

MEASURING WHAT WE MISJUDGE: EXPLORING THE POTENTIAL AND LIMITATIONS OF AI FOR FOOD WASTE QUANTIFICATION

Mengting Yu ^{1,*}, Luigi Palumbo ¹, Clara Cicatiello ², Tiziana Laureti ¹ and Luca Secondi ¹

¹ Department of Economics, Engineering, Society and Business Organization (DEIM), University of Tuscia, via del Paradiso 47, 01100 Viterbo, Italy

² Department of Innovation in Biological, Agro-food and Forestry Systems (DIBAF), University of Tuscia, via S. Camillo de Lellis, snc, 01100 Viterbo VT, Italy

Article Info:

Received:
6 March 2026
Revised:
17 September 2026
Accepted:
22 September 2026
Available online:
30 September 2026

Keywords:

Food waste quantification
Food waste tracking
Food recognition
Artificial intelligence
AI image analysis
Large Language Model

ABSTRACT

Food waste remains a critical yet insufficiently explored global challenge, and accurate quantification is essential for developing effective mitigation strategies. Traditional food waste measurement methods face multiple limitations, prompting growing interest in alternative approaches such as AI-based technologies. This exploratory study evaluates the potential of AI image analysis for food waste quantification under real-world conditions. The method involves comparing AI-estimated and human-estimated food waste weights against a benchmark of scale-measured weights to assess estimation accuracy and error patterns. Our findings indicate that while AI technologies show promise as scalable quantification tools, portion effect on estimation errors and variability in generalization across models underscores the need for thoughtful selection, and the performance could be substantially improved through further model training with larger and more diverse datasets. Overall, this study highlights both the potential and present limitations of AI-driven food waste quantification, emphasizing the need for continued research using more advanced models and expanded training data.

1. INTRODUCTION

Food loss and waste (FLW) remain among the most pressing challenges in global food systems. In Europe, more than 58 million tonnes of food are wasted annually (Eurostat, 2025) – a figure that contrasts sharply with the fact that nearly 42 million people cannot afford a quality meal every second day (Eurostat, 2023). While FLW occurs across the entire food value chain, increasing attention has been directed towards the consumption stage, where waste is generated after food has reached the consumers. In the European Union, the three largest contributors to food waste (FW) are households (53%), the manufacture of food products and beverages (19%), and restaurants and food services (11%) (Eurostat, 2025). Globally, an estimated 1.05 billion tonnes of food were wasted, accounting for 19% of food available to consumers at the household, food services, and retail levels (UNEP, 2024). These figures underscore the urgency of developing more accurate and scalable solutions for monitoring and mitigating FW, particularly in the later stages of the supply chain, namely the consumption stage (Dhir et al., 2020; Kaur et al., 2021).

FW quantification is a crucial first step with respect to the urge to develop and implement FW mitigation strate-

gies (Corrado et al., 2019; Vardopoulos et al., 2024). The traditional approaches include self-reported surveys, manual waste tracking, and compositional analyses, which have provided valuable insights but face limitations of accuracy, scalability, and cost-efficiency, which might lead to misjudgments on FW-related problems (Corrado et al., 2019; Vardopoulos et al., 2024). Recent advances in artificial intelligence (AI) and computer vision offer new opportunities to overcome these limitations. The advancement of AI unlocks a new era in food science towards more data-driven decision-making. Among various AI technologies, Large Language Models (LLMs) have witnessed successes in food safety detection and monitoring, nutritional analysis, recipe development, and diet planning (Fridolfsson et al., 2025; Ma et al., 2024). Reference to the success of using LLMs in food images (e.g. nutritional content estimation, quality inspection) (Fridolfsson et al., 2025; Ma et al., 2024), LLMs powered with a computer-vision system are capable of automatically identifying, classifying, and quantifying FW in real time, reducing reliance on subjective reporting and enabling continuous monitoring in households, food services, and retail contexts. By generating high-resolution and standardized data, AI technologies have the potential to transform FW research and practice, supporting

 * Corresponding author:
Mengting Yu
email: mengting.yu@unitus.it

more targeted activities for FW reduction with data-driven evidence (Kakani et al., 2020; Sigala et al., 2025).

However, AI in FW research remains at an early stage. Empirical research in FW quantification with AI technologies is rare in the current literature. There is also a limited understanding of how AI-enabled monitoring can improve the accuracy and efficiency of FW quantification, support real-time decision-making, and enhance the effectiveness of reduction strategies across different contexts of consumption.

Building on these gaps, this study explores the application of AI-based technologies, particularly using LLMs, for FW quantification in terms of their potential and limitations. Specifically, it addresses the following research questions:

- Is AI-based technology a reliable tool for FW quantification compared to traditional methods (e.g. subjective evaluation, manual measurement)?
- What is the level of errors when using AI for FW quantification?
- How can AI-enabled data streams be integrated into broader strategies to quantify FW at scale?

By addressing these questions, the study contributes to the literature by advancing the use of AI technologies as robust tools for FW quantification – as far as we know, the first attempt to systematically evaluate the capabilities of the open-source LLMs in FW identification and tracking. This study examines the accuracy of FW measurement by comparing subjective and objective weight data, with a particular focus on the potential of AI-based image analysis to enhance reliability. This approach contributes to advancing methodological tools for large-scale FW tracking, improving data quality, and supporting more effective strategies for FW reduction.

In the following context, Section 2 provides a background on the existing literature about FW quantification and AI for food science. Section 3 explains the research design and methodology. We extend the discussion with our analysis results in Section 4. Lastly, a conclusive summary closes this exploratory study while pointing out the limitations and directions for future work in Section 5.

2. LITERATURE REVIEW

2.1 Food waste and food waste quantification

We have been standing at a critical turning point since 2022: on the one hand, nearly one-fifth of the food available to consumers at the retail, food service, and household levels is wasted (UNEP, 2024); on the other hand, 733 million people in the world continue to suffer from hunger and 2.33 billion are in danger of food insecurity (FAO, 2024). Besides, FW contributes to a less sustainable global environment, in particular, greenhouse gas emissions, water resources, land, and biodiversity (Stancu et al., 2016). FW quantification, or FW accounting, forms the baseline of tackling any FW issues: identification of the most relevant waste streams (Corrado et al., 2019). Methodologies used for FW quantification vary from direct measurement (e.g. weighing, digital trackers) to surveys and observations. Each ap-

proach has its pros and cons, contributing differently to the advancement of FW research – direct measurement provides more details and comparable estimates while others shed lights on FW generation from behaviors of consumers and businesses (Falasconi et al., 2015; Derqui et al., 2016; Vardopoulos et al., 2024). The measured or estimated results of FW quantity share a high inconsistency rate. Lack of standardized framework and methodology, and the complexity of food systems are the major challenges (Comis et al., 2025; Filimonau & De Coteau, 2019); however, different methods to measure and achieve FW quantity contribute to the discrepancy as well (Corrado et al., 2019).

2.2 Methodological approaches for food waste quantification

Methodological approaches for FW quantification discussed in this paper give exclusive attention to the methods for collecting primary data, for the reason that we aim to evaluate an alternative method or tool for primary FW data collection. Direct and non-direct measurement methods are both widely used in the existing literature; furthermore, a combination of methods is employed to enhance the robustness of the research (Filimonau et al., 2022; Vardopoulos et al., 2024).

