Journal of Food Protection 89 (2026) 100829

Contents lists available at ScienceDirect

Journal of Food Protection

journal homepage: www.elsevier.com/locate/jfp

Research Paper

A Retrieval-Augmented Natural Language Interface for Data Description and Meta-Analysis in the Pathogens-in-Foods (PIF) Database ☆

Lucas Ribeiro Silva, Ursula Gonzales-Barron, Vasco Cadavez *

CIMO, LA SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

ARTICLE INFO

Keywords:
Data visualization
Food safety
Meta-analysis
Natural language interface
Open-weight language models
Tool-using agents

ABSTRACT

Food-safety occurrence databases are increasingly important for surveillance, evidence appraisal, and quantitative risk assessment, yet their routine analytical use remains constrained by the need for database literacy and statistical programming. Building on the curated and harmonized Pathogens-in-Foods (PIF) database, we developed and evaluated a retrieval-augmented natural-language interface designed to support grounded querying and reproducible evidence synthesis. The system includes two complementary modes: an Open Chat Mode for exploratory, tool-mediated interrogation of the database and a Guided Meta-Analysis Mode that couples structured user input to a deterministic R-based analytical pipeline. Evaluation included four compact language models: Phi-4 Mini (3.8B), DeepSeek-R1 Tool-Calling (14B), Cogito (14B), and Qwen 3 (8B), together with Gemini 2.5 Pro as a larger proprietary baseline model. Within a 10-query benchmark, all models achieved 100% tool selection accuracy and retrieval correctness; for the five argument bearing queries, all models also achieved 100% argument extraction F1-score, indicating reliable grounding of database operations for the evaluated query set. In a guided case study on Toxoplasma in meat and meat products (153 records from 65 studies), the system achieved 100% numerical concordance and high visual informativeness; the highest report quality index was 93% with Qwen 3 (8B). Performance differences across models arose primarily from the factual precision and economy of their written interpretations rather than from failures in tool execution. These findings support hybrid, evidence-grounded analytical interfaces built on curated data resources and deterministic statistical backends as practical tools for accelerating surveillance-oriented evidence synthesis in food protection.

Foodborne disease remains a major public–health and food-systems challenge. An estimated 600 million people develop foodborne illness each year and approximately 420,000 die, with children under five years of age, bearing a disproportionate share of this burden (World Health Organization, 2020). Beyond its direct health effects, unsafe food imposes substantial economic losses through healthcare expenditure, lost productivity, and trade disruption (World Health Organization, 2020). These realities sustain the need for food-safety surveillance systems that do more than store data: they must also support timely, transparent, and analytically defensible use of evidence.

One persistent obstacle is that occurrence data for biological hazards in foods remain scattered across primary studies, described with heterogeneous terminology, and laborious to reassemble for each new assessment. This fragmentation slows evidence synthesis and raises the technical threshold for conducting secondary analyses that should, in principle, be routine in food microbiology and risk assessment. The practical value of systematic review and multilevel meta-analysis for integrating such evidence has already been shown for Portuguese meats and meat products, poultry meat, retail fruits and vegetables, and sheep- and goat-milk products (Xavier et al., 2014; Gonçalves-Tenório et al., 2018; Silva et al., 2017; Gonzales-Barron et al., 2017). The Pathogens-in-Foods (PIF) database was developed to address this upstream problem by centralizing published prevalence and concentration data for pathogenic bacteria, viruses, and parasites in foods, extracted through systematic review and organized in harmonized data structures under controlled terminologies (Gonzales-Barron et al., 2025).

The foundational PIF article is central to the present work because it defines both the evidentiary basis and the governance logic on which the intelligent agent depends. It presents PIF as an open-access European resource designed for structured reuse of hazard-occurrence evidence and aligned with FAIR data stewardship (Gonzales-Barron et al., 2025;

☆ This article is part of a special issue entitled: 'VSI: 13th ICPMF' published in Journal of Food Protection.

* Corresponding author.
E-mail address: vcadavez@ipb.pt (V. Cadavez).

https://doi.org/10.1016/j.jfp.2026.100829
Received 21 April 2026; Accepted 6 June 2026
Available online 9 June 2026

Wilkinson et al., 2016). Its workflow for evidence identification, screening, structured extraction, validation, and harmonization is consistent with contemporary systematic review reporting and with the group's broader methodological work on meta-analytical integration of food safety evidence (Page et al., 2021; Gonzales-Barron et al., 2021). Its data quality layer further emphasizes fit-for-purpose structures and explicit controls over completeness, consistency, and representational conformity (Gonzales-Barron et al., 2025; Wang and Strong, 1996; Kahn et al., 2016). In practical terms, PIF already solves the difficult problem of assembling a usable evidence base from a highly heterogeneous literature.

However, a curated and FAIR-oriented database does not automatically become easy to use in day-to-day analytical practice. For many food microbiologists, surveillance analysts, and risk assessors, moving from evidence retrieval to descriptive analysis or meta-analysis still requires familiarity with schema structure, variable harmonization, and statistical workflows. That practical barrier is evident in the complexity of systematic-review and multilevel meta-analytic procedures already applied to food safety evidence synthesis, including recent work on sporadic toxoplasmosis and broader observational risk factor integration (Thebault et al., 2021; Gonzales-Barron et al., 2021). Put differently, PIF improves evidence availability, but it does not by itself eliminate the friction involved in interrogating that evidence quickly and correctly. Natural Language Interfaces (NLIs) offer a plausible way to reduce this friction by allowing users to query structured resources using ordinary language rather than formal query syntax (Katsogiannis-Meimarakis et al., 2023). Recent advances in Retrieval-Augmented Generation (RAG) and compact instruction-tuned language models have made this approach more feasible for specialized scientific settings. In a food safety context, retrieval grounding is particularly important because it constrains the model to curated evidence rather than unsupported model memory, while smaller models may enable local, lower-cost deployment (Lewis et al., 2020; Fan et al., 2024; Klesel and Wittmann, 2025).

Against this background, the present study should be understood as an analyst-facing extension of the PIF infrastructure rather than as a generic conversational AI application. Whereas the original PIF publication established the curated occurrence-data resource and its FAIR/CCC governance logic (Gonzales-Barron et al., 2025), and earlier work from the same methodological line demonstrated how heterogeneous food safety evidence can be integrated through a systematic review and multilevel meta-analysis (Gonzales-Barron and Butler, 2011; Gonzales-Barron et al., 2021), the present study addresses the layer required to make that resource operationally accessible. We designed and evaluated a retrieval-augmented natural-language interface that enables users to (i) interrogate the PIF database in plain language, (ii) retrieve correctly grounded evidence subsets, and (iii) generate reproducible descriptive and meta-analytical outputs without direct MongoDB programming. The objective was not simply conversational access, but a practical reduction in the time and technical expertise needed to perform surveillance-oriented querying and evidence synthesis in food protection.

