LLM Pipeline for Mapping Heterogeneous Data: A Case Study in Food Classification

Kevin Nils Röhl,[1] Rainer Alt,[2] Jan Wirsam[1]

[1] HTW Berlin University of Applied Sciences, Berlin, Germany
roehl@htw-berlin.de, wirsam@htw-berlin.de
[2] Leipzig University, Leipzig, Germany
rainer.alt@uni-leipzig.de

DOI
https://doi.org/10.18690/um.fov.4.2025.30

ISBN
978-961-286-998-4

Accurate food classification is essential for ensuring compliance with dietary regulations, nutritional standards, and sustainability guidelines, but it remains challenging due to fragmented data and semantic complexity. This study presents a pipeline leveraging large language model (LLM) embeddings, ontology mapping, and human-in-the-loop validation to enhance food classification in institutional food services. The pipeline achieves high accuracy in dietary-group mapping (precision 0.94, recall 0.91, F1-score 0.92), though precise FoodEx2 code matching remains challenging. A confidence-based validation strategy effectively balances automated processes with expert oversight to manage ambiguity. The proposed approach enables digital transformation of traditionally fragmented food service systems, enhancing transparency, operational efficiency, and alignment with dietary and public health guidelines. Future research should deploy this pipeline in operational canteen settings to refine embedding techniques, enhance accuracy, and support sustainable nutrition management.

Keywords:
AI pipeline,
LLM embeddings,
classification,
food,
ontology

University of Maribor Press

1 Introduction

The rapid advancement of Emerging Digital Technologies (EDT), including Artificial Intelligence (AI), data pipelines, and automation, is reshaping industries by enabling data-driven decision-making, and process optimization (Oluwaseun Badmus *et al.*, 2024). From healthcare and finance to supply chain management and public services, these technologies offer increased efficiency (Bialas *et al.*, 2023; Elias *et al.*, 2024; Khedr and S, 2024; Kulal *et al.*, 2024). However, their widespread adoption presents significant challenges, particularly regarding data interoperability, ethical AI governance, and transparency (Danks and Trusilo, 2022).

In institutional food services, these challenges manifest in heterogeneous data formats, fragmented digital ecosystems, and the complexity of integrating AI-driven solutions into existing infrastructure (Wolfert *et al.*, 2023; Agrawal *et al.*, 2025; Dhal and Kar, 2025). Ensuring sustainable and nutritionally balanced food offerings in canteens requires a structured approach to data harmonization and regulatory alignment (Gaitán-Cremaschi and Valbuena, 2024). One approach to this is FoodEx2, which is a comprehensive food classification system developed by the European Food Safety Authority and is used to categorize and standardize food products, ingredients, and food-related data (EFSA, 2015). By utilizing food ontologies like FoodOn, food items can be systematically classified and connections between food, health, and the environment established (Dooley *et al.*, 2018). Unlike isolated ontologies, FoodOntoMap links multiple sources with semantic tags, enhancing interoperability (Popovski *et al.*, 2019). This ontology mapping addresses the lack of annotated food corpora and named-entity recognition systems, supporting research on food systems, human health, and sustainability (Popovski, Seljak and Eftimov, 2020). Food classification is essential for evaluating national and international dietary guidelines, as all food-based dietary guidelines (FBDG) rely on food groups. Across 90 countries, FBDG emphasizes core principles such as dietary variety, prioritizing fruits, vegetables, and legumes, and limiting sugar, fat, and salt. In detail, the recommendations on dairy, red meat, and fats can vary (Herforth *et al.*, 2019). A systematic review by Leme et al., 2021 found that, on average, only 40% of individuals across both high-income and low- to middle-income countries meet their national dietary recommendations. Addressing this issue requires scalable tools to evaluate the nutritional quality of the food supply and support the development of effective public health interventions. Despite the challenges, existing research shows promising AI approaches in extracting food data and estimating nutritional values (Hu, Ahmed and L'Abbé, 2023; Harris *et al.*, 2025). To further address these challenges, knowledge graphs, NER and machine learning integration have emerged as a promising solution (Cudré-Mauroux, 2020). Ontologies provide a structured framework for standardizing food-related data, allowing AI-driven systems to harmonize diverse nutritional databases, procurement records, and sustainability tracking tools (Popovski *et al.*, 2019; Shirai *et al.*, 2021; Min *et al.*, 2022).

This study applies the EDT framework to enable business integration and digital transformation within institutional food services, exemplified through company canteens (Serrano-Santoyo *et al.*, 2021). The developed pipeline, leveraging Large Language Models (LLM), addresses fragmented data, inconsistent food classification, and regulatory compliance, thereby enhancing transparency and operational efficiency. By aligning food offerings with dietary guidelines and public health goals, the pipeline supports the nutritional and operational transformation to an adaptive digital ecosystem.