Direct measurement includes physical weighing and volumetric evaluation. According to the different methodologies applied, the measured weights or volumes are in different unit formats, such as per meal/tray, of each type of food component, or per person/entire sample size (Vardopoulos et al., 2024). However, these methods are highly resource-demanding, including cost, time, expertise and direct access to the research targets, consequently limiting the scalability of the research results (Corrado et al., 2019).

Waste registration and questionnaires are used to measure FW quantity at both business and individual (or household) levels indirectly. The commonly used techniques to register FW quantity or conduct a FW diary are visual observation and self-reporting. Compared to direct measurement methods, they are more approachable operationally and financially, repeatable, and practical for a large-scale implementation, despite the drawbacks of data quality due to inaccurate data entry, subjective enlistment, sample bias, and other concerns (Corrado et al., 2019; Vardopoulos et al., 2024).

2.3 AI for food science

Digital innovations have been largely integrated into the research and applications in food science and food systems, offering impressive positive outcomes which cannot be achieved by non-digital approaches. Taking a few examples, real-time tracking and visualized reports to the key stakeholders, and nudging effects at an immediate and long-term level on individuals. AI, in particular, LLMs among various AI technologies, have been catching eyes in recent years for their high reputation in data analysis and interpretation (Kakani et al., 2020; Ma et al., 2024; Yu et al., 2025). LLMs are a group of advanced models which are developed and trained to generate text by predicting the next word based on an input text. Powered by deep learn-

ing algorithms, these models are powerful in processing natural languages, predicting, making decisions, and performing other complex text-based tasks (Ma et al., 2024). The big names of LLMs with different developers include but not limited to ChatGPT by OpenAI, Gemini by Google, Llama by Meta, Claude by Anthropic, DeepSeek by DeepSeek, and Qwen by Alibaba. Taking OpenAI as an example, ChatGPT is only one model developed for natural language processing tasks such as text and code generation, data analysis and visualization, image analysis and creation, deep searching, and so on (OpenAI, n.d.). However, developed by the same company, ChatGPT is outperformed by another model, GPT-4o, in the context of visual and audio processing tasks (OpenAI, n.d.). Despite the tailored LLMs for specific tasks, the general performance and capabilities of an LLM depend on model complexity, data volume, and computational power, although all models share a similar fundamental backbone. From this perspective, LLMs stand for large-scale pre-trained language models (Ma et al., 2024). One of the straightforward indicators is the parameter size, which, to some extent, differentiates LLMs from other small language models. In the context of AI, parameters refer to the internal variables learned during training that determine how the model processes inputs and generates outputs. The number of parameters represents a model's capacity, as larger models typically possess greater representational power and can capture more nuanced patterns. The LLMs exemplified before have model versions ranging from billions to trillions parameters (Ma et al., 2024).

LLMs have been experimented with and quickly adapted in many fields of food science, such as predictive analysis for food safety, regulatory compliance automation, food quality control and inspection, quality improvement according to consumer feedback analysis, efficiency improvement of the whole food supply chain, food fraud detection, recipe generation, and nutritional analysis for diet planning (Kakani et al., 2020; Ma et al., 2024). Which, the last two applications highlight the capabilities of LLMs in food image processing, and therefore, the potential application to FW research. In the context of recipe generation, it can capture the information by reading food images and generate recipes to meet specific dietary requirements or preferences. On the other hand, LLMs can understand complex dietary data by extracting key information about food components and quantities, so that personalized nutrition advice and diet plans can be generated (Lo et al., 2024; Ma et al., 2024; Fridolfsson et al., 2025). Lo et al. (2024) achieves mean absolute errors lower than 5% even in challenging conditions when comparing the estimated results by GPT-4V to those of dietitians. Not only are common food items correctly identified, but also regional ones with cultural and geographic features. The estimation results of the portion size are close to those evaluated by humans, although the model is not designed to process depth or spatial information in the food images and bears a certain level of estimation errors. Leading LLMs like ChatGPT and Claude demonstrate an accuracy level of food weight estimation from the images, close to the self-reporting method

in another recent work in dietary assessment (Fridolfsson et al., 2025).

LLMs have shown notable capabilities in food and food component recognition in the research field of nutrition and diet. In the meantime, the application of weight estimation has achieved performance no worse than traditional self-reported methods; however, unlike the traditional methods, LLMs process the information without subjective bias, and they are free from social desirability effects or participant burden. LLMs can support tasks like automation at a large scale and continuous passive monitoring, hence offering efficiency and scalability. This turns LLMs into a potential and valuable tool for other scientific research tasks, including research in FW. Meanwhile, limitations of LLMs persist, at least for the current stage. The accuracy of the volume estimation from 2D food images needs real-world validation.

2.4 AI for food waste

The academic discussion around digital innovations in FW has been expanding, but to our best knowledge, the implementation of AI technologies in FW is very limited. Kitro, the Swiss startup, integrates a self-developed AI model with the IoT solution for FW management. Through grounded data analysis, operational optimization for food businesses is achieved, and benefits the users with up to 55% of FW reduction and 5% cost saving (Martin-Rios et al., 2020). However, the AI technology used by Kitro is only a small friction. Kitro's AI-empowered solution is designed specifically for food service providers with commercial purposes; it is limited in generalization. In an earlier attempt with a small sample size (n=54), a leading LLM, GPT-4o by OpenAI, is used to process the food images taken in the real-world environment. It shows remarkable capability in food recognition and component identification, and a significant linear discrepancy between the AI-estimated and self-reported FW weights (Yu et al., 2025). To our best knowledge, this paper takes the first experiment and comparison study on different open-source LLMs in FW identification and quantification with the purpose of accuracy evaluation.

3. DATA COLLECTION AND METHODOLOGY

3.1 Research design and data collection

This exploratory study was designed to evaluate and compare the accuracy of food weight estimation and food identification generated by LLMs and human judgment using food images captured in real-world environments. The study was conducted within the framework of the PERFETTO project, financed by the European Union, which aims to reduce consumer FW in the restaurants through behavioral nudges – such as doggy bags and half-portion – and employs AI-assisted monitoring technologies within a large-scale field experiment in 40 restaurants in the Umbria region of Italy. The primary objective was to assess the reliability and practical applicability of AI-based estimation by benchmarking its outputs against scale measurements and comparing them to human estimates. To this end, a dual data collection framework was implemented, consist-

ing of food images, scale-based measurements, AI-generated estimates, and human-based assessments.

For the human estimation component, purposive sampling was employed, targeting academic researchers and specialists in waste management, sustainability, and economics. As a non-probability sampling strategy, purposive sampling offers operational feasibility and analytical relevance for exploratory research (Vehovar et al., 2016). The rationale for targeting this participant group was that their professional activities are often associated with waste quantification, waste source accounting, and numerical assessment, making them a relevant stakeholder group for studies related to FW estimation. However, participants were not recruited on the basis of expertise in food portion-size or food weight estimation. To further expand the participant pool and improve coverage across relevant disciplines, a snowball recruitment approach was adopted, whereby initial participants were invited to refer other qualified experts (Ting et al., 2025). We should note the potential selection bias and limitation of the sample representativeness. Information about participants' familiarity with food types, cultural background, demographic characteristics, and prior training in volumetric estimation was not collected. Consequently, the influence of these factors on human estimation performance could not be assessed and should be considered when interpreting the results.

