AGRICULTURAL DATA AND AI AT THE USDA

US Department of Agriculture | Local and Regional Foods Division Research Branch

Ashley Chu, Data Fellow | *UC San Diego, Cognitive Science & Machine Learning*

Galiba Anjum, Data Fellow | *Cornell University, Computer Science & Information Science*

Keywords:

Financial analytics, automated reporting, grants awards dashboards, data cleaning, visualization

Summary:

To periodically report grant spending, Ashley designed and developed dashboards and visualizations summarizing Grants Division data in **Tableau** and **Python**. They wrote **SQL** queries and Python code to pull the data from the database in the Enterprise Data Analytics Platform and then parse, clean and analyze it for reporting. The work culminated in an automated report in the form of a **Quarto** notebook to continuously explain more detailed views of the obligations, disbursements, and trends over time.

AGRICULTURAL DATA AND AI AT THE USDA

US Department of Agriculture | Local and Regional Foods Division Research Branch

Ashley Chu, Data Fellow | *UC San Diego, Cognitive Science & Machine Learning*

Galiba Anjum, Data Fellow | *Cornell University, Computer Science & Information Science*

Keywords:

Data cleaning, Python, geographic blocking, data standardization, record linkage

Summary: Galiba standardized and consolidated over 200,000 Local Food Purchase Assistance program records submitted by all U.S. states, territories, and tribal partners through quarterly spreadsheets. Using Python libraries such as pandas and fuzzywuzzy, they built automated scripts for entity resolution by normalizing text, applying fuzzy similarity scoring, and implementing state-level geographic blocking to improve match accuracy and reduce processing time. Location fields were validated and standardized using U.S. Census Bureau datasets. Purchase and distribution records were then linked to track the flow of goods from vendors to recipients. Excel pivot tables were used for quality assurance to identify and correct any remaining errors or inconsistencies in the mapping table. This resulted in a clean and consistent that will support accurate program analysis and reporting in the future.

AGRICULTURAL DATA AND AI AT THE USDA

US Department of Agriculture | Local and Regional Foods Division Research Branch

Ashley Chu, Data Fellow | *UC San Diego, Cognitive Science & Machine Learning*

Galiba Anjum, Data Fellow | *Cornell University, Computer Science & Information Science*

Keywords:

Retrieval-augmented generation, large language models, classification, natural language processing

Summary:

To streamline the **NEPA environmental review process**, Ashley developed a **retrieval-augmented generation model** and **pipeline** that could ingest and reference policy documents and guidelines to determine what level of review a construction project proposal requires. They wrote **Python** scripts to chunk and process the reference texts and embed them into a vector database. The relevant policy texts could then be retrieved and used as context for the **LLM** to quote and use in the classification and explanation it generates.

AGRICULTURAL DATA AND AI AT THE USDA

US Department of Agriculture | Local and Regional Foods Division Research Branch

Ashley Chu, Data Fellow | *UC San Diego, Cognitive Science & Machine Learning*

Galiba Anjum, Data Fellow | *Cornell University, Computer Science & Information Science*

Keywords:

Data visualization, R programming, geospatial mapping, survey analytics, stakeholder feedback

Summary: Ashley and Galiba are analyzing data from the 2019 - 2020 National Agritourism Survey, conducted by the University of Vermont in collaboration with the USDA, which includes responses from over 1,600 U.S. farm, ranch, and vineyard operators. Using R and Python, they are developing reproducible geospatial and statistical visualizations - such as interactive maps to highlight regional participation patterns and comparative charts to reveal relationships between key business factors - to uncover trends, challenges, and opportunities in agritourism. These preliminary visuals are iteratively refined based on stakeholder feedback to ensure they are accessible to producers, partners, and the public, with the goal of informing research, outreach, and policy to strengthen agritourism operations.

AGRICULTURAL DATA AND AI AT THE USDA

Local and Regional Foods Division Research Branch
U.S. Department of Agriculture
Kamran Zendehdel — Branch Chief

coding it forward >

USDA

ASHLEY CHU
UC San Diego
Cognitive Science

GALIBA ANJUM
Cornell University
Computer Science

The Local and Regional Foods Division (LRFD) supports local food producers, markets, and communities by providing easy access to the **ideas, innovations and research** necessary to grow and sustain a productive business.