2 Literature

2.1 Data Pipelines

Data pipelines play a critical role in integrating, processing, and analyzing data across multiple sources. They enable data flow between disparate systems, ensuring that information is structured and ready for analysis. In sectors where data originates from heterogeneous and unstructured sources, traditional data pipelines often struggle to maintain interoperability (Foidl *et al.*, 2024). Unlike conventional extract, transform and load systems, modern data pipelines leverage artificial intelligence and automation to handle real-time ingestion, data standardization, and transformation (Kolluri, 2024). In institutional food services, information is often scattered across different systems, including Enterprise Resource Planning (ERP) software, supplier databases, nutritional databases, and menu management tools. These datasets frequently exist in incompatible formats such as Excel spreadsheets, PDFs, CSV files, and proprietary ERP exports. The lack of a standardized data exchange format makes integration difficult, requiring complex preprocessing before meaningful analysis can take place (Zadeh *et al.*, 2018). Greater access to real-world data further complicates the evolution of data schemas across platforms (Zhang *et al.*, 2022).

Current data pipelines struggle with inconsistencies, quality issues, and inefficiencies in dynamic environments. While machine learning can automate schema adaptation and improve data quality, their use in scalable, cross-platform workflows is still limited. More research is needed to develop AI-driven automation for better interoperability, quality assurance, and regulatory compliance (Santhosh Bussa, 2024).

2.2 Ontology Mapping for Data Integration

Ontologies provide structured vocabularies that enable semantic interoperability between disparate datasets. In domains where data consistency is critical, ontology-based classification ensures that different terms referring to the same entity are aligned under a common framework (Gruber, 1995). While ontologies have traditionally been developed as rule-based systems, recent advances in artificial intelligence and natural language processing have enhanced their usability, enabling automated ontology mapping across multiple data sources (Wei and Li, 2025).

Machine learning models, particularly deep learning and transformer-based architectures have significantly improved the automation of entity recognition and classification. LLMs are now capable of extracting structured knowledge from unstructured text, making them valuable tools for ontology mapping in data pipelines (Ciatto *et al.*, 2025). With advanced LLMs from companies like OpenAI, ontology mapping with embedding models like text-embedding-3-large has significantly improved in accuracy. Embeddings represent concepts as high-dimensional vectors, capturing semantic similarities beyond lexical differences. This enhances alignment in complex domains like biomedicine and regulatory compliance, where terminologies evolve across institutions. By reducing reliance on rule-based mappings, embeddings improve scalability while maintaining high precision in entity linking (Sousa, Lima and Trojahn, 2025; Taboada *et al.*, 2025). In the healthcare sector, AI-driven Fast Healthcare Interoperability Resources mapping has enabled seamless integration of patient records across hospitals, insurance providers, and government agencies (Li *et al.*, 2023). Similarly, in supply management, AI-powered ontology mapping has improved logistics efficiency by standardizing product categories across global supplier networks (Regal and Pereira, 2018).

Despite the existence of standardized food classification systems like FoodEx2 and ontologies, their practical implementation remains limited, and current policies still fail to drive the necessary transformation toward a sustainable food system. To overcome fragmented approaches and address systemic crises, a new research and policy agenda is needed that strengthens cross-sectoral governance and effectively integrates food policies (Edwards, Sonnino and López Cifuentes, 2024). AI-based mapping offers a potential solution to those challenges by automatically linking raw ingredient data to a predefined classification system, reducing manual labor and improving data consistency (Hua *et al.*, no date; Goel and Bagler, 2022; Youn, Li and Tagkopoulos, 2023).

2.3 Digital Integration for Sustainable Food Systems

Digital tools like blockchain, AI, and digital twins can boost sustainability in the food value chain by improving efficiency and resource use. Yet, outdated infrastructure, regulatory hurdles, and fragmented data limit their integration and impact (Michel *et al.*, 2024). Achieving meaningful business integration in institutional food services requires overcoming siloed policy approaches by connecting policies across sectors and governance scales. Effective integration involves moving away from isolated processes toward interconnected governance frameworks, addressing interactions among health, environmental, and socio-economic sectors (Edwards, Sonnino and López Cifuentes, 2024). By systematically combining fragmented data from diverse stages of the food supply chain, AI and digitalization provide advanced predictive capabilities. Furthermore, integrating data-driven approaches fosters transparency, trust, and compliance with sustainability objectives throughout the food value chain (Marvin *et al.*, 2022).

