FDA Insights: Simple interactive access to FDA information for public, industry, clinical research Using customized LLMs and RAG

Chetan “Aryan” Paul,
Ravichandran Sarangan*,
Ramprasad Venkatraman*
Center for Veterinary Medicine (CVM) FDA and Leidos

Abstract

The FDA is a science-based regulatory agency that uses scientific research to ensure the safety and effectiveness of the products it regulates. FDA disseminates public information, industry guidance and scientific research in a variety of ways, including FDA website, guidance documents, and compliance actions like product recalls, import refusals. Empowering public, industry and research communities with tools to access accurate and up-to-date FDA-information in a simple interactive manner will help enhance and streamline communication and promote FDA’s core mission to protect, and advance public health.

FDA Insights is an intelligent, interactive solution designed to simplify access to FDA-information. This solution uses multi-channel and multi-format content from FDA-official sources like FDA website, guidance documents, and data dashboards to provide precise and contextually relevant responses to user queries. FDA Insights current capabilities include FDA-regulated food products, compliance actions including recalls and import refusals and industry guidance from FDA Center of Veterinary Medicine. The solution is readily scalable to include additional FDA-regulated products like drugs, biologics, devices, and tobacco. This solution can securely integrate with agency’s regulatory review process to assist with the timeliness, and effectiveness of reviews, ensuring compliance with existing laws (User Fee Acts, FSMA), CFRs, policies and guidance.

Introduction

To protect the public's health and to ensure the safety, efficacy, and security of a variety of products, FDA uses multiple channels to provide accurate information for various stakeholders like public, industry and scientific research communities. The FDA-information is available across diverse sources like FDA public website, industry guidance documents, public meetings and webinars, and data dashboards.

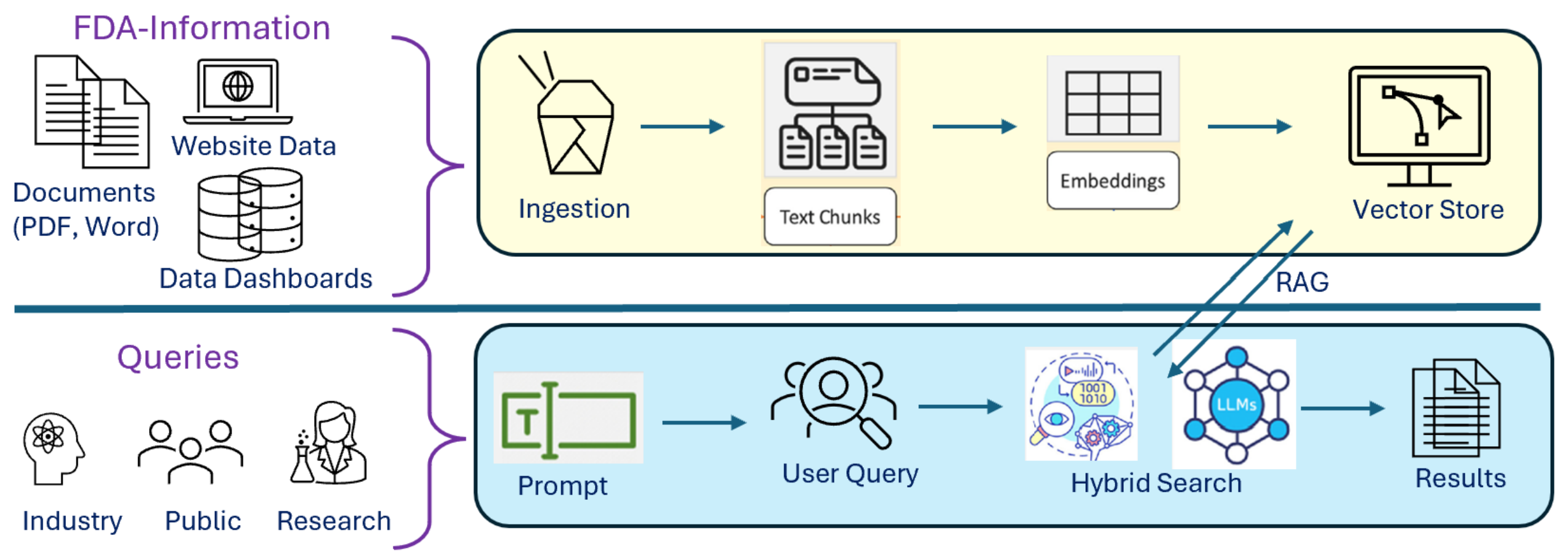
We face the challenge of an overabundance of information related to public health. Using traditional internet search methods to access specific information is prone to fetch potentially unverified information that is not validated, endorsed and distributed by FDA. Some of this information may be false and potentially harmful, making it more difficult to identify verified facts from trusted sources, such as the FDA.

