FoodSafeSum: Enabling Natural Language Processing Applications for Food Safety Document Summarization and Analysis

Juli Bakagianni[1], Korbinian Randl[2*], Guido Rocchietti[3,4*], Cosimo Rulli[3], Franco Maria Nardini[3], Aron Henriksson[2], Salvatore Trani[3], Anna Romanova[5], John Pavlopoulos[1,2,6†]

[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Department of Computer and Systems Sciences, Stockholm University, Sweden
[3]ISTI-CNR, Italy [4]Department of Computer Science, University of Pisa, Italy
[5]SGS Digicomply, Switzerland [6]Archimedes, Athena Research Center, Greece

Abstract

Food safety demands timely detection, regulation, and public communication, yet the lack of structured datasets hinders Natural Language Processing (NLP) research. We present and release a new dataset of human-written and Large Language Model (LLM)-generated summaries of food safety documents, plus food safety related metadata. We evaluate its utility on three NLP tasks directly reflecting food safety practices: multilabel classification for organizing documents into domain-specific categories; document retrieval for accessing regulatory and scientific evidence; and question answering via retrieval-augmented generation that improves factual accuracy. We show that LLM summaries perform comparably or better than human ones across tasks. We also demonstrate clustering of summaries for event tracking and compliance monitoring. This dataset enables NLP applications that support core food safety practices, including the organization of regulatory and scientific evidence, monitoring of compliance issues, and communication of risks to the public.

1 Introduction

The increasing prevalence of food safety incidents has recently attracted the interest of researchers, including the Natural Language Processing (NLP) community. Understanding the causes of these incidents and addressing contamination (Zhou et al., 2020) is a pressing challenge, prompting initiatives that are reshaping the landscape of food safety research.[1] Additionally, governments worldwide have established regulatory frameworks to oversee the production, distribution, and consumption of food products, enabling the development of compliance-checking systems (Hassani, 2024).

Food safety-related information, however, is fragmented across multiple sources, including regulatory reports, scientific publications, news articles, and government guidelines, making it challenging to extract, analyse, and act upon critical insights efficiently. The rapid growth of these textual resources presents an urgent need for automated, scalable NLP solutions to facilitate information retrieval, risk assessment, and decision-making in food safety (Goldberg et al., 2022). Key attributes such as the topic of a document (e.g., contaminants, labelling, policies) or the hazard it reports (e.g., pathogens, residues, food fraud) are essential for structuring this information, yet, progress is hindered by the lack of publicly available and machine-actionable datasets that provide such annotations for NLP research. While exceptions exist (Randl et al., 2024; Goldberg et al., 2022), we are still missing a comprehensive resource that integrates diverse sources of food safety information.

To this end, we introduce FOODSAFESUM, **a novel dataset** that contains manual and Large Language Model (LLM)-generated summaries of food safety-related documents collected from diverse sources. The dataset includes manually annotated metadata, such as relevant topics and document types, and automatically extracted hazard annotations, enabling structured analysis of food safety risks. Using this dataset, we experiment with **three core NLP tasks** that mirror real-world food safety needs (examples shown in Table 1): 1) multilabel classification (MLC), where food safety-specific topics (e.g., contaminants, sustainability, methods and manufacturing) reflect how professionals categorize and monitor documents; 2) information retrieval (IR), which supports practitioners in efficiently locating relevant reports from summaries or titles, and extends to topic-based retrieval aligned with food safety taxonomies; and 3) question answering (QA) using a retrieval augmented generation (RAG) approach, designed to address practical

*Equal contribution.
†Corresponding author:annis@aueb.gr
[1]See for example the EU-funded EFRA Project.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16786–16804
November 4-9, 2025

Task	Example Input	Expected Output
MLC	Summary: The European FCM Working Group's meeting protocol outlines the way forward to amend the FCM plastic Regulation (EU) No. 10/2011, including the upcoming 16th amendment, indicating future developments in food contact materials regulation.	Topics: {"Contaminants, residues and contact materials", "Sustainability", "Policies and Laws"}
IR	Summary: Following an increase in walnut-related allergy cases, the Japanese Customer Affairs Agency is implementing mandatory labelling of walnuts, transitioning from the current voluntary practice, to ensure consumer safety and awareness.	Title of Retrieved Document: "Walnut to become mandatory allergen declaration in Japan"
QA	Query: When should the new methodology to support harmonisation of pesticide use in EU Member States be available? a) By the end of 2021, b) By the end of 2022, c) In the second half of 2021, d) In January 2020	Correct Answer: "b) By the end of 2022"

Table 1: Examples of the food safety NLP tasks of MLC, IR, and QA.

queries in the domain, such as identifying the cause of contamination, checking compliance with regulations, or clarifying risks for public health communication. These experiments demonstrate that automatically generated summaries can perform comparably to or even better than human-written ones in supporting downstream tasks.

In addition, we develop **an exploratory use case** of text clustering, which groups related events across document types (e.g., legislative updates appearing in both regulations and news). This enables applications such as event tracking, compliance monitoring, and early-warning systems for emerging food safety threats.

2 Related Work

To date, publicly available datasets for NLP research in the food safety domain remain scarce. Several agencies, including the U.S. Food Safety and Inspection Service (FSIS)[2] and the European Food Safety Authority (EFSA),[3] provide diverse food safety data — such as monitoring results, occurrence data, and exposure assessments — that are primarily intended for regulatory and scientific use. While these datasets are valuable, their formats often require adaptation for NLP applications (e.g., text extraction, annotation), underscoring the need for dedicated NLP-focused food safety resources. To our knowledge, the only publicly available dataset specifically designed to advance NLP research in the food safety domain is the *Food Recall Incidents* dataset (Randl et al., 2024, 2025). This dataset comprises 7,546 recall announcement titles, along with associated metadata (e.g., publication dates) and manual annotations of hazards and products at both coarse and fine granularity.

Recent studies have begun to leverage NLP to address various food safety challenges (Song and Pei, 2024; Hu et al., 2022; Goldberg et al., 2022). An early study by Zhang and El-Gohary (2016) employed rule-based *information extraction* from regulatory documents to automate compliance checking. Maharana et al. (2019); Tao et al. (2023) tackled *food risk report detection*, focusing on unsafe food reports from Amazon reviews and Twitter data, respectively. In addition, Xia et al. (2022) leveraged lexicon-based *sentiment analysis* and LDA-based *topic modelling* on social media to analyse public opinions on food safety, while Xiong et al. (2023) performed *food safety news classification* into four distinct food safety event categories.

Summarization plays a crucial role in extracting key insights from lengthy and complex food safety documents. Prior research has explored summarization across various domains, including legal and medical texts, using extractive, abstractive, and hybrid techniques (Gupta and Lehal, 2010; Supriyono et al., 2024). Large-scale datasets, such as *BillSum* (Kornilova and Eidelman, 2019) and *EUR-Lex-Sum* (Aumiller et al., 2022), have facilitated advancements in legal summarization, while models incorporating RAG and domain-aware adaptations have improved summary quality and relevance (Hou et al., 2024; Edge et al., 2024).

3 The FOODSAFESUM Dataset

FOODSAFESUM is a structured dataset derived from 2,091 food safety-related documents collected

[2] https://www.fsis.usda.gov/
[3] https://www.efsa.europa.eu/en

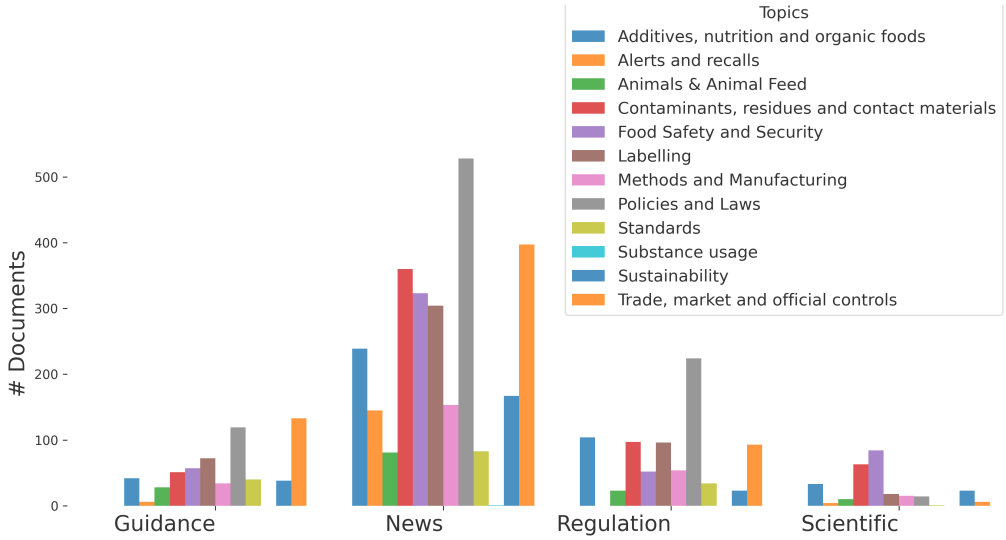


Figure 1: Topic distribution per source type.

by SGS Digicomply[4] on May 24, 2024. These documents are sourced from news platforms, legal and regulatory agencies, food safety authorities and organizations, government guidance portals, scientific journals, and research platforms, where SGS food safety experts continuously monitored them and retrieved content relevant to food safety hazards. They span the years 2002–2023, with approximately 65.7% originally in English and the remainder in 33 other languages. Non-English documents were automatically translated using Google Translate or DeepL and subsequently curated by SGS experts.

