

---

# RAG vs FINE-TUNING: PIPELINES, TRADEOFFS, AND A CASE STUDY ON AGRICULTURE

---

## Microsoft

Angels Balaguer, Vinamra Benara, Renato Cunha, Roberto Estevão, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, Ranveer Chandra

## ABSTRACT

There are two common ways in which developers are incorporating proprietary and domain-specific data when building applications of Large Language Models (LLMs): Retrieval-Augmented Generation (RAG) and Fine-Tuning. RAG augments the prompt with the external data, while fine-Tuning incorporates the additional knowledge into the model itself. However, the pros and cons of both approaches are not well understood. In this paper, we propose a pipeline for fine-tuning and RAG, and present the tradeoffs of both for multiple popular LLMs, including Llama2-13B, GPT-3.5, and GPT-4. Our pipeline consists of multiple stages, including extracting information from PDFs, generating questions and answers, using them for fine-tuning, and leveraging GPT-4 for evaluating the results. We propose metrics to assess the performance of different stages of the RAG and fine-Tuning pipeline. We conduct an in-depth study on an agricultural dataset. Agriculture as an industry has not seen much penetration of AI, and we study a potentially disruptive application - what if we could provide location-specific insights to a farmer? Our results show the effectiveness of our dataset generation pipeline in capturing geographic-specific knowledge, and the quantitative and qualitative benefits of RAG and fine-tuning. We see an accuracy increase of over 6 p.p. when fine-tuning the model and this is cumulative with RAG, which increases accuracy by 5 p.p. further. In one particular experiment, we also demonstrate that the fine-tuned model leverages information from across geographies to answer specific questions, increasing answer similarity from 47% to 72%. Overall, the results point to how systems built using LLMs can be adapted to respond and incorporate knowledge across a dimension that is critical for a specific industry, paving the way for further applications of LLMs in other industrial domains.

**Keywords** GPT-4 · Agriculture · Retrieval Augmented Generation · Fine-tuning

## 1 Introduction

Over the past few years, artificial intelligence and natural language processing have seen significant advancements, leading to the development of powerful large language models (LLMs) such as the Generative Pre-trained Transformer (GPT). The technology driving LLMs, including advanced deep learning techniques, large-scale transformers, and vast amounts of data, have propelled their rapid evolution. Models like GPT-4 (OpenAI, 2023) and Llama 2 (Touvron et al., 2023b) have demonstrated exceptional performance across numerous tasks and domains, often without specific prompts. These models surpass their predecessors and hold immense potential in various fields like coding, medicine, law, agriculture, and psychology, closely approaching human-level expertise (Bubeck et al., 2023; Nori et al., 2023; Demszky et al., 2023). As LLM research continues, it is critical to identify their limitations and address the challenges of developing more comprehensive artificial general intelligence (AGI) systems. Moreover, the machine learning community must move beyond traditional benchmarking datasets and evaluate LLMs in ways that closely resemble human cognitive ability assessments.

The adoption of Artificial Intelligence (AI) copilots across various industries is revolutionizing the way businesses operate and interact with their environment. These AI copilots, powered by LLMs, provide invaluable assistance in

data processing and decision-making processes. In healthcare, for example, AI copilots are being leveraged to predict patient risks and improve diagnostic accuracy (Kim et al., 2023; Thirunavukarasu et al., 2023; Alowais et al., 2023). In manufacturing, they aid in enhancing operational efficiency, reducing downtime, and improving product quality (Vanti, 2023; Li et al., 2023). In the realm of finance, AI copilots help in fraud detection, risk management, and investment decision-making (AI4Finance-Foundation, 2022; Solutions, 2022). By harnessing the power of AI copilots, industries can drive innovation, optimize performance and gain a competitive edge.

Despite these advancements, the application of AI in specific fields such as agriculture is still limited due to a lack of specialized training data. While AI has been used to derive insights from satellite imagery and sensor data in agriculture (Vasisht et al., 2017; Chandra et al., 2022; Microsoft, 2021; Zhao et al., 2023; Kumar et al., 2021; Sharma et al., 2023), the technology is still slowly being adopted by farmers. While GPT-4 and Bing are powerful tools for finding information, they may not provide the best solutions for farmers who have very specific questions about their crops and livestock. These questions often require knowledge of local conditions, specific varieties, and up-to-date data that might not be readily available through general search engines (Silva et al., 2023). As an example, Table 1 compares the answers from GPT-4 and an agronomist expert to the same query asked for three different U.S. states. While an expert would provide contextualized answers grounded on the states specific climate and agriculture tradition, LLMs provide a generic answer that, although correct, is not as precise for each state as the expert answer.