Materials and methods

PIF as the evidentiary backbone

The PIF Intelligent Agent was developed on top of the PIF database (Gonzales-Barron et al., 2025). In that underlying resource, occurrence data for biological hazards in foods are assembled through systematic review, screened for relevance and methodological adequacy, extracted into harmonized structures, and curated under controlled terminologies suitable for downstream synthesis and risk assessment. This design is consistent with current reporting guidance for systematic reviews and with pathogen-specific evidence syntheses previously developed by the group for meats, poultry, produce, dairy products, and parasitic hazards (Page et al., 2021; Xavier et al., 2014; Gonçalves-Tenório et al., 2018; Silva et al., 2017; Gonzales-Barron et al., 2017; Thebault et al., 2021). The same framework adopts FAIR-oriented data stewardship and explicit quality controls so that the stored evidence is both reusable and analytically fit for purpose (Wilkinson et al., 2016; Wang and Strong, 1996; Kahn et al., 2016).

For the present study, PIF therefore functioned as a curated evidentiary layer rather than as a generic data store. The intelligent agent did not answer questions from parametric model memory alone; instead, it operated over structured occurrence data whose provenance, harmonization rules, and quality controls had already been formalized in the PIF framework and in the broader evidence-synthesis strategy underlying this line of work (Gonzales-Barron et al., 2025; Gonzales-Barron et al., 2021). For food-protection applications, this distinction is important because the reliability of the system depends jointly on the language-model layer and on the methodological rigor of the underlying evidence resource.

System architecture and operating modes

The PIF Intelligent Agent was implemented as a distributed architecture designed to separate user interaction, language-model reasoning, and deterministic statistical computation. As shown in Figure 1, the system comprised three layers: (i) a frontend layer for user interaction, (ii) a backend layer containing the intelligent-agent and statistical services, and (iii) a data layer for storage and retrieval. Relative to the database/back-end/front-end structure already described for the base PIF platform (Gonzales-Barron et al., 2025), the present system adds a natural-language access and reporting layer directed at evidence synthesis. (see Figs. 2–6).

The frontend layer was implemented as a single-page application served through Flask. It supported two interaction modes.

1. **Guided Meta-Analysis Mode:** This mode used a finite-state workflow to guide users through dataset selection, preprocessing, descriptive analysis, and pooled prevalence estimation. The objective was to standardize user input and produce reproducible analytical reports with minimal technical intervention.
2. **Open Chat Mode:** This mode allowed free-form exploratory interaction in natural language. User requests were translated into constrained tool calls for database interrogation and downstream analytical tasks.

The backend layer consisted of two coordinated microservices. The intelligent-agent core, implemented in Python, handled natural-language understanding, prompt construction, tool orchestration, and model invocation. The meta-analysis and reporting engine, implemented in R through the Plumber framework, performed data preprocessing, model fitting, figure generation, and report assembly. This separation was intentional: probabilistic models handled intent interpretation and narrative drafting, whereas quantitative outputs were produced by deterministic statistical code.

The data layer combined MongoDB and ChromaDB. MongoDB stored the structured PIF occurrence dataset, whereas ChromaDB stored vectorized schema information used to support retrieval-augmented interpretation of user requests. This arrangement enabled the system to map user expressions such as "chicken" to the relevant controlled vocabulary and database fields before execution.

Prompting strategy and model configuration

The core reasoning functions were provided by Small Language Model (SLMs) and one larger external baseline model. To reduce hallucination and improve reproducibility, the system used a constrained prompting strategy managed by the intelligent-agent core. Three

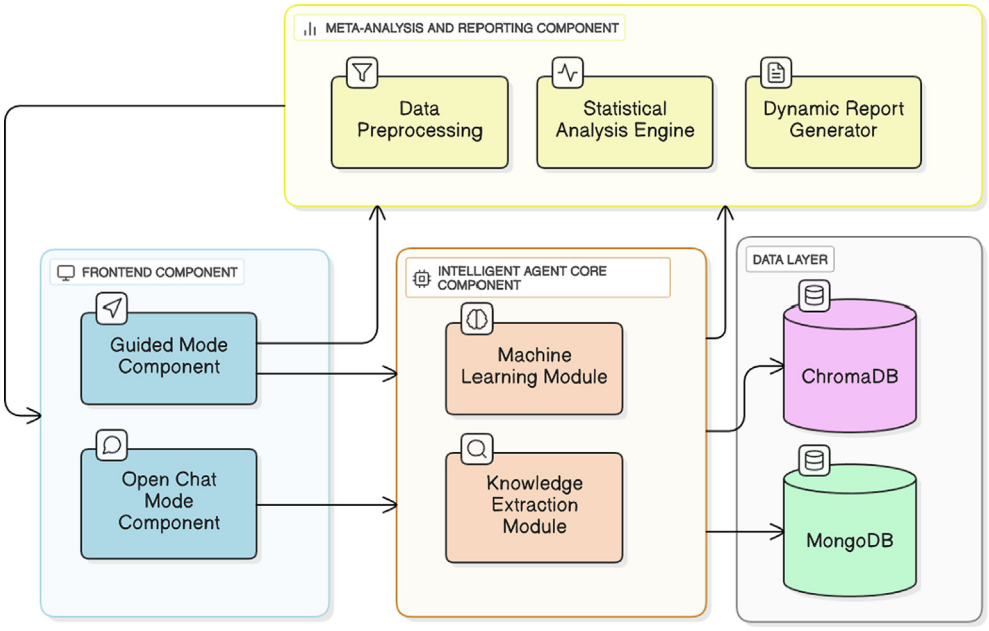

Figure 1. High-level system architecture illustrating the layered interaction between the frontend, backend (intelligent agent core and meta-analysis/reporting), and data components.

design elements were central. First, **role-specific system prompts** constrained the model according to the task. For example, during entity extraction the model was instructed to behave as a deterministic parser and to return valid JSON only. Second, **structured context injection** ensured that analytical prompts were grounded in data retrieved from MongoDB or computed by the R backend rather than in the latent model memory. Third, **zero-shot output constraints** were used for summarization tasks to control length, register, and formatting.

The open weight models were run locally using the Ollama framework. To prioritize factual stability over stylistic variation, inference parameters were fixed at temperature 0.0 and top-p 0.1 (Holtzman et al., 2019; Chang et al., 2024). Gemini 2.5 Pro was included as a state-of-the-art external baseline and was evaluated under the same task structure and scoring framework, but not under the same local deployment conditions. This distinction should be considered when interpreting practical deployment implications.

Analytical pipeline for prevalence synthesis and report generation

All quantitative analyses were executed in R through a deterministic backend, independently of the language model used for interaction or narrative generation. This architecture ensured that numerical results were reproducible and not contingent on probabilistic model behavior. Language models were therefore used to interpret and communicate results, but they did not perform the statistical computations themselves and did not modify the analytical outputs generated by the R pipeline.

After retrieval from MongoDB, records were preprocessed to standardize field names, harmonize analytical variables, and construct the inputs required for descriptive summaries and prevalence synthesis. For meta-analytical workflows, effect sizes and corresponding sampling variances were computed with the `escalc` function from the metafor package (Viechtbauer, 2010), using the predefined prevalence analysis settings implemented in the production pipeline. Random-effects models were then fitted with `rma.mv`, allowing estimation of between-study heterogeneity while accounting for sampling error and, where applicable, hierarchical data structure. When the moderator variables were specified by the user, the workflow extended to multilevel meta-regression using the same deterministic analytical backend.

