Machine Vision Models for Food Recognition and Portion Estimation

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Software Engineering
June 2026
Sharmin Sultana

Supervisors:
Tuomas Mäkilä
Rehan Khalil

UNIVERSITY OF TURKU
Department of Computing

SHARMIN SULTANA: Machine Vision Models for Food Recognition and Portion Estimation

Master of Science (Tech) Thesis, 88 p., 4 app. p.
Software Engineering
June 2026

Accurate monitoring of dietary intake is essential for public health and nutritional research. Traditional approaches are time-consuming and erroneous. An alternative strategy is provided by automated food recognition and portion estimation systems, but their dependability is yet unclear. This study investigates the potential of multimodal large language models for automating food item identification and nutritional estimation from meal images in a real cafeteria environment, with a focus on how structured contextual information affects model performance.

The study employs an experimental comparative methodology using 192 meal images from a university cafeteria research environment in Finland (Flavoria). Two LLM families, which are GPT and Gemini, were evaluated across four analyses covering three contextual conditions: no context, a daily restaurant menu as context(Sodexo), and the Finnish national food composition database as context(Fineli). The performance was evaluated using mean absolute error, mean absolute percentage error, and error distribution analysis against measured ground truth nutritional data.

Key findings of this study reveal that both models produced significant estimation errors, without any real contextual grounding. The menu-based context seemed to improve performance quite a bit for both models, but the Fineli database context came across as more helpful for Gemini than for GPT, for carbohydrate estimation. Visually ambiguous food items, such as sauces, were consistently harder to estimate than structured dishes. Food name inconsistency is also a practical issue needing some post-processing normalisation.

This study contributes empirical evidence on the feasibility and limitations of multimodal LLMs for food recognition in food service settings. The results suggest that while contextual grounding meaningfully improves model outputs, current general-purpose LLMs are not yet sufficiently accurate for precise nutritional monitoring. Future work directions include retrieval-augmented generation combining menu and database context, and evaluation on larger and more diverse datasets.

Keywords: multimodal large language models, food recognition, nutritional estimation, contextual information, portion size estimation, dietary assessment

Acknowledgements

I would like to begin by expressing my sincere gratitude to my supervisor, Tuomas Mäkilä, for his invaluable guidance, patience, and constant support throughout this journey. His insightful feedback and encouragement at every stage of this thesis made a profound difference in shaping this work. I am also grateful to him for providing access to the OpenAI and Gemini API keys, as well as covering the computational costs essential for the experimental work carried out in this study.

I am also genuinely thankful to Rehan Khalil for his valuable technical advice and consultation throughout this study. His willingness to offer guidance and prompt troubleshooting insights whenever challenges arose were greatly appreciated.

Finally, in the spirit of academic transparency and embracing modern research practices, I would like to acknowledge the use of AI-assisted writing tools during the preparation of this thesis. Tools such as Consensus and NotebookLM supported the literature search process by helping identify and organise relevant papers, while Claude was used to refine phrasing, enhance clarity, and improve the narrative flow of certain sections. I believe acknowledging this honestly reflects the evolving landscape of academic writing and research methodologies.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The accurate tracking of food consumption is vital for an individual's health, nutrition studies, and sustainable food services. Knowing people's diets, the amount of food they consume, and the conditions under which they are eating gives lots of information on the habits of the people, the nutrients they are getting, and the health of the population. Authentic and reliable data on food consumption is also important for restaurants and catering services to optimize meal production and reduce food waste [1].

Though collecting proper dietary information is significant, it has remained challenging for years because traditional approaches for collecting that information, such as maintaining food diaries or weighing manually, are time-consuming, laborious, and likely to have human errors, and people tend to misreport or underreport their food intake, and this leads to incomplete and biased information [1]. Consequently, researchers, healthcare professionals, and food service providers have been looking for more automated and scalable solutions to the problem of dietary monitoring for a long time [1].

The rapid growth of digitalization and artificial intelligence has been a critical factor in overcoming these difficulties and has opened up new avenues for research and development in recent years such as, computer vision, the field of AI that helps

machines to understand and evaluate visual data [2]. Computer vision systems can recognize food items and even calculate the calories by analyzing pictures of meals and applying machine learning algorithms. These methods, based on images, not only cut down the need for human input but also provide a more convenient and precise way of collecting dietary information [3].

Deep learning, especially CNNs (convolutional neural networks), turned out to be the core part of modern computer vision-based food recognition. Deep learning models are trained with a very large amount of labeled training data to learn how to detect even the most complex patterns and the finest textures in the images. When models are applied to food analysis, they can recognize different dishes, classify food types, and even estimate the weights of the portions in some cases [3]. Researchers developed various methods (datasets and architectures) to enhance the recognition accuracy while still dealing with the difficulties of visual similarities between food classes and their variations in presentation. Studies such as [4], carried out in controlled conditions have shown very promising results, suggesting that the adoption of deep learning systems may provide very high accuracy even for common food items and also showed that a camera-based setup in a self-service lunch restaurant could not only accurately identify meal components but also give the right weights, which emphasizes the viability of AI-based food analysis in real food service scenarios. Such research is an example of how AI can be integrated into complex daily food-related situations to extract very helpful data about consumer behavior.

Nevertheless, these technological developments have not yet closed the gap of automated food recognition's widespread use in real-world circumstances due to various technical challenges, as well as not the least hard practical ones. The technical difficulties are such that the nature of food images is complex by definition. The obstacles to precise identification are very diverse: for example, a dish may consist of many items, all having different shapes, and it may be hard to tell which one is

which; lighting conditions may change as well. Although CNN models can achieve high performance on controlled laboratory datasets, they may still fail to work when applied to different food types, cuisines, or presentation styles of the same food. The accuracy of estimating the size of the portion is, in fact, wholly dependent on the camera angle, distance, and calibration, and all of that can be highly variable in unregulated environments [5]. Moreover, the demand for huge amounts of varied and high-quality datasets for training and testing, along with the need for their annotation, comprises a major roadblock in the path of model performance enhancement efforts [5].