cif>

OVERVIEW OF PROJECTS

Grants Division Dashboards and Reporting

Local Food Purchase Assistance Cooperative Agreement Program Data Wrangling

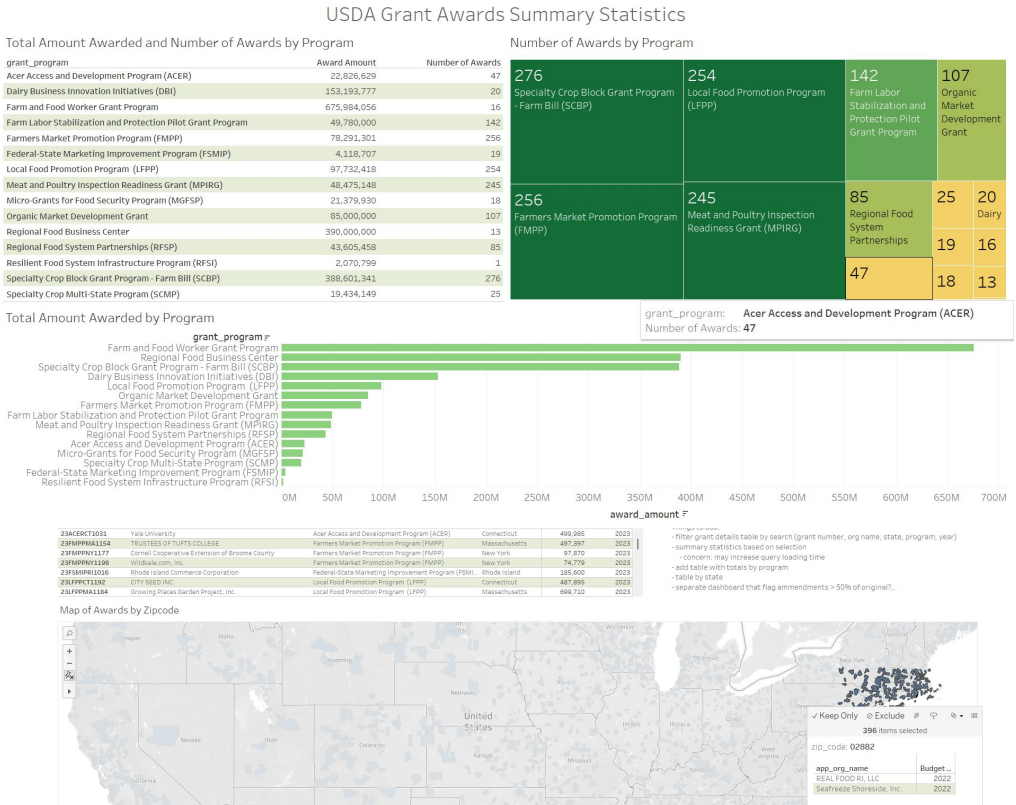
LLMs and RAG for Qualitative Data

Agritourism Data Visualizations

cif>

GRANTS DIVISION DASHBOARDS

- Write SQL queries to pull data from the USDA's Enterprise Data Analytics Platform into Tableau and Python notebooks → data wrangling
- Create proof-of-concept dashboards for:
  - Summary statistics to report awards by program, by year, etc
  - Maps of grants over the years
- Automated report in Quarto notebook
- More detailed view of spending over time track obligations, disbursements and trends



LFPA DATA WRANGLING PIPELINE

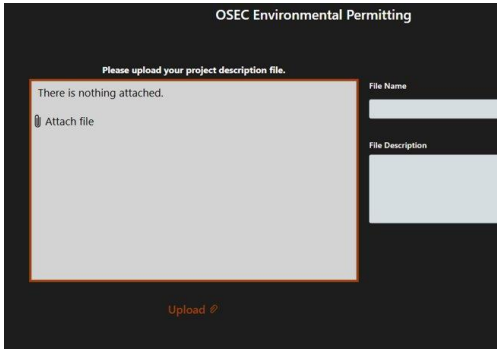
The LFPA dataset (2022–2024), manually submitted by local partners across all 50 states and territories, had minimal data entry standards - leading to data-quality issues which complicated aggregation and analysis of the data