Graphical outputs, including forest plots, funnel plots, bar plots, ridge plots, and exploratory text visualizations, were generated programmatically in R from the same analytical objects used for numerical reporting. Final reports were assembled from parameterized templates that integrated deterministic statistical outputs with model-generated narrative interpretation. Accordingly, the analytical validity of each report depended on the R pipeline, whereas the quality of the written interpretation depended on the language model.

System evaluation framework

The system was evaluated with a dual-track design intended to distinguish functional correctness in tool-mediated interaction from the quality of the resulting analytical communication. This distinction was necessary because correct database access and correct statistical executions do not, by themselves, guarantee scientifically appropriate narrative interpretation.

Models and execution environment. Experiments were conducted on a high-performance computing node running Ubuntu 24.04.3 LTS. Hardware specifications are summarized in Table 1. Four compact language models were deployed within the local tool-calling environment: Phi-4 Mini (3.8B), DeepSeek-R1 Tool-Calling (14B), Cogito (14B), and Qwen 3 (8B). Gemini 2.5 Pro was evaluated as an external baseline model. Model identity, prompting structure, and execution settings were kept fixed throughout the benchmark to support consistent cross-model comparison.

Evaluation endpoints. Evaluation endpoints were organized according to the analytical layer being assessed: functional correctness of model orchestration, fidelity of the deterministic analytical backend, and quality of the resulting analytical communication (Table 2).

Benchmark design and evaluation protocol. Open Chat Mode was evaluated with a curated benchmark of 10 representative natural-language queries spanning metadata inspection, schema understanding, descriptive retrieval, prevalence estimation, concentration summaries, food-source identification, and temporal trend analysis (Table 3 and Table 4). The benchmark was designed to assess functional correctness across distinct analytical intents rather than

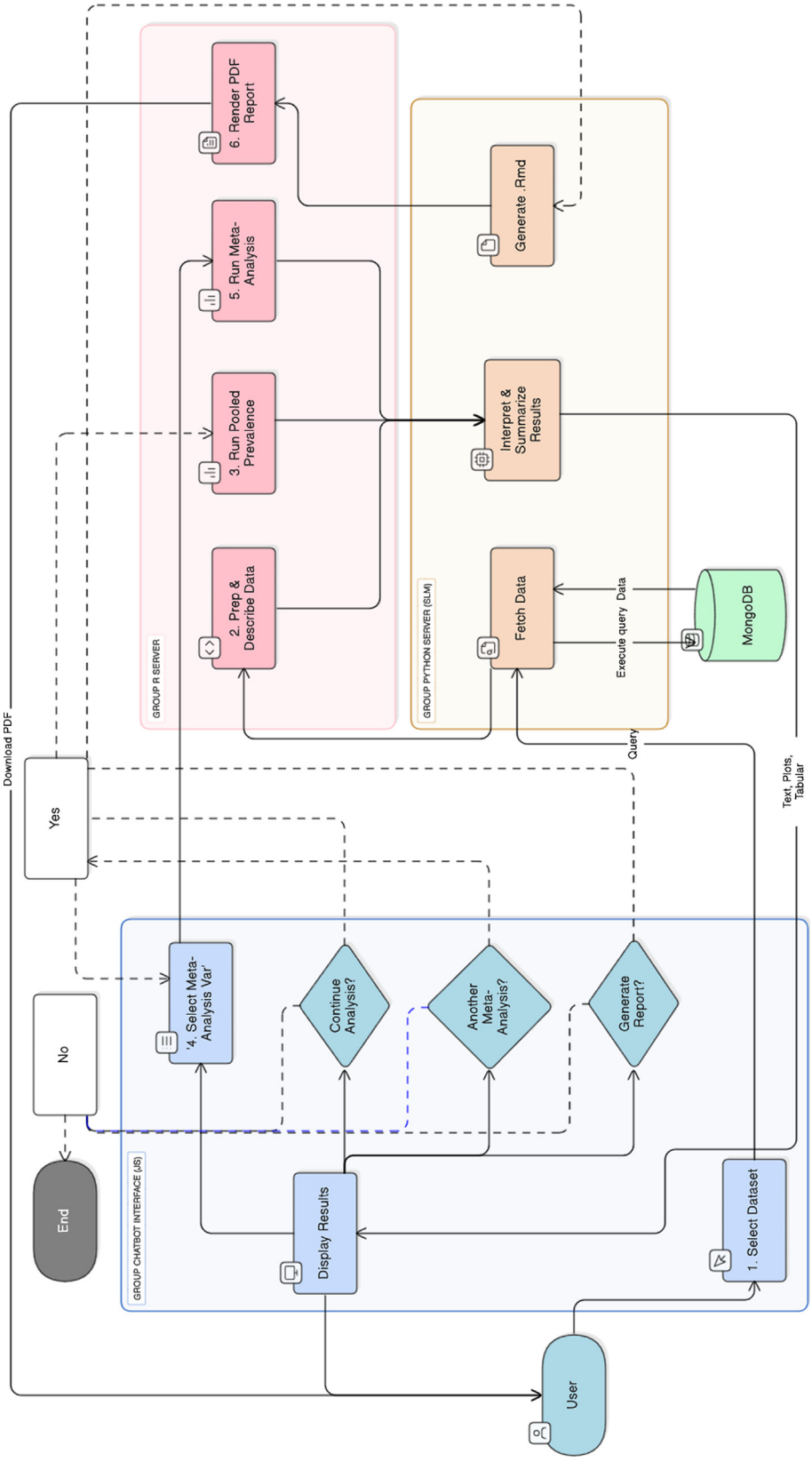

Figure 2. Workflow for the guided meta-analysis mode.

broad conversational ability. For each query, outputs were evaluated for correct tool selection, extraction of analytical arguments, and retrieval of the expected data subset. (see Table 5).

Argument-extraction F1-score evaluation was performed only for benchmark questions requiring explicit analytical parameter extraction (Q1 and Q7–Q10). Queries involving schema inspection, metadata

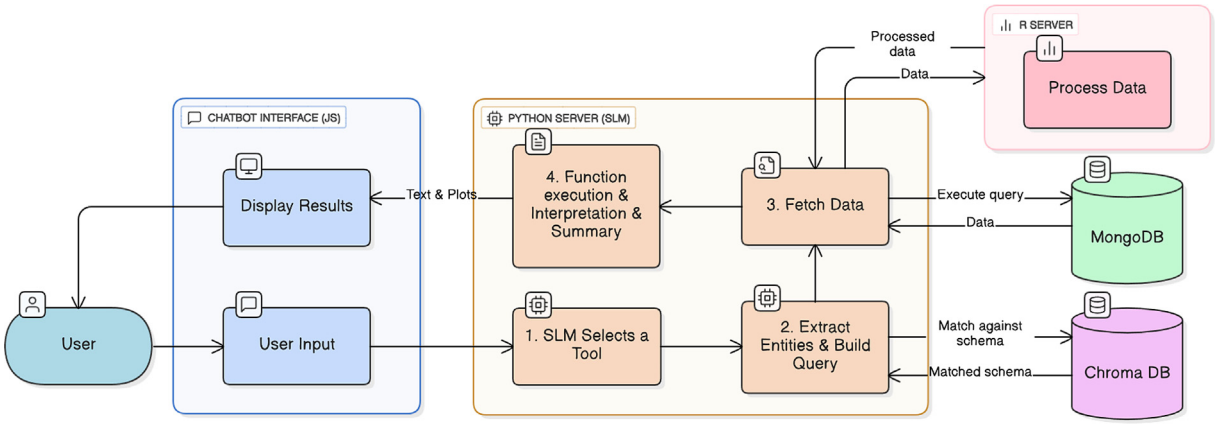

Figure 3. Workflow for the open chat mode.