Apart from the technical aspects, practical matters such as ease of use, confidentiality of data, and needing to coexist with the existing systems also have to be figured out. Though AI technologies can guess what is on a plate, the real benefit of such predictions lies in their utilization. To illustrate, the results of automated recognition could be directly connected with the nutritional information databases to calculate calorie and nutrient intake, or they could even become part of restaurant management systems to provide real-time feedback on food production and waste. However, this extent of automation will not be possible without the perfect interaction of recognition models, data infrastructures, and user interfaces, which is still a minimally explored area in present research [3]. Against this backdrop, a newer class of AI models has begun to attract research attention for food recognition tasks. Multimodal large language models, such as those in the GPT and Gemini families are capable of processing images and text together within a single general-purpose model [6], [7]. Unlike CNN-based systems that require large amounts of domain-specific labelled data and retraining whenever the food environment changes, multimodal LLMs can identify food items, estimate portions, and reason about nutritional content from an image without any task-specific fine-tuning. A particularly useful property of these models is their ability to incorporate structured contextual information such

as a restaurant's daily menu or a national nutritional database directly as text input, allowing their outputs to be grounded in external knowledge without any changes to the model itself.

In the Flavoria research restaurant at the University of Turku, operated by Sodexo Oy, two such sources of contextual information are naturally available [8], [9]. The first is the Sodexo daily menu, which specifies the exact dishes served on any given day. The second is Fineli, the Finnish national food composition database maintained by the Finnish Institute for Health and Welfare, which contains nutritional information for over 4,000 food items [10]. Whether providing this contextual information to multimodal LLMs improves their food recognition and nutritional estimation performance — and by how much — is a question with direct practical relevance for automated dietary monitoring in institutional food service settings, and forms the central motivation for this thesis.

1.2 Problem Statement

For nutrition-based research, monitoring health and everyday dietary assessment, it is important to know what people eat and in what quantity. Though the traditional way of collecting dietary information provides quite accurate information, these methods are time-consuming and can have errors while inputting the information into any system or database [1]. Moreover, they require continuous human involvement and significant effort, which limits their use in large-scale or everyday settings.

Nowadays, multimodal large language models have improved and made it possible to analyse images together with text [6]. This opens the possibility of estimating meal components, portion sizes, and nutritional values directly from food images. However, in practice, the reliability of these LLMs is still unclear. There are multiple scenarios in which the results can be affected [11]. First of all, different LLMs can provide different results. Secondly, their output can vary depending on the additional

information or context given to the model, such as the description of the menu and the nutritional database. Another challenge is that different combinations of models and contextual information may give different results. For now, there is limited experimental evidence or research regarding which combination performs best and how their outputs compare to carefully measured ground-truth data [12].

The problem addressed in this thesis is to evaluate how well multi-modal LLMs can estimate meal components, portion weights, and nutritional values from images, and to understand how the choice of model and contextual information affects their performance. The goal is to evaluate the models by comparing their outputs against ground-truth measurements obtained under controlled conditions.

By conducting a structured comparison and validation process, this thesis aims to provide a clear understanding of the strengths and limitations of LLM-based automated food recognition. The results can help determine whether such systems are suitable for practical nutritional analysis and under what conditions they perform best.

1.3 Research Questions

The main research question of this thesis is:

> *How well can multimodal large language models recognise food items and estimate nutritional values from meal images in a real world environment, and how does the provision of contextual nutritional information affect their performance?*

To address this question in a structured way, the following sub-research questions are defined:

RQ1. How accurately can multimodal LLMs identify food items and estimate

nutritional values from cafeteria meal images without any external contextual information?

RQ2. How does providing a focused daily restaurant menu as contextual information affect the food recognition and nutritional estimation accuracy of multimodal LLMs?

RQ3. How does providing a comprehensive national nutritional database as contextual information affect model performance compared to focused menu-based context?

RQ4. How do different multimodal LLM families and generations compare in food recognition and nutritional estimation performance across different contextual conditions?

1.4 Research Objective

The primary objective of this thesis is to evaluate whether general-purpose multimodal large language models can serve as a viable approach to automated food recognition and nutritional estimation in a real-world environment and to understand the extent to which structured contextual information can affect their outputs. This is pursued through a systematic experimental comparison across four analysis conditions, using two LLM families evaluated against objectively measured ground truth nutritional data from the Flavoria lunch line.

A secondary objective is to understand and document the practical challenges involved in building an automated LLM-based food recognition pipeline — including prompt design, context injection, output normalisation, and evaluation methodology. These aspects are important for anyone attempting to build similar systems in practice.

This thesis is not intended to produce a ready-to-deploy system for monitoring food. It is not even trying to tackle the problem of fully automatic nutritional analysis. What the thesis is trying to do is to provide a realistic and empirical assessment of what current multimodal LLMs are capable of and what they are not in the context of monitoring food, for the purpose of dietary monitoring. The findings will be useful for people who are interested in AI-based food monitoring and for people who are considering using LLMs in monitoring food in institutional food services and others.

1.5 Thesis Structure

The following chapters of this thesis are organised as follows. Chapter 2 reviews the relevant literature that covers automated food recognition systems, the development of multimodal large language models, the role of contextual information in food analysis, and previous research in the Flavoria environment. The conclusion of the chapter is drawn by identifying the research gaps that this thesis addresses.

Chapter 3 describes the research methodology in detail, which includes the dataset, the models selected, the experimental design across four analyses, the contextual data preparation for Sodexo and Fineli conditions, the prompt design and context injection strategy, the automated inference pipeline, and the evaluation metrics used.

The results from the four experimental analyses presented in Chapter 4 are discussed in turn, outlining the findings in terms of the accuracy of calorie and macronutrient estimation, the performance of individual dishes, and patterns of error for both models, across all conditions. Chapter 4 also addresses the challenge of food name inconsistencies and summarizes the main findings from the experiments.

Chapter 5 presents the results of the experiments conducted with the models. The results are explained in detail, discussed in relation to existing literature and possible limitations, and the problems encountered while conducting the research

are discussed.

Chapter 6 concludes the work of the thesis. The research questions are answered and possible future work directions are discussed such as, retrieval-augmented generation, larger and more diverse datasets, and the implementation of an automated name normalisation approach.

2 Literature Review

2.1 Introduction to Automated Food Recognition and Nutritional Monitoring

Accurate and reliable measurement of dietary intake is crucial for public health research, clinical nutrition and the management of many diet-related diseases and conditions, including obesity, type 2 diabetes and heart disease. Traditional tools used for dietary assessment include the 24-hour dietary recalls (24HR), food frequency questionnaires (FFQ) and weighed food records (WFR) [1]. All of these methods have serious limitations and are subject to bias. Participants' memories of their dietary intake may be unreliable, recall accuracy often declines as the day progresses, and long-term recall is particularly difficult, especially for less common foods [1].