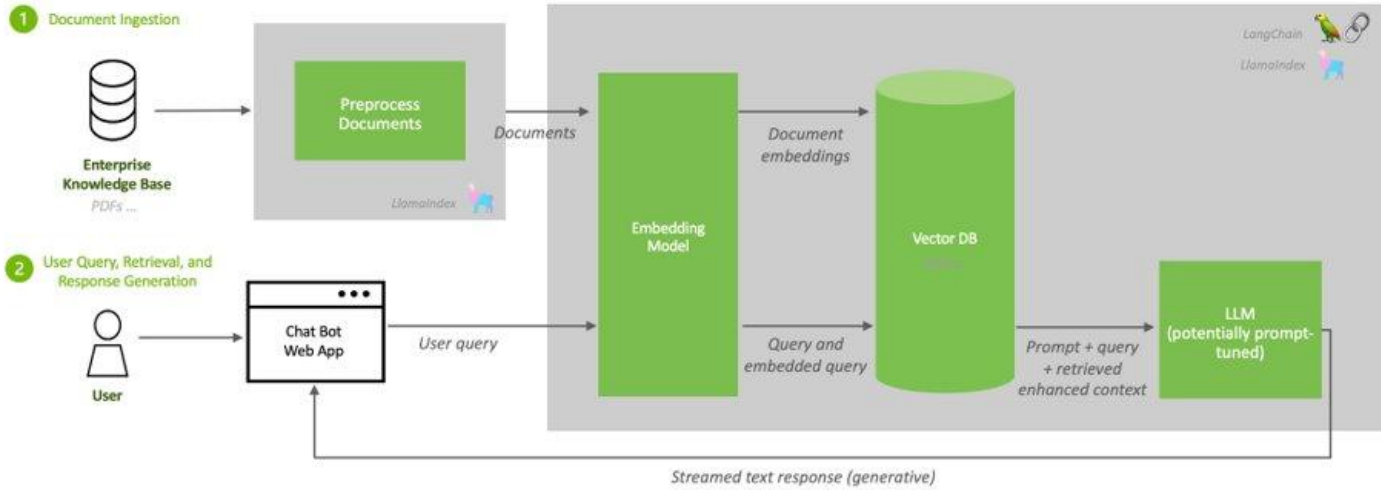
Core Tasks	Approach	Result
Standardize Organization & Producer Names	Used Python libraries for basic normalization of text and applied matching to consolidate variant entries.	Consolidated all duplicates into standardized names.
Geospatial Validation & Normalization	Standardized ZIP codes to 5-digit format and validated city/state consistency via U.S. Census Bureau data.	Corrected ZIPs and validated place assignments.
Link Vendors and Distributors	Performed joins between Purchases and Distributions datasets to align vendor and distributor identities.	Linked vendor and distributor records accurately

cif>

RETRIEVAL AUGMENTED GENERATION

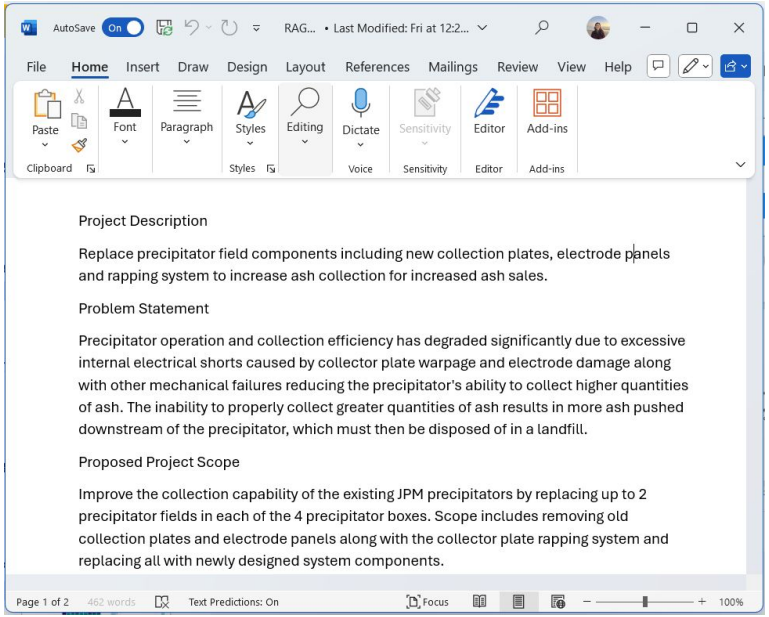
- Grants Division chatbot ↔ classification model for USDA Accelerator Hackathon
- Power App interface ↔ HTTP request via Power Automate ↔ notebooks and models in Databricks
- Embedding model: GTE Large En
- LLM: Claude Sonnet 3.7