The documents were manually annotated by SGS experts with document type, covering news articles, regulations, governmental guidance, and scientific articles, as well as with one or more relevant topics. The topics span 12 high-level categories that reflect key thematic areas in food safety, including regulation and policy, scientific and technical aspects, consumer-focused issues, and broader systemic concerns. The hazard annotations were automatically extracted using a controlled vocabulary derived from the Food Recall Incidents dataset (Randl et al., 2024). Within the food safety domain, a hazard refers to any substance, object, or type of non-compliance detected or reported in a post.[5] This includes biological agents (e.g., Salmonella), chemical substances (e.g., pesticide residues, allergens), physical contaminants (e.g., metal fragments), regulatory violations (e.g., exceeding permitted levels of additives), and food fraud. These metadata enable structured analyses of food safety risks and support downstream NLP tasks.

FOODSAFESUM also contains multiple document versions: manual summaries (MANUAL-S) and titles (MANUAL-T) created by SGS experts, as well as LLM-generated summaries (LL70B-S) and titles (LL70B-T) using the instruction-tuned `meta.llama3-70b-instruct-v1:0` model (via Amazon Bedrock), where S and T denote summary and title versions, respectively. In summary, the publicly released dataset provides for each document: (i) manual annotations of document type and topics, (ii) automatically extracted hazard annotations, (iii) manual summaries and titles, and (iv) LLM-generated summaries and titles. The original full documents (D) are included for internal experiments but are not publicly released. The dataset is publicly available on Zenodo.[6]

3.1 Dataset Statistics

The dataset exhibits an imbalanced distribution across document types. Most documents are news articles (1,320), followed by regulations (352) and governmental guidance (288), while scientific articles are the least represented (131).

Similarly, the topic distribution is imbalanced, with samples in the smallest class amounting to 16% of the size of the largest class "Policies and laws," which is the underlying topic for 44% of the documents (see Appendix A.1) .

An in-depth examination of the topic distribution

[4] https://www.digicomply.com/

[5] https://www.fao.org/4/w5975e/w5975e07.htm

[6] https://doi.org/10.5281/zenodo.17140845

by document type is presented in Figure 1, showing imbalances within each document type. Except from "Alerts and Recalls", which is absent in regulations, and "Substance Usage", which appears in only one news article, the remaining topics are discussed across all document types. "Policies and Laws" and "Trade, Market, and Official Controls" are dominant, except in scientific articles, which are predominantly focused on topics such as "Food Safety and Security" and "Contaminants, residues and contact materials".

A document can be assigned to multiple topics. All topics co-occur except "Standards" and "Alerts and Recalls", which never appear together, likely because the former focuses on prevention, while the latter addresses immediate risks. "Labelling" most often appears with "Policies and Laws", reflecting labelling on a regulatory basis, or individually indicating labelling practices without broader policy discussions.[7]

We also performed an analysis of the hazards mentioned in the data. We extracted information about hazard names, as defined in Randl et al. (2024). Using exact matching, we identified 85 of the 128 hazards in the list.[8]. Figure 8 in the Appendix presents a lower-triangular heatmap showing the Spearman correlation between document types based on the frequencies of hazard mentions. Overall, all correlations are strongly positive, ranging from 0.73 to 0.87, suggesting a consistent distribution of hazard mentions across document types. The highest correlation occurs between "Regulations" and "Guidance" (0.87), indicating a strong similarity in their hazard mention frequency distributions, given their similar normative role and target audiences. In contrast, the lowest correlation (0.73) is observed between "Scientific" and "News" articles, possibly due to differing content priorities—scientific articles focus on specialized or emerging hazards, whereas news tends to highlight high-profile or public-interest topics.

3.2 Text Statistics

We analyse all document versions (D, MANUAL-S, LL70B-S, MANUAL-T, LL70B-T) to compare expert-provided and LLM-generated content, and to assess whether the compressed versions (S and T) can support the downstream tasks as effectively

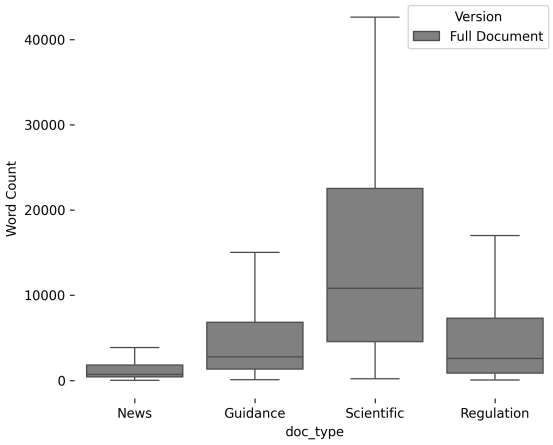


Figure 2: Boxplots of word counts in full documents across different document types.

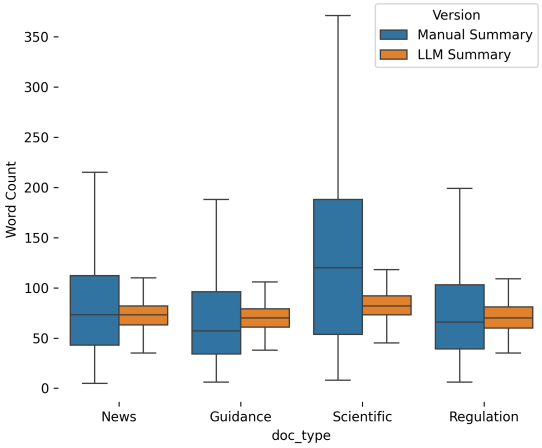


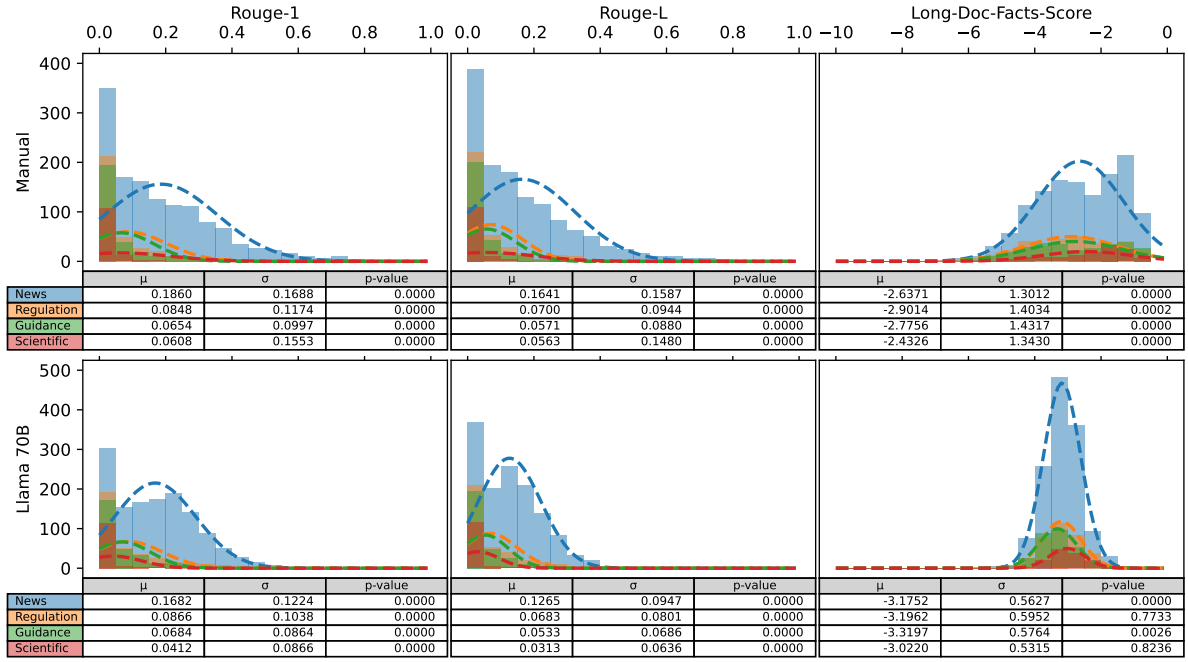
Figure 3: Boxplots of word counts in summaries across different document types.

as the . For the Llama3-70b-generated content, we truncated texts exceeding 7,500 tokens to fit within model's 8,192-token context length, leaving space for the instruction prompt (see Table 8 in the Appendix), and set the maximum output length to 1,024 tokens. Figure 2 shows the word counts for all document versions.

Scientific documents are longer than the other types. Figure 3 shows word counts of summaries, which are an order of magnitude shorter. MANUAL-S (in blue) tend to be lengthier compared to LL70B-S. Notably, however, the LLM-generated versions are more standardized and less influenced by the variability inherent in the original texts; i.e., from brief news articles to extensive scientific reports.

[7] A heatmap presenting co-occurrences can be found in Appendix A.2. Hence, an analysis of their co-occurrence can reveal topic associations.

