Scalable Curation of Toxic Elements Surveillance Data Using a Human-in-the-Loop AI-Assisted Pipeline

Wade Huang, Victoria Incorvati, and Lauren Posnick Robin

OFCSDSI/OPMA/DCC, College Park

Introduction

- The FDA's Toxic Elements Program (TEP) routinely tests foods for toxic elements including arsenic, lead, cadmium, and mercury.
- Curating surveillance data at scale can present significant consistency and efficiency challenges.
- To address this, we developed a human-in-the-loop agentic AI workflow that supports data cleaning, standardized food naming, and commodity and product-category assignment, while preserving human oversight throughout.

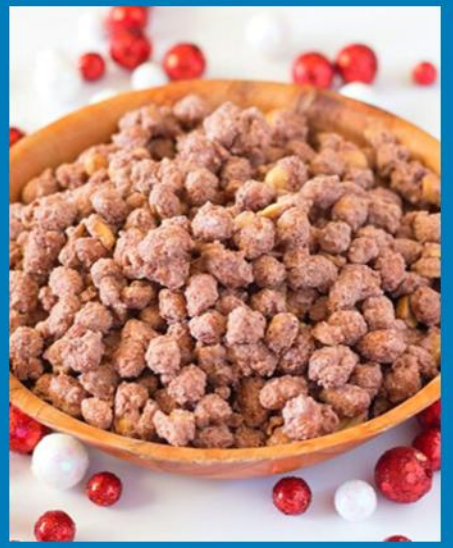
Record Example

LLM Model Inputs

Compliance Management System (CMS) Record

Sample Description:

Collection Remarks:



Product Description: CACAHUATE GARAMPINADO

Product Code Description: Candy Bar or Pieces

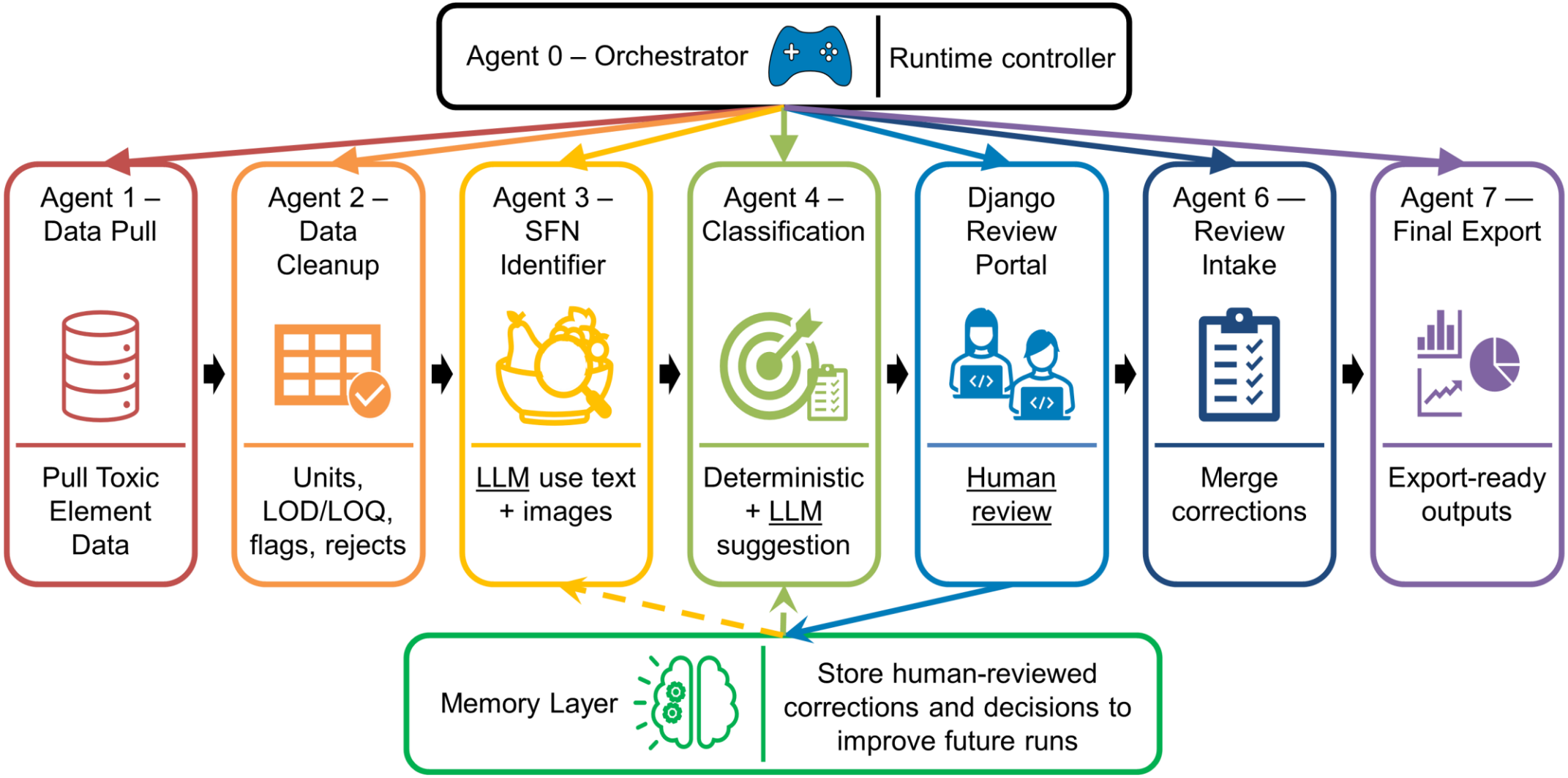
LLM Model Output

Agent 3 Output – SFN: *Caramelized Peanut*

Agent 4 Output – Commodity: *Hard Candy*

Category: *Candy/Chocolate*

Agent Workflow



Agent 3

Agent 3 – Standardized Food Name Identifier

- Uses a classifier to generate systematic SFNs based on the CMS records. Classifier training and testing are described below.

Classifier Training

1. Used 11,000+ manually assigned FDA food names from 2017–2024.
2. Claude Sonnet 4.6 interpreted product identity from CMS/HFDP fields and package images (**Figure 1**).
3. Generated structured attributes such as product family, base product, processing state, and form.
4. Developed SFN naming rules based mainly on product family and base product.
5. Compared generated SFNs with curated names to calculate confidence.
6. Assigned SFNs to six confidence tiers, from exact matches to newly constructed names.
7. Trained a classifier to assign SFNs and confidence tiers to new products.

Classifier Testing

1. The AI-assisted workflow and outputs were evaluated using 1,024 FDA-collected products from October 1, 2025 through May 31, 2026.
2. The developed classifier was used to generate the SFNs and sort the SFNs into six confidence based tiers.

Agent 4



Agent 4 – Commodity and Category Classification

- Uses pre-existing product codes developed for the annual toxic elements Sample Collection Operation Planning Effort (SCOPE) to match each SFN with a “commodity” and “category” classification.
- Sends unmatched products to Claude Sonnet 4.6 to propose appropriate commodity category.