In this paper, we introduce a new focus: the creation of AI copilots for industries that require specific contexts and adaptive responses, such as the agriculture industry. We propose a comprehensive LLM pipeline to generate high-quality, industry-specific questions and answers. This approach involves a systematic process comprising the identification and collection of relevant documents encompassing a wide range of agricultural topics. These documents are then cleaned and structured to facilitate the generation of meaningful Q&A pairs using the base GPT model. The generated pairs are subsequently evaluated and filtered based on their quality. Our goal is to create a valuable knowledge resource for an specific industry, with a case study in agriculture with the goal of ultimately contributing to the advancement of this crucial field.

The proposed pipeline aims to generate domain-specific questions and answers catering to professionals and stakeholders in an industry where answers from a copilot are expected to be grounded by relevant industry-specific factors. In the case of our agriculture study, we are aiming to generate geography-specific answers. For this, our starting point is an agriculture dataset, which is fed into three main components: Q&A generation, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), and fine-tuning process. The Q&A generation creates question and answer pairs based on the information available in the agriculture dataset, while RAG uses it as a knowledge source. The generated data is then refined and used to fine-tune several models while their quality is evaluated using a combination of proposed metrics. Through this comprehensive approach, we aim to harness the power of LLMs for the benefit of the agriculture industry and its stakeholders.

In this paper, we make several noteworthy contributions to the understanding of LLMs in the agriculture domain. These contributions can be itemized as follows:

- **Comprehensive evaluation of LLMs:** we conducted an extensive evaluation of large language models, including LLaMa2-13B, GPT-4 and Vicuna (Zheng et al., 2023), in answering agriculture-related questions. This was done using benchmark datasets from major agriculture producer countries. Our evaluation included the complete fine-tuning and RAG pipeline, each with its own set of metrics. The findings from this evaluation provide a crucial baseline understanding of the performance of these models within the agricultural context. Moreover, we conduct evaluations that demonstrate the impact of spatial shift on the knowledge encoded by existing LMs and the improvements offered by spatially-scoped fine-tuning. In our analysis, GPT-4 consistently outperformed other models, however the cost associated with its fine-tuning and inference needs to be taken into consideration.
- **Impact of retrieval techniques and fine-tuning:** we investigated the impact of retrieval techniques and fine-tuning on the performance of LLMs. The study revealed that both RAG and fine-tuning are effective techniques for improving the performance of LLMs. RAG was shown to be highly effective in instances where data is contextually relevant, such as in the interpretation of farm data, while also leading to more succinct responses than the base model. Fine-tuning, on the other hand, was found to be useful in teaching the model new skills specific to the agricultural domain, and providing more precise and succinct responses. However, the high initial cost due to the extensive work required to fine-tune the model on new data is a significant consideration.
- **Implications for potential uses of LLMs in different industries:** this research serves as a pioneering step in establishing a pipeline for the application of RAG and fine-tuning techniques in LLMs, fostering innovation and collaboration across multiple industries. With our initial focus on agriculture, we demonstrated how these strategies can lead to more efficient models starting from the Q&A generation process. The insights gained