3 Methodology

This study follows a Design Science Research approach to develop an AI-driven data pipeline for classification and mapping. The research focuses on structuring unstructured data, aligning it with predefined taxonomies, and integrating mechanisms for human oversight to ensure regulatory traceability. The study applies principles from the Exploratory Framework for Ethical and Regulatory Implications of EDT in Table 1, to examine how AI-driven classification can be validated and mapped to food guidelines. The methodology involves the development of a crosswalk-based framework to establish interoperability between classification systems.

Table 1: Framework for Ethical and Regulatory Implications of EDT

Observed Conditions	Research Questions	Desired Conditions
Recipe data in PDFs is unstructured, making it difficult to analyze and categorize.	How can LLMs be leveraged to classify and map food data from unstructured text?	AI should support, rather than replace, expert decision-making in food classification and dietary assessments.
Current methods for classifying foods against nutritional guidelines are manual, time-consuming	How can regulatory elements be integrated into AI-driven food classification?	The pipeline should improve efficiency while ensuring transparency

3.1 Architecture

The LLM pipeline for heterogeneous data in Figure 1 supports transparent and regulatory-compliant AI decisions through ontology mapping, iterative learning, and structured expert oversight. By integrating human-in-the-loop validation, the pipeline ensures alignment with policy and ethical governance principles. Its architecture facilitates business integration by transforming fragmented food-service data into a unified digital ecosystem, shifting from manual, disconnected processes toward integrated compliance and sustainability management. The pipeline's adaptable design further allows implementation across multiple domains, including public health, sustainability initiatives, and environmental monitoring, where regulatory tracking and accuracy are critical.

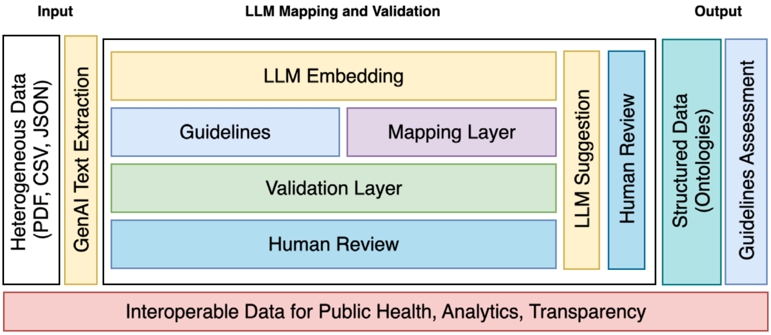
Input
LLM Mapping and Validation
Output
Heterogeneous Data (PDF, CSV, JSON)
GenAI Text Extraction
LLM Embedding
Guidelines
Mapping Layer
Validation Layer
Human Review
LLM Suggestion
Human Review
Structured Data (Ontologies)
Guidelines Assessment
Interoperable Data for Public Health, Analytics, Transparency

Figure 1: LLM Pipeline for Heterogeneous Data Mapping

Unstructured data from diverse sources (e.g., text documents, PDFs, databases) can be extracted and standardized using Natural Language Processing and entity recognition models. AI models generate semantic embeddings that allow for ontology-based classification, linking extracted data to structured taxonomies or regulatory databases. The use of crosswalk files ensures consistency and facilitates interoperability between national and international guidelines. This approach allows the pipeline to adapt to country-specific regulatory frameworks while maintaining cross-domain applicability. AI-generated classifications undergo a confidence-based validation process, where low-confidence mappings are flagged for human review. This ensures that automated classification aligns with national guidelines, balancing efficiency with expert oversight. Validated classifications feed into an iterative learning system, improving ontology mapping accuracy over time. This continuous refinement makes the system adaptive to policy changes.

3.2 Use-Case Canteen Recipe Ingredient Classification

OpenAIs LLM embedding "text-embedding-3-large" is used to match extracted entities to predefined taxonomies, selected for its semantic performance and ease of integration. Human validation processes refine mappings and address uncertainties. The pipeline architecture is implemented in the context of a company canteen setting, where ingredient names can be extracted from PDFs, and mapped to Planetary Health Diet (PHD) groups and FoodEx2 term codes.