FDA Insights solution addresses the above challenge by providing a simple and interactive mechanism for public, industry and scientific community to access accurate and up-to-date information retrieved from FDA-information sources only. This solution ingests multi-channel, multi-format content from FDA-information sources like FDA website, industry and public guidance documents and data dashboards. The ingested data is pre-processed and prepared for retrieval using powerful natural processing techniques like text chunking, embeddings and vector store. The user queries input as prompts are processed by a Retrieval-Augmented Generation (RAG) framework that combines large language models (LLMs) with retrieval from ingested data as “reference context” to generate precise and contextually relevant responses.

Materials and methods

FDA Insights core architecture, is built upon our similar pilot for U.S. Customs and Border Petrol (CBP) Trusted Travelers Program (TTP) that uses a RAG framework by combining large language base model operating on reference data sets. The dataset’s robustness and relevance is validated using custom prompts and Reinforcement Learning (RL) of the RAG framework.

RAG combines the strengths of large language models with dynamic information retrieval, allowing for the continual integration of the most current and relevant data into the decision-making process. Through RAG framework, the input prompt of LLMs is enriched and customized to meet the unique requirements of providing accurate and results within the context of FDA-information sources only. Another important advantage of this framework is its ability to integrate future LLMs to stay current with evolving technology. The FDA Insights architecture is depicted below:



The top row represents the **“Data ingestion and processing”** pipeline. We used the following FDA data sources for FDA Insights:

- FDA Website (Food): www.fda.gov/food
- FDA Industry Guidance (CVM): Registering with CVM Electronic Submission System https://www.fda.gov/media/70064/download and Acceptable file types for CVM eSubmitter https://www.fda.gov/media/120368
- FDA Data Dashboard: (Compliance) Recalls and Import refusals data

All the FDA-information sources listed above are in public domain, but to extend the solution to support agency’s regulatory review process it’s imperative that sensitive regulatory data and proprietary information in the review process is not put through public LLMs. For security and confidentiality, the source data is securely maintained in an appropriate FedRAMP designated FDA’s AWS cloud S3 storage. The embeddings are also calculated and stored in a secure vector store in AWS cloud.