To address the limitations of conventional dietary data collection methods, researchers have been investigating automated image-based dietary assessment (IBDA), and to achieve accurate dietary data, image-based dietary assessment methods utilize digital cameras and smartphone applications to capture photographs of meals consumed, which are then analyzed by a computer program to identify the types of food consumed and the quantities [2]. IBDA offers numerous advantages, including reduced self-reporting burden, elimination of selection and recall bias, and the ability to collect continuous dietary data over time without requiring individuals to have extensive nutritional knowledge [2]. The study [2] conducted a systematic scoping

review of AI-based dietary assessment from food images published since 2015, a period during which major advancements in automated food image analysis utilizing deep learning techniques have occurred. They found that while significant improvements have been achieved in food identification, portion size estimation remains challenging, particularly in real-world settings where individuals consume food in complex manners and from various angles.

Automated monitoring of human food intake in settings such as workplace cafeterias and self-service lunch restaurants offers new potential for dietary assessment while also presenting a specific set of technical challenges [3]. Because of the structured, semi-controlled nature of such settings where the same predefined menu items are offered, portion sizes are usually uniform within a day, and the eating environment typically remains unchanged, which increases the opportunities for automated monitoring of dietary intake as compared to other settings [3]. For monitoring food intake using mobile technologies, including computer vision-based methods for food identification and caloric estimation, system performance sharply decreases when moving from a controlled laboratory environment to real-life conditions where people eat more naturally, hence robust evaluation in real settings is of utmost importance [3]. One of the most comprehensively instrumented cafeteria research platforms is the Flavoria research restaurant at the University of Turku in Finland, which supports multidisciplinary food research with weighing sensors, RFID tracking technology, and overhead camera systems [4].

Recent advances in multimodal large language models (MLLMs) that process both visual and textual information at the same time have enabled novel approaches for automated food image recognition and dietary data collection [6]. Task-specific visual models such as convolutional neural networks (CNNs) require large amounts of labelled training data in the specific domain of interest; in contrast, MLLMs can acquire broad visual and semantic knowledge and generalize to zero-shot or

few-shot food classification to new domains without any domain-specific fine-tuning on labelled training data [6], [7]. This thesis experimentally evaluated that to what extent MLLMs can be used for the identification of food items and portion size estimation of meal images from the Flavoria cafeteria, and assess how the inclusion of context (the Sodexo daily menu and the Finnish national food composition database Fineli) affects model performance.

The rest of this chapter is organized as follows. Section 2.2 overviews methods for food image analysis using deep learning and computer vision techniques, with a focus on CNN-based recognition systems and challenges for portion size and weight estimation. Section 2.3 introduces multimodal language models, reviews current literature and applications on food recognition and dietary assessment, and discusses prompt engineering for zero-shot visual tasks. Section 2.4 elaborates on the theoretical and empirical basis for context-augmented food recognition, leveraging examples from menu-based contextualization, national food composition databases (such as Fineli in Finland), and context injection approaches. Section 2.5 discusses suitable evaluation metrics for food image classification and nutrient estimation tasks. In Section 2.6, we describe the Flavoria research environment and present previous research conducted on the platform. Finally, in Section 2.7, we discuss research gaps and outline the corresponding research questions addressed by this thesis.

2.2 Deep Learning and Computer Vision in Food Analysis

2.2.1 CNN-Based Food Recognition Systems

With the rapid development and application of deep learning in food image recognition since approximately 2014, convolutional neural networks (CNNs) have achieved results

on the large-scale visual classification task [13]. This section aims to give an overview of the development of deep learning for food image recognition, with a focus on architectures such as ResNet and EfficientNet [14] that have been used to achieve accuracy rates greater than 90 percent on benchmark food image recognition datasets, including the more challenging Food-101 dataset that consists of 101,000 food images that are categorized into 101 different food types [13]. The results obtained by these deep CNN-based approaches have significantly outperformed traditional visual feature-based approaches and have shown great potential in applying computer vision for automated food classification [13].

Among the CNN models proposed for food recognition, the ResNet family (residual networks) was one of the first to occupy a prominent place, mainly due to the introduction of the residual connections, which made it possible to train much deeper networks and helped to overcome the vanishing gradient problem. The paper [4] applied a ResNet-101-based backbone to food identification in the Flavoria self-service lunch restaurant, and attained a macro-averaged F1 score of 0.907 for 130 classes of food items. With enough training data and sufficient quality, CNNs can attain high recognition accuracies in food identification in cafeteria settings. Liu [13] studied the use of ResNet-50 and EfficientNet-B0 in food classification tasks, and results revealed that EfficientNet-B0 outperformed ResNet-50 with an accuracy of 75.14% when both networks were similarly trained. EfficientNet's compound scaling that increases depth, width, and input size of the network simultaneously, is the key factor of such improvements. In addition, the paper [15] showed that for the Food-11 dataset, an average classification accuracy of 96.40% could be attained by combining EfficientNet-B7 with Convolutional Block Attention Module (CBAM) and with appropriate data augmentation. This result indicates that attention mechanisms can also improve the food recognition accuracy by suppressing the influence of irrelevant information and concentrating on discriminative features specific to food items.

Recently, there has been growing interest in using transformer-based architectures to tackle food recognition problems that were typically treated using convolutional neural networks (CNNs) previously, and the study of [16] presented a hybrid architecture consisting of a conventional convolutional neural network (CNN) coupled with vision transformers for real-time food image classification. Vision transformers facilitated the network to exploit complex long-range spatial relations, which could be pertinent for separating many closely looking foods whose distinguishing characteristics could stem from their textures and structural organization.

While CNN-based approaches have achieved strong results for food classification and even weight estimation in the food image classification benchmark datasets, there are many practical challenges to apply food recognition models to real-world environments [4]. Firstly, despite the improvements in learning general representations, there is often a large performance drop when applying models trained on benchmark datasets to environments that operate very differently (e.g., overhead cafeteria cameras taking multi-food images under highly varying illumination conditions), as the distributions of training and deployment images may differ drastically [3]. Secondly, foods from various cultures and restaurants are challenging for the current models to recognize. Lastly, the challenge of annotating large datasets with ground truth weights (manually weighing plates of food or seeking nutritional expertise) is often much harder than annotating a classification task [5]. Motivated by these new challenges to food image analysis, this thesis investigates whether general-purpose multimodal pre-trained language models (LLMs) can effectively perform food recognition and portion estimation of images without any task-specific training and how contextual information affects their performance.

2.2.2 Portion Size and Weight Estimation from Images

Estimating food portion sizes and weights from meal images is one of the most challenging sub-problems in automated dietary assessment [5]. Food item identification is a classification problem that can be tackled by supervised learning method and estimating food weight requires recovering three-dimensional volumetric information from two-dimensional images, which is fundamentally a problem without explicit depth information or calibrated objects of known volume [5]. The recovery of accurate portion size information is a major bottleneck for IBDA systems to migrate from research to practical applications for nutrition monitoring. The paper [5] presents a comprehensive survey on single-view and multi-view image-based food recognition and volume estimation approaches.