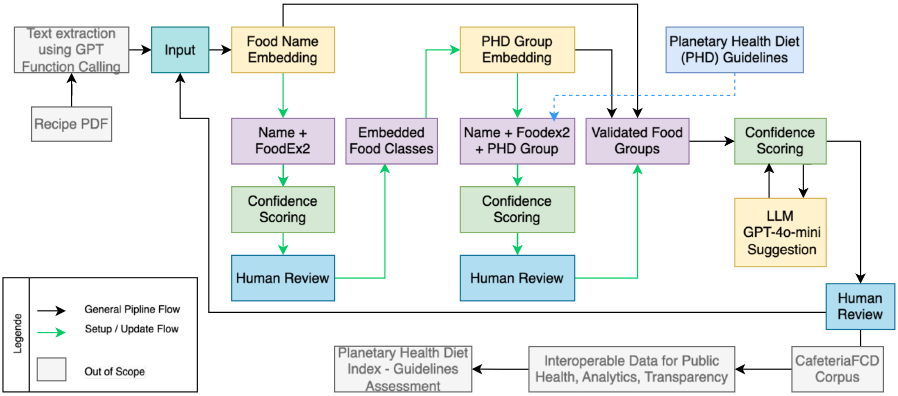

Figure 2: Company Canteen Pipeline for PHD Mapping

3.3 Validation

To validate the embeddings mapping, a test set of FoodEx2 names were manually labeled with PHD categories from the 430 subgroups of the FoodEx2 hierarchy code and applied to the underlying names. The first and second hierarchy has been removed to reduce misleading general food category matching, resulting in 4,416 entries. Flavors, vitamins, chemical elements, and composite dishes, which could belong to multiple food groups, were labeled as "unknown". The final validation file consists of 2,750 entries, including FoodEx2 codes and PHD group classifications. To assess the effectiveness of the pipeline, F1-score and recall metrics were used to evaluate the PHD group embeddings generated by the "text-embedding-3-large" model against the validated file. A second test was conducted using zero-shot prompting with gpt4o-mini for mapping PHD groups to the FoodEx2 validation names. Another validation for ingredient data, was performed by mapping FoodEx2 name embeddings to given ingredients from FRIDA, the national food database from Denmark. FoodEx2 codes from FRIDA have been filtered based on the "unknown" codes list earlier, resulting in 935 entries with food names and their validated FoodEx2 codes. Embeddings have been created for all entries using "text-embedding-3-large" to compare them with the embedded FoodEx2 names. Cosine similarity was applied to determine the accuracy of this mapping.

4 Results

Figure 3 presents the classification results for a food group matching task using an embedding model ("text-embedding-3-large") and GPT-4o-mini. The GPT-4o-mini model exhibits higher performance across all food groups. The "text-embedding-3-large" model achieves a precision of 0.84, recall of 0.78, and an F1-score of 0.79, while GPT-4o-mini attains a precision of 0.94, recall of 0.91, and an F1-score of 0.92. In the "text-embedding-3-large" classification, the food categories "eggs" (precision: 0.17, recall: 1.00) and "unsaturated fats" (precision: 1.00, recall: 0.06) represent the lowest-performing cases, with eggs displaying low precision but high recall, and unsaturated fats exhibiting high-precision but low recall. Additionally, starchy vegetables show low precision (0.51) and recall (0.53). In contrast, dairy (precision: 0.94, recall: 0.84) is among the highest-performing categories. The GPT-4o-mini model demonstrates its lowest classification performance in the "unsaturated fats" (precision: 0.79, recall: 0.62) and "saturated fats" (precision: 0.71, recall: 0.78) categories, while achieving the highest precision and recall for "fish and seafood" (precision: 0.99, recall: 0.98) and "eggs" (precision: 0.97, recall: 0.97).

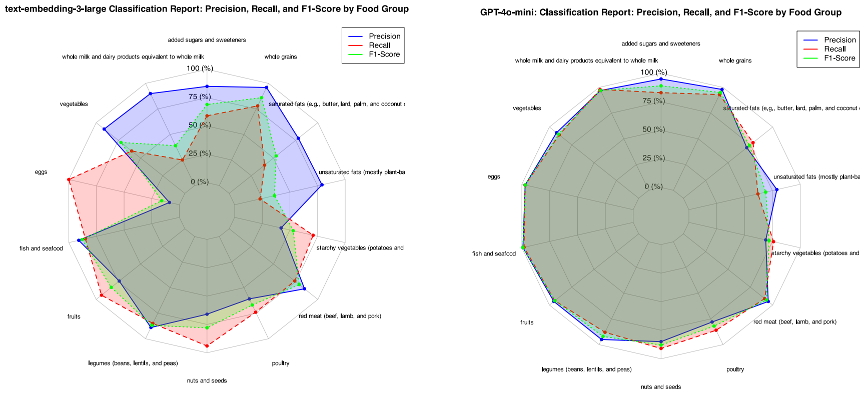

Figure 3: Food Group Classification Report

The similarity score distribution in Figure 4 shows a distinction between correct and incorrect classifications of FoodEx codes, with correct matches generally having higher similarity scores. Incorrect classifications tend to cluster at lower values, with some overlap between the distributions. The median similarity score for correct matches is higher than for incorrect ones.