State	What is the best times to plant trees and shrubs in _____ ?
<i>GPT-4 answers</i>	
Arkansas	The best time to plant trees and shrubs in Arkansas is during the <b>fall and early spring seasons</b> . Planting in the fall (October to November) allows the roots to establish themselves before the hot summer temperatures arrive. Early spring (February to March) is also a good time to plant, as it gives the plants a chance to grow and establish themselves before the summer heat sets in.
Connecticut	The best time to plant trees and shrubs in Connecticut is during the <b>early spring and fall seasons</b> . Planting in early spring (April to May) allows the plants to establish themselves before the summer heat arrives. Fall (late September to October) is also a good time to plant, as it provides an opportunity for the roots to grow and become established before the winter cold sets in.
Georgia	The best time to plant trees and shrubs in Georgia is during the <b>fall and early spring seasons</b> . Planting in the fall (October to November) allows the roots to establish themselves before the hot summer temperatures arrive. Early spring (February to March) is also a good time to plant, as it gives the plants a chance to grow and establish themselves before the summer heat sets in.
<i>Bing Chat answers</i>	
Arkansas	The best time to plant trees and shrubs in Arkansas is generally in the <b>fall</b> . More specifically, September through October is often recommended. This allows enough time for the plants to establish strong roots before the ground starts to freeze. However, <b>spring</b> can also be a good time to plant when soil temperatures are mild.
Connecticut	The best time of the year to plant most trees and shrubs is in the <b>fall</b> . The cooler temperatures are easier on the trees and shrubs than the harsh summer heat, and the autumn rain helps them establish a stronger root system.
Georgia	The best time to plant trees and shrubs in Georgia is during the <b>fall and winter months</b> . Trees planted in November and December have time to produce new roots before having to supply water to lots of thirsty leaves next spring
<i>Expert's answers</i>	
Arkansas	The best times to plant trees and shrubs are in the <b>spring and fall</b> . Spring planting is recommended because the soil is typically moist and warm, allowing for easy digging and root growth.
Connecticut	The best time to plant shrubs in shady areas is during <b>late winter or early spring</b> in well-prepared soil.
Georgia	The best time to plant trees and shrubs is in the <b>fall, specifically during the dormant season</b> , which is when the plant has just entered dormancy or is in full dormancy.

Table 1: Comparison of GPT-4, Bing Chat, and expert’s answers for an example query (“What is the best times to plant trees and shrubs?”), considering three locations (*Arkansas*, *Connecticut*, and *Georgia*). GPT-4 is not able to incorporate location-specific knowledge, providing the same answer in all cases without considering geographical particularities. Enhanced by a search engine, Bing Chat does a better job adapting the answer to each location, but the expert’s answers are still more precise.

from this study could be applied to other sectors, potentially leading to the development of more efficient AI models for a variety of applications. For instance, one potential application could be in the development of AI copilots for different industries, where the ability to provide accurate, relevant and succinct responses to user queries is paramount.

The remainder of this work is organized as follows. Section 2 presents the methodology in detail, including the data acquisition process, information extraction procedure, question and answer generation, and fine-tuning of the model. We then describe in Section 3 the dataset used in the study, which includes data from the USA, Brazil, and India. In Section 4, we outline the metrics used to evaluate the effectiveness of the proposed methodology, focusing on both question and answer evaluation. Section 5 presents a comprehensive evaluation of various models and their performance in generating question-answer pairs within the context of agricultural data, using these for RAG on GPT-4, Vicuna (Zheng et al., 2023) and Llama2 13B (Touvron et al., 2023a,b) and fine-tuning of GPT-4 and Llama2 13B. Finally, we conclude with a summary of the main findings and possible directions for future research in Section 6.

## 2 Methodology

The methodology proposed in this paper revolves around a pipeline designed to generate and evaluate question-answer pairs for building domain-specific copilots. The proposed pipeline is shown in Figure 1.

The pipeline begins with data acquisition, described in Section 2.1. The initial focus is on gathering a diverse and curated dataset pertinent to the industry domain. This includes sourcing data from various high-quality repositories such as government agencies, scientific knowledge databases, and proprietary data, if needed. The details of potential data sources and the types of documents selected are exemplified and further elaborated in Section 3.

Following data acquisition, the pipeline proceeds to extract information from the collected documents. This step is crucial as it involves parsing complex and unstructured PDF files to recover the underlying content and structure. This process, detailed further in Section 2.2, employs robust text extraction tools and machine learning algorithms to recover textual, tabular, and visual information, while also identifying the semantic structure of the documents and possible cross-relations in them.

The next component of the pipeline is question and answer generation. The objective here is to generate contextually grounded and high-quality questions that accurately reflect the content of the extracted text. The methodology employs a framework to control the structural composition of both inputs and outputs, thereby augmenting the overall efficacy of response generation from language models. This part of the process is detailed in Section 2.3.

