DEEP CROP

AI ANALYSIS OF SDS'S & PRODUCT LABELS

AUTHORS

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

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.

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 non-uniformity is a key challenge - but also a major opportunity for innovation.

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

RELATED LITERATURE

- 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

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)

- **I**nitially 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:

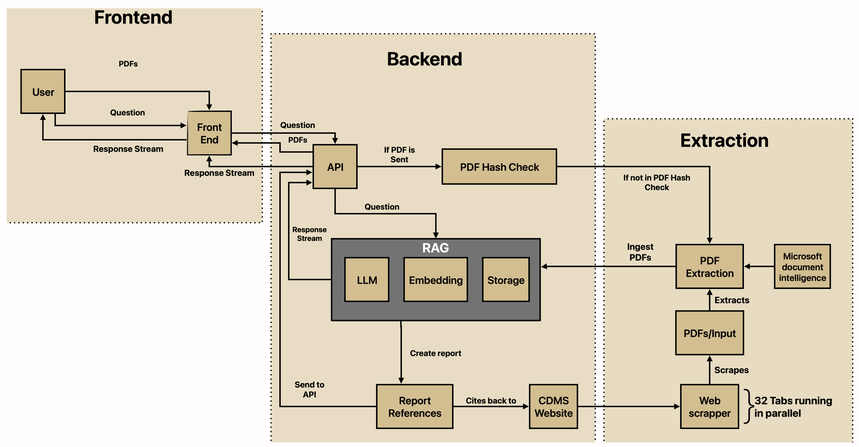
- **I**mprovement in response accuracy by leveraging both semantic similarity and document relationships

Final Product Development

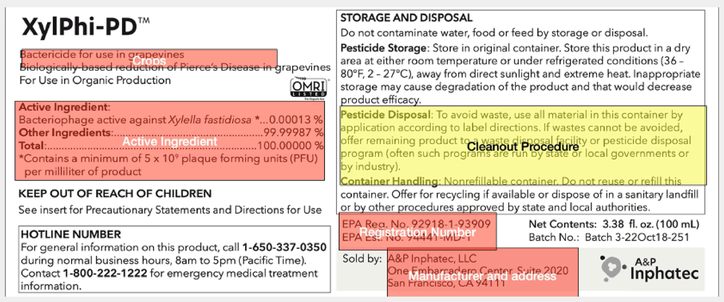
- Started with Svelte for its simplicity and performance
- Switched to React due to better documentation, community support, and library integration
- React enabled faster development and improved scalability of the UI

PROJECT ARCHITECTURE

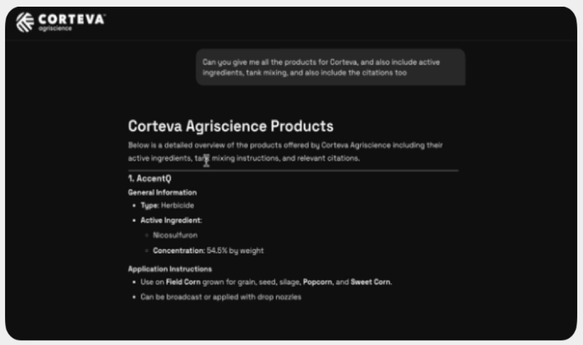
- Initially explored a Graph RAG approach but ran into scalability and complexity issues
- Developed a custom hybrid RAG combining vector-based retrieval with graph-based context
- This improved response accuracy by leveraging both semantic similarity and document relationships



TEST / DEVELOPMENT RESULTS



Closed Sourced	Open Sourced		
Azure	Marker	MarkerOCR	PyMuPDF
97.2%	70.8%	47.2%	50.9%



FINAL PRODUCT

Features

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

CONCLUSION

- 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