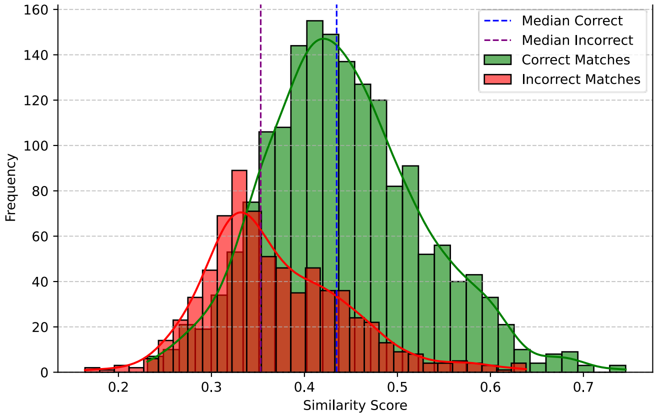

Figure 4: Distribution of Similarity Scores for PHD Category Matches

Table 2 provides two examples of false classifications in the text-embedding model. The model misclassified "emperors" as eggs (similarity score: 0.257) and "sea asters" as fish and seafood (similarity score: 0.372).

Table 2: False PHD Category Labels

Food Name	Correct Category	Incorrect Category	Cosine Similarity
Emperors	Fish and seafood	Eggs	0.257
Sea asters	Vegetables	Fish and seafood	0.372

4.1 FoodEx2 Code Matching Results

Figure 5 visualizes the distribution of similarity scores for correct and incorrect FoodEx2 code matches using OpenAI's "text-embedding-3-large" model. Correct matches are shown in green (median 0.73), while incorrect matches are shown in red (median 0.66). The graph on the right displays the accuracy of FoodEx2 code matching based on cosine similarity across the closest three matches. The 1st match has an accuracy of 46.28%, the 2nd match has 13.19%, and the 3rd match has 5.74%.

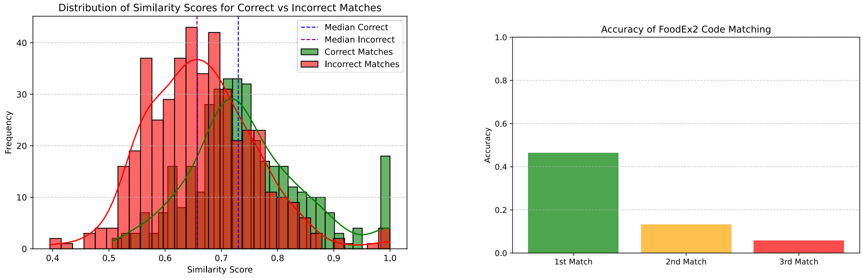

Figure 5: Embedding Matches of FoodEx2 and FRIDA Food Names

5 Discussion

The developed LLM-based classification pipeline demonstrates significant potential for aligning food categorization with national and international dietary guidelines, such as the PHD. By integrating crosswalk files, category assessment, and confidence-based validation, the system provides a structured approach to food classification. The ability to automate category assignment using GPT-4o-mini has shown promising results, with the model successfully aligning food items to broader dietary groups with high accuracy. This suggests that LLM-assisted classification can support large-scale food policy assessments, improving transparency in food system monitoring. However, while category-level classification performs well, the exact matching of FoodEx2 codes remains a challenge due to the nuances in food naming conventions. The results showed that 46.28% of food items were correctly classified in the first match, while additional codes were identified in the second and third matches. A small part of the matching is shown in Table 3 in the appendix. This highlights that even when the top match is incorrect, valid classifications frequently appear in the nearest similarity ranks, demonstrating that embeddings capture meaningful relationships between food names even if perfect alignment is not always achieved.

5.1 Balancing Precision

Overlapping similarity scores between correct and incorrect classifications of FoodEx2 codes, complicate establishing a clear decision threshold. Especially when food names differ subtly due to preparation methods or regional variations. For example, categories like "eggs" and "unsaturated fats" illustrate how semantic ambiguity can significantly impact precision and recall.

This underscores the need for a strategic balance between automation accuracy and expert validation. Setting similarity thresholds too strictly reduces classification errors but increases manual effort, while lower thresholds boost automation but risk misclassification. Thus, confidence-based human oversight is essential for balancing precision effectively. Ensuring this balance is critical for reliable dietary assessment, effective regulatory compliance, and successful digital transformation in nutritional management and business integration.