Data collection was conducted across three distinct periods between 2025 and 2026. The dataset comprises two main sources: image batch i with food images, scale-measured weights, and image description collected directly by the research team in both household and out-of-home dining settings, and image batch ii supplementary data contributed by a group of academic experts, who followed detailed guidelines provided by the researchers. In addition to collecting the same data as image batch i, this batch also included weight estimation of 9 images by these experts. The scale-measured food weights ranged from 17 g to 621 g in image batch i and from 5 g to 550 g in image batch ii.

All food images were captured in free environments to reflect real-world consumption contexts. For all observations in image batch i, a digital kitchen scale (Brand: SUNLUX MDC005; CE certificated) with weighing precision of 1 g and maximum weighing capacity of 5 kg, was used to record the precise weight of food in the images, while a smartphone camera was used to capture corresponding images. Before each weighing action, simple check and zero calibration was conducted according to the manufacturer's instructions. To reflect real-world settings, images

were collected under unconstrained conditions rather than following a strictly controlled protocol. Photographers were not instructed to use specific camera angles, distances, or lighting conditions, but only instructed to utilize household kitchen scales and smart phones following the simple guidelines. It was suggested but not mandatory to take photos from multiple angles and depths, or photograph a wide diversity of food types, portion sizes, containers, and culinary culture. No reference objects of known dimensions were included in the images. This design was intended to evaluate the baseline capability of LLMs to perform weight estimation and food recognition from ordinary user-generated photos without task-specific training or additional guiding information. However, details on the scales used to measure weights of image batch ii was not collected. Some measurement heterogeneity may exist within the benchmark dataset. Given the exploratory nature of this study, the scale-measured weights were nevertheless considered the most objective reference available for evaluating human and AI estimations under real-world conditions.

This systematic data collection procedure generated paired datasets comprising food images, ground-truth weight measurements, AI estimations, and human estimates, thereby enabling direct and objective comparison across different weight measurement methods. Table 1 summarizes the overall data collection design and dataset structure, while Figure 1 presents representative image examples, which primarily depict food items, with containers and dining surfaces forming the dominant background elements.

3.2 AI model implementation

7 open-source LLMs were employed to evaluate and compare the accuracy of AI-based food weight estimation and food identification. As summarized in Table 2, these models were developed by different organizations and vary substantially in their parameter sizes, reflecting differences in model scale, architectural complexity, and representational capacity. Each model was used to process the same set of images and pre-designed prompts to ensure the comparative study. The standard prompts ensure a fair evaluation of the LLM models. Sophisticated prompts with additional details might improve the performance of results, but they might fail to reflect the real-world scenarios (Fridolfsson et al., 2025). Therefore, the prompts were deliberately kept simple to establish a baseline evaluation. Minimal contextual guidance was provided solely to reduce unrealistic estimations, for example the two weight ranges which were derived from the scale-recorded weight ranges of image batch i and image batch ii.

TABLE 1: Summary of data collection design and dataset structure.

Data collection period	Apr-Jul 2025	Sep-Oct 2025	Jan-Feb 2026
Data contributors	Researchers	Academic experts	Academic experts and specialists
Key data collected and collection methods	Food images, Scale-measured food weight	Food images, Scale-measured food weight, Human estimates	Human estimates
Food types	First main dish, Second main dish, Pizza, Sandwich, Side dish, Dessert, Beverage, Soup, Pastry		
Cuisine Culture	Italian, Western (International), Asian, Indian, Middle Eastern		

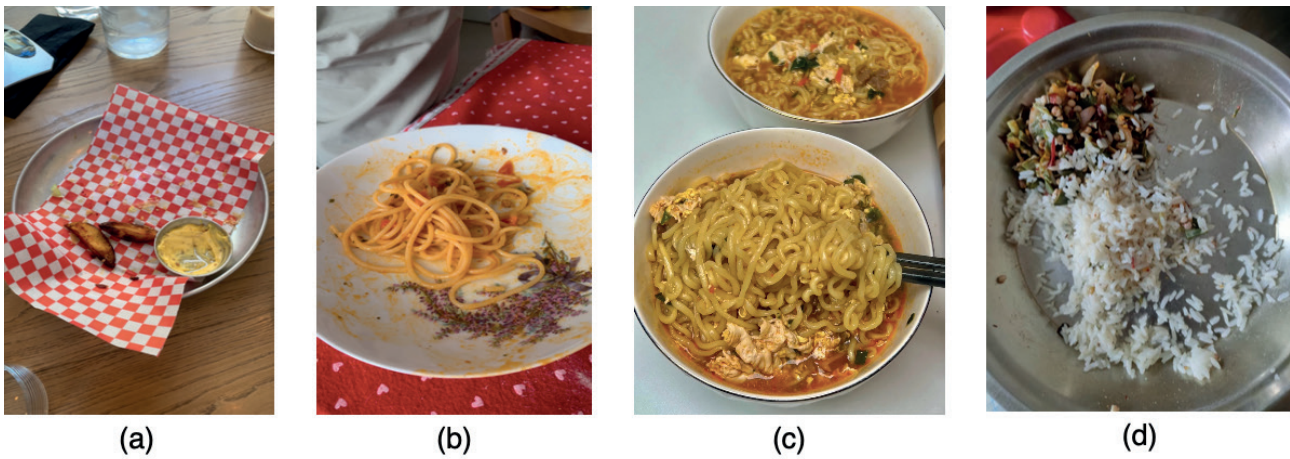


FIGURE 1: Examples of food images collected.

"You are a food waste analyst. You received an image of food leftovers for your research. Given the image, provide the following information: the type of the main food in the image; the estimated weight of the food in the image in grams. In general, food leftover weights are between 10 grams and 750 grams. Aim to be as accurate as possible in your weight estimation." – Prompt applied to the analysis of image batch i.

"You are a food waste analyst. You received an image of food leftovers for your research. Given the image, provide the following information: the type of the main food in the image; the estimated weight of the food in the image in grams. In general, food leftover weights are between 4 grams and 750 grams. Aim to be as accurate as possible in your weight estimation." – Prompt applied to the analysis of image batch ii.

By systematically testing models spanning diverse origins and parameter ranges, this study provides a comparative assessment of their effectiveness in performing multimodal tasks involving natural language understanding and image-based food weight quantification.

Open-source LLMs were intentionally selected to enable a transparent and reproducible evaluation of baseline model performance for FW quantification. In contrast to proprietary models (e.g. GPT-4 and Claude), open-source frameworks provide full access to model code and architecture, enable local deployment, and allow flexible system configuration, thereby ensuring experimental control, enhanced data privacy, and avoiding vendor lock-in. These characteristics make open-source models particularly suitable for systematic benchmarking studies in which models are assessed in their out-of-the-box configurations using a standardized prompt. Moreover, these models facilitate future development, including fine-tuning, domain-specific adaptation, and customization. While these capabilities are not exploited in the present study, they provide a strong basis for extending this work toward more practical systems. In this regard, open-source models are valuable for subsequent projects in FW quantification, continuous monitoring, and targeted intervention development, enabling the construction of more accurate, context-sensitive, and application-oriented AI systems.

TABLE 2: List of open-source LLMs for food image analysis.