Determining appropriate portion sizes is a complex problem. The task is further complicated in a cafeteria environment. Firstly, because food is often served on a common plate and thus obscured by other items [4]. Furthermore, many different types of food exist, some of which are amorphous in nature (e.g., sauces, soups, stews, and salads), making it difficult to accurately estimate volume based solely on the visual appearance of an image. The caloric density of these foods can vary significantly and should be evaluated on a per-preparation basis [4]. Konstantakopoulos et al. showed that foods are recognized using image processing techniques by systems utilizing AI and deep learning methods [17], with studies indicating that volume and weight estimation is performed at lower levels of accuracy than food classification. Specifically, the mean absolute error of weight estimation per item ranged from 20g to 80g, depending on the food type being estimated and the parameters of the imaging setup [17]. Sarapisto et al. [4] achieved high-quality results in mean absolute weight estimation errors per food item of around 15 grams for daily menu items in the Flavoria setting. However, such results are still not accurate enough to monitor the nutritional intake of individual people on an everyday basis. The main

remaining error-source is the difficulty in estimating the 3D volume of food items from two-dimensional images with heavy occlusions. While it is possible to obtain depth information from depth-sensing cameras, or even use some reference objects of known size in the scene, for the time being such resources are not available in the Flavoria settings. In a related work, Zhao et al. [18] presented a novel framework that combines classification and regression techniques to estimate nutrients of food items, achieving good results on the Nutrition5k dataset, including mean absolute errors for calorie estimation lower than those achieved by purely classification-based methods. Modeling the food spatial extent, provided by the segmentation masks, gives an important inductive structure to this weight estimation problem.

While most nutrition modeling studies focus on the absolute accuracy of weight or nutritional estimates, it is important to recall that errors in weight estimation tend to amplify hierarchically. In other words, percentage errors in weight estimation are magnified by the density of the food multiplied by the number of grams for which an estimate is required in order to calculate nutritional content. As a result, small percentage errors in weight estimation translate into larger percentage errors in the calculation of nutritional content, and even larger errors in calorie estimates for energy-dense foods such as fats and oils [18]. These errors have important implications for the models presented in this thesis, as both absolute weight estimation accuracy and the accuracy of subsequent nutritional estimation are critical performance criteria. An additional criterion relates to the relative difficulty of estimating different macronutrient categories, for example that carbohydrate is generally the easiest to estimate, while fat and energy are the hardest. This is largely due to the fact that fat content is visually invisible and also that fat yields more than twice the energy per gram of protein or carbohydrate [17].

2.3 Multimodal Large Language Models for Image Understanding

2.3.1 Overview of Multimodal LLMs: Model Families and Generational Progression

Multimodal large language models are one of the main recent breakthroughs in artificial intelligence. One general-purpose model can now process both visual and textual inputs and reason over them jointly, without the need of task-specific fine-tuning. The paper [6] provided a survey of recent advances of multimodal large language models. Every multimodal LLM model consists of three components: 1) a visual encoder, e.g. convolutional neural networks, that maps input images of arbitrary length to high-dimensional features; 2) a modality alignment module that maps visual features to language model's semantic space; 3) A large autoregressive language model backbone that generates output texts conditioned on both visual and linguistic inputs. They show that in one single model, MLLMs can be applied to a wide range of visually grounded tasks such as image captioning, visual question answering, scene understanding, and object recognition tasks, without the need of changing architectures for different tasks. In addition, the work of [7] showed that models also have ability of multi-step visual reasoning, optical character recognition and even generated human-level, coherent, natural language narratives given a complex visual scene. Unlike older task-specific models that could only handle narrow tasks and needed separate fine-tuning for each one, MLLMs offer a more flexible and general approach to visual understanding [7]. This thesis experiments with two major MLLM model families which include OpenAI's GPT models and Google DeepMind's Gemini models to examine their performance across two consecutive generations of each model family. The study demonstrates two main aspects of model development in

this field through its experimental design which tracks both existing research progress and future research directions. Initially, GPT-4.1 mini and Gemini 2.5 Pro were used for zero-shot food classification without any contextual augmentation. The goal was to establish which model family had the better baseline for food classification before progressing to the full experiment design. This was possible because, by April 2025, OpenAI had released GPT-4.1 mini, which improved instruction following and multimodal reasoning over previous variants of GPT-4. Conversely, Google DeepMind had reached the forefront of Google's multimodal capability with Gemini 2.5 Pro, released in March 2025 [19][20]. The results of this initial experiment informed the choice of and setup for the remainder of the experiments in this thesis.

In order to validate the results obtained using the mainstream language models available in the market, two successor models that are far more capable were used as the main tools to support the experiments in this thesis. The two models, GPT-5.4 and Gemini 3.1 Pro Preview, were employed across all three experimental scenarios: zero-shot, Sodexo menu-augmented and Fineli database-augmented. The first successor model GPT-5.4 is a much more advanced version, available through OpenAI API as part of the GPT-5 frontier models. While it is primarily a text model, it is capable of vision, reasoning and following instructions [20]. The improvement in multimodal understanding is reflected in its multimodal improvements that are relevant to the visual food classification and nutritional estimation tasks of this thesis. The second model Gemini 3.1 Pro Preview is Google's most advanced model for complex tasks as of February 2026. The official model card describes this model as having native multimodal support for text, images, audio, video and even code in a unified architecture. It can handle very long context input length of up to one million tokens [21], which is critical for the context injection experiments for this thesis where a lot of structured information from the Fineli database are injected into the language models together with food images in a single call. Moreover, Gemini

3.1 Pro Preview differs from the vision-and-language models. It processes visual and textual inputs within a single transformer network [21]. This native multimodality potentially enables better performance in leveraging both vision and language for disambiguating foods that are difficult to recognize, such as sauce-covered dishes, by taking advantage of textual cues in menus.

The experimental design for this thesis uses two different AI systems which include GPT-4.1 mini and Gemini 2.5 Pro, and GPT-5.4 and Gemini 3.1 Pro Preview to lead its research work. First, the study enables researchers to test whether general-purpose multimodal reasoning improvements from one generation to another produce actual performance enhancements in food recognition and nutritional estimation tasks which have not been studied before in dietary assessment research. The main research results of the thesis demonstrate the capabilities of the most advanced MLLM systems which existed in the market during the research period, thus providing results that remain relevant and current for professionals who want to use these systems in actual institutional food service operations.