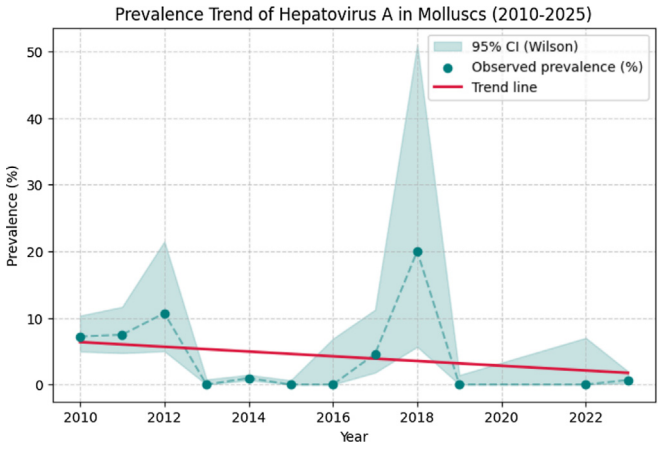

Figure 4. Temporal trend analysis of Hepatovirus A prevalence in molluscs (2010–2025), generated by the Open Chat Mode. The figure displays the observed annual prevalence rates (circles) with 95% confidence intervals and the linear regression trend line.

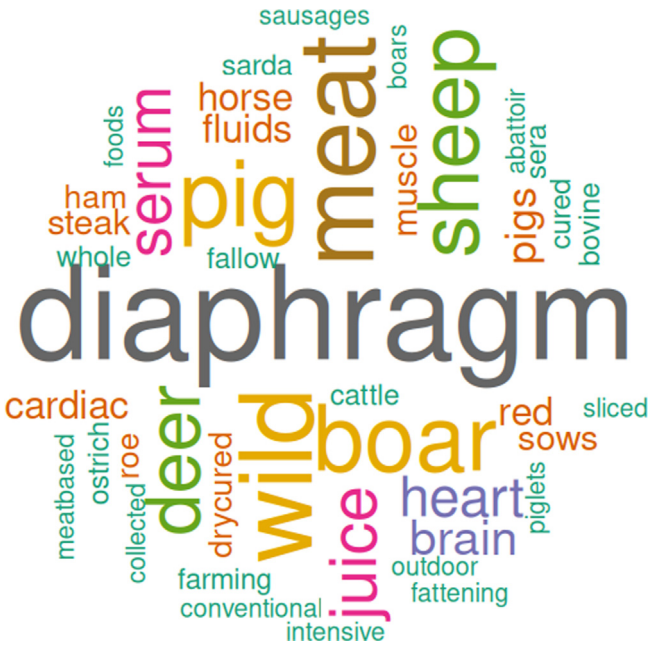

Figure 5. Word cloud of food types.

listing, or dataset summarization (Q2–Q6) did not require argument extraction and were therefore excluded from F1-score computation. These queries were evaluated solely for tool-selection accuracy. Thus, the F1 denominator comprised only the total number of expected canonical field-value arguments across Q1 and Q7–Q10; Q2–Q6 contributed neither true positives, false positives, nor false negatives to the argument extraction F1-score.

Argument-extraction performance was evaluated by comparing the extracted entities against the reference analytical arguments defined in Table 4. Precision was defined as the proportion of extracted entities matching the expected analytical arguments, whereas recall measured the proportion of expected entities successfully extracted from the query. The reported F1-score corresponds to the harmonic mean of precision and recall across all extraction-required benchmark queries.

Entity matching was evaluated at the semantic field level rather than through strict character-level equivalence. Extracted entities were considered correct when they resolved unambiguously to the intended database concept and analytical field, including normalized lexical variants corresponding to the same controlled-vocabulary entry. Minor linguistic variations, pluralization differences, and synonymous expressions were therefore accepted when mapped to the same schema entity.

To reduce ambiguity during extraction, the prompting strategy enforced a deterministic entity-priority hierarchy in which highly specific attributes (e.g., ready-to-eat status, packaging status, sampling stage, or country) were extracted before food-item resolution. Extracted entities were subsequently normalized against the schema definitions stored in the vector database, enabling synonym-aware matching and reducing field-assignment ambiguity. During benchmark evaluation, no observable field-mapping errors were identified.

The benchmark should nevertheless be interpreted as a focused proof-of-concept evaluation designed to assess representative surveillance-oriented analytical intents rather than as a comprehensive stress test of open-domain conversational robustness. More challenging scenarios involving ambiguous phrasing, typographical errors, negation, compositional constraints, synonym variability, and multi-turn interactions remain important directions for future evaluation.

Guided Meta-Analysis Mode was evaluated through an end-to-end proof-of-concept case study on *Toxoplasma* in meat and meat products comprising 153 records from 65 studies. This case was selected because it required execution of the complete analytical workflow, including dataset characterization, prevalence synthesis, subgroup analysis, figure generation, and draft report production.

For the qualitative endpoints of interpretation quality and visual informativeness, outputs were scored independently on a 1–5 scale by three food-safety experts, and mean scores were used for model comparison. Because formal inter-rater reliability statistics were not

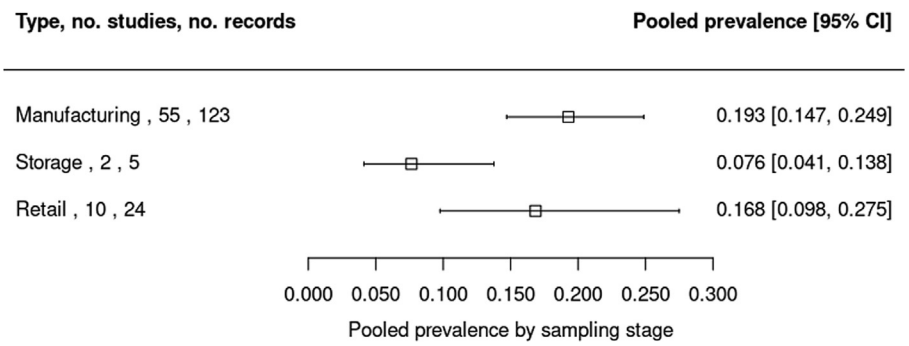

Figure 6. Forest plot of pooled prevalence by sampling stage.