Model	Developer	Release Date	Approx. Parameters	Deployment
Granite Vision 3.2 2B	IBM	2025 February	2 billion	LM Studio
Internvl3 2B Instruct	OpenGVLab	2025 April	2 billion	LM Studio
Llama 4 Scout 17B 16E Instruct	Meta	2025 April	16 billion	Groq
Moondream 2B 2025-04-14	Moondream	2025 April	2 billion	LM Studio
Qwen 2.5 VL 3B Instruct	Alibaba Qwen	2025 March	3 billion	LM Studio
Qwen3 VL 32B Instruct	Alibaba Qwen	2025 October	32 billion	Together.ai
Smolvlm2 2.2B Instruct	Huggingface	2025 February	2.2 billion	LM Studio

Abbreviations: Granite Vision 3.2 2B – Granite; Internvl3 2B Instruct – Internvl3; Llama 4 Scout 17B 16E Instruct – Meta-LLaMA; Moondream 2B 2025-04-14 – Moondream; Qwen 2.5 VL 3B Instruct – Qwen2.5; Qwen3 VL 32B Instruct – Qwen3; Smolvlm2 2.2B Instruct – SmolVLM2

Note:

Granite: <https://huggingface.co/ibm-granite/granite-vision-3.2-2b>

Internvl3: <https://huggingface.co/OpenGVLab/InternVL3-2B>

Meta-LLaMA: <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct>

Moondream: <https://huggingface.co/moondream/moondream-2b-2025-04-14-4bit>

Qwen2.5: <https://huggingface.co/Qwen/Qwen2.5-3B-Instruct>

Qwen3: <https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct>

SmolVLM2: <https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct>

3.3 Dataset structure

Following data cleaning, processing, and elaboration, two structured datasets were constructed with estimated weights, weight estimators, scale-measured weights, and image descriptions for analysis.

The first dataset – sub dataset – comprises human-based food weight estimations of 9 selected food images. A total of 52 participants completed this estimation task through an online survey administered via Google Form, in which the same 9 images were presented in a fixed order. As noted in Section 3.1, participants, those were invited to the survey, also contributed to construct the second image batch. To avoid familiarity effects, the image selection for the estimation task was therefore restricted to the first image batch. The selected images were chosen to provide a representative subset of image batch i in terms of food type (e.g. pastry, savory, snack, starter or soup, main meal), photo-taking conditions and visual complexity. Specifically, the images capture a range of scenarios in free environments, including meals photographed both at home and outside the home, foods presented in different types of containers, and foods that were either partially consumed or untouched. This design was intended to balance respondent burden and data quality while ensuring sufficient variability. Nevertheless, limitations and potential sources of bias arising from this subset selection should be acknowledged. For instance, the selected images do not adequately represent the diversity of food cultures, which could compromise the verification outcome. Furthermore, the images selected for human estimation task only weights from 19 g to 127 g. Consequently, larger portion sizes are underrepresented in the evaluation. This is partially attributable to the theme of the online survey, “Food (Plate) Waste Weight Estimation”, which favored the selection of images with partially consumed food and, therefore, generally smaller remaining portions.

The second dataset – full dataset – consists of the complete set of 91 food images collected in real-world environments. These images were subsequently processed by the 7 selected LLMs, each independently generating food weight estimates, thereby producing 7 distinct sets of AI-generated outputs – a total of 637 observations.

The resulting integrated dataset enables multi-dimensional comparative analysis across three estimation modalities: (i) precise scale-based measurements as objective benchmarks, (ii) human visual estimation, and (iii) AI estimation by multiple LLMs. This comprehensive data structure supports robust evaluation of both estimation accuracy and inter-model variability, while allowing systematic assessment of the reliability and generalizability of AI-LLM-based food waste measurement under unconstrained, real-world conditions.

3.4 Data analysis and statistical methods

For both the human-based and AI-generated weight estimation datasets, estimation errors were first computed as the difference between estimated food weights and the corresponding scale-based measurements, which served as the ground-truth reference in this study. Multiple sta-

tistical indicators were calculated to evaluate estimation errors and dispersion, including standard deviation (SD), mean absolute error (MAE), mean absolute percentage error (MAPE), mean error (ME), and mean percentage error (MPE). MAE quantifies the average magnitude of the estimation error in absolute terms, providing a direct indication of how large the errors are, while ME reflects the direction of the errors and thus indicates the tendency towards overestimation or underestimation. In contrast, MAPE and MPE express errors as percentage-based metrics. They are more sensitive to small reference values and may amplify the apparent magnitude of errors associated with small portions. SD characterizes the variability of the errors and therefore provides insight into the consistency of the estimation pattern. These metrics were used to conduct systematic comparisons between human and AI estimation performance. In addition, a scatter plot was conducted using ME provided by survey participants and by all LLMs, respectively, to examine potential systematic patterns and biases.

Following the evaluation of AI-human performance on food weight estimation, inter-model evaluation across the 7 LLMs was subsequently performed. Scatter plots were employed to investigate and visualize potential error patterns, model-specific biases, and inter-model variability. Together, these quantitative and visual approaches provide a comprehensive framework for evaluating estimation performance and uncovering underlying trends.

In addition to food weight estimation, the accuracy of food recognition derived from image analysis was systematically evaluated. For each image, all LLMs generated not only quantitative weight estimates but also descriptive textual outputs identifying the food items present. These descriptions were subjected to a structured, line-by-line verification process conducted independently by the research team, using the image descriptions recorded at the data collection phase as reference and following a predefined evaluation protocol to ensure consistency and minimize subjective bias.

The verification protocol comprised multiple criteria: (1) exact keyword matching between model-generated labels and ground-truth food items; (2) category-level matching when exact identification was not achieved (e.g., classifying pain au chocolat as pastry rather than requiring exact naming); (3) partial item matching was considered failed recognition (e.g., identifying only “mushroom” when mushroom and chicken were present); (4) mismatches arising from cultural interpretation were assessed contextually, such that cross-cultural approximations (e.g., Chinese “jianbing” identified as a burrito) were accepted, whereas semantically inconsistent matches (e.g., pizza) were rejected; (5) generic meal-type descriptions (e.g., “breakfast” or “brunch”) were treated as failed recognition when not accompanied by correct food item identification; and (6) variations in cooking method were disregarded in recognition assessment, while misidentification of sauces, condiments, or liquids was excluded from error evaluation. This multi-criteria framework enables systematic assessment of semantic recognition accuracy while accounting for linguistic variability, cultural interpretation, and practical

tolerance thresholds in real-world food image analysis.

Inferential and agreement analyses were conducted alongside descriptive statistics. Preliminary examinations of absolute error distributions using histogram and Shapiro-Wilk test indicated that the absolute errors of all models distributed significantly from normality. Accordingly, non-parametric statistical methods were adopted for subsequent analyses. A Friedman test was first applied to assess overall differences in absolute estimation errors across models, followed by post-hoc Wilcoxon signed-rank tests with Holm correction to identify pairwise differences (Granato et al., 2013). To further examine performance structure, Wilcoxon rank-sum test and Spearman correlation (Granato et al., 2013) analysis were used to assess the relationship between portion size and estimation error. In contrast, formal inferential testing for human–AI comparisons were not performed due to the limited size of the subset sample, which restricts statistical power. Finally, inter-rater agreement for the food recognition evaluation protocol was assessed using Cohen's kappa statistic (Cooper et al., 2006; Wani et al., 2024), applied to a randomly selected subset of 100 observations drawn from the full dataset. The subset included 12 or 13 images from each LLM outputs, thereby providing balanced coverage across all models. The two volunteering raters, those are not involved in this work, carried out assessment independently according to the predefined six-criterion protocol and image descriptions. This allows for comprehensive model comparison, targeted diagnostic assessment, and validation of the subjective evaluation protocol.