[8] We did not include the hazards "other" and "processing"



Manual	Rouge-1 μ	Rouge-1 σ	Rouge-1 p-value	Rouge-L μ	Rouge-L σ	Rouge-L p-value	Long-Doc-Facts-Score μ	Long-Doc-Facts-Score σ	Long-Doc-Facts-Score p-value
News	0.1860	0.1688	0.0000	0.1641	0.1587	0.0000	-2.6371	1.3012	0.0000
Regulation	0.0848	0.1174	0.0000	0.0700	0.0944	0.0000	-2.9014	1.4034	0.0002
Guidance	0.0654	0.0997	0.0000	0.0571	0.0880	0.0000	-2.7756	1.4317	0.0000
Scientific	0.0608	0.1553	0.0000	0.0563	0.1480	0.0000	-2.4326	1.3430	0.0000

Llama 70B	Rouge-1 μ	Rouge-1 σ	Rouge-1 p-value	Rouge-L μ	Rouge-L σ	Rouge-L p-value	Long-Doc-Facts-Score μ	Long-Doc-Facts-Score σ	Long-Doc-Facts-Score p-value
News	0.1682	0.1224	0.0000	0.1265	0.0947	0.0000	-3.1752	0.5627	0.0000
Regulation	0.0866	0.1038	0.0000	0.0683	0.0801	0.0000	-3.1962	0.5952	0.7733
Guidance	0.0684	0.0864	0.0000	0.0533	0.0686	0.0000	-3.3197	0.5764	0.0026
Scientific	0.0412	0.0866	0.0000	0.0313	0.0636	0.0000	-3.0220	0.5315	0.8236

Figure 4: Similarity of summaries with the original documents according to established metrics: ROUGE-1 and ROUGE-L (F-measure) (Lin, 2004) and Long-Doc-Fact-Score (Bishop et al., 2024). Although almost none of the distributions are strictly normal according to a Shapiro-Wilk test (see p-values), mean μ and standard deviation σ can still give a rough estimation of the distributions as shown in the plots.

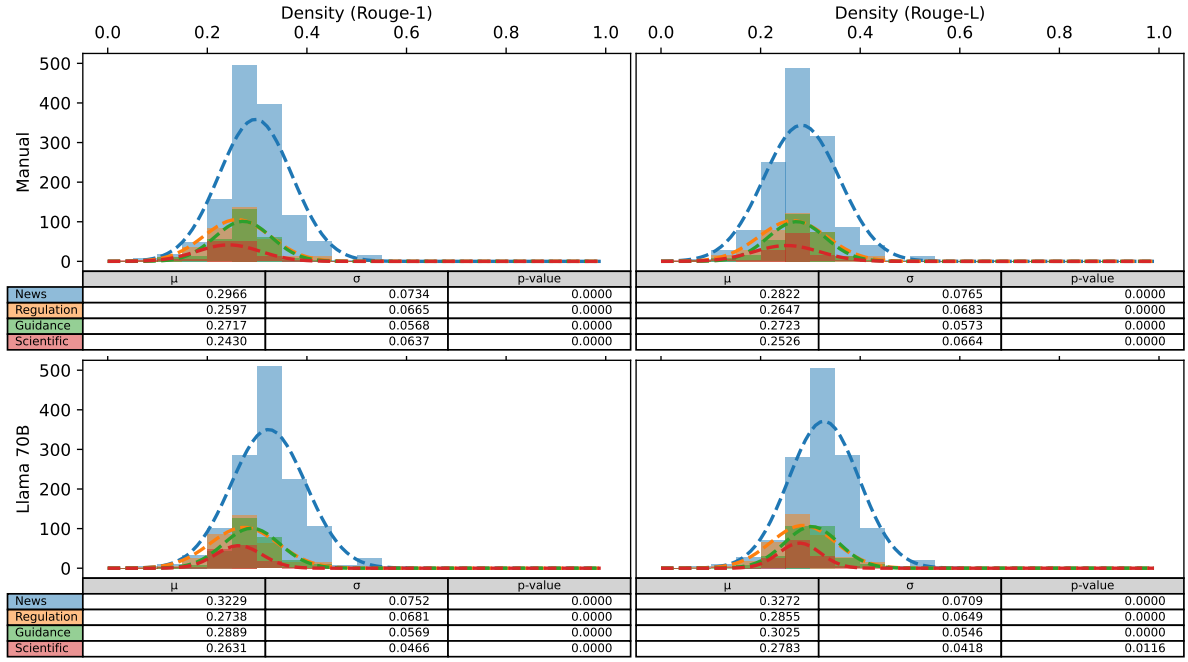
4 Summarization Analysis

4.1 Traditional Summarization Quality

Although it may disregard synonymy and abstraction, we report (see Figure 9 in the Appendix) ROUGE-1 (token overlap) and ROUGE-L (longest subsequence matching) in their F-measure variants for completeness and comparability (Lin, 2004). Comparing LL70B-S to the MANUAL-S "ground truth," we see ROUGE-1 normally distributed (Shapiro-Wilk test at $\alpha = 0.05$) around 0.41 to 0.42 for "News," "Regulations," and "Guidance". LL70B-S for "Scientific" documents also achieve a mean ROUGE-1 of 0.42 but the scores are not strictly normally distributed. The mean ROUGE-L values for all types are between 0.28 and 0.30, but not normally distributed. Despite not following a normal distribution, the mean values represent their respective populations. These scores suggest a moderate information overlap between LL70B-S and MANUAL-S.

A more reliable metric is the Long-Doc-Fact-Score (LDFScore) introduced by Bishop et al. (2024). Based on dense representations (Yuan et al., 2021, BARTScore) and passage sampling, LDFScore is particularly suited for evaluating abstractive summaries of long documents like ours. Notably, LDFScore is defined on a logarithmic scale, where $-\infty$ indicates no similarity and 0 denotes complete similarity. LDFScore provides a more nuanced perspective than ROUGE: while overall similarity remains low, Figure 4 reveals that MANUAL-S exhibit higher variance than LL70B-S. Furthermore, the LDFScore distribution for MANUAL-S compared to D shows two peaks: one aligned with the peak of LL70B-S and another at a slightly higher similarity level. This suggests greater variability in the manual summarization process. Unlike Llama, which follows a fixed summarization procedure, human experts may be employing diverse strategies, leading to more varied results.

The high p-values for the regulation and scientific categories in the LDFScore distribution for LL70B-S summaries (Figure 4) suggest approximate normality, unlike the guidance and news categories. This pattern likely reflects the more consistent and standardized structure of regulatory and scientific documents, which allows the Llama model to generate summaries with more uniform factual alignment. In contrast, guidance and news texts tend to be more heterogeneous in tone, format, and content, leading to greater variation in the generated summaries. This effect is observed only in the summaries of Llama, which applies



	Density (Rouge-1) μ	Density (Rouge-1) σ	Density (Rouge-1) p-value	Density (Rouge-L) μ	Density (Rouge-L) σ	Density (Rouge-L) p-value
Manual – News	0.2966	0.0734	0.0000	0.2822	0.0765	0.0000
Manual – Regulation	0.2597	0.0665	0.0000	0.2647	0.0683	0.0000
Manual – Guidance	0.2717	0.0568	0.0000	0.2723	0.0573	0.0000
Manual – Scientific	0.2430	0.0637	0.0000	0.2526	0.0664	0.0000
Llama 70B – News	0.3229	0.0752	0.0000	0.3272	0.0709	0.0000
Llama 70B – Regulation	0.2738	0.0681	0.0000	0.2855	0.0649	0.0000
Llama 70B – Guidance	0.2889	0.0569	0.0000	0.3025	0.0546	0.0000
Llama 70B – Scientific	0.2631	0.0466	0.0000	0.2783	0.0418	0.0116

Figure 5: Density of Summarized Information in the original document (D): We report the fraction of snippets in the original document for which the combined similarity with the summary reaches at least 50% of the total similarity, as measured by ROUGE-1 and ROUGE-L (F-measure) (Lin, 2004). Although none of the distributions strictly follow a normal distribution, as confirmed by a Shapiro-Wilk test (see p-values), the mean (μ) and standard deviation (σ) still provide a useful approximation, as illustrated in the plots.

a fixed summarization strategy across document types, whereas human experts adapt to each source, reducing variability and leading to consistently non-normal distributions across categories.

We additionally experimented with FactCC (Kryscinski et al., 2020), a BERT-based document-sentence metric for factual consistency checking. Since FactCC cannot process entire long documents, we restrict this analysis to source documents shorter than 400 tokens (i.e., cases without the need for truncation). In this setting, LL70B-S attain a higher mean score (0.58) than MANUAL-S (0.49), suggesting that LL70B-S summaries align more closely with the surface form of short documents.

Figure 4 also shows the similarity between D and each derived S (paired), using ROUGE. The ROUGE scores are generally low (ranging from 3% to 18%), which is expected given that we summarize lengthy Ds into relatively short Ss (see Table 7). The low values of ROUGE-L indicate that long common subsequences between D and S are improbable. It should be highlighted here that the matching does not require the tokens to appear consecutively in both documents

4.2 Information Loss during Summarization

As demonstrated in the previous section, summarizing lengthy documents inevitably results in some loss of information. In this section, we assess how concentrated the retained information is within D. To do so, we divide the original text into snippets of up to 20 tokens, ensuring that each snippet starts at a sentence or paragraph boundary. We maximize the number of snippets within the 20-token window while avoiding sentence fragmentation or mixing paragraphs.

Figure 5 presents the fraction of original document snippets where the combined similarity with the summary reaches at least 50% of the total similarity, as measured by ROUGE-1 and ROUGE-L (F-measure) (Lin, 2004). Intuitively, a value of 0.5 indicates that summarized information is evenly distributed throughout the document. Lower values suggest that the summary primarily derives from a specific subset of snippets. In our case, 50% of the summarized information is concentrated within 24% to 32% of the original document. This suggests that a significant portion of the Ds is omitted during summarization. The extent of this omission varies by document type: while regulatory and scientific documents tend to discard larger portions

of text, news articles and guidance literature retain a greater proportion of the original content. We argue that this reflects the nature of highly standardized documents, such as those in scientific and regulatory literature, which often contain sections that are irrelevant for summarization (e.g., author lists, bibliographies, and formal disclaimers).