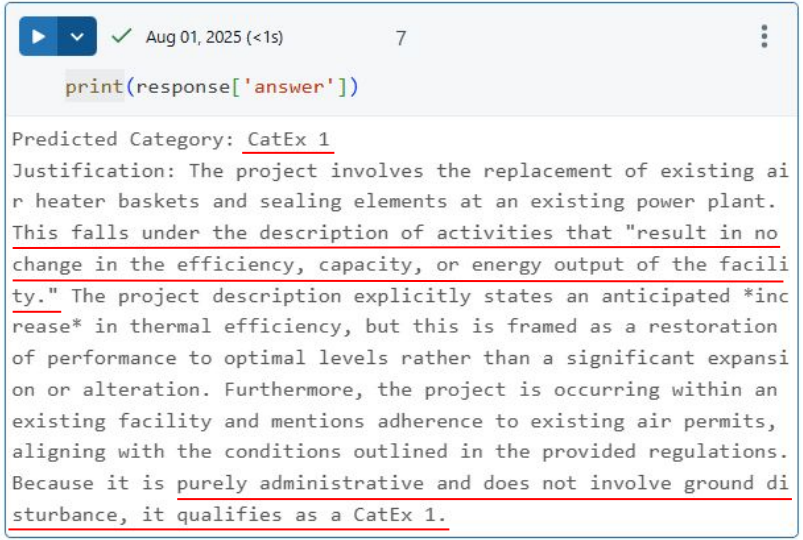


RETRIEVAL AUGMENTED GENERATION

Input

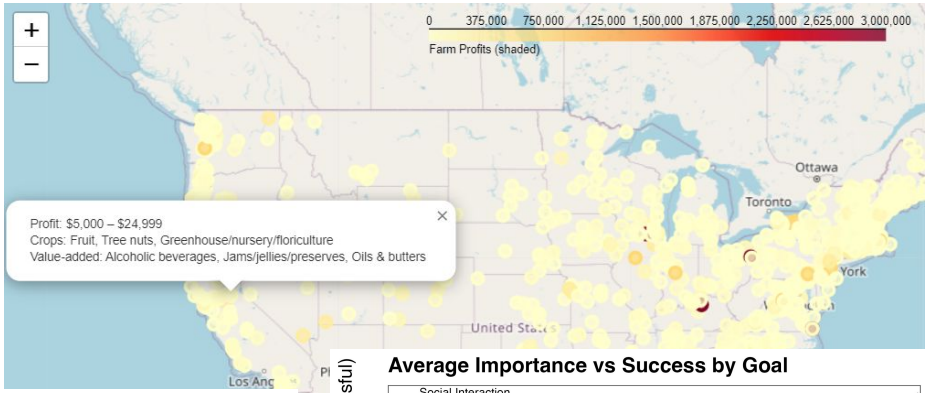


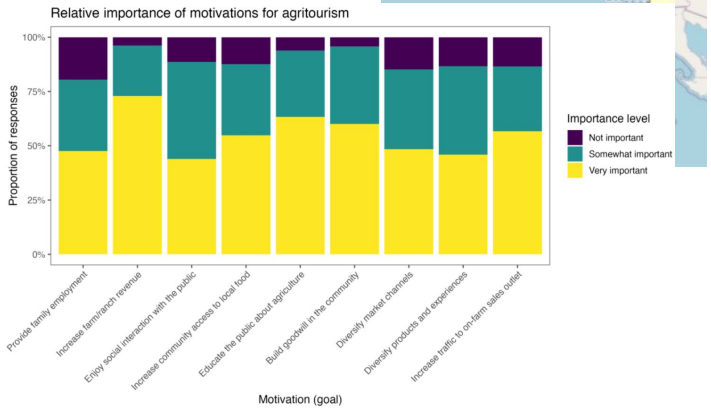
Output

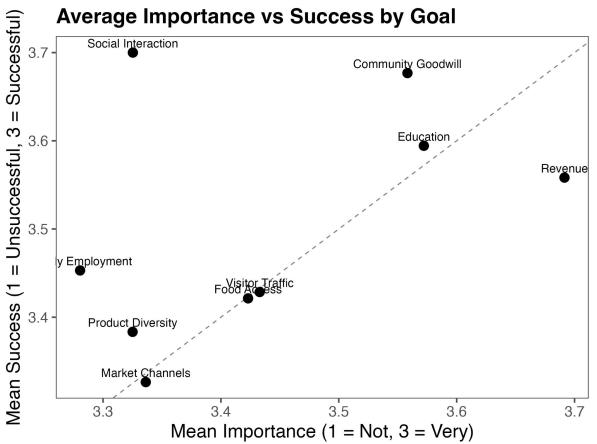


AGRITOURISM STUDY

- Cooperative agreement with University of Vermont to research agritourism
- Analyze and visualize survey data to understand motivations, successes, demographics, challenges, and profitability
- Report for producer reference







*These are preliminary visualizations and subject to change before final publication.

LEARNINGS AND TAKEAWAYS

- Unpredictability & Adaptability
  - Glimpse into what it's like working in the federal government
- Learned so much about how data is used at the USDA across all four projects
  - Possibilities of AI to streamline processes + summarize/reference qualitative data
  - Leverage data to make an impact + better support producers across the country

cif>

THANK YOU!

Special thanks to Sara Elazan, Elliot Hohn, Kamran Zendehdel, Yuyang Zhong, Cassie Rubio, and everyone we had the privilege of working with this summer!

cif>