4. RESULTS AND DISCUSSION

Since MAPE and MPE impose greater penalties on estimation errors associated with smaller portion sizes, thereby amplifying relative inaccuracies for images with low-weight food items. The sub dataset used for the human–AI comparative analysis contains food weights representing a relatively small portion size range – larger portion sizes were underrepresented in this comparison. Therefore, for results reporting, MAE was adopted as the primary performance indicator, supported with ME and SD for the insights of potential error patterns. Other metrics were provided only in the supplementary document.

4.1 Human-AI performance on food weight quantification

The following analyses are based on the subset of 9 selected images, for which both human and AI weight estimates were available. Based on MAE, more than half human participants outperformed AI models for small-portion food weight estimation tasks, with the top-performing LLM (Qwen3) ranking 31st. This result indicates that, under unconstrained real-world conditions, human visual judgment currently demonstrates superior accuracy on smaller portions compared to state-of-the-art open-source LLMs. This finding highlights a substantial performance gap between human cognition and AI-based inference in complex visual estimation tasks, particularly when precise quantitative judgment is required. This is aligned with the study of

Liu et al. (2025), who discovered that the accuracy of using the AI method to estimate the food weight – liquid food in a hospital environment – was lower than the visual estimation method. However, we must acknowledge that this comparison was conducted using on a subset of food images with small portion weights. Therefore, larger portion sizes were insufficiently evaluated. Additionally, all LLM outputs were generated without any task-specific fine-tuning or additional data training and relied solely on a standardized, minimal prompt, thereby reflecting the models' baseline performance for fairness evaluation in this study.

The SD metric reveals a markedly different consistency of estimation errors, with lower values indicating a more stable error pattern. At the individual level, 6 participants exhibited lower SD values than the top-performing LLM (Qwen3), however, when considering the overall performance across all participants, the mean SD of human estimations (SD=35.9) was higher than that of Qwen3 (SD=19.9), suggesting that, on average, Qwen3 tended to follow more systematic and structured error patterns than the human participants as a group. This behavior reflects a model-specific tendency toward consistent over- or under-estimation of the food weights across our sample dataset, such as systematic overestimation by Qwen3 (ME=36) and underestimation by Moondream (ME=-69.6). Compared to humans, whose estimations exhibit greater variability across individuals and images, AI models demonstrate stronger internal consistency, albeit with persistent directional bias.

Interestingly, all LLMs except Moondream overestimated the food weights, while humans tended to underestimate the weight in our study (Table 2) – the average ME value of all human estimations is negative (-11.6). The general overestimation of LLMs and underestimation of humans contrasts with the findings of the authors' earlier study, which used ChatGPT-4o of OpenAI for the same food-weight-estimation task (Yu et al., 2025) and a comparative study by Liu et al. (2025) – AI methods give underestimations while humans overestimate.

Furthermore, within this evaluated range, the superior human performance suggests that humans may possess a comparative advantage in visually estimating small food portions and dish sizes, a pattern also reflected in Figure 2. In contrast, current LLMs appear less robust in handling small quantities. However, given the restricted weight range evaluated, this finding should be interpreted as specific to the evaluated portion-size range rather than as evidence across the full dataset.

4.2 Inter-model evaluation on food weight quantification

Table 3 reports the comparative performance of the 7 LLMs using MAE, ME, and SD, based on the analyses of the full dataset unless otherwise specified. Qwen3 demonstrates the lowest estimation error (MAE=68.6), followed by Meta-LLaMA (MAE=110). These models outperform the remaining LLMs in absolute error terms, indicating comparatively stronger capacity in food weight estimation from images. Fridolfsson et al. (2025) report that proprietary

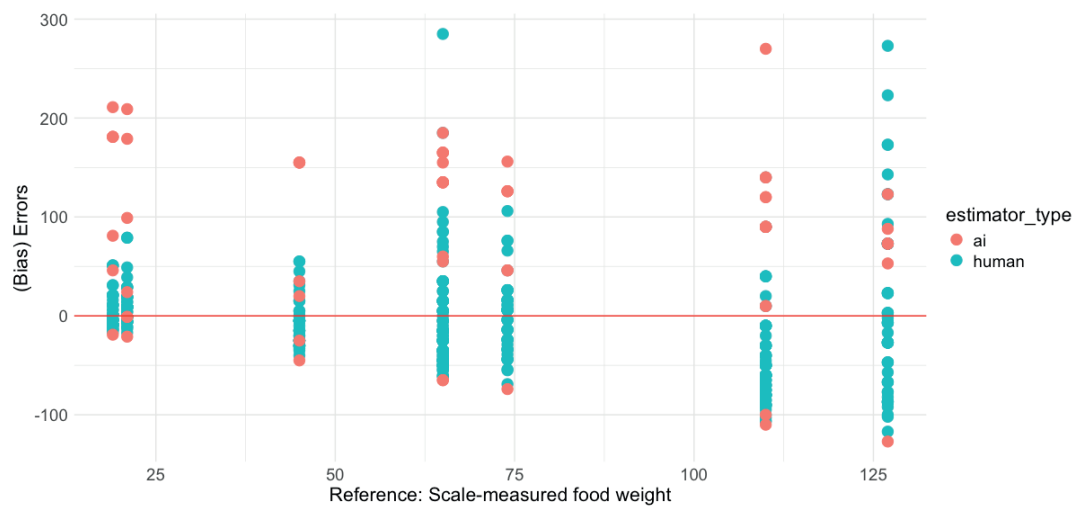


FIGURE 2: Scatter plot of human-ai estimation (bias) errors. Note: Bias Errors (BE) of each survey participant (blue points) and LLMs (red points) were plotted with the benchmarking reference weights measured by the scale.

models (e.g. ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) achieve a MAPE of approximately 36-37% in FW estimation tasks. In comparison, the best-performing model (Qwen3) yields a MAPE of 66.5%, highlighting considerable higher estimation error under the present experimental conditions.

TABLE 3: Performance summary of food weight estimation by LLMs and human.

Model	MAE	ME	SD
(Based on the full dataset of 91 images)			
Granite	131.2	87.2	78.8
Internvl3	122.5	-12.3	106.7
Meta-LLaMA	110.0	64.4	80.0
Moondream	269.5	3.3	640.8
Qwen2.5	120.9	54.3	80.3
Qwen3	68.6	15.0	68.2
SmolVLM2	144.4	77.8	77.9
All LLMs	138.2	41.4	259.4
(Based on the sub dataset of 9 images)			
Granite	134.9	134.9	36.1
Internvl3	101.8	96.0	60.5
Meta-LLaMA	96.6	96.6	47.4
Moondream	69.6	-69.6	37.8
Qwen2.5	115.1	114.9	53.5
Qwen3	36.0	36.0	19.9
SmolVLM2	176.1	176.1	50.8
All LLMs	103.1	82.1	59.7
All humans	37.1	-11.6	35.9

Notes: The values of MAE, ME, and SD, which were calculated for all LLMs and all human participants, were calculated based on the pooled 7 x 91 estimates (637 observations of the full dataset for LLMs), 7 x 9 estimates (63 observations of the sub dataset for LLMs), and 52 x 9 estimates (468 observations of the sub dataset for humans). Abbreviations: MAE – Mean Absolute Error; ME – Mean Error; SD – Standard Deviation.