Subsequently, the pipeline generates answers for the formulated questions. The methodology employed here leverages Retrieval-Augmented Generation (RAG), which combines the power of retrieval and generation mechanisms to create high-quality answers. The answer generation process is discussed further in Section 2.4.

Finally, the pipeline fine-tunes the models with the Q&A pairs. The optimization process, discussed in the Section 2.5, employs methods like Low Rank Adaptation (LoRA) (Hu et al., 2021) and ensures a comprehensive understanding of the content and context of the scientific literature, making it a valuable resource for various domains or industries.

In the following sections, we will delve deeper in each components of the pipeline, highlighting their objectives, inputs and outputs, and reasoning behind why they were added to the pipeline.

### 2.1 Data Acquisition

The initial focus of the pipeline is to gather an assorted and well-curated dataset that captures information of interest to an industry. This enables the generation of questions and answers, forming the foundation for refining the models to produce more precise and pertinent responses. For this step, we seek data sources that contain high-quality, authoritative information on the topic of interest. For example, in agriculture, this includes agricultural and environmental government agencies, scientific knowledge repositories, and agronomist exams databases. It is also important that the information extracted to be aligned with the grounding that will be provided to the model. For instance, in the case of agricultural data, we sourced guidelines and procedures that were geography-specific, i.e. with a shared location among documents. We provide further details on the sources and type of documents selected for this step in Section 3.

With authoritative sources defined, web scraping tools come into play to gather the required data. We employed web scraping frameworks, including Scrapy (Zyte, 2023) and BeautifulSoup (BeautifulSoup, 2023), to parse through the websites, uncovering all available documents and downloading the relevant files.

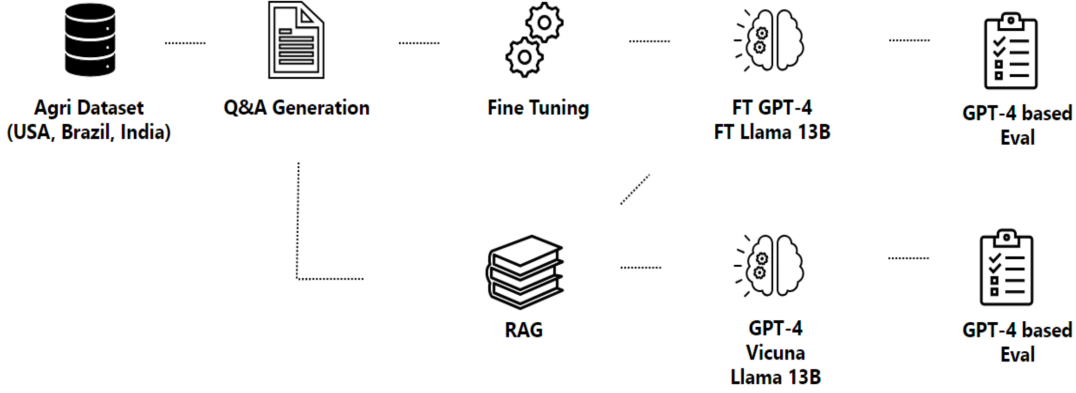


Figure 1: Methodology pipeline. Domain-specific datasets are collected, and the content and structure of the documents are extracted. This information is then fed to the Q&A generation step. Synthesized question-answer pairs are used to fine-tune the LLMs. Models are evaluated with and without RAG under different GPT-4-based metrics.