Table 1
Hardware specifications of the experimental testbed

Component	Specification
Operating system	Ubuntu 24.04.3 LTS
CPU	AMD EPYC 9554P (32 Cores)
System memory (RAM)	125.79 GB
GPU	NVIDIA GRID A100D-16C
GPU memory (VRAM)	16 GB
NVIDIA driver version	550.127.05
CUDA version	12.4
Disk space (SSD)	252 GB

estimated, these qualitative scores should be interpreted as structured expert judgment rather than as a validated psychometric scale.

Expert scoring rubric and composite report-quality assessment. For the qualitative endpoints, expert scoring was guided by a structured rubric designed to evaluate the scientific usefulness of model-generated outputs for food-safety interpretation. Interpretation Quality and Coherence (IQC) was assessed across four dimensions: factual accuracy, logical coherence, completeness, and proportionality of interpretation to the underlying evidence. Visual Informativeness (VI) was assessed across three dimensions: graphical clarity, interpretability, and correspondence between the figure and the underlying analytical output. Each dimension was scored on a 1–5 scale, where 1 indicated poor performance and 5 indicated excellent performance. Experts were instructed to score outputs on the basis of their scientific adequacy and practical usefulness for surveillance-oriented evidence synthesis rather than on stylistic preference alone.

Table 3
Curated benchmark questions and their corresponding target tools

ID	Natural-language question	Target tool (intent)
Q1	Load dataset on Salmonella in Portugal	`dataset_load`
Q2	What columns are there in the dataset?	`dataset_list_columns`
Q3	What does the column RTE mean?	`dataset_column_description`
Q4	List unique values of column PackStatus.	`dataset_unique_values`
Q5	How many records for each category?	`dataset_records_by_group`
Q6	Provide a summary of the dataset.	`dataset_summary`
Q7	What is the prevalence of Listeria in cheese in Portugal?	`get_prevalence_by_food`
Q8	What is the average concentration of Staphylococcus in composite dish in Turkey?	`get_concentration_stats`
Q9	Which foods are most associated with Norovirus?	`get_food_sources_for_pathogen`
Q10	Has Hepatovirus A prevalence in molluscs increased from 2010 to 2025?	`trend_analysis`

For each model, mean IQC and VI scores were calculated across raters. To support end-to-end comparison of guided reports, an exploratory Report Quality Index (RQI) was additionally calculated as a secondary summary measure. The RQI combined standardized scores for interpretive quality, numerical concordance, and visual informativeness using a predefined weighting scheme:

Table 2
Evaluation endpoints for the PIF Intelligent Agent

Evaluation layer	Endpoint	Type	Definition/measurement method
Functional correctness (Open Chat Mode)	Tool-selection accuracy	Quantitative	Proportion of queries for which the model selected the correct tool/function
Functional correctness (Open Chat Mode)	Argument-extraction F1-score	Quantitative	Accuracy and completeness of extracted analytical arguments relative to the reference specification
Functional correctness (Open Chat Mode)	Retrieval correctness	Quantitative	Whether the resulting data subset matched the expected subset for the query
Interpretive quality (Open Chat Mode)	Interpretation Quality and Coherence (IQC)	Qualitative	Expert scoring of factual accuracy, logical coherence, completeness, and proportionality of the generated text (1–5 scale)
Interpretive quality (Guided Mode)	Interpretation Quality and Coherence (IQC)	Qualitative	Expert scoring of factual accuracy, logical coherence, completeness, and proportionality of the generated text (1–5 scale)
Analytical fidelity (Both Modes)	Numerical concordance	Quantitative	Agreement between reported numerical results and the deterministic R outputs
Communication quality (Both Modes)	Visual Informativeness (VI)	Qualitative	Expert scoring of clarity, interpretability, and correctness of generated figures (1–5 scale)
Composite report quality (Guided Mode)	Report Quality Index (RQI)	Composite	Exploratory summary index used to compare end-to-end report quality across models

Table 4
Reference analytical arguments expected for the benchmark questions

ID	Expected Entities (JSON argument)
Q1	`agent": Salmonella", country_sampling": Portugal", ...`
Q2	N/A
Q3	N/A
Q4	N/A
Q5	N/A
Q6	N/A
Q7	`agent": Listeria", food_item": cheese", country_sampling": Portugal", ...`
Q8	`agent": Staphylococcus", food_item": composite dish", country_sampling": Turkey", ...`
Q9	`agent": Norovirus", ...`
Q10	`agent": Hepatovirus A", food_item": molluscs", ...`

$$RQI = 0.40 \times IQC_{std} + 0.40 \times NC_{std} + 0.20 \times VI_{std} \quad (1)$$

where IQC_{std} and VI_{std} are the standardized 0–100 versions of the mean expert scores, obtained as:

$$IQC_{std} = 100 \times \frac{IQC_{mean} - 1}{4} \quad (2)$$

$$VI_{std} = 100 \times \frac{VI_{mean} - 1}{4} \quad (3)$$

and NC_{std} is the percentage numerical concordance between the reported values and the deterministic R outputs. This weighting scheme was defined a priori to give equal prominence to interpretive fidelity and numerical correctness, while assigning a lower but still relevant contribution to graphical communication. The RQI was used as an exploratory comparative index and should not be interpreted as a validated psychometric measure.

Statistical analysis of evaluation outcomes. Quantitative evaluation endpoints were summarized as proportions or means, as appropriate. For benchmark-derived endpoints based on discrete query outcomes, results were reported descriptively and interpreted in relation to the fixed benchmark rather than as estimates of general conversational performance. Qualitative expert scores were summarized as mean values across raters for each model and endpoint. The composite RQI was used as a secondary summary measure of end-to-end report performance and was interpreted alongside, rather than instead of, its component metrics.

Given the exploratory scope of the benchmark and the limited number of test queries, inferential comparisons between models were treated cautiously. The evaluation was therefore designed primarily to characterize functional reliability and practical differences in interpretive quality across models, rather than to support a broad population-level claims about model superiority.

Results

Functional performance of evaluated models in Open Chat Mode

Within the evaluated 10-query benchmark, the Open Chat Mode showed that all five models correctly executed the structured steps required for the evaluated benchmark queries. Tool selection accuracy, argument extraction F1-score, and retrieval correctness were each 100% for every model tested (Table 6).

Operationally, this means that each model correctly routed the benchmark questions to the appropriate tool and supplied the required arguments for dataset interrogation and downstream analysis. For example, queries such as “Has *Hepatovirus A* prevalence in molluscs increased from 2010 to 2025?” were consistently mapped to the `trend_analysis` function, while food-pathogen-country combinations were extracted without observable argument drift. Within this benchmark, the retrieval-grounded architecture therefore prevented misrouting and maintained alignment between user request, database subset, and analytical task.

These results should, however, be interpreted as evidence of reliable performance on a constrained benchmark rather than as exhaustive validation of all possible user inputs. The benchmark was intentionally small and representative, designed to test whether the architecture could robustly support canonical surveillance-oriented queries under the evaluated conditions.

Interpretive performance in Open Chat Mode

Uniform functional performance did not translate into uniform scientific usefulness. When outputs were evaluated for IQC, marked differences emerged among models (Table 7). Phi-4 Mini produced the weakest overall performance (mean 2.73), whereas Cogito, Qwen 3, and Gemini 2.5 Pro all achieved mean scores above 4.5, indicating substantially stronger factual precision and summary quality.