5 Empirical Analysis

To assess the quality of the FOODSAFESUM dataset's summaries relative to each other and to the original documents,, we experiment with three NLP tasks - IR, MLC, and QA - which serve as proxies for real world food safety applications. In addition, we demonstrate a use case of text clustering, to identify related events and recurring issues. These tasks, along with the clustering use case, are discussed in the following sections. We represented the data using contextual embeddings from Stella 1.5b (Zhang et al., 2025),[9] opting for a model that achieves an optimal trade-off of performance on the MTEB Leaderboard,[10] memory usage, and the maximum token capacity. Approximately 0.4% of the DS exceeded the model's maximum token limit (131,072) and were truncated.

5.1 IR

Retrieving the original document given a summary can serve as a practical evaluation of summary representativeness, especially in real-world settings where users such as regulators and researchers may rely on summaries to locate full reports. While not a standard summarization evaluation metric, retrieval performance helps ensure that summaries retain enough distinctive information to identify their source documents.

This task evaluates the effectiveness of summaries to capture enough information from their corresponding full document to serve as a reliable identifier. We compute contextual embeddings for both summaries and full documents, then rank the full documents by their cosine similarity to each summary. In other words, a well-crafted summary should be most similar to its source document among all candidates. We evaluate the performance using Mean Reciprocal Rank (MRR) (Baeza-Yates and Ribeiro-Neto, 1999), which measures how high the correct document in the ranked list on average,[11] and Recall at a given cutoff K (R@1, R@3, R@5), which, in this case, indicates if the correct document is returned within the top-K ranked results.

		MRR	R@1	R@3	R@5
S	MANUAL-S	0.891	0.846	0.927	0.947
	LL70B-S	**0.958**	**0.935**	**0.978**	**0.986**
T	MANUAL-T	**0.893**	0.843	**0.937**	**0.952**
	LL70B-T	**0.893**	**0.846**	0.932	0.950

Table 2: MRR and Recall@K for S- and T-based document identification. Highlighted in bold are the best scores per metric for S and T.

Table 2 demonstrates the near-perfect performance of LL70B-S in retrieving their corresponding full documents. With an MRR of 0.958 and leading recall scores (R@1, R@3, and R@5), LL70B-S effectively encapsulates the core content of full documents, ensuring that its embedding closely matches the source document. This confirms that a well-crafted summary can serve as a reliable identifier of its full document. Additionally, title-length summaries (MANUAL-T and LL70B-T) generally outperform manual summaries (MANUAL-S) across all metrics—except for R@1, where MANUAL-T lags slightly—suggesting that more concise representations also capture key document content effectively. In Table 9 in the Appendix, we report further results with different embedding models.

5.2 QA via RAG

In regulatory and investigative contexts, stakeholders often need clear answers to specific questions based on policy documents, recall notices, or technical summaries. We formulate this task as QA using RAG, focusing on multiple-choice and yes/no question types. The system retrieves relevant passages from full documents, human-written summaries, or LLM-generated summaries, and generates an answer grounded in the retrieved content. This setup enables us to evaluate how well different textual representations support factual and context-sensitive decision-making.

We evaluate the effectiveness of the RAG approach for the full documents (D) and the summaries (S), using a synthetic benchmark derived

[9] https://huggingface.co/billatsectorflow/stella_en_1.5B_v5

[10] https://huggingface.co/spaces/mteb/leaderboard

[11] MRR is 1.0 when the correct document is always in the first position of the rank.

from the Ds of our corpus. The dataset comprises 27 multiple-choice and 34 yes/no questions across diverse user roles (e.g., expert, consumer). The 61 QA pairs are automatically generated and then validated by annotators. See also Appendix D.

For retrieval, we split each document into passages of 150 tokens. Splitting by paragraph was not ideal because the texts of the D document version were extremely fragmented, with an average of 679 paragraphs per text and an average paragraph length of only 25 words. We then indexed these passages using a BM25 index. The RAG pipeline retrieves the top-k relevant passages and provides this context to LLaMA-3-70B for answer generation. The LLM is instructed to answer the question using only the provided context. The augmented generation instruction is provided in Appendix D.3. We evaluate final answers using accuracy and assess the retrieval step separately with the information retrieval metrics that were used in §5.1.

		MRR	R@1	R@3	R@5	R@10
S	MANUAL-S	0.602	0.541	0.656	0.689	0.705
	LL70B-S	**0.824**	**0.803**	**0.820**	**0.836**	**0.902**
D		0.813	0.754	**0.836**	0.902	0.934

Table 3: MRR and Recall@K for S and D document versions for the QA RAG-based system.

Table 3 presents the retrieval results for each document version, using the QA benchmark questions and retrieving passages accordingly. We observe that LL70B-S outperforms both MANUAL-S and D across MRR and Recall@k metrics.

		@1	@2	@3	@4	@5
S	MANUAL-S	0.393	0.410	0.410	0.443	0.443
	LL70B-S	**0.574**	**0.557**	**0.590**	**0.541**	**0.525**
D		**0.738**	**0.771**	**0.771**	**0.771**	**0.803**

Table 4: Accuracy@k for the topk passages augmenting the prompt (for k=1 to 5)

Table 4 presents the accuracy of the QA RAG-based system for top-k values ranging from one to five. We observe that LL70B-S outperforms MANUAL-S in both retrieval and accuracy metrics and it is comparable to D in the retrieval metrics, the D version achieves higher accuracy in the QA evaluation. LL70B-S clearly captures the subject of the question, i.e., the adequate key phrases, it is not as informative as D to answer the questions.

5.3 MLC

Food safety issues span multiple regulatory and risk-related domains, including contamination types, recall reasons, and legislative actions. This task concerns the classification of documents, titles, and summaries across topics. As many cases involve overlapping concerns, we frame this classification as a multilabel problem.[12]

Models and Evaluation Manually annotated topics serve as ground truth labels to evaluate performance across different document versions. The data was uniformly split into training and testing sets. We employed three common machine learning classification algorithms – Logistic Regression (LR), Random Forests (RF), and Support Vector Machines (SVM).[13] Given the dataset's imbalance across topics, we evaluated the models using micro-F1 ($F1_\mu$), macro-F1 ($F1_M$), and samples-F1 ($F1_S$).

		Classifier	$F1_\mu$	$F1_M$	$F1_S$
S	MANUAL-S	LR	0.611	0.549	0.600
		RF	0.428	0.294	0.376
		SVM	0.581	0.521	0.572
	LL70B-S	LR	**0.621**	**0.565**	**0.618**
		RF	0.481	0.350	0.438
		SVM	0.600	0.549	0.595
T	MANUAL-T	LR	**0.583**	**0.542**	**0.571**
		RF	0.442	0.308	0.386
		SVM	0.571	0.521	0.558
	LL70B-T	LR	0.573	0.515	0.567
		RF	0.476	0.343	0.430
		SVM	0.556	0.501	0.549
D		LR	**0.652**	**0.594**	**0.641**
		RF	0.510	0.348	0.465
		SVM	0.615	0.554	0.604

Table 5: micro-F1 ($F1_\mu$), macro-F1 ($F1_M$), and samples-F1 ($F1_S$) for the MLC task on summaries (S) or titles (T), created manually or by LLAMA70B, and on the full document (D).

Results Table 5 shows that LL70B-S consistently outperforms MANUAL-S across all algorithms in terms of F1 scores. In contrast, MANUAL-T outperforms LL70B-T across all metrics. Lastly, D achieves the highest F1 scores. However, the gap between D and the Ss is relatively small, with around a 5% improvement for D across all metrics compared to LL70B-S, and about a 7% improvement compared to MANUAL-S. This indicates that

[12] Appendix C comprises results for the task of Search, which can be considered as the inverse of MLC.

[13] We used using the scikit-learn library.

while longer and more detailed documents generally yield better classification performance, LLM-generated summaries (LL70B-S) still perform well, especially when compared to MANUAL-S. Per-topic performance and results using TF-IDF are shown in the Appendix (Tables 12 and 11).

5.4 Use Case: Text Clustering

Different FOODSAFESUM documents may be referring to the same entity yet from a different perspective or source. To mine such associations, we opted for data clustering. By clustering the full documents and their compressed versions (titles and summaries), we can aggregate diverse posts (e.g., from BBC, WHO, and regulatory agencies) into coherent clusters representing a single food safety event, such as a specific regulation.

We applied the DBSCAN clustering algorithm to the embeddings and tuned the ϵ parameter, which defines the density threshold for clustering—the maximum distance between two points for them to belong to the same cluster. Analysing pairwise similarities, revealed two modes, with the lowest one being near zero (see Figure 10 in the Appendix). Therefore, to define the optimal threshold, we varied $\epsilon \in [0.02, 1]$ and chose the value that maximizes the Silhouette score (Rousseeuw, 1987). This approach balances clear cluster formation with minimal noise.

We assessed clustering quality with homogeneity scores (Rosenberg and Hirschberg, 2007), by using metadata as our ground truth (i.e., topics, hazards, document type, and source name). While document type and source name capture mainly stylistic similarities, topics and hazards indicate common events. Table 6 presents the results. For summaries, MANUAL-S clusters show perfect homogeneity in source name and document type, meaning they primarily capture stylistic similarities. However, they perform poorly in grouping documents by topics and hazards, suggesting that the clustered texts do not necessarily refer to the same events. In contrast, LL70B-S clusters achieve high homogeneity in topics and hazards, meaning they better capture event-based similarities, even if source name and document type vary. For titles, we also observe differences, but the varying number of clusters (K) between the two methods complicates direct comparison.