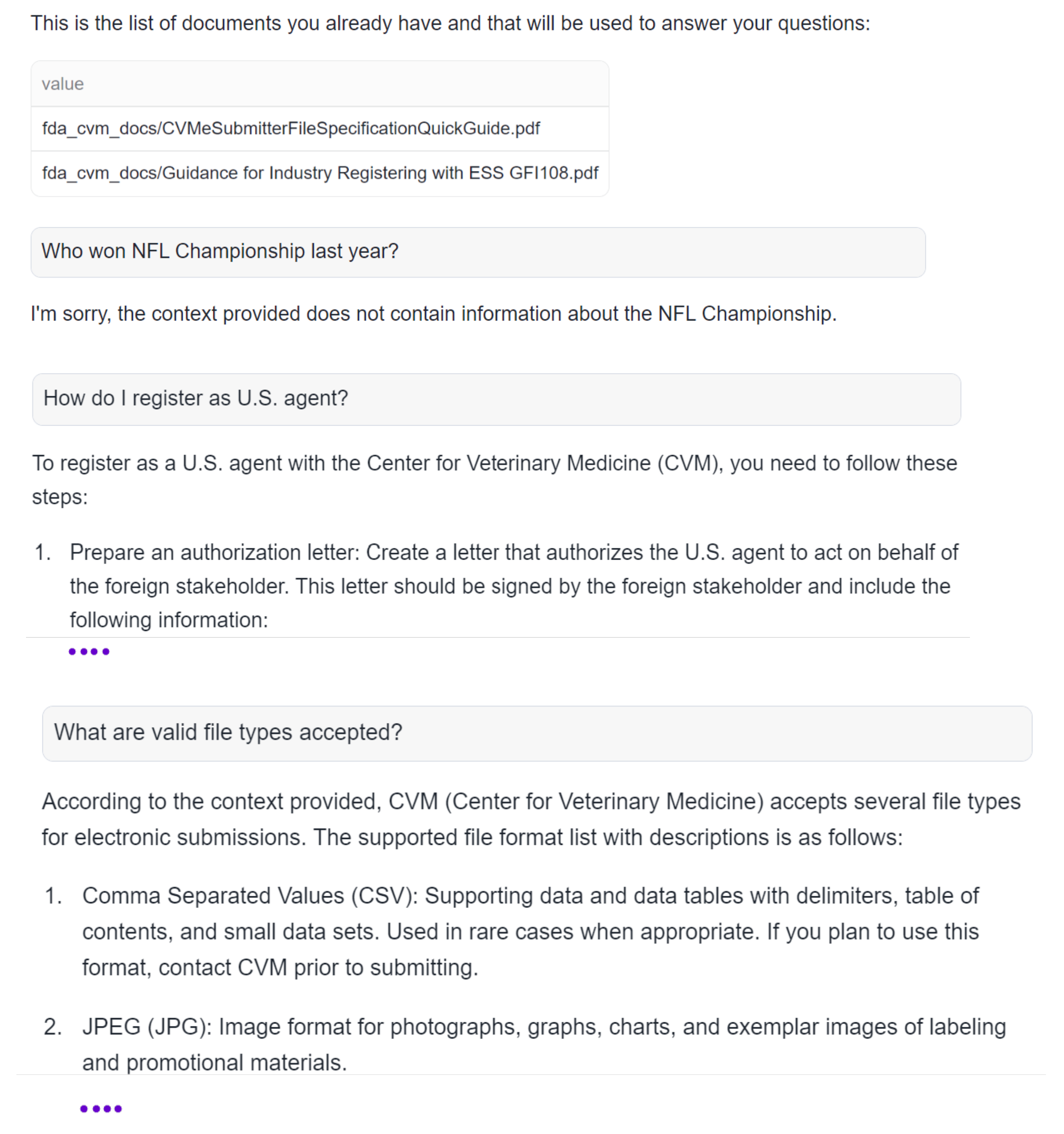
The bottom row represents the **“Query processing”** pipeline. It provides a simple interactive interface to accept the user queries as “prompts” that are processed by a large language model and uses the RAG framework to provide accurate and contextually relevant responses based on the FDA data sources provided. The responses also provide traceable evidence from the source data as links and references to ensure transparency and explainability,

We implemented the above architecture using the following technology stacks to perform a cost-benefit analysis in alignment with FDA’s IT Strategy principles and goals of mission driven business, agility and informed decision making.

- **Open-source software (OSS):** Python, ChromaDB (vector store), Azure OpenAI
- **Snowflake fully-managed services:** Cortex AI, Cortex Analytics, Warehouse
- **AWS fully-managed services:** Bedrock, Knowledge Base (vector store), Titan Text Embeddings and Llama (LLMs), S3 (secure data storage)

Results and discussion

We evaluated the FDA Insights solution through query prompts within the context of information provided in the data sources and intentionally out of context queries. The responses generated were validated against the data sources and for out of context queries, the FDA Insights tool did not hallucinate and appropriately refused to provide any response. The below prompts demonstrate the capabilities of FDA Insights solution:



This is the list of documents you already have and that will be used to answer your questions:

value
fda_cvm_docs/CVMeSubmitterFileSpecificationQuickGuide.pdf
fda_cvm_docs/Guidance for Industry Registering with ESS GFI108.pdf

Who won NFL Championship last year?