2.3.2 LLMs in Food Recognition and Dietary Assessment

Integration of multiple modalities, such as images and text, has become an interesting direction for improving the performance of food image recognition and dietary assessment based food recording systems using multimodal language models (LLMs). Empirical studies conducted in the last few years have revealed both promising achievements and limitations of current deep learning models [12], [22] . Although only early explorations have been conducted on leveraging multimodal LLMs for dietary assessment, a recent work by Lo et al. [22] have systematically evaluated the capabilities of GPT-4V for the food image recognition and dietary assessment tasks without any domain-specific fine-tuning. Combining images and text (captured by wearable cameras in a semi-free-living setting) as inputs, Lo et al. [22] have

obtained high food recognition accuracy up to 89.8 percent, outperforming some existing CNN-based specialized systems for certain food categories. Interestingly, their results have shown that GPT-4V can learn implicit scale references from context objects like cutlery and plates on the table to achieve improved portion estimation accuracy. Furthermore, fine-tuning with culturally specific context information, such as specifying food be from a particular country or region, has been able to significantly enhance the recognition accuracy for corresponding local cuisine. These findings indicate that multimodal LLMs possess great potential for facilitating food image recognition tasks and dietary assessment applications.

In paper [12], they conducted a comparative evaluation of three MLLMs, including ChatGPT-4o, the literature-inspired Claude 3.5 Sonnet, and the market-leading Gemini 1.5 Pro. In their work, they used a variety of test data consisting of about 1,400 photos of individual food items as well as meals for three different portion sizes. For evaluating the quality of weighed food intake tools that employ MLLMs, it is crucial to know both the accuracy in terms of weight estimation, and the accuracy in terms of nutrition estimation. Thus, the authors calculated mean absolute percentage error (MAPE), Pearson correlation coefficients and performed a Bland-Altman analysis. This analysis showed that ChatGPT-4o and Claude 3.5 Sonnet reached MAPE values for weight estimation of 36.3% and 37.3%, respectively. For Gemini 1.5 Pro, MAPE values ranging from 64.2% to 109.9% were obtained. Importantly, an underestimation of consistent magnitude was observed for all three MLLMs, and this underestimation grew with increasing portion size. The findings of Gärtner et al. [12] provide a useful basis of what achievable accuracy one can expect for MLLMs in general. While certain models, in specific tasks and usages, can achieve higher accuracy, it is clear from these results that general-purpose MLLMs do not yet achieve the level of accuracy that is required for clinical nutritional monitoring. But they surely reach accuracy levels that are sufficient for several other application

scenarios involving foods.

Romero-Tapiador et al. [11] applied MLLM(s) to dietary assessment and evaluated six vision-language models, including commercial closed-source models such as ChatGPT, Gemini, and Claude, and open-source models Moondream, DeepSeek, and LLaVA. Experiments performed using a dataset consisting of images of packaged food products annotated by dietitians with a wide range of nutritional information showed that commercial closed-source models performed better than open-source ones for dietary assessment, while none of the models assessed reached the accuracy of a dietitian for all categories. The most difficult aspect to tackle for all systems was portion size estimation, particularly for foods with amorphous shapes, such as sauces, soups, and mixed dishes. The research found three specific items which validate the documented difficulties described in the CNN-based studies [5], [17] to measure sauces and soups and combined food items. The Flavoria cafeteria serves exactly this type of dish because it includes cream sauces and composite salads, which serve as the most difficult visual estimation challenges.

The paper [23] implemented a pipeline using EfficientNet-B4 as visual backbone and Gemini LLM for food image based nutritional estimation on a culturally specific Chinese food dataset. It was observed that visual recognition errors propagate semantically through the pipeline to produce erroneous nutritional estimates. This study highlights that the quality of food recognition at identification stage affects the accuracy of nutritional estimation most. The quality of nutritional database for estimation does not affect it much. Min et al. [24] demonstrated that LLMs can be made for food specific tasks by domain specific fine-tuning on large food related corpus. In this study, FoodSky model achieved 83.3% accuracy on Chinese National Chef Examination after food corpus fine-tuning. However, they require large domain specific datasets which is out of scope for Finnish cafeteria foods studied in this thesis. Huang et al. [25] showed that domain-specialised LLMs substantially outperform

general-purpose LLMs on nutrient estimation. For the specific food dataset used in the study, macro-nutrient accuracy increased from 0.43–0.63% for general models to 0.91–0.97% for fine-tuned FoodyLLM. While general-purpose MLLMs can serve as strong zero-shot baselines for food image classification, significant performance gains can be achieved through domain adaptation. These results offer opportunities for future work and are not further explored in this thesis.

2.3.3 Prompt Engineering and Zero-Shot Capabilities

The design of input prompts is critical for the overall performance of MLLM in visual recognition tasks, especially zero-shot settings where no task-specific fine-tuning is needed. The paper [26] presented a meta-prompting framework for zero-shot visual recognition, in which category-specific prompts were automatically generated by asking LLM questions about visual attributes of target classes. The performance gains of up to 19.8% over standard CLIP-based zero-shot classification were obtained by averaging a number of prompts generated by different LLM models. They observed that the performance of zero-shot MLLM in food recognition strongly depends on how prompts are designed, and therefore, the design of prompts should be a critical research axis in the future studies.

With respect to dietary assessment, prompt engineering includes several dimensions, such as 1) the granularity of the food identification prompt (e.g., identifying a food on the Multiple Food Tracking plate versus a specific portion of a food), 2) output constraints (e.g., generating a number for the portion of a food versus a written description), 3) the inclusion of chain-of-thought prompts (e.g., having the model explain its step-by-step thinking about a food's energy), and 4) the prompt's ability to elicit uncertainty estimates (e.g., soliciting a range for the energy of a food). Recent work by Tanabe et al. [27] found that fine-grained estimation prompts, such as having the model first predict visual dimensions of a serving (e.g., height,

diameter) and then use those dimensions to calculate volume and estimate weight, substantially improved MLLM performance on the food energy estimation task. This form of "chain-of-thought" prompting enables the model to apply its training in a step-wise, highly structured manner, similar to how a trained nutritionist might approach the task, yet without direct ability to measure depth or scale.

The paper [22] showed that zero-shot, context-augmented, and database-augmented prompting with contextual information about the cultural origins of food can substantially improve the recognition performance of pre-trained visual models on generic food images. Even within a fixed food image, prompting with the daily menu at Flavoria's restaurant, operated by Sodexo, should improve the performance of food recognition, in line with the hypothesis of this thesis. The three approaches represent the three axes of the experimental design in this thesis, differing mainly in how additional context is incorporated into the image-input of a pre-trained MLLM.