The qualitative differences were not trivial. DeepSeek-R1 Tool-Calling hallucinated a prevalence estimate of 4.98% for *Listeria*, where the correct system-derived value was 2.32%. Phi-4 Mini returned a generic template instead of a usable answer for the *Staphylococcus* concentration query. By contrast, Cogito and Qwen 3 handled the *Hepatovirus A* trend question with appropriate restraint, correctly describing the fitted trend as a non significant decrease rather than overstating the result.

Taken together, these findings show that the major source of variability in Open Chat Mode was not tool execution, but interpretive

Table 5
Structured expert scoring rubric used for qualitative evaluation of model-generated interpretations and figures

Domain	Dimension	Score 1	Score 3	Score 5
IQC	Factual accuracy	Contains major factual errors, unsupported claims, or misstates analytical results	Mostly accurate but includes minor imprecision or incomplete linkage to results	Fully accurate, evidence-grounded, and consistent with the analytical outputs
IQC	Logical coherence	Interpretation is disorganized, contradictory, or difficult to follow	Reasonably organized but with some weak transitions or uneven reasoning	Clear, logically structured, and easy to follow throughout
IQC	Completeness	Omits key findings, qualifiers, or relevant analytical context	Covers the main findings but lacks some important detail or qualification	Covers all major findings, key qualifiers, and relevant contextual interpretation
IQC	Proportionality to evidence	Overstates certainty, exaggerates implications, or underplays limitations	Generally balanced but with occasional overstatement or insufficient caution	Appropriately cautious, proportional to the data, and explicit about limitations where relevant
VI	Graphical clarity	Figure is cluttered, unclear, or difficult to read	Figure is usable but has some clarity or formatting limitations	Figure is clear, readable, and effectively communicates the intended result
VI	Interpretability	Figure requires substantial effort to understand or lacks clear meaning	Figure is interpretable but not maximally informative	Figure is immediately interpretable and supports analytical understanding
VI	Correspondence to analytical output	Figure is inconsistent with the reported result or misleading	Figure is broadly consistent but could be more precisely aligned with the result	Figure accurately and faithfully represents the underlying analytical output

Table 6
Core functional performance metrics for the five evaluated models in Open Chat Mode. Scores of 100% were obtained for all models within the evaluated benchmark. Argument extraction F1-score was calculated only for Q1 and Q7-Q10; Q2-Q6 were N/A for argument extraction and excluded from F1 averaging

Metric	Phi-4 Mini (3.8B)	DeepSeek-R1 tool-calling (14B)	Cogito (14B)	Qwen 3 (8B)	Gemini 2.5 Pro
Tool-selection accuracy	100%	100%	100%	100%	100%
Argument-extraction F1-score	100%	100%	100%	100%	100%
Retrieval correctness	100%	100%	100%	100%	100%

Table 7
Mean IQC scores by question for each evaluated model in Open Chat Mode (1–5 scale)

Question	Phi-4 Mini (3.8B)	DeepSeek-R1 tool-calling (14B)	Cogito (14B)	Qwen 3 (8B)	Gemini 2.5 Pro
"Provide a summary of the dataset."	2.67	3.67	4.67	4.33	4.67
"What is the prevalence of Listeria…?"	3.00	1.67	4.33	4.67	4.67
"What is the average concentration…?"	1.33	4.67	4.67	4.33	4.67
"Which foods are most associated…?"	3.33	4.00	4.67	4.33	4.67
"Has Hepatovirus A prevalence increased…?"	3.33	4.33	5.00	5.00	4.33
Overall Mean	**2.73**	**3.67**	**4.67**	**4.53**	**4.60**
Standard Deviation	**0.83**	**1.18**	**0.24**	**0.30**	**0.15**

fidelity. Once the retrieval and analysis stages were successfully grounded, the practical differentiator became whether the model could accurately verbalize what the computed output meant. Representative high- and low-quality excerpts are provided in Supplementary Material S2.

Guided meta-analysis mode

The Guided Meta-Analysis Mode was evaluated through an end-to-end case study on *Toxoplasma* in meat and meat products, comprising 153 records from 65 studies. This use case was selected because it required the system to execute a complete evidence-synthesis workflow, including dataset characterization, prevalence exploration, subgroup analysis, figure generation, and narrative reporting. The results presented in this section should therefore be interpreted as evidence from a specific application scenario rather than as a general validation of the approach across all meta-analytical contexts.

Across all models, the deterministic analytical layer produced a numerical concordance score of 100%. In practical terms, all reported statistical values in the generated reports—including pooled prevalence estimates, heterogeneity statistics (I^2, τ^2), and subgroup summaries—matched the manually verified outputs of the R pipeline exactly. Within this case study, the numerical backbone of the report remained fixed and reproducible despite variation in narrative generation across models.

The complete automatically generated report is provided as Supplementary Material S1.

Visual output quality. The figures generated in the guided report were rated highly overall, with a mean Visual Informativeness (VI) score of 4.83/5 (Table 8). Experts judged most visualizations to be clear, scientifically conventional, and appropriate for rapid interpretation of the evidence structure and meta-analytical outputs within this use case.

Exploratory figures such as the word clouds received the highest ratings because they provided an immediate overview of food types and assay categories represented in the dataset. Forest plots stratified by sampling stage, packaging status, temperature, and ready-to-eat status were also consistently scored highly, indicating that the system could produce interpretable analytical graphics beyond simple descriptive displays in this application context. The only visibly lower score was for the pooled prevalence analysis by year, suggesting that temporal synthesis plots may benefit from further refinement in layout or annotation.

Table 8
VI ratings for figures generated in the automated report (*Supplementary Material S1*), averaged across three expert evaluators

Figure in supplementary report	Mean rating (1–5)
Figure 1: Word cloud of food type	5.00
Figure 2: Word cloud of assay type	5.00
Figure 3: Pooled prevalence analysis by year	4.00
Figure 4: Pooled prevalence analysis by sampling stage	5.00
Figure 5: Pooled prevalence analysis by pack status	5.00
Figure 6: Pooled prevalence analysis by temperature of sampling	5.00
Figure 7: Pooled prevalence analysis by ready-to-eat status	5.00
Figures 8, 11, 14, 17: Number of records (bar plots)	5.00
Figures 9, 12, 15, 18: Prevalence distribution (ridge plots)	4.50
Figures 10, 13, 16, 19: Funnel plots	4.83
Overall Mean	**4.83**

Interpretation quality and coherence in guided mode. As in Open Chat Mode, the main variability in Guided Mode arose in the written interpretation of otherwise stable analytical outputs. IQC scores differed across both models and report sections (Table 9). Phi-4 Mini showed the weakest overall performance (mean 2.75), whereas Qwen 3 achieved the highest overall score (4.37), followed closely by Cogito and Gemini 2.5 Pro (both 4.23).

