

Sameeksha Desai, Sriman Donthireddi, Jay Vatti, Jason Chan, Lakshya Chaudhry, Khoa Nguyen, Nick Loyd, Nandika Yadav



## INTRODUCTION

Labels and Safety Data Sheets (SDSs) contain vital information about a crop product's composition, usage, and safety protocols. However, differences in formatting and structure layout make automated data extraction from these document difficult. Building cost-effective and efficient tools to handle this nonuniformity is a key challenge - but also a major opportunity for innovation.

## METHODOLOGY

#### **Extraction Research & Tools**

#### Goal?

- Reliable data extraction is critical for building accurate databases and chatbot responses
- Standard methods often fail due to scanned PDFs and inconsistent layouts

#### Our approach:

- Evaluated tools: PyMuPDF, Marker, and Marker OCR (open-source), and Azure Document Intelligence (closed-source)
- Tested on a diverse sample of PDF pages
- Assessed the result base on text quality, layou t preservation, and relevance

#### **Retrieval-augmented generation (RAG)**

- Initially explored a Graph RAG approach but ran into scalability and complexity issues.
- Developed a custom hybrid RAG combining vector-based retrieval with graph-based context

### **Final Result:**

• Improvement in response accuracy by leveraging both semantic similarity and document relationships

### **Final Product Development**

- Started with Svelte for its simplicity and performance
- Switched to React due to better documentation, community support, and library integration
- React enabled faster development and improved scalability of the UI

Bactericide for use in grapevines  Biologically-based reduction of Pierce's Disease in grapevines    For Use in Organic Production  Image: Constant of Constant of Pierce's Disease in grapevines    Active Ingredient:  Bacteriophage active against Xylella fastidiosa *0.00013 %    Other Ingredients:  99.99987 %    Total:  Active Ingredient    *Contains a minimum of 5 x 10° plaque forming units (PFU)  Pesticide Disparation active industry).	inate water, food or feed by storage ige: Store in original container. Store oom temperature or under refrigera , away from direct sunlight and extre ause degradation of the product and cy. osal: To avoid waste, use all materia cording to label directions. If wastes product to a waste disposal facility n such programs are run by state or	e this p ited co eme he d that v l in this canno
Biologically-based reduction of Pierce's Disease in grapevines  area at either reduction    For Use in Organic Production  Storage may carried in the product of the produc	oom temperature or under refrigera , away from direct sunlight and extre ause degradation of the product and cy. osal: To avoid waste, use all materia cording to label directions. If wastes	ited co eme he d that v in this canno
Active Ingredient: Bacteriophage active against <i>Xylella fastidiosa</i> *0.00013 % Other Ingredients: Total: *Contains a minimum of 5 x 10° plaque forming units (PFU) per milliliter of product	osal: To avoid waste, use all materia cording to label directions. If wastes	canno
Other Ingredients:		
*Contains a minimum of 5 x 10 <sup>9</sup> plaque forming units (PFU) per milliliter of product by industry).	Cleanout Procedure	or pe
per milliliter of product		locol a
Container Han	dling: Nonrefillable container. Do n	
	r for recycling if available or dispos	
See insert for Precautionary Statements and Directions for Use or by other pro	ocedures approved by state and loc	al auth
HOTLINE NUMBER For general information on this product, call 1-650-337-0350	92918-1-93909 Net Conten 9441-902-97 Batch No.:	
during normal business hours, 8am to 5pm (Pacific Time), Sold by: A&P I		
Contact 1-800-222-1222 for emergency medical treatment One San F	Finharcadero Center Suite 2020 Manufacturer and address rancisco, CA 94111	