5.2 Limitations and Future Research Directions

The results indicate specific challenges in food category classification, particularly in groups like "eggs" and "unsaturated fats", where semantic ambiguity and overlapping similarity scores reduce classification accuracy. This highlights the need for further refinement in category-level matching, especially for food names overlapping multiple categories. A potential improvement would be to introduce finer-grained subcategories within broader food groups and use crosswalk files to systematically link them to higher-level categories. This could improve classification accuracy by allowing AI models to recognize context-specific variations.

A key limitation of this study is the reliance on FRIDA FoodEx2 codes as the validation dataset. As demonstrated in Table 3, a direct name match for boiled potatoes does not correspond to the expected code due to FRIDA's assignment of a distinct FoodEx2 code for the boiling process, despite the existence of a code for boiled potatoes. Further cleaning by removing or separating cooking processes could improve the matching. This discrepancy constrains validation accuracy, underscoring the need for a more comprehensive and systematically curated validation dataset. More broadly, this issue reflects fundamental challenges within food systems ontology, where inconsistencies in classification, granularity, and contextual interpretation can limit interoperability and data harmonization across different databases.

For FoodEx2 classification, a more effective strategy could also involve ranking the top five similarity-based matches instead of relying solely on the highest-scoring option. Presenting multiple candidate classifications would allow an AI-driven decision layer to assess contextual relevance and select the most appropriate match. This approach could improve classification robustness, particularly in cases where food names have slight variations due to preparation methods, processing techniques, or regional terminology differences. Further research could also explore reasoning models, few-shot approaches, and smaller models for local usage. Reasoning models and Few-shot learning approaches could help resolve semantic ambiguities by leveraging contextual understanding, improving classification in complex categories like "eggs" and "unsaturated fats". Smaller, locally deployable models could enhance data security and privacy by processing sensitive food classification data on-device. This approach minimizes data transmission risks, ensures compliance with privacy regulations, and allows organizations to retain control over proprietary datasets. Further research should implement the LLM pipeline in operational canteens, with active involvement of stakeholders, for evaluating the pipeline adoption in institutional food services.

6 Conclusion

This study explored an LLM food classification pipeline to address technical and regulatory challenges in institutional food services. The developed LLM pipeline, combining embeddings, crosswalk mapping, and human validation, improves dietary assessment, data standardization, and regulatory compliance. While GPT-4o-mini achieved high accuracy at the dietary group level, precise FoodEx2 code matching remains challenging due to semantic ambiguity and naming variations. A confidence-based validation ensures transparency and accountability by balancing automation with human oversight, supporting expert decision-making. It demonstrates AI's potential to transform fragmented processes into integrated digital ecosystems through practical and ethical interventions. Future research should focus on evaluating the pipeline within operational canteen environments, engaging stakeholders directly, investigating LLM embeddings, reasoning models, and locally deployable AI models to enhance data privacy, nutritional quality, and sustainability outcomes.

References

Agrawal, K. *et al.* (2025) 'AI-driven transformation in food manufacturing: a pathway to sustainable efficiency and quality assurance', *Frontiers in Nutrition*, 12. Available at: https://doi.org/10.3389/FNUT.2025.1553942.

Bialas, C. *et al.* (2023) 'A Holistic View on the Adoption and Cost-Effectiveness of Technology-Driven Supply Chain Management Practices in Healthcare', *Sustainability 2023, Vol. 15, Page 5541*, 15(6), p. 5541. Available at: https://doi.org/10.3390/SU15065541.

Ciatto, G. *et al.* (2025) 'Large language models as oracles for instantiating ontologies with domain-specific knowledge', *Knowledge-Based Systems*, 310, p. 112940. Available at: https://doi.org/10.1016/J.KNOSYS.2024.112940.

Cudré-Mauroux, P. (2020) 'Leveraging Knowledge Graphs for Big Data Integration: The XI Pipeline', *Semantic Web*, 11(1), pp. 13–17. Available at: https://doi.org/10.3233/SW-190371.

Danks, D. and Trusilo, D. (2022) 'The Challenge of Ethical Interoperability', *Digital Society 2022 1:2*, 1(2), pp. 1–20. Available at: https://doi.org/10.1007/S44206-022-00014-2.

Dhal, S.B. and Kar, D. (2025) 'Leveraging artificial intelligence and advanced food processing techniques for enhanced food safety, quality, and security: a comprehensive review', *Discover Applied Sciences*, 7(1). Available at: https://doi.org/10.1007/S42452-025-06472-W.