In terms of bias patterns, most LLMs continue to exhibit systematic overestimation of food weights, consistent with findings from the smaller subsample analysis. Internvl3 constitutes an exception with underestimations. The persistence of overestimation across models indicates structural challenges in estimating weight from two-dimensional images, particularly under heterogeneous real-world conditions.

Notably, SD values do not align perfectly with overall accuracy performance. While Qwen3, the top-performing model in this exploratory work, achieved the lowest SD, indicating the most consistent weight estimation pattern, other relatively accurate models did not necessarily exhibit lower SD values. This finding suggests that lower estimation errors does not always imply greater consistency in the distribution of errors.

Inferential analyses supported the descriptive comparisons of model performance. The Friedman test revealed significant differences in absolute estimation errors among the LLMs ($p < 0.0005$), indicating substantial heterogeneity in estimation performance. Post-hoc pairwise Wilcoxon signed-rank tests with Holm correction further showed that Qwen3 consistently achieved significantly lower errors than all other models (all $p < 0.0005$) (more details reported in the supplementary document). These findings correspond the descriptive results and confirm the superior estimation performance of Qwen3 in this inter-model evaluation.

To further examine the distributional patterns of estimation errors, a scatter plot (Figure 3) was used to visualize the relationship between the scale-measured food weights and AI-estimated weights across all LLMs. Except for Moondream with some outliers, all other models demonstrate a consistent downward tendency in the error distribution. This pattern indicates a systematic shift in estimation bias, where smaller portion sizes tend to be overestimated, while larger portion sizes are more likely to be underestimated. Such results suggest the presence of scale-dependent estimation bias in LLM-based weight inference from images.

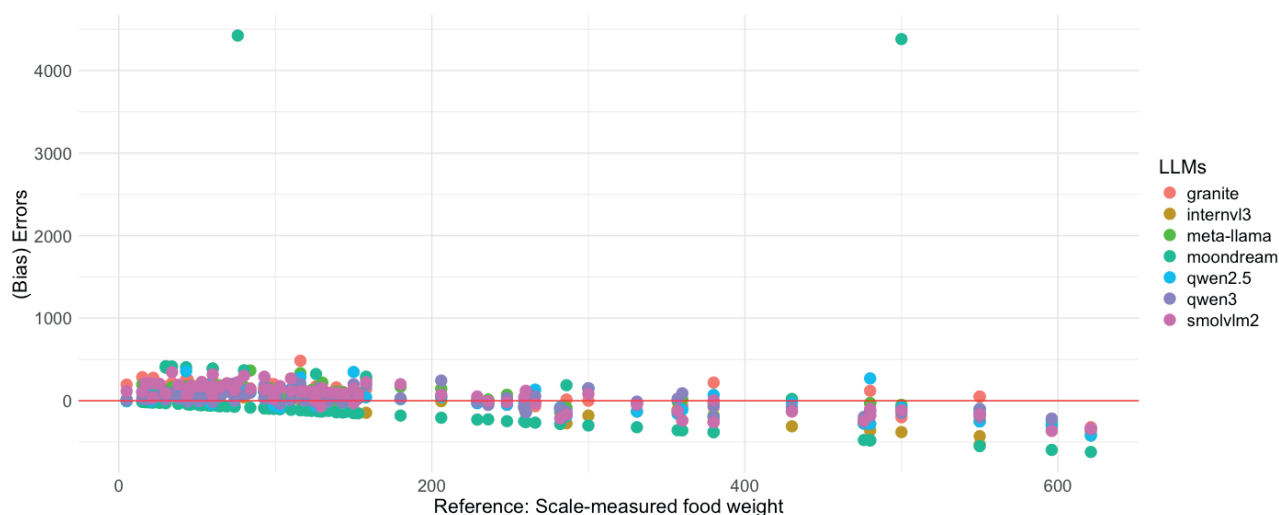


FIGURE 3: Scatter plot of AI estimation (bias) errors. Note: Bias Errors (BE) of each LLM were plotted with the benchmarking reference weights measured by the scale.

Given the observed relationship between estimation errors and portion sizes, model performance was further compared across small and large portions which were divided according to the average of scale-measured weight (160 g). Based on MAE, small portions exhibited lower absolute errors, with mean errors approximately 65 g smaller than those observed for large portions. In contrast, MAPE values were 173.3% higher for small portions, indicating substantially poorer accuracy. Wilcoxon tests confirmed that differences in both MAE and MAPE between the two portion-size groups were statistically significant (both $p < 0.005$). Moreover, the significant positive Spearman correlation observed for Qwen3 ($\rho = 0.41$, $p < 0.0005$) indicates that absolute errors tended to increase with portion size. These findings provide evidence that estimation performance is scale dependent. It also suggests that although LLMs achieved lower errors, but relative errors increased considerably for smaller portions, probably due to small denominators in the sample. This pattern partially agrees with Fridolfsson et al. (2025), who observed decreasing accuracy with increasing portions, but through the calculation of MAPE (20% to 30% approximately lower for small portions) rather than MAE.

Beyond portion size effects, dataset composition and image heterogeneity may also influence estimation performance. With the exception of Granite and Qwen2.5, which showed only minor MAE differences of 3.7 g and -5.7 g, respectively, when moving from the smaller subset to the full dataset, the remaining LLMs exhibited larger changes in MAE. In addition, all models exhibited increased SD values, indicating less consistent error patterns. Moondream showed the largest increases in both MAE (199.9) and SD (603). These results suggest that increasing image variability may negatively affect both accuracy and error consistency, raising concerns regarding the robustness and generalization of certain models.

Overall, the inter-model evaluation highlights substantial variation in predictive accuracy, bias structure, and robustness across LLMs. These findings underscore the importance of careful model selection and further mod-

el-specific calibration when deploying AI-LLM-assisted systems for FW quantification in real-world environments.

4.3 Inter-model evaluation on food recognition

Beyond quantitative differences in estimation accuracy, substantial qualitative variation was observed in the food recognition outputs generated by different LLMs. Models differed markedly in the depth and granularity of their image-based descriptions. Some LLMs produced highly detailed outputs, identifying specific food items, individual ingredients, cooking methods, and even distinguishing culinary styles or cultural origins – a similar pattern demonstrated in our earlier study (Yu et al. 2025). In contrast, others generated relatively basic descriptions, often limited to single-word labels or broad food category classifications (Table 4). The accuracy rates of the 7 LLMs range from 62.6% to 83.5%, with an average accuracy of 69.1%, which is notably lower than the performance observed (+98% accuracy rate) in our previous study using ChatGPT-4o for food recognition tasks (Yu et al. 2025). This difference suggests that model performance may vary depending on task complexity, dataset characteristics, model architecture and capacity.

Importantly, descriptive sophistication does not uniformly translate into higher weight estimation accuracy. While Qwen3 demonstrates strong performance in both detailed food recognition and relatively accurate weight estimation, other models achieve comparatively better numerical estimation despite simpler descriptive outputs or lower accuracy in food identification, such as Internvl3 and Qwen2.5. This pattern suggests that accurate semantic recognition and accurate food-weight estimation rely on partially distinct model capacities rather than a single underlying competence. Table 4 presents representative examples illustrating differences in descriptive depth and their relationship with estimation performance.

In addition, model accuracy was compared across the full dataset and its sub dataset in Table 5. At the aggregate level, difference in overall average accuracy was observed, with accuracy increasing slightly from 69.1% for the full dataset to 74.2% for the smaller subset sample. Internvl3, Me-

TABLE 4: Summary of LLMs in food recognition with their output examples and accuracy rates.