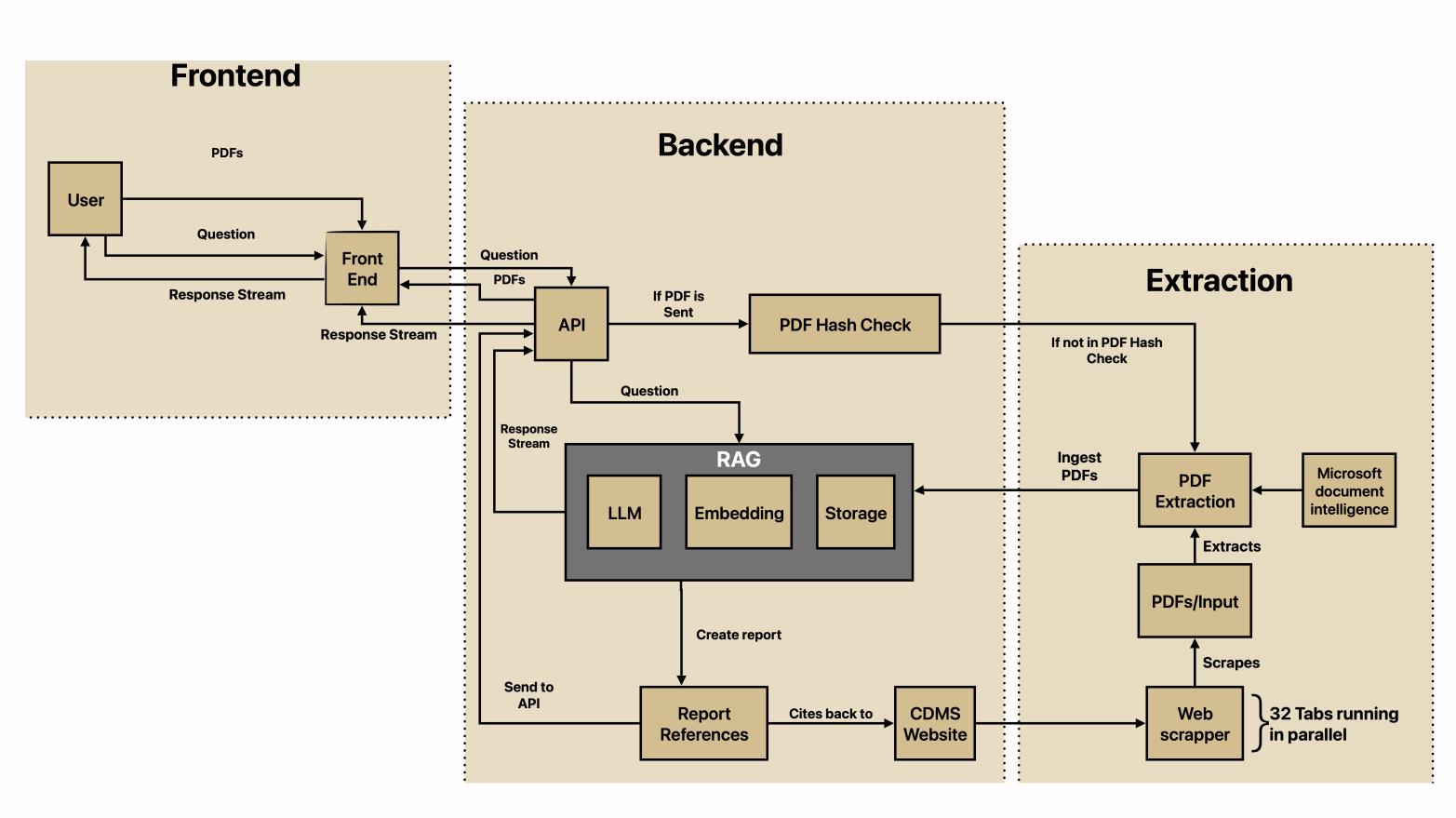
# AI ANALYSIS OF SDS'S & PRODUCT LABELS

## OBJECTIVE

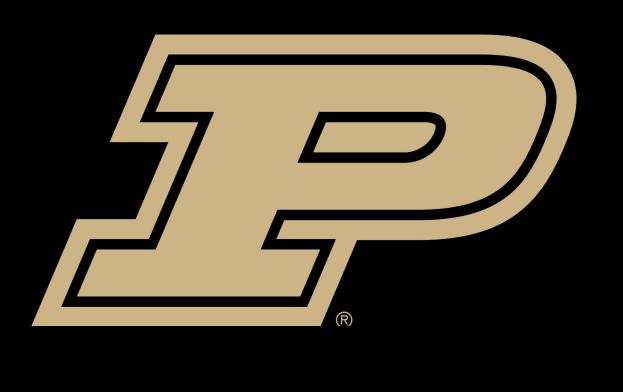
- A web scraper to collect Labels and SDSs from multiple sources • A structured database to support analytics, visualizations, and reporting • A chatbot that enables users to query the data and receive accurate,
- real-time answers