2.4 Contextual Information in Food Recognition

Classical computer vision as well as recent multimodal LLM-based approaches for food visual recognition have demonstrated that providing structured contextual information can significantly reduce the task complexity and improve performance [22]. In daily institutional food service settings, a rich, structured source of contextual information is naturally available, such as information in the daily menu (e.g., menu items, descriptions), a nutritional database that structures information about the foods actually served (e.g., nutrients, ingredients), as well as the physical serving environment. The key insight here is that all of this information can be harnessed to improve food recognition in practical scenarios that multimodal LLMs are well-suited to handle, processing natural language context to generate input representations [22].

2.4.1 Menu-Based Context in Food Identification

Inspired by the CNN-based literature, the paper [4] explored the use of the daily menu information as a contextual prior for food recognition and developed two alternative architectures for learning with menu information. For the menu information to be learned with a ResNet-based food recognition architecture, sarapisto et al. [4] concatenated a binary vector indicating the presence of each menu item to the corresponding fully connected layers and learned feature weights based on a binary vector indicating the presence of each menu item. Among the different models explored, the model in which a learned weighting vector of the binary menu vector is multiplied to the feature channels of the penultimate convolutional layer, referred to as MenuProd, reached the highest macro F1 score of 0.907. Interestingly, the improvement in performance of incorporating menu context into food classification comes from specifically reducing the number of false positives for off-menu food names like additional salad or grain with different ingredients, and the improvement is particularly significant for classes that are visually similar.

In the thesis, menu-based contextualisation is achieved in a completely different way. The daily Sodexo menu which covers lunch options such as home food, vegetarian, and special diet options, is incorporated directly into the system prompt as structured natural language text. This allows the model to constrain its food identification outputs to the items available on that specific day, forming the basis for in-context learning. The menu items for each lunch line including home food, were managed by Sodexo Oy catering company responsible for Flavoria's restaurant, and every day a different daily menu was produced for the home food, the various types of vegetarian and the special diets option including home food. Hence, the Sodexo Oy catering company produced a lot of menu options for customers [8]. The menu for the day was always produced as part of the restaurant's normal operations, and therefore required no extra effort to acquire data for the automated food recognition.

In the menu-context experimental condition, we test the hypothesis that model operators will decrease their hallucination of off-menu items when searching capacity limitations are imposed using foods available on the current day. In this condition, menu content will help to solve visual uncertainties while nutritional values will become more precise by requiring actual food composition to be tested against existing menu items. This hypothesis has been previously proposed for both CNN-based [4] and LLM-based [22] food recognition systems, and this thesis provide the first empirical evaluation in the Flavoria environment using multimodal LLMs.

2.4.2 Nutritional Databases (Fineli)

National food composition databases (FCDBs) are required for any dietary assessment system to translate the results of food identification to meaningful nutritional information. FCDBs typically list the average nutrients, per 100 g of food sales weight, after cooking loss and edible fraction, as consumed within local culinary practices and serve as the authoritative reference for generating accurate nutritional information for population dietary surveys and clinical dietetic practice [28]. The quality, accuracy and coverage of the FCDBs utilised by a dietary data system are independent of the accuracy of food identification and weight estimation steps [17].

Fineli is a national food composition database of Finland, containing data on the nutritional composition of over 4 000 foods and food products. Information on 55 different nutritional components is available. The data are constantly updated by the Finnish Institute for Health and Welfare (THL) [10]. Fineli is the principal database used in nutrition research and monitoring of health. One of the largest nutrition surveys conducted in Finland using Fineli was the Finnish National Dietary Survey (FinDiet 2017) that assessed the usual dietary intake of energy and several nutrients and foods from a 24-hour dietary recall of adult and elderly Finnish individuals [28]. A food recognition system used in the dietary data collection in the food service

settings in Finland would be supported best by Fineli, the Finnish national food composition database, given the culture and types of foods and meals most commonly served in this type of setting, reflecting well known Finnish cuisines and the produce available in the country.

Nutritional databases have been integrated with AI-based food recognition systems in several distinct ways in recent literature. In this paper [29], they propose the DietAI24 framework that combines multimodal food classification with nutritional lookup from a large nutritional database using a retrieval-augmented generation (RAG) architecture. Specifically, they employed a large language model (LLM) to first recognize visual inputs of mixed dishes into different foods, and then employed the recognized foods to perform a database lookup from the Food and Nutrient Database for Dietary Studies (FNDDS) to obtain the corresponding nutritional information. They used the retrieved nutritional information in the computation of the total nutritional estimates for the mixed dishes. They reported [29] that results are obtained, with a 63% reduction in the mean absolute error of estimating the weight of foods in mixed dishes compared to existing baselines. This substantial improvement in nutritional estimation accuracy can be attributed to the fact that grounding the nutritional estimates of LLMs in authoritative database entries substantially outperforms relying on the nutritional information that the LLMs have internally learned and generalized from their knowledge, despite their vast training data. The reason for this improvement possibly stems from the model's inherently imprecise, outdated, or even culturally biased nutritional values for particular foods, where nutritional values in nationally specific databases, such as Fineli that follows the conventional ways of preparing foods in the Finnish cuisine, are not well generalized in the training data of general-purpose models.

We incorporate Fineli, a database maintained by the Finnish National Institute for Health and Welfare, as contextual information source. The nutritional information

from Fineli is hand-copied into the system prompt in unprocessed, structured text form. We subsequently extract the nutritional compositions for the foods from the Sodexo daily menu and use this information to produce portion weight and nutrient estimates. The ability to include this high-quality, authoritative database allows us to inspect the gain of nutritional database information beyond zero-shot and menu-only performance. The results demonstrate that incorporating external nutritional data provides a strong improvement in nutritional estimate accuracy.

2.4.3 Context Injection Strategies and Their Effect on Model Behaviour

Another crucial aspect is how the contextual information is integrated into the overall architecture of the LLM inference pipeline. In this thesis, the structured contextual information of the Sodexo menu and the Fineli nutritional data are incorporated into the system through the input prompt. This information is then fed into the network before the input image and affect how the LLM generates the food identifications and the corresponding nutritional estimations by specifying prior beliefs that the network employs for reasoning [16].