I'm sorry, the context provided does not contain information about the NFL Championship.

How do I register as U.S. agent?

To register as a U.S. agent with the Center for Veterinary Medicine (CVM), you need to follow these steps:

1. Prepare an authorization letter: Create a letter that authorizes the U.S. agent to act on behalf of the foreign stakeholder. This letter should be signed by the foreign stakeholder and include the following information:

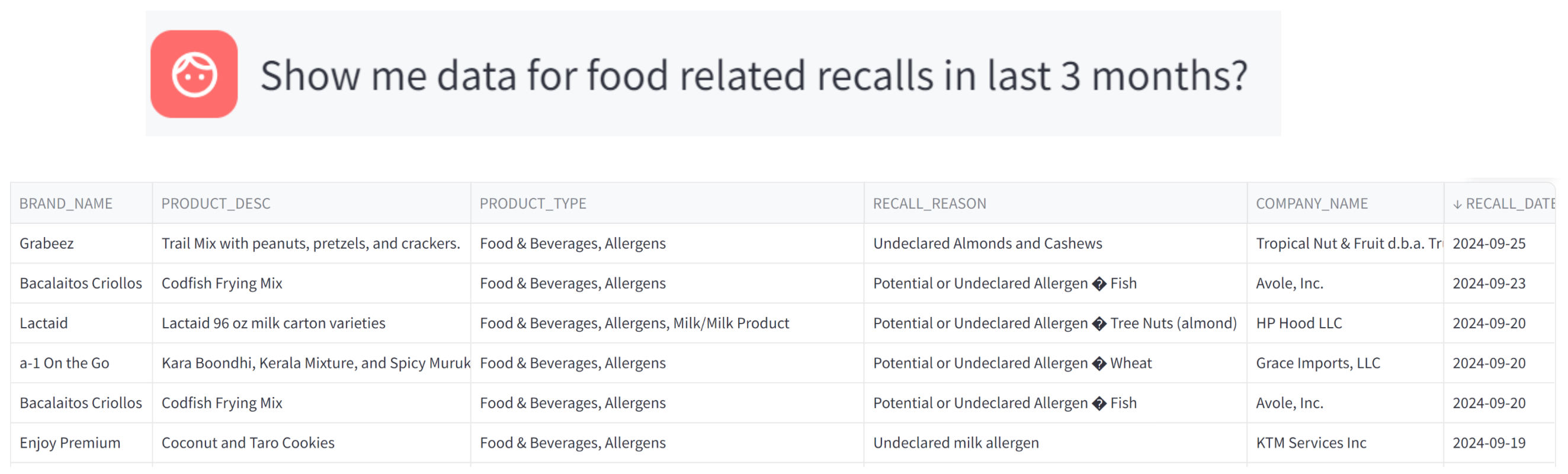
••••

What are valid file types accepted?

According to the context provided, CVM (Center for Veterinary Medicine) accepts several file types for electronic submissions. The supported file format list with descriptions is as follows:

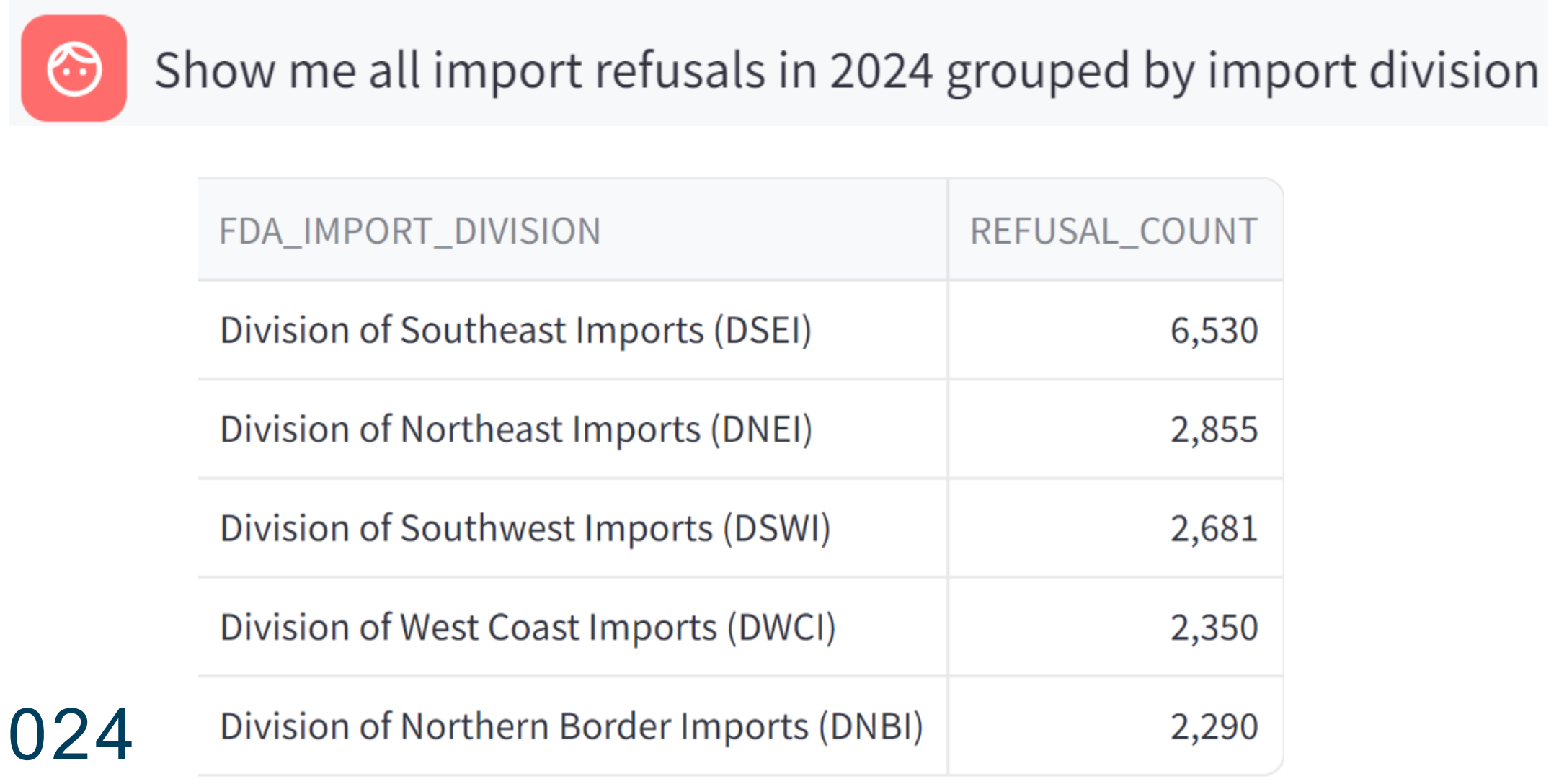
1. Comma Separated Values (CSV): Supporting data and data tables with delimiters, table of contents, and small data sets. Used in rare cases when appropriate. If you plan to use this format, contact CVM prior to submitting.
2. JPEG (JPG): Image format for photographs, graphs, charts, and exemplar images of labeling and promotional materials.

