Li, J., Meng, Z., Sadak, F., Hadi, M. U., & Karkee, M. (2024). Multi-Modal LLMs in Agriculture: A Comprehensive Review. <https://doi.org/10.36227/techrxiv.172651082.24507804/v1>
- Singh, N., Wang’ombe, J., Okanga, N., Zelenska, T., Repishti, J., K, J. G., Mishra, S., Manokaran, R., Singh, V., Rafiq, M. I., Gandhi, R., & Nambi, A. (2024). Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers. <http://arxiv.org/abs/2409.08916>
- Hoda, A., Gulati, A., Jose, S., & Rajkhowa, P. (2021). Sources and Drivers of Agricultural Growth in Bihar (pp. 211–246). [https://doi.org/10.1007/978-981-15-9335-2\\_8](https://doi.org/10.1007/978-981-15-9335-2_8)
- Marañón, Boris. (2006). Human Development Report 2006, Human Development Report Office, Tension Between Agricultural Growth and Sustainability: The El Bajío Case, Mexico.
- FAO. <https://www.fao.org/kenya/fao-in-kenya/kenya-at-a-glance/en/>
- Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350-361. DOI: 10.1016/j.neucom.2017.01.026
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *arXiv preprint*. DOI: 10.48550/arXiv.1911.02116

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449-12460. DOI: 10.48550/arXiv.2006.11477