The paper [16] explored different forms of structured prompting for the LLM nutritional assessment task. They found that the double-step prompting strategy significantly improved the performance of the LLM relative to all single-step alternatives, including the original input and output structure. Specifically, the double-step approach follows a double-step processing procedure where the LLM first classifies images of foods into being nutritious or non-nutritious based on standard nutritional indicators, and subsequently calculates the nutrition values for each food. This approach leads to higher agreement with nutrition experts compared to single-step prompts, because it facilitates a process of reasoning that is similar to how a trained nutritionist would process information. The system prompt used in this thesis follows

a similar structured approach where the LLM is first instructed to identify visual food images into specific food items; then it constrained to select only the foods listed in the menu. Finally, the LLM is asked to estimate the portion weight of each food item selected and compute its nutrition values using the Fineli nutritional data provided in the input.

While general-purpose LLMs have been shown to learn rich contextual representations that enhance food classification, a number of studies indicate that the injection of contextual information into food classification models does not always have uniformly positive effects on their accuracy, and can even introduce additional biases in nutritional estimates [6]. Analyzing the effect of Sodexo menu context on two separately trained models, this thesis finds that injection of such context causes one of the models to have a marked positive bias in estimating the calorie and macronutrient content of meals it classifies, while having more modest and well-balanced increases for the other model. While models are specifically designed to augment visual food image classification with contextual and culinary information, previous work [24], notes that general-purpose LLMs tend to over-rely on contextual information relative to images, with model behavior being particularly inconsistent when applying domain-specific nutritional context. This thesis systematically characterizes the effects of menu context on models, observing biases that reflect trade-offs the models make between their prior knowledge of nutrition and externally-procured contextual information. These findings provide important insights into the potential limitations and limitations of context-augmented food classification. Furthermore, they establish a basis for further work on issues that arise when combining LLMs and vision for food classification, complementing prior work on the positive uses of context in this domain.

In a longer study (DietAI24) [29], the benefits of database grounding over model-internal knowledge were found to be most significant for foods with high nutritional

specificity, i.e. foods with nutritional values that are significantly different when prepared in various ways or in different varieties, where general nutritional priors included in the model weights do not suffice to cover all cases adequately. On the other hand, the benefits are moderate for foods such as plain cooked rice or boiled potatoes with consistent nutritional values. The results of the Fineli context condition in this thesis can therefore be interpreted with a view to the benefits being more significant for nutritionally varied dishes such as composite sauces and mixes of various types of foods (salads).

2.5 Evaluation Metrics for Food Recognition and Nutritional Estimation

Selecting appropriate evaluation metrics for automated food recognition and nutritional estimation methods is a non-trivial decision. Metrics define what aspects of performance are measured and reported, such as average error, direction of error, robustness to outliers, and others. Different metrics are more or less appropriate for different types of applications, and for different types of prediction tasks [12].

For the food identification sub-task in FoodNet Challenge, the F1 score (F1) which is the harmonic mean of precision and recall, was used as the performance metric. Macro-averaged F1 score calculates the F1 score of each class first and then finds the average (without using class weights) that ensures rare food categories have an equally important influence on the overall score as common food categories. On the other hand, micro-averaged F1 score first aggregates the prediction outcomes for all the instances and then calculates the F1 score, and as a result, scores of prevalent food categories have a greater influence on the averaged F1 score [4]. The paper [4] used both macro and micro F1 scores to report results of their deep learning CNN-based food identification system.

In order to assess the quality of nutritional estimations for the nutrition estimation task, two different error measures are used within this thesis. The mean absolute error (MAE) is used to measure the average absolute difference between estimated nutritional value and real value of weight, macronutrients or energy per serving size in units of the estimated value. As outlying data points have less influence on the MAE than on the mean squared error, the MAE is a robust and easily interpretable measure of error. The mean absolute percentage error (MAPE) on the other hand expresses the error as a percentage of the true value which means that the scale of the different nutritional values does not matter. Therefore MAPE is more suitable to compare models in order to find out which one is performing best on a certain task even though the distribution of serving sizes and therefore the distribution of the ground truth values differ.

In the work of [12], they utilised MAPE, Pearson correlation and Bland-Altman plots to compare the performance of three models for the nutritional content estimation. In this thesis, MAE and MAPE were used as the primary measures of nutritional estimation accuracy, and signed error analysis was performed to investigate if the models presented in this work consistently overestimated or underestimated nutrients, whether this was to a greater or lesser extent than random chance. The paper [18] employed a proportional MAE (PMAE) metric that normalises by the target value, which allows for comparison of estimation performance between nutrient labels that have different scales.

When evaluating the performance of food recognition systems based on MLLM for learning visual models, it is important to take into account how dish name inconsistency is handled. Unlike traditional CNN-based classifiers that output a probability distribution over a fixed vocabulary of class names, MLLMs generate natural language food names. Those food names are often variable, and as the models are based on generative models, the same food can be identified with different types

of names and wording order. Moreover, dishes that look similar may receive food names that describe them differently. Since many evaluation methods compare model predicted names to ground truth categories based on string matching, this name inconsistency often results in lower performance for MLLMs than what is achieved by corresponding CNNs. In this work, a manual name normalisation mapping is created to allow the evaluation of food recognition performance by mapping each model's output of food names to a consistent set of categories for a fair evaluation. This name normalisation is consistently applied throughout all the experiments, ensuring that performance differences obtained with different models, data and methods are meaningful and directly related to the properties of the models and their training data.

2.6 Previous Research in the Flavoria Environment

The Flavoria research platform at the University of Turku is a multidisciplinary research and development environment for measuring exposure to foods. The studies are conducted in an operational self-service lunch restaurant and café. The research activities are led by the Nutrition and Food Research Centre at the University of Turku, and in collaboration with the Faculty of Technology, Faculty of Social Sciences, and Turku School of Economics. The restaurant operations are contracted to Sodexo Oy. The daily menu of each lunch line on each day is known for each study day. This is a key aspect for the food recognition experiments presented in this thesis [9].

In the Flavoria restuarant, the physical setup of the lunch lines includes RFID readers and weighing sensors installed next to each food station, which allows to record the weight of every food item that is distributed that each customer takes. In MyFlavoria mobile application [30], food items and corresponding portion sizes are recorded during the lunch visit. For each customer, the composition of meals, portion sizes and the bio-waste during the complete lunch visit is recorded in the

session identifier linked to the RFID-tag of the tray. In this work [4], a sensor-based ground truth collection system is used to record weight estimates for individual food items of every customer. Expert visual estimates are not as robust as the high quality ground truth available, which is used for a high-quality comparative evaluation of the results of the MLLM approach in this thesis.

2.7 Identified Research Gaps and Summary

The preceding literature review shows multiple research gaps which define all the boundaries of the present thesis. The research design of the study requires identification and explanation of research gaps which will be presented in the following section according to their degree of importance.