Dooley, D.M. *et al.* (2018) 'FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration', *npj Science of Food 2018 2:1*, 2(1), pp. 1–10. Available at: https://doi.org/10.1038/s41538-018-0032-6.

Edwards, F., Sonnino, R. and López Cifuentes, M. (2024) 'Connecting the dots: Integrating food policies towards food system transformation', *Environmental Science & Policy*, 156, p. 103735. Available at: https://doi.org/10.1016/J.ENVSCI.2024.103735.

EFSA (2015) 'The food classification and description system FoodEx 2 (revision 2)', *EFSA Supporting Publications*, 12(5), p. 804E. Available at: https://doi.org/https://doi.org/10.2903/sp.efsa.2015.EN-804.

Elias, O. *et al.* (2024) 'The evolution of green fintech: Leveraging AI and IoT for sustainable financial services and smart contract implementation', *https://wjarr.com/sites/default/files/WJARR-2024-2272.pdf*, 23(1), pp. 2710–2723. Available at: https://doi.org/10.30574/WJARR.2024.23.1.2272.

Foidl, H. *et al.* (2024) 'Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers', *Journal of Systems and Software*, 207, p. 111855. Available at: https://doi.org/10.1016/J.JSS.2023.111855.

Gaitán-Cremaschi, D. and Valbuena, D. (2024) 'Examining purchasing strategies in public food procurement: Integrating sustainability, nutrition, and health in Spanish school meals and social care centres', *Food Policy*, 129, p. 102742. Available at: https://doi.org/10.1016/J.FOODPOL.2024.102742.

Goel, M. and Bagler, G. (2022) 'Computational gastronomy: A data science approach to food', *Journal of Biosciences*, 47(1). Available at: https://doi.org/10.1007/S12038-021-00248-1.

Gruber, T.R. (1995) 'Toward principles for the design of ontologies used for knowledge sharing?', *International Journal of Human-Computer Studies*, 43(5–6), pp. 907–928. Available at: https://doi.org/10.1006/IJHC.1995.1081.

Harris, J. *et al.* (2025) 'Evaluating Large Language Models for Public Health Classification and Extraction Tasks'.

Herforth, A. *et al.* (2019) 'A Global Review of Food-Based Dietary Guidelines', *Advances in Nutrition*, 10(4), pp. 590–605. Available at: https://doi.org/10.1093/ADVANCES/NMY130.

Hu, G., Ahmed, M. and L'Abbé, M.R. (2023) 'Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods', *American Journal of Clinical Nutrition*, 117(3), pp. 553–563. Available at: https://doi.org/10.1016/J.AJCNUT.2022.11.022.

Hua, A. *et al.* (no date) 'NUTRIBENCH: A Dataset for Evaluating Large Language Models on Nutrition Estimation from Meal Descriptions'. Available at: https://mehak126.github.io/nutribench.html (Accessed: 6 February 2025).

Khedr, A.M. and S, S.R. (2024) 'Enhancing supply chain management with deep learning and machine learning techniques: A review', *Journal of Open Innovation: Technology, Market, and Complexity*, 10(4), p. 100379. Available at: https://doi.org/10.1016/J.JOITMC.2024.100379.

Kolluri, S. (2024) 'Automating Data Pipelines with AI for Scalable, Real-Time Process Optimization in the Cloud', *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(6), pp. 2070–2079. Available at: https://doi.org/10.32628/CSEIT242612405.

Kulal, A. *et al.* (2024) 'Enhancing public service delivery efficiency: Exploring the impact of AI', *Journal of Open Innovation: Technology, Market, and Complexity*, 10(3), p. 100329. Available at: https://doi.org/10.1016/J.JOITMC.2024.100329.

Leme, A.C.B. *et al.* (2021) 'Adherence to Food-Based Dietary Guidelines: A Systemic Review of High-Income and Low- and Middle-Income Countries', *Nutrients*, 13(3). Available at: https://doi.org/10.3390/NU13031038.

Li, Y. *et al.* (2023) 'Enhancing Health Data Interoperability with Large Language Models: A FHIR Study'. Available at: https://arxiv.org/abs/2310.12989v1 (Accessed: 21 March 2025).

Marvin, H.J.P. *et al.* (2022) 'Digitalisation and Artificial Intelligence for sustainable food systems', *Trends in Food Science & Technology*, 120, pp. 344–348. Available at: https://doi.org/10.1016/J.TIFS.2022.01.020.

Michel, M. *et al.* (2024) 'Benefits and challenges of food processing in the context of food systems, value chains and sustainable development goals', *Trends in Food Science & Technology*, 153, p. 104703. Available at: https://doi.org/10.1016/J.TIFS.2024.104703.