The largest differences appeared in sections that required synthesis across multiple subgroup meta-analyses, where lower-performing models tended to become verbose, numerically loose, or stylistically awkward. Higher-performing models were better at preserving magnitude, direction, and caution when describing heterogeneity, subgroup contrasts, and publication-bias diagnostics. Within this case study, these findings suggest that the statistical pipeline remained stable, whereas the narrative layer was more sensitive to model choice.

Representative excerpts from both stronger and weaker guided reports are supplied in Supplementary Material S2.

Overall report quality. The composite Report Quality Index (RQI), which integrates interpretive quality, numerical concordance, and visual informativeness, placed Qwen 3 highest at 93%, with Cogito and Gemini 2.5 Pro close behind at 91% (Table 10). Phi-4 Mini and DeepSeek-R1 Tool-Calling lagged primarily because a weaker narrative interpretation reduced overall report quality despite the shared deterministic analytical backbone.

Taken together, the Guided Mode results demonstrate the feasibility of producing statistically correct and visually strong draft reports

Table 9
Mean IQC scores by report section for each evaluated model in guided meta-analysis mode

Section	Phi-4 Mini (3.8B)	DeepSeek-R1 tool-calling (14B)	Cogito (14B)	Qwen 3 (8B)	Gemini 2.5 Pro
1. Dataset Overview	2.67	3.67	4.33	4.00	4.58
2.2 Prevalence by sampling stage	3.67	3.17	4.17	4.50	4.17
2.3 Prevalence by pack status	3.33	4.07	4.83	3.77	4.17
2.4 Prevalence by temperature	2.60	4.50	4.17	4.50	4.17
2.5 Prevalence by RTE Status	3.07	4.17	4.50	4.50	4.17
2.6 Prevalence by country	2.43	4.17	4.83	4.17	3.83
3.1 Meta-analysis (food subcategory)	2.00	2.67	4.50	4.17	4.17
3.2 Meta-analysis (food class)	2.50	2.77	4.13	4.73	4.17
3.3 Meta-analysis (organ sampled)	2.67	2.73	3.47	4.50	4.58
3.4 Meta-analysis (assay detection)	1.50	2.83	3.50	4.83	4.25
3.5 Meta-analysis (mean concentration)	3.77	4.07	4.10	4.43	4.25
Overall Mean	**2.75**	**3.53**	**4.23**	**4.37**	**4.23**

Table 10
Final standardized component scores and computed RQI for the Guided Mode case study for each of the five evaluated models

Metric/index	Phi-4 Mini (3.8B)	DeepSeek-R1 tool-calling (14B)	Cogito (14B)	Qwen 3 (8B)	Gemini 2.5 Pro
Numerical concordance, NC_{std}	100.00	100.00	100.00	100.00	100.00
Visual Informativeness, VI_{std}	95.75	95.75	95.75	95.75	95.75
Interpretation Quality, IQC_{std}	43.75	63.25	80.75	84.25	80.75
Report Quality Index (RQI)	**76.65%**	**84.45%**	**91.45%**	**92.85%**	**91.45%**

with limited user intervention in this specific *Toxoplasma* case study. However, the findings also indicate that scientific usefulness remained dependent on human review of the narrative layer. Broader conclusions regarding robustness, generalizability, and performance across different pathogens and food categories scenarios will require evaluation in future multicase studies.

Discussion

Relevance for food-safety surveillance and risk assessment. The principal contribution of this study is practical rather than purely technical. Curated occurrence databases have clear value for food protection only when analysts can retrieve, summarize, and statistically interpret the underlying evidence without excessive technical overhead. In that sense, the PIF Intelligent Agent is not best understood as a generic conversational system, but as an analyst-facing interface for grounded evidence synthesis. By converting natural-language requests into constrained database operations and delegating quantitative work to a deterministic R backend, the system supports a workflow that is directly relevant to surveillance, rapid evidence appraisal, and risk assessment.

Read alongside the original PIF publication, this article represents a downstream extension of an already curated evidence infrastructure. The 2025 PIF paper addressed the upstream challenge of assembling harmonized hazard-in-food occurrence data under FAIR principles and explicit quality controls (Gonzales-Barron et al., 2025; Wilkinson et al., 2016). That upstream effort is methodologically aligned with the group's earlier occurrence meta-analyses in meats, poultry, produce, dairy matrices, and parasitic–hazard assessment, which showed both the utility of pooled prevalence synthesis and the recurring difficulty of reusing heterogeneous primary data (Xavier et al., 2014; Gonçalves-Tenório et al., 2018; Silva et al., 2017; Gonzales-Barron et al., 2017; Thebault et al., 2021). The present study addresses the next operational challenge: making that curated evidence accessible through low-friction analytical interaction. This distinction is important because food-safety infrastructures often fail not because evidence is absent, but because the path from data availability to usable analysis remains too technically demanding for routine practice.

Why the hybrid architecture mattered. Within the scope of this proof-of-concept evaluation, the results support a clear design principle for AI-assisted scientific analytics in food safety: language models are most useful when they interpret user intent and draft narrative output, whereas deterministic software remains responsible for the quantitative analysis itself. In the present study, all five evaluated models achieved perfect functional performance on the 10-query Open Chat benchmark, and the Guided Mode achieved perfect numerical concordance in the *Toxoplasma* case study. Although these results derive from a limited benchmark, they indicate that the critical safeguard was not model size per se, but the architectural separation between language interpretation and statistical computation. For food-protection applications, where numerical reproducibility is essential, that separation is likely to be more important than marginal differences in conversational fluency.

Interpretive fidelity, not tool execution, differentiated models. A central result of the study is that correct tool execution did not guarantee equally useful scientific communication. All tested models retrieved the correct evidence subsets and triggered the appropriate analytical functions, yet their written outputs differed substantially in factual precision, coherence, and rhetorical economy. This distinction has practical consequences. In a food-safety setting, a system that computes the right prevalence estimate but states it incorrectly, overinterprets a non-significant trend, or produces a diffuse summary still creates an interpretive burden for the user. The strongest-performing models were therefore not merely those that completed the workflow, but those that remained close to the computed evidence and expressed conclusions in a restrained scientific register.

Implications for local deployment. The performance of Qwen 3 (8B) is operationally encouraging because it suggests that smaller, locally deployable models may be sufficient for tightly structured evidence-synthesis tasks in food safety. This matters for institutions that handle sensitive surveillance data, operate under constrained budgets, or prefer on-premises analytical infrastructures. The present findings do not imply that larger proprietary models lack value; rather, they indicate that open-weight models can be competitive when the task is retrieval-grounded, domain-bounded, and supported by deterministic analytical tools. For many applied food-safety settings, that may be the more realistic deployment scenario.

Limits of the current evidence. The findings should be interpreted as strong proof-of-concept evidence, not as comprehensive validation. First, the Open Chat benchmark contained only 10

representative queries, which is sufficient to demonstrate feasibility but not to characterize the full distribution of user inputs likely to arise in practice. Second, the Guided Mode was assessed using a single case study on *Toxoplasma* in meat products, so generalizability across hazards, matrices, and analytical contexts remains to be demonstrated. Third, although three food-safety experts evaluated qualitative outputs, formal inter-rater reliability statistics was not estimated. Fourth, the study did not include direct usability testing with intended end users, nor did it compare latency, computational cost, or task performance against existing manual workflows. Finally, the observed performance depends partly on the current structure of the PIF database, the curated schema exposed to the agent, and the prompt design used to mediate tool calling. These dependencies do not weaken the proof of concept, but they do define the boundaries of the present evidence.