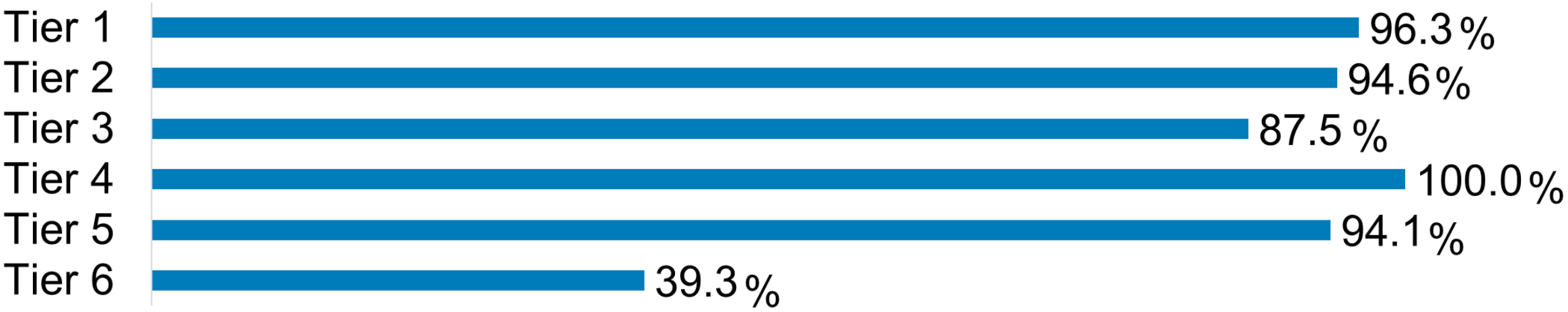
Human-in-the-Loop



Review Portal

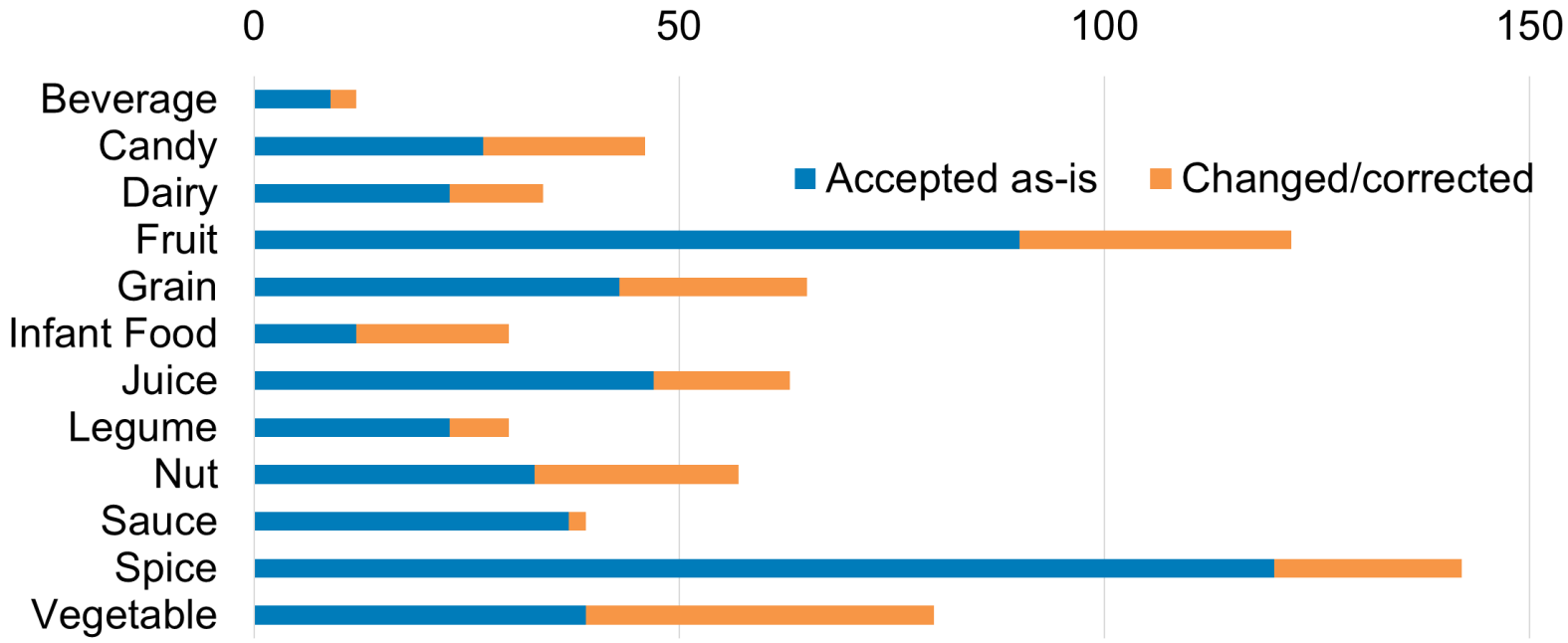
- The portal displays analytical records and all AI-generated results on the same page for FDA SME review.
- For analytical results, SME reviewers evaluate records that have been flagged for review.
- For AI-generated classification, reviewers evaluate all products.
- The portal allows reviewers to accept or revise analytical records and classification candidates or remove products from the dataset that do not meet pre-established criteria, e.g., removing candy from a vegetable dataset.

Results



Agent 3 in the model classifies SFNs into different “tiers” based on how well the SFNs align with manually curated names. In the test phase, we compared the tiering results to human review. For tiers 1-5, from most to least compatibility, human acceptance ranged from 85 to 100 percent with no trend observed. Tier 6 had low human acceptance (35%). The low acceptance rate for Tier 6 suggests that the tiering paradigm in Agent 3 is effective at identifying mismatches.

Results



Agent 3 also classifies samples to product groups based on CMS records. During the testing phase, the SFN classifications were compared with SME classifications to evaluate performance. In this **Figure**, the SFNs are grouped by product groups to present the performance across the groups. Some product groups, including sauces, spices, and legumes, had strong acceptance rates, suggesting they may be well represented in the training data or have low naming complexity. Other groups, such as infant foods, vegetables, and nuts, required more corrections, which indicates greater naming complexity in these categories and potential areas for training data expansion.

Conclusion

- This study shows that a human-in-the-loop agentic AI workflow can improve the consistency and traceability of product naming, enabling systematic collation of FDA surveillance data.
- A built-in tiering system for naming confidence effectively concentrated uncertain cases for expert review.
- Overall, 85.3% of naming candidates were accepted without modification.

- Future work will add more details to SFN naming system, incorporate product properties such as processing state, and expand the memory layer to reduce review burden over time.

Acknowledgements

- HFP/OFCSDSI/OPMA/DAI/HAAB for assistance on setting up Amazon Web Service working instance
  - Ernest Kwegyir-Afful
  - Nathan Smith
- HFP/OFCSDSI/OPMA/DCC for SFN consults
  - Sara McGrath
  - Rhoma Johnson
- HFDP team for data access