Min, W. *et al.* (2022) 'Applications of knowledge graphs for food science and industry', *Patterns*, 3(5), p. 100484. Available at: https://doi.org/10.1016/J.PATTER.2022.100484.

Oluwaseun Badmus *et al.* (2024) 'AI-driven business analytics and decision making', *World Journal of Advanced Research and Reviews*, 24(1), pp. 616–633. Available at: https://doi.org/10.30574/WJARR.2024.24.1.3093.

Popovski, G. *et al.* (2019) 'FoodOntoMap: Linking Food Concepts across Different Food Ontologies'. Available at: https://doi.org/10.5220/0008353201950202.

Popovski, G., Seljak, B.K. and Eftimov, T. (2020) 'FoodOntoMapV2: Food Concepts Normalization Across Food Ontologies', *Communications in Computer and Information Science*, 1297, pp. 413–426. Available at: https://doi.org/10.1007/978-3-030-66196-0_19/TABLES/5.

Regal, T. and Pereira, C.E. (2018) 'Ontology for Conceptual Modelling of Intelligent Maintenance Systems and Spare Parts Supply Chain Integration', *IFAC-PapersOnLine*, 51(11), pp. 1511–1516. Available at: https://doi.org/10.1016/J.IFACOL.2018.08.285.

Santhosh Bussa (2024) 'Evolution of Data Engineering in Modern Software Development', *Journal of Sustainable Solutions*, 1(4), pp. 116–130. Available at: https://doi.org/10.36676/J.SUST.SOL.V1.I4.43.

Serrano-Santoyo, A. *et al.* (2021) 'Ethical implications regarding the adoption of emerging digital technologies: An exploratory framework', *Progress in Ethical Practices of Businesses: A Focus on Behavioral Interactions*, pp. 219–239. Available at: https://doi.org/10.1007/978-3-030-60727-2_12/FIGURES/6.

Shirai, S.S. *et al.* (2021) 'Identifying Ingredient Substitutions Using a Knowledge Graph of Food', *Frontiers in Artificial Intelligence*, 3. Available at: https://doi.org/10.3389/FRAI.2020.621766.

Sousa, G., Lima, R. and Trojahn, C. (2025) 'Complex Ontology Matching with Large Language Model Embeddings'. Available at: http://arxiv.org/abs/2502.13619 (Accessed: 19 March 2025).

Taboada, M. *et al.* (2025) 'Ontology Matching with Large Language Models and Prioritized Depth-First Search'. Available at: http://arxiv.org/abs/2501.11441 (Accessed: 19 March 2025).

Wei, Y. and Li, X. (2025) 'Knowledge-enhanced ontology-to-vector for automated ontology concept enrichment in BIM', *Journal of Industrial Information Integration*, p. 100836. Available at: https://doi.org/10.1016/J.JII.2025.100836.

Wolfert, S. *et al.* (2023) 'Digital innovation ecosystems in agri-food: design principles and organizational framework', *Agricultural Systems*, 204, p. 103558. Available at: https://doi.org/10.1016/J.AGSY.2022.103558.

Youn, J., Li, F. and Tagkopoulos, I. (2023) 'Semi-Automated Construction of Food Composition Knowledge Base'. Available at: https://github.com/ibpa/SemiAutomatedFoodKBC. (Accessed: 6 February 2025).

Zadeh, A.H. *et al.* (2018) 'Cloud ERP systems for smalland-medium enterprises: A case study in the food industry', *Journal of Cases on Information Technology*, 20(4), pp. 53–70. Available at: https://doi.org/10.4018/JCIT.2018100104.

Zhang, J. *et al.* (2022) 'Best practices in the real-world data life cycle', *PLOS Digital Health*, 1(1), p. e0000003. Available at: https://doi.org/10.1371/JOURNAL.PDIG.0000003.

Appendix

Table 3: FRIDA food name embedding cosine similarity FoodEx2 matches

FRIDA Name	FoodEx2 Name 1	cosSim 1	Match 1	FoodEx2 Name 2	cosSim 2	Match 2	FoodEx2 Name 3	cosSim 1	Match 3
goat milk	goat milk	1,00	True	goat milk fat	0,84	False	yoghurt goat milk	0,78	False
potato boiled	potato boiled	1,00	False	potato baked	0,73	False	potatoes	0,63	True
semolina	semolina	0,99	False	wheat semolina	0,82	True	maize semolina	0,74	False
rice flour	rice flour	1,00	True	rice grain	0,77	False	rice rolled grains	0,72	False