Impact Appendix F presents summaries that belong to the same cluster yet originating from different sources. For instance, a draft of a standard that replaces GB 28050-2011 exists in two sources: 'WTO Documents online (Notifications' and in 'Food and Drug Administration Philippines'. These clusters can be used to identify similar summary pairs that refer to the same event, revealing associations across source and document types. Over time, such a data organisation can reveal entities that have attracted interest in across various document types and ones that have been disregarded.

		e	K	SIZE	H@T	H@H	H@S	H@D
S	MANUAL-S	0.02	7	2.29 (0.49)	0.650	0.679	1.000	1.000
	LL70B-S	0.02	6	2.00 (0.00)	0.728	0.915	0.829	0.657
T	MANUAL-T	0.02	11	2.64 (1.80)	0.771	0.779	0.905	0.961
	LL70B-T	0.06	67	2.97 (2.81)	0.781	0.736	0.909	0.751
D		0.10	123	2.63 (2.05)	0.821	0.801	0.944	0.798

Table 6: Homogeneity of clustering using ground truth according to topics (H@T), hazards (H@H), source (H@S), and document type (H@D). We report the threshold e, the number of clusters (K) and the average size (st.d.) per cluster per representation.

6 Conclusions

This work bridges a critical gap in NLP for food safety by releasing a machine-actionable dataset that is enriched with metadata and multiple summary versions, and by evaluating its utility on tasks directly aligned with food safety practice. Our experiments show that multilabel classification supports the organization of documents by food safety topics; information retrieval improves practitioners' access to relevant regulatory and scientific information; and retrieval-augmented question answering addresses practical needs, such as identifying contamination causes or verifying compliance. The exploratory use case of clustering further demonstrates how related events can be tracked across document types, enabling compliance monitoring and early-warning applications. Beyond highlighting the potential of modern LLMs for summarization, this study provides an open resource to foster research in automated food safety monitoring, risk detection, policy enforcement, and public health communication. We envision FOODSAFESUM as a foundation for developing food safety AI systems, policy-driven NLP, and decision support tools for regulators, industry stakeholders, and researchers.

Limitations

Faithfulness and factual accuracy LLM-generated summaries lead to better results across NLP tasks compared to human-written

ones. However, challenges remain in ensuring faithfulness and factual accuracy, particularly when summarizing technical or regulatory content. Future work should explore human-in-the-loop approaches to refine LLM-generated summaries while preserving efficiency.

Information loss Both manual and LLM-generated summaries inevitably lose information compared to the original documents. While our evaluations suggest that critical information is generally preserved, downstream tasks relying on fine-grained details (e.g., specific legal clauses or threshold values) may require access to full-text documents.

Linguistic challenges Complex regulatory language and domain-specific terminology increase the difficulty of classification and retrieval compared to general NLP benchmarks. Additionally, since 34.3% of the dataset is translated to English, noise may be introduced through imperfect translations. This can affect linguistic fidelity and model performance, particularly in tasks sensitive to nuances of legal or regulatory phrasing.

Dataset coverage While FOODSAFESUM integrates diverse food safety sources, some topics and document types remain underrepresented. This imbalance may limit model generalization across all areas of food safety and calls for future extensions to broaden coverage.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., USA.

Jennifer A. Bishop, Sophia Ananiadou, and Qianqian Xie. 2024. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10777–10789, Torino, Italia. ELRA and ICCL.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint*. ArXiv:2404.16130.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating diverse q&a benchmarks for rag evaluation with datamorgana. *arXiv preprint arXiv:2501.12789*.

David M Goldberg, Samee Khan, Nohel Zaman, Richard J Gruss, and Alan S Abrahams. 2022. Text mining approaches for postmarket food safety surveillance using online media. *Risk Analysis*, 42(8):1749–1768.

Vishal Gupta and Gurpreet Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2.

Shabnam Hassani. 2024. Enhancing legal compliance and regulation analysis with large language models. *arXiv preprint arXiv:2404.17522*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. CLERC: A Dataset for Legal Case Retrieval and Retrieval-Augmented Analysis Generation. *arXiv preprint*. ArXiv:2406.17186 [cs].

Ruofan Hu, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng, and Elke Rundensteiner. 2022. Tweet-fid: An annotated dataset for multiple foodborne illness detection tasks. *arXiv preprint arXiv:2205.10726*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Adyasha Maharana, Kunlin Cai, Joseph Hellerstein, Yulin Hswen, Michael Munsell, Valentina Staneva, Miki Verma, Cynthia Vint, Derry Wijaya, and Elaine O Nsoesie. 2019. Detecting reports of unsafe foods in consumer product reviews. *JAMIA open*, 2(3):330–338.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. CICLe: Conformal in-context learning for largescale multi-class food risk classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7695–7715, Bangkok, Thailand. Association for Computational Linguistics.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, Tony Lindgren, and Juli Bakagianni. 2025. SemEval-2025 task 9: The food hazard detection challenge. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2523–2534, Vienna, Austria. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Jinyi Song and Jiayin Pei. 2024. Mining text for causality: a new perspective on food safety crisis management. *Frontiers in Sustainable Food Systems*, 8:1491255.

Supriyono, Aji Prasetya Wibawa, Suyono, and Fachrul Kurniawan. 2024. A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, 7:100070.

Dandan Tao, Ruofan Hu, Dongyu Zhang, Jasmine Laber, Anne Lapsley, Timothy Kwan, Liam Rathke, Elke Rundensteiner, and Hao Feng. 2023. A novel foodborne illness detection and web application tool based on social media. *Foods*, 12(14):2769.

Lei Xia, Bo Chen, Kyle Hunt, Jun Zhuang, and Cen Song. 2022. Food safety awareness and opinions in china: A social network analysis approach. *Foods*, 11(18):2909.

Shufeng Xiong, Wenjie Tian, Vishwash Batra, Xiaobo Fan, Lei Xi, Hebing Liu, and Liangliang Liu. 2023. Food safety news events classification via a hierarchical transformer model. *Heliyon*, 9(7).

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Jiansong Zhang and Nora M El-Gohary. 2016. Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering*, 30(2):04015014.

Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Mark Achtman, Derek Brown, Marie Chattaway, Tim Dallman, Richard Delahay, Christian Kornschober, et al. 2020. The enterobase user's guide, with case studies on salmonella transmissions, yersinia pestis phylogeny, and escherichia core genomic diversity. *Genome research*, 30(1):138–152.

A Exploratory analysis

Here we present some analysis regarding the data composition.

A.1 Imbalance in Topic Categories

Figure 6 shows the imbalance in topics. The balance ratio (size of minority to size of majority class) is 0.16. The major class is Policies and Laws with over 800 documents while the category "Substance Usage", which was assigned to only one document, was excluded from this calculation.

A.2 Topic co-occurrence and correlation of texts

Figure 7 shows topic co-occurrences, normalized by the number of documents. Diagonal values indicate the percentage of documents where a topic appears alone. "Substance Usage" occurs in a single document, where it co-occurs with five other topics and is not included in Figure 7

A.3 Cross-Source Correlation of Hazard Mentions

Figure 8 presents a lower-triangular heatmap showing the Spearman correlation between document types based on the frequencies of hazard mentions.

A.4 Text statistics

Table 7 displays word counts and vocabulary size for documents (D), summaries (S), and titles (T), also for summary versions (MANUAL-S and LL70B-S) across various document types. The pronounced

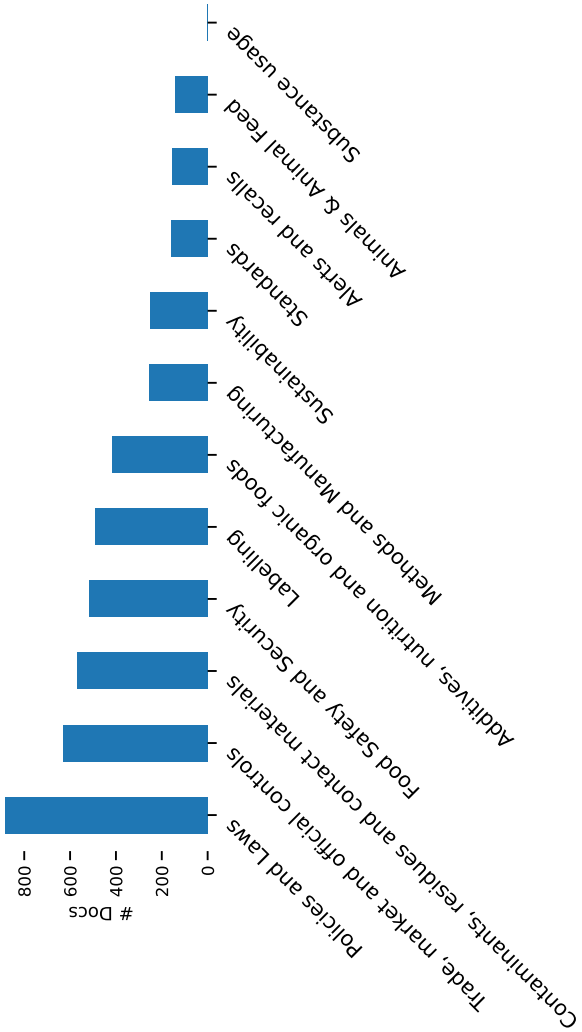


Figure 6: Topic distribution.

variability in D word counts (std: 20,009.06; see Table 7) reflects the diversity of the source documents — from brief news articles to extensive scientific reports.