Detailed level of output	Model	Examples of description outputs	Accuracy
High	Qwen3, Meta-Llama	Pasta with dark sauce (likely mushroom or truffle-based); leftover cooked meat with mushrooms; Stir-fried cabbage with carrots and diced meat (likely pork or ham); Tom Yum soup (or similar Thai-style spicy soup)	~78% - 92%
Medium	Internvl3, Qwen2.5,	ground meat with nuts and raisins; stir-fried vegetables with meat; Instant Noodles	~63% - 69%
Low	Granite, Moondream, SmoIVLM2	pasta; Pastry; breakfast;	~62% - 71%

ta-LlaMA, and Qwen3 showed minor changes in accuracy rate, indicating relatively stable and consistent performance across both datasets, therefore a degree of robustness to sample size variation. However, Granite, Moondream, and SmoIVLM2 increased their accuracy rates by approximately 11.8%, 14%, and 12.4%, respectively, when applied to the smaller dataset. In contrast, Qwen2.5 experienced a decline in performance, with its accuracy rate decreasing by 9.3%. Moreover, model accuracy in food recognition was investigated between small and large portion-size groups – an evaluation with Wilcoxon test which was conducted to disclose the potential relationship between portion size and weight estimation errors. The results showed no significant difference in the accuracy rates ($p=0.43$) between the two groups.

Inter-rater agreement, calculated with Cohen's k statistic ($k = 0.817$, $p<0.0005$), indicates almost perfect agreement between the two independent raters, supporting the consistency of the six-criterion food recognition evaluation protocol and the evaluated results in this work.

In short, LLMs handle the tasks of food weight estimation and item identification with different capacities. While aggregate performance appears relatively stable, model-specific variability remains, underscoring the importance of inter-model evaluation and cautious model selection in AI-based food recognition and identification.

5. CONCLUSIONS

This study provides early empirical evidence on the application and comparative evaluation of open-source LLMs for FW quantification and recognition, addressing a critical gap in AI-enabled FW measurement and tracking research. Despite rapid advancements in AI applications across food science, empirical validation of AI-based FW measurement remains limited. By comparing AI-estimates, human judgments, and scale measurements under real-world conditions, this study contributes to methodological understanding of AI-based FW quantification and its practical limitations.

The findings suggest that AI-based FW measurement system, while promising some operational advantages, have not yet achieved reliability, in particular, small-portion FW quantification tasks in comparison to human visual judgement. However, AI-LLMs show stronger internal consistency and more structured error patterns, indicating a trade-off between accuracy and consistency. This implies

that AI systems may currently function better as decision-support tools rather than standalone measurement instruments in FW monitoring.

Error analysis reveals systematic bias and model-specific patterns. Most models exhibited a tendency toward over-estimation of food weights, while human estimators slightly underestimated. Furthermore, weight estimation accuracy is portion-dependent, with better performance observed for smaller portion sizes. Model performance in weight estimation and food recognition also varies across architectures and dataset configurations. Some models demonstrate stable generalization capabilities, while others are highly sensitive to image heterogeneity and dataset expansion.

Despite these limitations, AI-enabled FW quantification provides important opportunities for scalable sustainability monitoring. Open-source LLMs offer advantages in transparency, reproducibility, and customizable deployment for longitudinal FW tracking. AI systems can potentially support real-time behavioral feedback, policy evaluation, and automated waste monitoring across food consumption environments.

TABLE 5: Performance summary of LLMs in food recognition.

Model	Accuracy
(Based on the full dataset of 91 images)	
Granite	65.9%
Internvl3	65.9%
Meta-LlaMA	76.9%
Moondream	63.7%
Qwen2.5	64.8%
Qwen3	83.5%
SmoIVLM2	62.6%
All LLMs	69.1%
(Based on the sub dataset of 9 images)	
Granite	77.8%
Internvl3	66.7%
Meta-LlaMA	77.8%
Moondream	77.8%
Qwen2.5	55.6%
Qwen3	88.9%
SmoIVLM2	75.0%
All LLMs	74.2%

From a managerial perspective, AI-based FW monitoring can further support operational optimization in the food service sector, including portion design, consumer behavior interventions, and digital feedback systems. However, practitioners should avoid over-reliance on AI estimates. In this context, hybrid human–AI monitoring frameworks may provide more reliable FW management solutions.

This study has several limitations. First, the sample characteristics, although reflective of real-world conditions, remains limited in terms of item and photo-taking environment diversity, culinary representativeness, and weight or portion differences. While evaluating human-AI estimation in food weights, the dominance of small-portion samples within the subset may have contributed to the observed performance gap, representing a potential limitation of this sub-sample and warranting cautious interpretation of comparative results. Additional limitation concerns participant characterization when evaluating the human estimations. Factors concerning participants' familiarity with the food items and volumetric estimation tasks were not collected. These factors may influence human's visual food weight estimation performance and therefore represent potential sources of unexplained variability in the results. Second, 2D image-based estimation inherently restricts volume and depth inference. Future research should explore multi-modal AI architectures integrating depth sensing, IoT-based monitoring, and domain-specific fine-tuning. Comparative studies between human and AI evaluation should pay special attention to the potential portion-size effects and use sample with good representativeness when it is feasible. Furthermore, researchers could investigate the extent to which individual characteristics influence estimation accuracy, explore other open-source and proprietary models, as well as cross-cultural dietary datasets, are also necessary to improve generalizability. Third, the nature of this exploratory study should be acknowledged. As an initial investigation into the capabilities of LLMs for FW quantification, the current study intentionally did not impose highly standardized protocols, predefined or specific optimization procedures. This design choice reflects realistic conditions and provides a preliminary understanding, therefore serves as a foundation for further methodological development. Future research could incorporate standardized image acquisition procedures, the use of reference objects or predefined photographic area (e.g. fixed dimension and depth) for food weight estimation, and model fine-tuning strategies (e.g. specific-task training, refined prompt) to enhance the overall performance of AI-empowered food image analysis.

In short, this study provides foundational evidence on AI-enabled FW quantification while highlighting the methodological, technical, and practical challenges that must be addressed before large-scale deployment.

ACKNOWLEDGEMENTS

This research was conducted within the project “PrE-venting consumer Food waste at resTaurantTs and fOod service” (PERFETTO), funded by the European Union (call: SMP-FOOD-2024-FW-STAKEHOLDERS-PJ) under Grant Agreement No. 101216783.

The authors would like to express their sincere gratitude to all volunteers who generously contributed their time, effort, and data to support the construction of the dataset used in this research. We particularly thank Dr. Joshi Disha (Indian Institute of Technology, Delhi), Dr. Bright Tyrone (Durban University of Technology), Dr. Lasaridi Kattia (Harokopio University of Athens), Dr. Soloha Raimonda (University of Latvia), Dr. Sigala Evangelia (Harokopio University of Athens), Dr. Dennis Dankwa (McGill University), Dr. Silvia Schütz (BOKU University), Dr. Yang Tianxue (Chinese Research Academy of Environmental Sciences), Dr. Zari Mohammed (King Abdulaziz University), Dr. Claudia Maurer (University of Augsburg), Dr. Sempira John Edison (Stellenbosch University) for their valuable participation and support.