••••



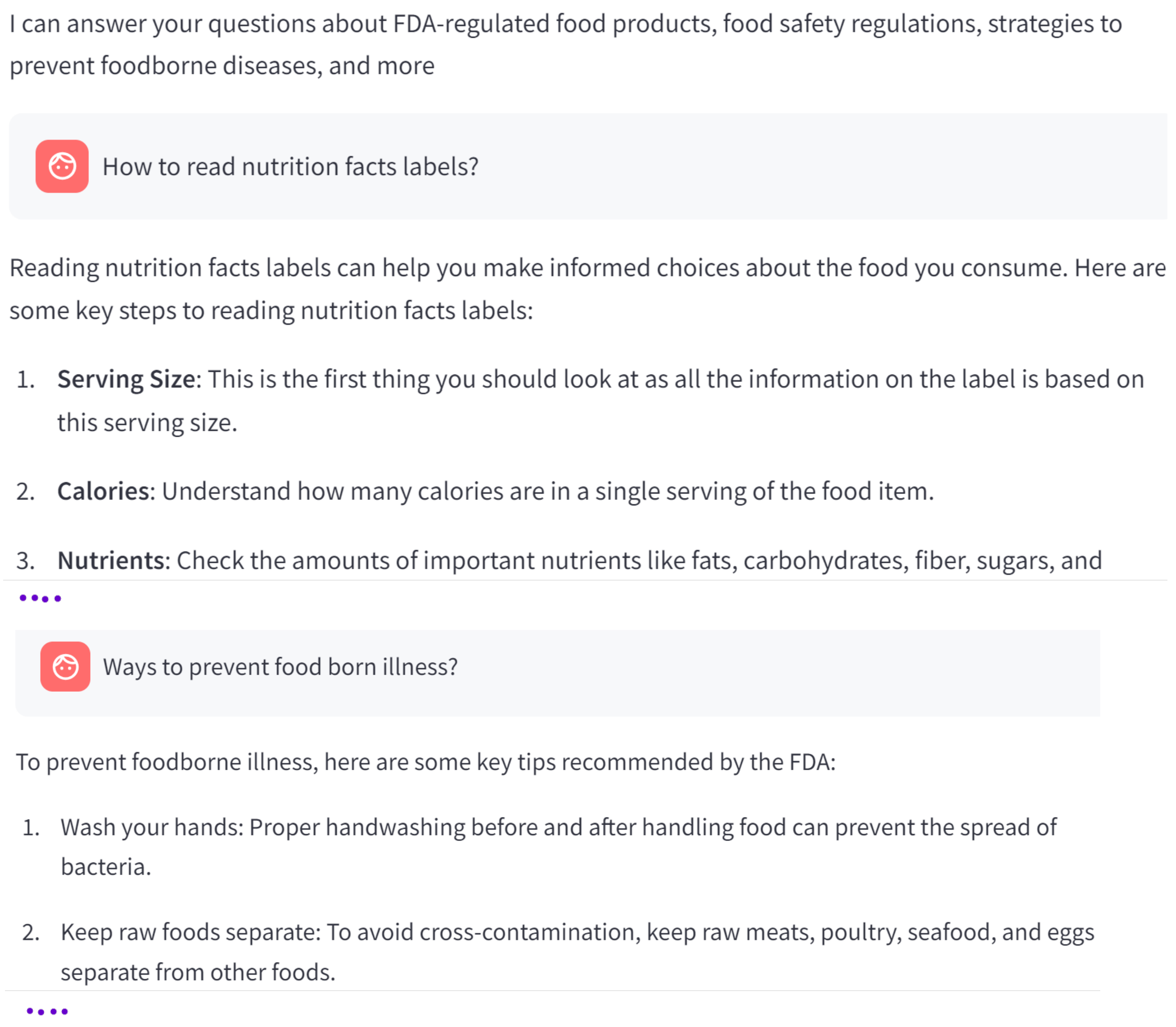
Show me data for food related recalls in last 3 months?