		Avg WC	Std WC	Avg VS	Std VS
D		6,448.09	20,009.06	826.97	1410.14
S	MANUAL-S	85.53	60.18	41.81	25.09
	LL70B-S	72.89	16.21	39.50	7.95
T	MANUAL-T	13.82	7.28	9.71	4.33
	LL70B-T	10.44	3.36	7.60	2.16

Table 7: Word count (WC) and vocabulary size (VS) statistics for documents (D), summaries (S), and titles (T).

B LLM Generated Summaries

In Table 8 we show the prompt used to generate synthetic summaries, in particular we rely on LLaMA-3-70B. As it can be observed, we directly ask the model to generate a json dictionary to be used in subsequent iteration of the research.

Instruction
You are an expert in food- related topics, including safety, nutrition, production, and regulations. I will provide you with a document, and your task is to analyze its content and generate a concise title and a summary. The title should capture the main theme or focus of the document. The summary should highlight the key points, such as the main topic, any significant issues or findings, and relevant details. Provide the output as a JSON dictionary with the following structure: {''title'': ''<Generated Title>'', ''summary'': ''<Generated Summary>''} Generate only the JSON dictionary and nothing else.

Table 8: Summarization instruction for LLM-prompting

In Figure 9 we report some further similarity analysis between the manual and the generated summaries.

Finally, in Table 9 we show some further results obtained when retrieving the original documents using both the title and the summary as input. To perform this evaluation we use two additional embedding models that showed good performances in different Information Retrieval areas such as conversational search, namely Dragon and Snowflake.

Embed. Model	Vers. Type	Doc. Version	MRR	R@1	R@3	R@5
Dragon	S	MANUAL-S	0.792	0.734	0.834	0.855
		LL70B-S	**0.877**	**0.832**	**0.909**	**0.935**
	T	MANUAL-T	0.798	0.733	0.845	0.877
		LL70B-T	**0.811**	**0.742**	**0.862**	**0.890**
Snowflake	S	MANUAL-S	0.856	0.809	0.894	0.907
		LL70B-S	**0.940**	**0.915**	**0.962**	**0.973**
	T	MANUAL-T	0.858	0.803	0.901	0.925
		LL70B-T	**0.857**	**0.801**	**0.903**	**0.926**

Table 9: MRR and Recall@K for document identification, grouped by Embedding model (*Dragon* and *Snowflake*) and version type. Best scores per metric and document version type are highlighted in bold.

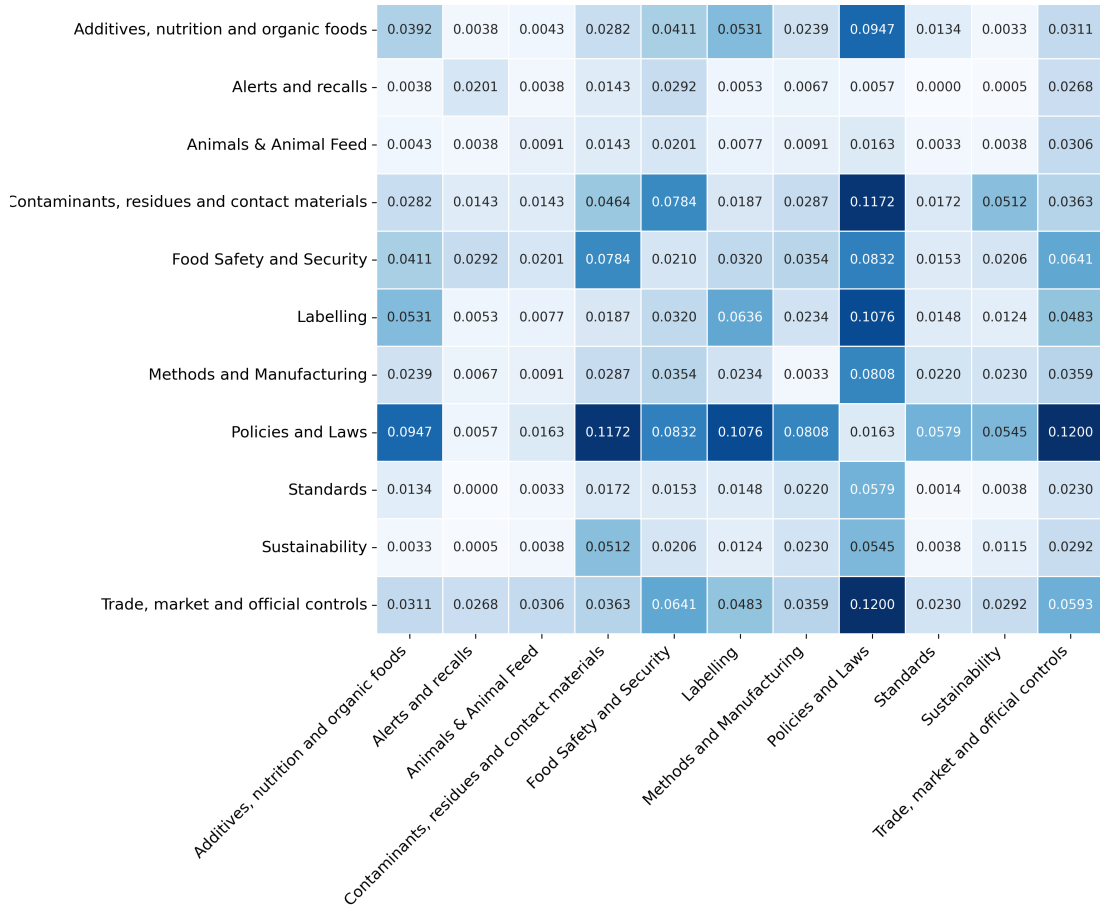


Figure 7: Topic co-occurrence

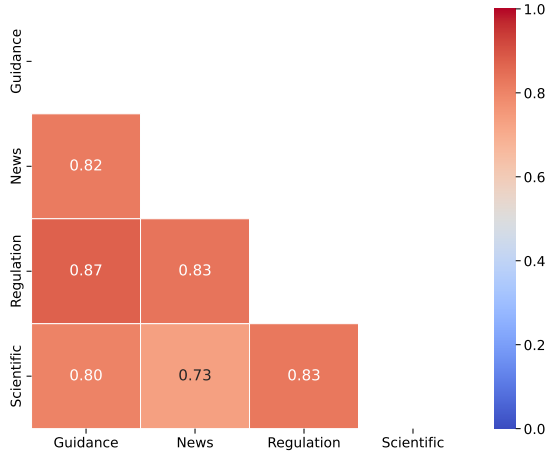


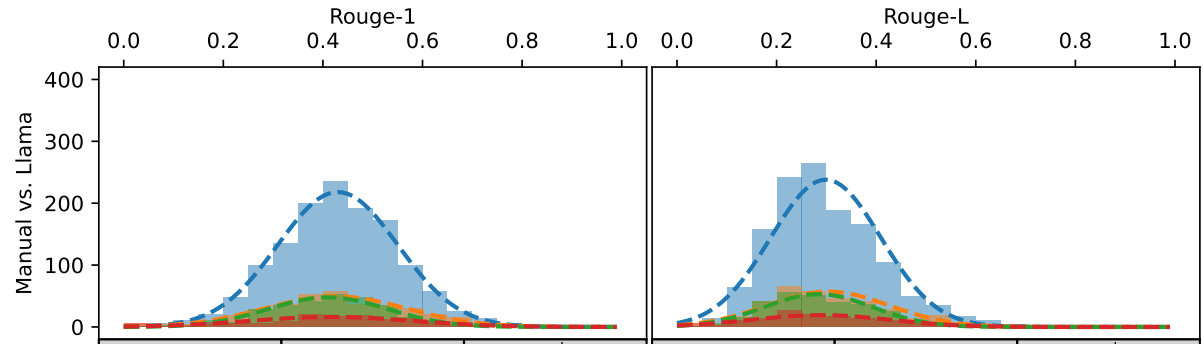
Figure 8: Lower triangular heatmap of Spearman correlation between document types when counting hazards mentioned in the respective documents.

C Search for topic-based document retrieval

Practitioners often rely on keyword-based search to filter relevant information from extensive archives. This task evaluates how well our dataset supports query-driven information extraction, allowing users to locate concise and relevant summaries based on predefined food safety topics. This is particularly useful in policy compliance checking, supply chain monitoring, and rapid risk assessments, where retrieving summaries by topic can expedite decision-making. We consider this as the inverse task of MLC (§5.3), assessing how well-detailed topic descriptions serve as queries for retrieving their corresponding documents.

In this setup, the topic labels are used as queries, where each query is a detailed description that outlines the focus of the corresponding topic. For instance, the query for the "Labelling" topic includes details about food labelling regulations, nutritional information disclosure, and ingredient transparency. To perform the retrieval, we use the Sentence-to-Passage (s2p) setting of the Stella 1.5b model, which is fine-tuned to generate retrieval-related query embeddings. These query embeddings are then compared to the document embeddings using cosine similarity.