Implications for future work. Future work should proceed in three directions. The first is broader technical validation using larger benchmark sets, less canonical question formulations, and a wider range of hazard-food combinations. The second is user-centered evaluation with microbiologists, food safety authorities, and risk assessors performing realistic tasks under time constraints. The third is analytical expansion, including additional evidence synthesis workflows and more explicit communication of uncertainty in generated reports. In methodological terms, this should build on the more general systematic review and meta-analysis strategy already developed for foodborne disease evidence integration (Gonzales-Barron et al., 2021) and on pathogen-specific syntheses that demonstrate how heterogeneous food safety evidence can be pooled across matrices, study designs, and exposure pathways (Xavier et al., 2014; Gonçalves-Tenório et al., 2018; Gonzales-Barron et al., 2017; Thebault et al., 2021). As the PIF platform itself continues to expand in scope and curation depth (Gonzales-Barron et al., 2025), the intelligent-agent layer should be tested against the newly incorporated evidence domains and more demanding cross-domain queries. These next steps will determine whether the system improves not only technical correctness, but also the quality and speed of real-world decision support in food protection.

Conclusions

A retrieval-augmented natural-language interface was developed for the PIF database and evaluated in two complementary modes: conversational querying and guided meta-analysis. Within the evaluated benchmark, all tested models achieved 100% tool-selection accuracy, argument-extraction F1-score, and retrieval correctness, indicating consistent grounding of database operations within the evaluated benchmark. In a guided case study on *Toxoplasma* in meat and meat products, the system achieved 100% numerical concordance and a highest Report Quality Index of 93% with Qwen 3 (8B).

Taken together, these results support hybrid architectures that separate language interpretation from deterministic statistical computation when deploying AI-assisted analytics in food safety. They also show that correct tool execution is not, by itself, a sufficient benchmark for scientific usefulness: interpretive fidelity remains the more discriminating criterion. Broader validation, formal usability testing, and additional food-safety case studies are now needed before routine operational deployment can be recommended.

Funding

This work was funded by the European Food Safety Authority (EFSA) through GP/EFSA/BIOHAW/2022/01 and GP/EFSA/BIOHAW/2023/05. The positions and opinions presented in this study are those of the authors and are not intended to represent the views or scientific works of EFSA.

Data availability

The software code for the PIF Intelligent Agent is available in the project repository: https://github.com/fsqanalytics/PIFChatBot-public. The data backbone used in this study is the Pathogens-in-Foods database described by Gonzales-Barron et al. (2025): https://github.com/fsqanalytics/PIFDataModel. Public information about the database and access portal is available through the official PIF website.

The PIFChatBot interface can be accessed through the main platform at: https://pathogens-in-food.vercel.app/, via the navigation path *Main Menu → Dashboard → AI PIF Explorer*. Access to this feature is restricted to users with an *advanced* role, which must be requested during account creation.

Access to the curated database used in this study may be subject to project governance and data-sharing conditions. The automatically generated case-study report is available as Supplementary Material S1, and representative text-output examples are provided as Supplementary Material S2.

CRediT authorship contribution statement

Ursula Gonzales-Barron: Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.**Lucas Ribeiro Silva**: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Vasco Cadavez**: Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the development context of the PIF database initiative and the expert input used to assess the quality of automated outputs in this study.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at: https://doi.org/10.1016/j.jfp.2026.100829.

References

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3). https://doi.org/10.1145/3641289.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 6491–6501). ACM.

Gonzales-Barron, U., & Butler, F. (2011). The use of meta-analytical tools in risk assessment for food safety. *Food Microbiology, 28*(4), 823–827. https://doi.org/10.1016/j.fm.2010.04.007.

Gonzales-Barron, U., Gonçalves-Tenório, A., Rodrigues, V., & Cadavez, V. (2017). Foodborne pathogens in raw milk and cheese of sheep and goat origin: A meta-analysis approach. *Current Opinion in Food Science, 18*, 7–13. https://doi.org/10.1016/j.cofs.2017.10.002.

Gonzales-Barron, U., Thébault, A., Kooh, P., Watier, L., Sanaa, M., & Cadavez, V. (2021). Strategy for systematic review of observational studies and meta-analysis modelling of risk factors for sporadic foodborne diseases. *Microbial Risk Analysis, 17*, 100082. https://doi.org/10.1016/j.mran.2019.07.003.

Gonzales-Barron, U., Faria, A. S., Thebault, A., Guillier, L., Mendes, L. R., Silva, L. R., Messens, W., Kooh, P., & Cadavez, V. (2025). Pathogens-in-Foods (PIF): An open-

access European database of occurrence data of biological hazards in foods. *Microbial Risk Analysis, 29*, 100342. https://doi.org/10.1016/j.mran.2025.100342.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751. https://doi.org/10.48550/arXiv.1904.09751..

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Generating Evidence & Methods to Improve Patient Outcomes), 4*(1), 1244. https://doi.org/10.13063/2327-9214.1244.

Katsogiannis-Meimarakis, G., Xydas, M., & Koutrika, G. (2023). Natural language interfaces for databases with deep learning. *Proceedings of the VLDB Endowment, 16* (12), 3878–3881. https://doi.org/10.14778/3611540.3611575.

Klesel, M., & Wittmann, H. F. (2025). Retrieval-augmented generation (RAG). *Business & Information Systems Engineering, 1–11*. https://doi.org/10.1007/s12599-024-00857-5.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71.

Silva, B. N., Cadavez, V., Teixeira, J. A., & Gonzales-Barron, U. (2017). Meta-analysis of the incidence of foodborne pathogens in vegetables and fruits from retail establishments in Europe. *Current Opinion in Food Science, 18*, 21–28. https://doi.org/10.1016/j.cofs.2017.10.001.

Gonçalves-Tenório, A., Silva, B. N., Rodrigues, V., Cadavez, V., & Gonzales-Barron, U. (2018). Prevalence of pathogens in poultry meat: A meta-analysis of European published surveys. *Foods, 7*(5), 69. https://doi.org/10.3390/foods7050069.

Thebault, A., Kooh, P., Cadavez, V., Gonzales-Barron, U., & Villena, I. (2021). Risk factors for sporadic toxoplasmosis: A systematic review and meta-analysis. *Microbial Risk Analysis, 17*, 100133. https://doi.org/10.1016/j.mran.2020.100133.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.

World Health Organization. (2020). Food safety. Accessed September 9, 2025. https://www.who.int/news-room/fact-sheets/detail/food-safety..

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Xavier, C., Gonzales-Barron, U., Paula, V., Estevinho, L., & Cadavez, V. (2014). Meta-analysis of the incidence of foodborne pathogens in Portuguese meats and their products. *Food Research International, 55*, 311–323. https://doi.org/10.1016/j.foodres.2013.11.024.