## 2.2 PDF Information Extraction

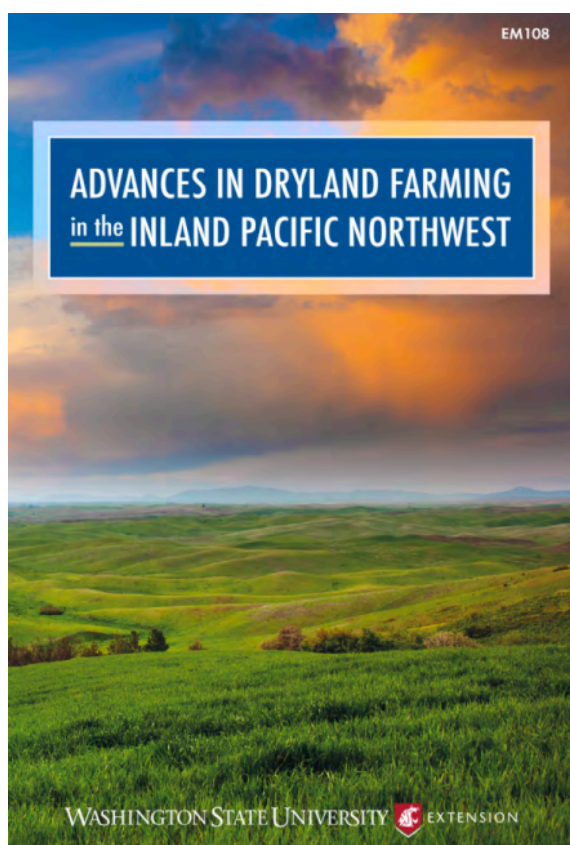
In our study, the extraction of information and text structure from the collected documents is critical to the quality of the subsequent steps. However, this is a challenging task as the primary purpose of PDFs is to accurately display a document across different systems, and not for easy information extraction. The underlying structure of a PDF file does not map onto the logical structure of a document, i.e., sections, subsection, and associated content. Additionally, with documents originating from various sources, we observe their layouts and formatting to be complex and lack standardization, often presenting a mixture of tables, images, sidebars, and page footers. We present in Figure 2 an example of a PDF file in our dataset.

With this in mind, the main objective of this step of the pipeline is to address the complexities inherent in processing data derived from a range of formatted PDF documents. This is achieved by leveraging robust text extraction tools and machine learning algorithms that employ advanced natural language processing techniques. The focus is not only in recovering the content of each file, but also its structure. Among other things, we are interested in discovering what are the sections and subsections, parsing the information presented in tables and diagrams, identifying cross-references within the document, and linking images with their caption and description. By retrieving the organization of the document, we can easily group information, reason over numerical data present in tables, and provide more consistent snippets of the text to the Q&A generation step. It is also very important that all available information is extracted from the document, with well-formed sentences.

There are multiple tools available online that extract information from PDFs (PDF2Text, 2023; PyPDF, 2023). However, many of them lack the ability to retrieve content in a structured way. For example, *pdf2text* is an open-source Python library offering methods to iterate over PDF’s pages and recover the textual information. We provide in Listing 1 the output of *pdf2text* over the document from Figure 2. The library is able to recover the textual information, but markers representing the beginning of a section or subsection are lost within the retrieved data, hindering our ability to reason over the document structure. Captions of tables and figures are also lost in conversion but sometimes contain critical information for the understanding of the document.

Considering this, we employed GROBID (GeneRation Of Bibliographic Data) (GRO, 2008–2023), a machine learning library specifically tailored for extracting and processing data from scientific literature in PDF format. The goal is to transform unstructured PDF data into structured data in the form of TEI (Text Encoding Initiative) format (Consortium, 2023), efficiently managing large volumes of files. The use of GROBID, trained on a vast corpus of scientific articles, enables the recognition of a wide array of document elements and extraction of associated bibliographic data. We illustrate its capabilities in Listing 2 with the output of GROBID for the document from Figure 2.

From the GROBID-generated TEI files, we extracted a subset of the sections of the TEI files comprising the document metadata (title, authors, abstract), sections, tables, figure references, bibliography, and the content itself. Crucially, this phase underscores the belief that the structure of the text is as important as its content. The final objective is to convert the TEI files into more manageable JSON files that preserve not only the content, but also the structure of the original