## PROJECT ARCHITECTURE

- Initially explored a Graph RAG approach but ran into scalability and complexity issues • Developed a custom hybrid RAG combining vector-based retrieval with graph-based context • This improved response accuracy by leveraging both semantic similarity and document
- relationships



sposal. product in a dry pnditions (36 - eat. Inappropriate would decrease	<b>Closed Sourced</b>		Open Sourced	
s container by ot be avoided, esticide disposal governments or se or refill this a sanitary landfill	Azure	Marker	MarkerOCR	PyMuPDF
A&P horities. AA horities. AA horities. AA horities. AA horities. A horities. A horities. A horities. A horities. A horities. horitie	97.2%	70.8%	47.2%	50.9%



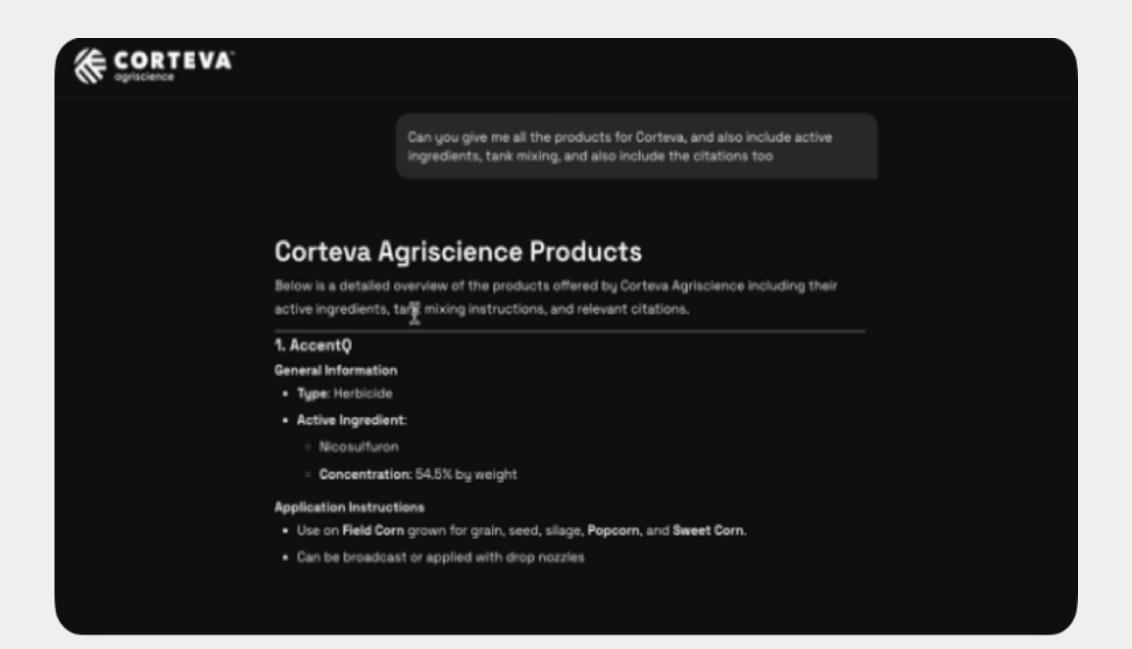
## **RELATED LITERATURE**

## TEST / DEVELOPMENT RESULTS

### Features

- Advanced Text Embedding: Converts extracted content into a custom RAG vector database for in-depth question-answering
- Structured Data Transformation: Transforms unstructured text into a structured format optimized for fast database querying
- Chatbot Interface: intuitive Q&A and on-demand insights
- Automated Research Report Generation: Generates comprehensive research reports with full citations back to original documents

We developed a chatbot that answers questions using a large collection of product labels and safety data sheets (SDS). Instead of manually searching through hundreds of PDFs, users can now get accurate answers within minutes. Using natural language processing, the chatbot efficiently retrieves relevant information, saving time and reducing the risk of missing critical details.



## ACKNOWLEDGEMENTS

Thanks to Data Mine staff, Bryce Castle and Maggie Betz, and our mentors Jason Sun, Sam Zukowski, Brandon Downer, Grant Kippenbrock, and Ryan Owen for their guidance and support throughout this year.

• Sharma, K., Kumar, P., & Li, Y. (2024). OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models. arXiv preprint arXiv:2412.15235. https://doi.org/10.48550/arXiv.2412.15235

## FINAL PRODUCT

• Bulk & Individual PDF Processing: Extracts information from web-scraped PDFs (e.g., CDMS website) either in bulk or per user request

• Developed a **chatbot** to answer questions using product **Labels** and **SDSs** • Uses natural language processing (NLP) to retrieve relevant information quickly • Eliminates the need to mannually search through hundreds of PDFs • Saves time and reduces the risk of missing critical details