BRAND_NAME	PRODUCT_DESC	PRODUCT_TYPE	RECALL_REASON	COMPANY_NAME	↓ RECALL_DATE
Grabeez	Trail Mix with peanuts, pretzels, and crackers.	Food & Beverages, Allergens	Undeclared Almonds and Cashews	Tropical Nut & Fruit d.b.a. Tr	2024-09-25
Bacalaitos Criollos	Codfish Frying Mix	Food & Beverages, Allergens	Potential or Undeclared Allergen � Fish	Avole, Inc.	2024-09-23
Lactaid	Lactaid 96 oz milk carton varieties	Food & Beverages, Allergens, Milk/Milk Product	Potential or Undeclared Allergen � Tree Nuts (almond)	HP Hood LLC	2024-09-20
a-1 On the Go	Kara Boondhi, Kerala Mixture, and Spicy Muruk	Food & Beverages, Allergens	Potential or Undeclared Allergen � Wheat	Grace Imports, LLC	2024-09-20
Bacalaitos Criollos	Codfish Frying Mix	Food & Beverages, Allergens	Potential or Undeclared Allergen � Fish	Avole, Inc.	2024-09-20
Enjoy Premium	Coconut and Taro Cookies	Food & Beverages, Allergens	Undeclared milk allergen	KTM Services Inc	2024-09-19



Show me all import refusals in 2024 grouped by import division

FDA_IMPORT_DIVISION	REFUSAL_COUNT
Division of Southeast Imports (DSEI)	6,530
Division of Northeast Imports (DNEI)	2,855
Division of Southwest Imports (DSWI)	2,681
Division of West Coast Imports (DWCI)	2,350
Division of Northern Border Imports (DNBI)	2,290



I can answer your questions about FDA-regulated food products, food safety regulations, strategies to prevent foodborne diseases, and more

How to read nutrition facts labels?

Reading nutrition facts labels can help you make informed choices about the food you consume. Here are some key steps to reading nutrition facts labels:

1. **Serving Size**: This is the first thing you should look at as all the information on the label is based on this serving size.
2. **Calories**: Understand how many calories are in a single serving of the food item.
3. **Nutrients**: Check the amounts of important nutrients like fats, carbohydrates, fiber, sugars, and

••••

Ways to prevent food born illness?

To prevent foodborne illness, here are some key tips recommended by the FDA:

1. Wash your hands: Proper handwashing before and after handling food can prevent the spread of bacteria.
2. Keep raw foods separate: To avoid cross-contamination, keep raw meats, poultry, seafood, and eggs separate from other foods.

••••

Conclusion

FDA Insights solution provides a simple and interactive interface to public, industry and scientific community to access accurate and up-to-date information retrieved from FDA-information sources only. This solution can also securely integrate into agency’s regulatory review process to assist with the timeliness, and effectiveness of reviews, and ensuring compliance with existing laws (User Fee Acts, FSMA), CFRs, policies and guidance

The data ingestion and processing pipeline is capable of multi-channel and multi-format content intake from FDA-official sources like FDA website, guidance documents, and data dashboards to provide precise and contextually relevant responses to user queries. For security and confidentiality, the source data is securely maintained in an appropriate FedRAMP designated FDA’s AWS cloud S3 storage. The embeddings are also calculated and stored in a secure vector store.

We noted that Open-source development requires python programming proficiency, provisioning and maintenance of infrastructure. Also, without effective storage management, ChromaDB can quickly run into scaling and performance bottlenecks.

The fully-managed services from AWS (Bedrock, Knowledge Bases) and Snowflake (Cortex AI and Analyst) alleviate infrastructure management burden, simplify the data ingestion and processing pipeline, provide a diverse range of base large language models (LLMs) out-of-box and can scale and perform in a secure FedRAMP cloud-native environment at a nominal cost. Managed-service providers like AWS and Snowflake also provide the ability to continually integrate with updated versions of LLMs to stay current with evolving technology.

As next steps, a cloud native localized solution using a base large language model trained, tuned and validated using Reinforcement learning and custom prompts can be implemented in FDA’s secure cloud environment to meet the required security needs of regulatory data and proprietary information.