The goal is to identify the documents that are most relevant to each topic. By evaluating re-



	μ	σ	p-value	μ	σ	p-value
News	0.4299	0.1208	0.4565	0.2988	0.1106	0.0000
Regulation	0.4252	0.1423	0.1040	0.3034	0.1230	0.0001
Guidance	0.4100	0.1202	0.2482	0.2840	0.1084	0.0000
Scientific	0.4181	0.1651	0.0117	0.2887	0.1391	0.0001

Figure 9: Rouge scores (F-measure) comparing LL70B-S summaries vs MANUAL-S. The ROUGE-1 distributions of "News," "Regulations," and "Guidance" are assumed to be normally distributed while there is evidence that the remaining histograms are not according to a Shapiro-Wilk test (see p-values). Nevertheless, mean μ and standard deviation σ can still give a rough estimation of the distributions as shown in the plots.

trieval performance we gain complementary insights into whether the topics truly capture the semantic essence of the documents. Evaluation is conducted using the Mean Average Precision (MAP), which measures the precision at multiple recall levels by averaging the precision scores at each relevant document's rank,[14] MRR (see §5.1), Precision at a given cutoff K (P@1, P@3), which measures the proportion of relevant documents among the top K retrieved,[15] and Normalized Discounted Cumulative Gain at cutoff 3 (nDCG@3), which measures the relevance of documents retrieved in the top 3 positions, giving more weight to higher-ranked relevant documents.[16]

		MAP	MRR	P@1	P@3	nDCG@3
S	MANUAL-S	0.356	**0.864**	**0.818**	0.636	0.673
	LL70B-S	**0.376**	0.773	0.636	**0.697**	**0.690**
T	MANUAL-T	**0.348**	**0.786**	**0.727**	**0.667**	**0.679**
	LL70B-T	0.345	0.571	0.364	0.515	0.481
D		0.381	0.849	0.818	0.788	0.786

Table 10: Results for the search task, using MAP, MRR, P@1,3, nDCG@3, for summaries (S), titles (T), and full documents (D); the best per S, T are shown in bold.

Table 10 shows that MANUAL-S are superior when it comes to placing a relevant document in the very first rank (higher MRR and P@1) even though LL70B-S provides better overall ranking quality across multiple top positions (i.e. better AP, P@3, and nDCG@3 metrics). D enhances overall retrieval performance across the ranking. For titles, MANUAL-T substantially outperforms LL70B-T across all metrics, highlighting that MANUAL-T is much more reliable for direct retrieval.

D QA Dataset Construction

To evaluate the RAG pipeline, we created a QA benchmark using 34 randomly selected documents from our corpus (each $\leq 2,500$ tokens). For every document, we used `gpt3.5-turbo` to generate two questions, yielding 68 QA pairs in total. To enable automatic evaluation via accuracy, all questions were constrained to be either multiple-choice or yes/no.

Each question was written from the perspective of a simulated user role—"consumer", "expert", "researcher", or "authority"—ensuring balanced representation across categories (see Appendix D.1 for role definitions). We adapted the instruction-based prompting paradigm from Filice et al. (2025) to produce the QA pairs (full prompt in the Appendix D.2).

After generation, the QA pairs were split among six annotators for review to ensure: (1) factual consistency, (2) grounding in the provided document (Es et al., 2024), and (3) meaningful differentiation between the two questions from the same document. Only validated QA pairs were retained, resulting in 61 entries in the final QA dataset.

[14] MAP is 1.0 when all relevant documents are ranked before any irrelevant ones across all queries.

[15] P@K reaches 1.0 when all K retrieved documents are relevant.

[16] nDCG@3 reaches 1.0 when the top 3 retrieved documents are perfectly ranked by relevance.

D.1 User Categories

For the QA generation process, we simulated four distinct user types to ensure coverage of different information needs. Below are the descriptions used to condition the generation process that were provided to the LLM instruction:

- **Consumer:** A regular consumer who uses the system to get basic food safety information, alerts on food recalls, advice on safe food handling, and guidance on avoiding contaminated products.
- **Expert:** A food safety expert who accesses detailed regulatory documents, incident reports, and compliance guidelines to manage food safety risks within their organization.
- **Researcher:** A researcher or scientist who uses the system to access food safety data, scientific articles, and risk assessments to support the development of new food safety interventions or technologies.
- **Authority:** A public health or food safety authority who uses the system to track emerging food risks, disseminate safety alerts and recalls to the public, monitor compliance across the food supply chain, and respond to food safety emergencies.

D.2 QA Prompt Template

We used the following prompt, adapted from the Filice et al. (2025), to generate two QA pairs from each selected document:

```
You are a user simulator that should generate two candidate questions for starting a conversation.

The two questions must be about facts discussed in the document you will receive. When generating the questions, assume that the real users you must simulate, as well as the readers of the questions, do not have access to this document. Therefore, never refer to the author of the document or the document themselves. Also, assume that whoever reads the questions will read each question independently.

The two questions must be diverse and different from each other. Return only the questions without any preamble.

### Each of the generated questions must reflect a user with the following characteristics:
- For the first question the user must be {user_category_1}.
- For the second question the user must be {user_category_2}.

### Each of the generated questions must have the following characteristics:
- It must {question_category}.

### Write each pair in a new line, in the following JSON format:
{''question'': <question>, ''choices'': <choices>, ''correct_answer'': <correct_answer>}.
```

D.3 Augmented Generation Prompt

```
Write exactly the text of the correct choice that answers the question using only the context below. Do not use any information that is not present in the context.

If the answer cannot be found in the context, respond only with: ''I don't know.''.

Context:
<passage_1>
...
<passage_k


Question:
<question>
Choices:
<multiple_choices_OR_yes_no>
Answer:
```

E MLC Reports and Baselines

In Table 12, we present the classification report detailing the results per topic for the best-performing system for each document version. D achieves the highest F1 scores for 55% of the topics, covering

	Doc. Version	Cl.	$F1_S$	$F1_\mu$)	$F1_M$
S	MANUAL-S	LR	0.307	0.387	0.256
	MANUAL-S	RF	0.426	0.488	0.360
	MANUAL-S	SVM	0.541	0.577	0.462
	LL70B-S	LR	0.370	0.432	0.309
	LL70B-S	RF	0.434	0.485	0.351
	LL70B-S	SVM	**0.553**	**0.588**	**0.505**
T	MANUAL-T	LR	0.295	0.369	0.242
	MANUAL-T	RF	0.405	0.474	0.344
	MANUAL-T	SVM	**0.520**	**0.568**	**0.472**
	LL70B-T	LR	0.328	0.394	0.270
	LL70B-T	RF	0.426	0.487	0.363
	LL70B-T	SVM	0.518	0.561	0.460
D		LR	0.305	0.365	0.216
		RF	0.452	0.519	0.381
		SVM	**0.537**	**0.581**	**0.490**

Table 11: micro-F1 ($F1_\mu$), macro-F1 ($F1_M$), and samples-F1 ($F1_S$) for the topic prediction task on summaries (S) or titles (T), created manually or by LLAMA70B, and on the full document (D). Texts are represented with TF-IDF.

both high-support topics (e.g., "Food Safety and Security") and low-support topics (e.g., "Standards"). For the remaining topics, LL70B-S achieves the highest F1 scores, which are comparable to those of D. This includes both high-support topics, such as "Policies and Laws", and low-support topics, such as "Animals & Animal Feed". MANUAL-S demonstrates moderate performance, outperforming the other approaches only in Trade, Market, and Official Controls, where topic support is high.

Finally, in Table 11, we establish a baseline for the task using a TF-IDF representation. Notably, LL70B-S outperforms all other document versions, including D, highlighting the advantage of its information-dense representation over the longer D documents in this setting.

F Clustering

In Figure 10 we report cosine distances to the second nearest neighbour for different embedding representations.

The summaries of Table 13 come from different sources but describe the same event: i.e., a draft of a standard that replaces GB 28050-2011. More examples are shown in Tables 14 to 16.

Label	Precision			Recall			F1-score			Support
	D	LL70B-S	MANUAL-S	D	LL70B-S	MANUAL-S	D	LL70B-S	MANUAL-S	
Additives, nutrition and organic foods	**0.70**	0.64	0.64	**0.72**	0.55	0.62	**0.71**	0.59	0.63	87
Alerts and recalls	0.76	**0.77**	0.74	**0.85**	0.77	0.65	**0.80**	0.77	0.69	26
Animals & Animal Feed	0.71	**0.80**	0.71	**0.61**	0.57	0.43	0.65	**0.67**	0.53	28
Contaminants, residues and contact materials	0.78	**0.82**	0.72	**0.82**	**0.82**	0.72	0.80	**0.82**	0.72	119
Food Safety and Security	**0.55**	0.44	0.50	**0.55**	0.45	0.46	**0.55**	0.44	0.48	103
Labelling	**0.82**	0.76	0.76	**0.72**	0.70	0.69	**0.76**	0.73	0.72	102
Methods and Manufacturing	**0.45**	0.37	0.41	0.27	**0.34**	0.32	0.34	**0.36**	**0.36**	56
Policies and Laws	0.64	**0.68**	0.65	**0.63**	0.62	0.62	0.64	**0.65**	0.64	185
Standards	**0.64**	0.56	0.59	**0.59**	0.56	0.59	**0.62**	0.56	0.59	27
Sustainability	0.70	0.62	**0.71**	**0.76**	0.71	0.61	**0.73**	0.67	0.66	49
Trade, market and official controls	**0.54**	0.53	0.52	0.53	0.53	**0.59**	0.54	0.53	**0.56**	116
micro avg	**0.66**	0.64	0.63	**0.64**	0.61	0.59	**0.65**	0.62	0.61	898
macro avg	**0.61**	0.58	0.58	**0.59**	0.55	0.53	**0.59**	0.57	0.55	898
weighted avg	**0.66**	0.64	0.63	**0.64**	0.61	0.59	**0.65**	0.62	0.61	898
samples avg	**0.68**	0.67	0.63	**0.69**	0.65	0.64	**0.64**	0.62	0.60	898

Table 12: Classification report for the topic prediction task using Logistic Regression as the classifier and contextual embeddings as the text representation. The systems differ based on the input text.