REFERENCES

- Comis, C., Principato, L., Zunica, S., Secondi, L., & Yu, M. (2025). Managing consumer food waste in the out-of-home context: From awareness to action. In *Reference Module in Food Science* (p. B9780443159763000623). Elsevier. <https://doi.org/10.1016/B978-0-443-15976-3.00062-3>
- Cooper, M., Ungar, W., & Zlotkin, S. (2006). 'An assessment of inter-rater agreement of the literature filtering process in the development of evidence-based dietary guidelines', *Public Health Nutrition*, 9(4), pp. 494–500. <https://doi.org/10.1079/PHN2005877>.
- Corrado, S., Caldeira, C., Eriksson, M., Hanssen, O. J., Hauser, H.-E., Van Holsteijn, F., Liu, G., Östergren, K., Parry, A., Secondi, L., Stenmarck, Å., & Sala, S. (2019). Food waste accounting methodologies: Challenges, opportunities, and further advancements. *Global Food Security*, 20, 93–100. <https://doi.org/10.1016/j.gfs.2019.01.002>
- Derqui, B., Fayos, T., & Fernandez, V. (2016). Towards a More Sustainable Food Supply Chain: Opening up Invisible Waste in Food Service. *Sustainability*, 8(7), 693. <https://doi.org/10.3390/su8070693>
- Dhir, A., Talwar, S., Kaur, P., & Malibari, A. (2020). Food waste in hospitality and food services: A systematic literature review and framework development approach. *Journal of Cleaner Production*, 270, 122861. <https://doi.org/10.1016/j.jclepro.2020.122861>
- Falasconi, L., Vittuari, M., Politano, A., & Segrè, A. (2015). Food Waste in School Catering: An Italian Case Study. *Sustainability*, 7(11), 14745–14760. <https://doi.org/10.3390/su71114745>
- Filimonau, V., & De Coteau, D. A. (2019). Food waste management in hospitality operations: A critical review. *Tourism Management*, 71, 234–245. <https://doi.org/10.1016/j.tourman.2018.10.009>
- Filimonau, V., Ermolaev, V. A., & Vasyukova, A. (2022). Food waste in foodservice provided in educational settings: An exploratory study of institutions of early childhood education. *International Journal of Gastronomy and Food Science*, 28, 100531. <https://doi.org/10.1016/j.ijgfs.2022.100531>
- Fridolfsson, J., Sjöberg, E., Thiwāng, M., & Pettersson, S. (2025). Performance Evaluation of 3 Large Language Models for Nutritional Content Estimation from Food Images. *Current Developments in Nutrition*, 9(10), 107556. <https://doi.org/10.1016/j.cdnut.2025.107556>
- Granato, D., De Araújo Calado, V. M., & Jarvis, B. (2013). Observations on the use of statistical methods in Food Science and Technology. *Food Research International*, 55, 137–149. <https://doi.org/10.1016/j.foodres.2013.10.024>
- International College, Krirk University, Bangkok, Thailand, Ting, H., College of Tourism, Hospitality and Transportation Management, Polytechnic University of the Philippines, Philippines, School of Management and Marketing, Taylor's University, Selangor, Malaysia, Memon, M. A., Faculty of Business, Sohar University, Sohar, Oman, National University of Sciences and Technology, Islamabad, Pakistan, Chandigarh University, Mohali, India, Thuramamy, R., Faculty of Business, Sohar University, Sohar, Oman, School of Management, Universiti Sains Malaysia, Penang, Malaysia, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia, Cheah, J.-H., & Norwich Business School, University of East Anglia, Norwich, United Kingdom. (2025). *Snowball Sampling: A Review and Guidelines for Survey Research*. *Asian Journal of Business Research*, 15(1), 1–15. <https://doi.org/10.14707/ajbr.250186>

- Kakani, V., Nguyen, V. H., Kumar, B. P., Kim, H., & Pasupuleti, V. R. (2020). A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research*, 2, 100033. <https://doi.org/10.1016/j.jafr.2020.100033>
- Kaur, P., Dhir, A., Talwar, S., & Alrasheedy, M. (2021). Systematic literature review of food waste in educational institutions: Setting the research agenda. *International Journal of Contemporary Hospitality Management*, 33(4), 1160–1193. <https://doi.org/10.1108/IJCHM-07-2020-0672>
- Liu, R., Tagi, M., Suzuki, Y., Shan, X., Hamada, Y., Ozaki, K., Kubota, M., Amano, S., Konishi, T., & Hirose, J. (2025). Usability and accuracy of an artificial intelligence-based automatic meal tray photography device for estimating liquid food intake in healthcare settings. *Clinical Nutrition ESPEN*, S2405457725031249. <https://doi.org/10.1016/j.clnesp.2025.11.131>
- Lo, F. P.-W., Qiu, J., Wang, Z., Chen, J., Xiao, B., Yuan, W., Giannarou, S., Frost, G., & Lo, B. (2024). Dietary Assessment With Multimodal ChatGPT: A Systematic Analysis. *IEEE Journal of Biomedical and Health Informatics*, 28(12), 7577–7587. <https://doi.org/10.1109/JBHI.2024.3417280>
- Ma, P., Tsai, S., He, Y., Jia, X., Zhen, D., Yu, N., Wang, Q., Ahuja, J. K. C., & Wei, C.-I. (2024). Large language models in food science: Innovations, applications, and future. *Trends in Food Science & Technology*, 148, 104488. <https://doi.org/10.1016/j.tifs.2024.104488>
- Martin-Rios, C., Hofmann, A., & Mackenzie, N. (2020). Sustainability-Oriented Innovations in Food Waste Management Technology. *Sustainability*, 13(1), 210. <https://doi.org/10.3390/su13010210>
- Sigala, E. G., Gerwin, P., Chroni, C., Abeliotis, K., Strotmann, C., & Lasaridi, K. (2025). Reducing food waste in the HORECA sector using AI-based waste-tracking devices. *Waste Management*, 198, 77–86. <https://doi.org/10.1016/j.wasman.2025.02.044>
- Stancu, V., Haugaard, P., & Lähteenmäki, L. (2016). Determinants of consumer food waste behaviour: Two routes to food waste. *Appetite*, 96, 7–17. <https://doi.org/10.1016/j.appet.2015.08.025>
- Vardopoulos, I., Abeliotis, K., & Lasaridi, K. (2024). A Systematic Informetric Analysis and Literature Review of Food Waste Quantification Studies in the Food Service Industry. *Sustainability*, 17(1), 103. <https://doi.org/10.3390/su17010103>
- Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Non-probability Sampling. In C. Wolf, D. Joye, T. Smith, & Y. Fu, *The SAGE Handbook of Survey Methodology* (pp. 329–345). SAGE Publications Ltd. <https://doi.org/10.4135/9781473957893.n22>
- Wani, Y.A., Arfiani, E.P., Tanuwijaya, L.K., Fajr'ina, N.H., Ekasari, A.P., Tobing, T.A.H.L. & Maghfiroh, D. (2024) Inter-rater agreement of trained and untrained observers in estimating platewaste using digital method. *Food Research*. 8, 34–37. [https://doi.org/10.26656/fr.2017.8\(S6\).3](https://doi.org/10.26656/fr.2017.8(S6).3)
- Yu, M., Cicatiello, C., Comis, C., Principato, L., & Secondi, L. (2025). Tracking post-consumption of restaurant food and leftovers: Innovative digital solution and outcomes from REGUSTO. *Quality & Quantity*. <https://doi.org/10.1007/s11135-025-02196-6>