Advances in Dryland Farming in the Inland Pacific Northwest	
Introduction .....	1
<i>Georgine Yorgey, Chad Kruger</i>	
Acknowledgements .....	13
Chapter 1: Climate Considerations .....	15
<i>Chad Kruger, Elizabeth Allen, John Abatzoglou, Kirti Rajagopalan, Elizabeth Kirby</i>	
Chapter 2: Soil Health .....	47
<i>Rakesh Awale, Stephen Machado, Rajan Ghimire, Prakriti Bista</i>	
Chapter 3: Conservation Tillage Systems.....	99
<i>Prakriti Bista, Stephen Machado, Rajan Ghimire, Georgine Yorgey, Donald Wysocki</i>	
Chapter 4: Crop Residue Management.....	125
<i>Haiying Tao, Georgine Yorgey, David Huggins, Donald Wysocki</i>	
Chapter 5: Rotational Diversification and Intensification.....	163
<i>Elizabeth Kirby, William Pan, David Huggins, Kathleen Painter, Prakriti Bista</i>	
Chapter 6: Soil Fertility Management .....	237
<i>Kristy Borrelli, Tai Maaz, William Pan, Paul Carter, Haiying Tao</i>	
Chapter 7: Soil Amendments.....	283
<i>Georgine Yorgey, William Pan, Rakesh Awale, Stephen Machado, Andy Bary</i>	
Chapter 8: Precision Agriculture .....	319
<i>Bertie Weddell, Tabitha Brown, Kristy Borrelli</i>	
Chapter 9: Integrated Weed Managment.....	353
<i>Ian Burke, Kendall Kahl, Nicole Tautges, Frank Young</i>	

## Introduction

*Georgine Yorgey, Washington State University  
Chad Kruger, Washington State University*

The Pacific Northwest is an important wheat production region. In 2015, the National Agricultural Statistics Service indicated that Washington, Idaho, and Oregon harvested more than 240 million bushels of wheat, worth an estimated \$1.3 billion. The major areas of production in the inland Pacific Northwest are shown below, and incorporate both irrigated and dryland acreage.



The Columbia Plateau ecoregion, commonly referred to by growers as the inland Pacific Northwest.

### Advances in Dryland Farming in the Inland Pacific Northwest

The area includes three major land resource areas with distinctive geologic features and soils as defined by the US Department of Agriculture: the Columbia Basin, the Columbia Plateau, and the Palouse and Nez Perce Prairies, all of which are within the Northwestern Wheat and Range Region. It also includes a small portion of dryland cropping in the North Rocky Mountains major land resource area, adjacent to the eastern edge of the Palouse and Nez Perce Prairies. In the dryland areas, which are the focus of this book, wheat is grown in rotation with crop fallow and much smaller acreages of other small grains, legumes, and alternative crops.

This area, identified from here forward by the more familiar term "inland Pacific Northwest," encompasses great diversity, characterized by some common overarching patterns of climate, geography, and agriculture. The inland Pacific Northwest extends eastward from the Cascade Mountain Range in Washington and Oregon into parts of northern Idaho. The landscape includes glacial deposits, coulees, channeled scablands, and rolling terrain with deep, fertile soil. The climate is semi-arid, with cool, wet winters and hot, dry summers.

Across the dryland wheat production areas, there are three major agroecological classes (AECs), with different patterns of cropping:

- Grain-Fallow AEC (defined as areas with greater than 40% fallow)
- Annual Crop-Fallow Transition AEC (with 10–40% fallow)
- Annual Crop AEC (with less than 10% fallow)

There is a considerable precipitation gradient across the region, with drier conditions in the rain shadow immediately east of the Cascades and wetter conditions further inland. The Grain-Fallow AEC is associated with lower precipitation areas, while the Annual Crop AEC is generally, but not always, associated with areas that receive higher levels of precipitation. These AECs, further described in Chapter 1, are dynamic and change as land use and land cover shift over time—with the potential to be influenced by climate, soils, terrain, land and commodity prices, and other factors. Because many recommendations in this book are specific to a farm's AEC characteristics, research results have been coded accordingly: Grain-Fallow ■; Annual Crop-Fallow Transition ▲; and Annual Crop ●. On the first page of each chapter, we have included a legend to help readers easily identify these symbols and the information most pertinent to their AEC.

Figure 2: Example of a document from Washington state present in our dataset. The diverse layouts of PDF files, which often include textual and visual data, pose a significant challenge in terms of extracting not just the content, but also the underlying structure.

Advances in Dryland  
 Farming in the Inland  
 Pacific Northwest  
 Georgine Yorgey and Chad Kruger, editorsFor Sanford Eigenbrode, in recognition of his resolute  
 effort  
 leading the REACCH project.  
 For Iris, Toby, Leah, Jocelyn, Alexis, and Zakkary, who symbolize  
 why this work is so important.College of Agricultural, Human, and Natural Resource Sciences  
 Use pesticides with care. Apply them only to plants, animals, or sites as listed on the label.  
 When  
 mixing and applying pesticides, follow all label precautions to protect yourself and others  
 around you. It is a violation of the law to disregard label directions. If pesticides are  
 spilled on skin or clothing, remove clothing and wash skin thoroughly. Store pesticides in  
 their original containers and keep them out of the reach of children, pets, and livestock.