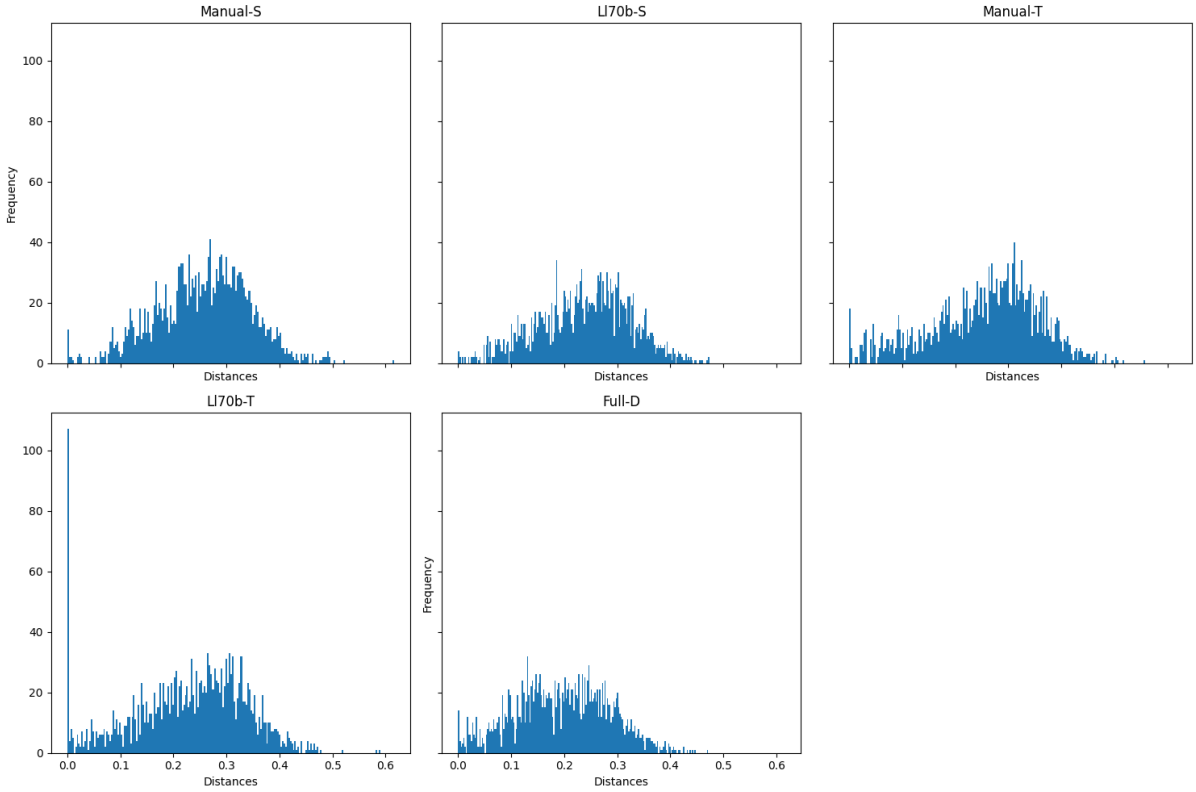


Figure 10: Histograms of cosine distances to the second nearest neighbour for different embedding representations. Each subplot corresponds to a specific embedding variant, where the x-axis represents the distances and the y-axis shows their frequency.

Source	**MANUAL-S**
WTO Documents online (Notifications)	The FDA Circular updates guidelines for assessing microbiological quality of processed food products, repealing FDA Circular No. 2013-010. The guidelines cover food establishments in the Philippines, providing definitions of terms, and outlining objectives, scope, and general and specific guidelines for ensuring food safety and compliance with Good Manufacturing Practices (GMP). The guidelines reference internationally recognized standards and methods for microbiological analysis.
Food and Drug Administration Philippines	The FDA Circular updates the guidelines for assessing the microbiological quality of processed food products in the Philippines, repealing FDA Circular No. 2013-010. The guidelines cover food establishments engaged in the manufacture, trade, and distribution of processed food products, and provide definitions of terms, objectives, and scope. The guidelines aim to ensure food safety and compliance with Good Manufacturing Practices (GMP) by providing reference criteria for specific food commodities and internationally recognized references for microbiological analysis.

Table 13: Summaries from the same cluster and referring to the same event yet from different sources

Source	LLAMA-3-70B-S	MANUAL-S
WTO Documents online (Notifications)	Food Standards Australia New Zealand (FSANZ) is calling for submissions on a proposal to amend the Australia New Zealand Food Standards Code to provide energy (kilojoule) labelling information on alcoholic beverages. The proposal aims to improve consumer awareness of energy content in alcoholic drinks. Submissions are invited until 6pm (Canberra time) on 27 February 2023. The proposal is part of a broader effort to improve nutrition labelling and reduce obesity and related health issues in Australia and New Zealand.	Description of content: The proposal considers amending the Australia New Zealand Food Standards Code to require energy (kilojoule/kilocalorie) labelling information on packaged alcoholic beverages containing 0.5% or more ABV sold in Australia and New Zealand. Deadline Feb 27/2023
Food Standards Australia New Zealand (FSANZ)	Food Standards Australia New Zealand (FSANZ) is calling for submissions on a proposal to amend the Australia New Zealand Food Standards Code to provide energy (kilojoule) labelling information on alcoholic beverages. The proposal aims to improve consumer awareness of energy content in alcoholic drinks. Submissions are invited until 6pm (Canberra time) on 27 February 2023. The proposal is part of a broader effort to improve nutrition labelling and reduce obesity and related health issues in Australia and New Zealand.	Food Standards Australia New Zealand (FSANZ) has assessed a proposal to consider amending the Australia New Zealand Food Standards Code to provide energy (kilojoule) labelling information on alcoholic beverages and has prepared a draft food regulatory measure. Pursuant to section 61 of the Food Standards Australia New Zealand Act 1991 (FSANZ Act), FSANZ now calls for submissions to assist consideration of the draft food regulatory measure.

Table 14: Summaries in Cluster 2 referring to the same event yet from different sources

Source	LLAMA-3-70B-S	MANUAL-S
WTO Documents online (Notifications)	The FDA Circular updates guidelines for assessing microbiological quality of processed food products, repealing FDA Circular No. 2013-010. The guidelines cover food establishments in the Philippines, providing definitions of terms, and outlining objectives, scope, and general and specific guidelines for ensuring food safety and compliance with Good Manufacturing Practices (GMP). The guidelines reference internationally recognized standards and methods for microbiological analysis.	Repealing Philippines FDA Circular No. 2013-010. The draft Circular aims to provide updated guidelines for the assessment of microbiological quality of certain processed food products and help ensure that food manufacturers comply with Good Manufacturing Practices (GMP).
Food and Drug Administration Philippines	The FDA Circular updates the guidelines for assessing the microbiological quality of processed food products in the Philippines, repealing FDA Circular No. 2013-010. The guidelines cover food establishments engaged in the manufacture, trade, and distribution of processed food products, and provide definitions of terms, objectives, and scope. The guidelines aim to ensure food safety and compliance with Good Manufacturing Practices (GMP) by providing reference criteria for specific food commodities and internationally recognized references for microbiological analysis.	The Philippine Food and Drugs Administration issued the draft for the "Updated Guidelines for the Assessment of Microbiological Quality of Processed Food Products Repealing FDA Circular No. 2013-010 "Revised Guidelines for the Assessment of Microbiological Quality of Processed Foods" for public comments. The reference criteria for specific food commodities as stated in Tables 1 to 15 now includes pre-packaged tubed and cubed ice, ethnic milk-based confectioneries, and food supplements. Also specified in the draft that the methods used for the enumeration or detection of specified microorganisms shall be those that have been internationally established such as FDA BAM, ICMSF, Codex Alimentarius Commission Guidelines, etc. All comments shall be made on or before 26 June 2021

Table 15: Summaries in Cluster 3 referring to the same event yet from different sources

Source	LL70B-S	MANUAL-S
Cluster 4:		
TURKEY MINISTRY OF AGRICULTURE AND FORESTRY	This document outlines amendments to the Turkish Food Codex Regulation on Food Labeling and Consumer Information, focusing on labeling requirements for food products. Key changes include specifications for font sizes, placement of information on labels, and prohibitions on misleading expressions. The amendments aim to ensure accurate and clear labeling, protecting consumers from confusion. Food business operators must comply with these changes by December 31, 2023.	The Ministry of Agriculture and Forestry published a draft amendment to Turkey's Food Codex Regulation stating that any expressions, names and images that can mislead the cosumers are not allowed to be used on food labels. The requirements for location of some elements of the label , e.g. product name, brand, as well as the requirements for font size of the labelling mandatory information are also specifiied.
T.C. GIDA TARIM VE HAYVANCILIK BAKANLIĞI - REPUBLIC OF TURKEY MINISTRY OF FOOD, AGRICULTURE AND LIVESTOCK	This document outlines amendments to the Turkish Food Codex Regulation on Food Labeling and Consumer Information, focusing on labeling requirements for food products. Key changes include specifications for font sizes, placement of information on labels, and prohibitions on misleading expressions. The amendments aim to improve consumer information and prevent confusion. Food business operators must comply with the new provisions by December 31, 2023.	On the surface area where the brand is written, some statements permitted for descriptive name and the formatting provisions Enter in force April 1/2022

Table 16: Summaries in Cluster 4 referring to the same event yet from different sources