2017 Washington State University WSU Extension bulletins contain material written and produced  
 for public distribution. Alternate  
 formats of our educational materials are available upon request for persons with disabilities.  
 Please contact Washington State University Extension for more information.  
 You may order copies of this and other publications from WSU Extension at 1-800-723-1763 or  
<http://pubs.wsu.edu>.

...

5871Introduction  
 Georgine Yorgey, Washington State University  
 Chad Kruger, Washington State University  
 The Pacific Northwest is an important wheat production region. In 2015, the  
 National Agricultural Statistics Service indicated that Washington, Idaho,  
 and Oregon harvested more than 240 million bushels of wheat, worth an estimated 1.3 billion.  
 The major areas of production in the inland

Listing 1: Text extracted from the PDF presented in Figure 2 using PDF2text. Structural markers, such as the “Introduction” section header, end up lost within the parsed text.

PDFs. This approach ensures a comprehensive understanding of the content and context of the scientific literature, making it a valuable resource for various domains or industries.

## 2.3 Question Generation

The initial focus of this section of the pipeline is to manage the inherent complexity and variability of natural language when generating questions from the extracted text. We aim to generate contextually grounded and high-quality questions that accurately reflect the content of the extracted text. For this, we employ the Guidance framework (Gui, 2023), whose primary advantage lies in its capacity to provide unparalleled control over the structural composition of both inputs and outputs, thereby augmenting the overall efficacy of response generation from language models. This degree of control results in outputs that are not only more precise, but also exhibit enhanced coherence and contextual relevance. The framework’s capability to amalgamate generation, prompting, and logical control into a singular, unified process closely parallels the inherent mechanisms of language model text processing. Moreover, the unique feature of Guidance that enables the direction of language models via context-specific prompts, contributes to a heightened level of semantic relevance in the resultant text. In our case, this ensures the questions will carry semantic relevance to the source text while taking into account the context extracted from the JSON files.

First, we augment the content and structure of available documents by explicitly adding supporting tags from the text. We formulated prompts to extract a list of locations and agronomic topics mentioned in each section of the document (e.g., if that section refers to crops, cattle, or diseases), as exemplified in Listing 3, and task the LLM model to answering them based on the data extracted from the JSON files. The aim is to use of the the additional information, including locations and mentioned topics, to ground the generation process, enhancing the relevance of the questions and guiding the model to cover a wide range of topics and challenges.

```

{
  "grobid_version": "0.7.3",
  "grobid_timestamp": "2023-07-04T13:05+0000",
  "language_code": "en",
  "citations": [
    {
      "authors": [
        {
          "given_name": "J",
          "surname": "Abatzoglou",
          "name": "J T Abatzoglou"
        },
        {
          "given_name": "T",
          "surname": "Brown",
          "name": "T J Brown"
        }
      ],
      "index": 0,
      "id": "b0",
      "date": "2012",
      "title": "A Comparison of Statistical Downscaling Methods Suited for Wildfire
        Applications",
      "journal": "International Journal of Climatology",
      "volume": "32",
      "issue": "5",
      "pages": "772-780"
    },
    ...
    {
      "title": "Introduction",
      "content": [
        "Climate has always had a dominant influence on dryland production in this
          region, shaping crop choices, agronomic management systems, and
          conservation efforts. Though farmers are already highly skilled managers
          in the context of variable temperature and precipitation patterns,
          climate change is expected to add uncertainty, stretching the limits of
          existing management systems. Projected climate change also brings new
          urgency to questions of sustainability, as changing seasonal climate
          patterns may exacerbate conditions that have been historically linked to
          major soil erosion events in the region.",
        ],
      "refs": []...
    }
  ]
}

```

Listing 2: JSON file extracted from the PDF of Figure 2. We use GROBID (GRO, 2008–2023), which can extract the content and some of the structure (e.g., section and subsection organization) from the document.