First, while studies applying multimodal deep learning models for dietary data collection are rapidly emerging, none of the previous studies have validated their models in real-life settings. Most studies have used standardized international images of model foods. These images are primarily suitable for data collection in specific cultures and countries [11], [25], and there is an evident need for models that are more versatile. In this thesis, the performance of a MLLM is evaluated on real-world images of served dishes in institutional food service settings. The images of dishes served in food service center Flavoria were used as input for the model. Images with properly organised and images with semi-organised composition were used as input for the model. Model's predictions were validated by comparing them to objectively measured ground truth data. The model's performance was also evaluated when it has access to information from a national food composition database Fineli. This study is the first attempt to validate performance of MLLM in Finnish institutional food service settings.

Second, the study has examined three distinct evaluation methods through which researchers tested multiple LLM approaches to food recognition in actual cafeteria

operations. The two research studies from [22] and [29] established distinct values through their respective methods which examined cultural context and database grounding as separate elements of the contextualisation problem. The present thesis integrates these two contextualisation strategies within a unified experimental framework applied to the same dataset and models which enables evaluation of the incremental value delivered by each contextual information source.

Third, the first comparative assessment of a range of pre-existing LLM generations, including the earlier, smaller models (GPT-4.1 mini and Gemini 2.5 Pro) against the more recent, larger variants capable of food recognition for dietary assessment, is provided in this study. Most current research evaluates the performance of the models available at a specific time point. In this thesis, we systematically assess the ability of models across LLM generations to analyse food images, using a consistent methodology to inform model selection for future dietary monitoring applications over time.

Fourth, a key issue that has been encountered and explored within this thesis is that of differential context integration; that is, it has been shown that different architectures process the same contextual information in differing ways, and that some models exhibit an over-reliance on provided context with resultant overestimation of performance. This phenomenon has not been previously addressed within the food recognition literature, and its identification and consequent analysis within this thesis reveals a number of important empirical results that have practical implications for the design of dietary monitoring systems that incorporate provided context using MLLM architectures.

Fifth, while issues related to accuracy such as precision and recall have received considerable attention, another problem, related to how such models are evaluated namely, the food name inconsistency problem, where generative LLMs tend to generate alternative, yet semantically equivalent (different lexicalisations of) food

names has been somewhat overlooked in dietary assessment. While in the conducted research a systematic name normalisation approach was developed and applied, it also needs to be available for use in future evaluations of such generative models for food recognition.

In this chapter, the developments from classical computer vision to deep learning-based CNNs and to the recent emergent multimodal LLMs for automated food image recognition and dietary assessment are summarized. We characterise the inherent challenges for portion size estimation and weight prediction from 2D images, discuss the characteristics and possibilities of MLLMs for food recognition, and review the empirical results obtained with contextual augmentation using menu information and with the Finnish national food composition database Fineli. We also discuss evaluation metrics for food recognition and nutritional assessment. Based on the found research gaps, this thesis investigates the ability of multimodal LLMs, with and without structured contextual augmentation to perform food item identification as well as estimation of nutritional information in a real life setting, i.e., in a Finnish cafeteria. Furthermore, the differences in performance between model generation and contextual information is also discussed.

3 Research Methodology

3.1 Research Design and Approach

This thesis is a quantitative, experimental, and comparative study. The goal of the study is to evaluate to what extent multimodal large language models (MLLMs) can identify foods and calculate their portion sizes and nutritional values from images of meals, and how providing additional external contextual information affects the model. Due to the nature of the study, an experimental-comparative approach was followed by creating a controlled experiment, by running an automated pipeline, and comparing numerical results across different models and conditions. This thesis fits best as an experimental-comparative methodology [31]. The study is mainly quantitative-based, errors having been evaluated using error metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), with predictions compared against ground truth nutritional values. However, the study also includes a limited qualitative element focusing on the behaviour of the models, any systematic bias observed, and the way in which the models differ when utilising different contextual information. This study also has exploratory characteristics. The effect of injecting structured nutritional information into the system prompt of a multimodal language model has not been studied before, and thus, this work explores how different types of knowledge affect the behavior, bias, and consistency of the models in a real-world setting, especially in a real cafeteria environment using

Finnish food data.

3.2 Dataset and Ground Truth

3.2.1 Data Source and Image Dataset

The meal image dataset used in this study was obtained from an ongoing research project conducted at Flavoria, a research restaurant at the University of Turku, Finland. Flavoria is a sensor-equipped self-service lunch restaurant that serves as a multidisciplinary food research purposes and operated by Sodexo Oy [8], [9]. A total of 192 meal images were recorded in Flavoria during one lunch service. The images were available to the researcher at the end of December 2025 and are accessed from the University of Turku network via EduVPN.

All images are assigned a unique image id and a url to the image and contain a meal image with identical components that are used in the other experiments. The core food items are the meatballs, cooked potato, cream sauce, and green salad. These four items form the basis of the food classification and the subsequent nutrition estimation in this thesis. Data utilised in this work was not collected specifically for this thesis. It had already been collected as part of ongoing research in the Flavoria platform [9]. The role of the researcher in the work described within this document has been to design and implement a pipeline to process the data and apply analysis using multimodal language models (LLMs).

3.2.2 Ground Truth Data Structure

For each image in the dataset, it contained ground truth nutritional information stored in a structured JSON object called lunchline_data_json. The ground truth in this work is obtained from the sensor-based Flavoria lunch line infrastructure. Each food item is RFID-tagged, customers receive trays that are tracked by the system

using RFID readers, and weighing sensors are mounted above each food station to exactly measure the amount of each food that is dispensed to a customer [8]. This data contrasts to the nutritional values obtainable from an expert, an online database search, or other sources and provides the user with real, objectively measured values.

The ground truth JSON for each analysis contains the food item names for all the recipes that were part of the mixed meal, weights of individual dishes, total weight of the meal, and the total nutritional values of the meal. The nutritional values are stored as a list of nutrient objects where each nutrient object is identified by a type code. This thesis uses the type codes ENER for energy in kilocalories, PROT for protein in grams, CHOT for carbohydrates in grams, and FAT for fat in grams. The total weight of the meal is stored as a separate numeric field called totalWeight. The analysis evaluation phase extracted these values through programmatic methods.

3.2.3 Plate Capture Conditions

In total, 192 images were recorded under two different plate presentation conditions, referred to as properly organised and semi-organised. In properly organised images, all the food items were spread out separately on the plate, each dish was well separated and defined with minimal overlapping areas between individual food portions.

The semi-organised condition extends to more difficult images where food items are less distinctly separated, resulting in a higher number of false positives, and may even overlap with adjacent objects. Moreover, some sauces spread across the surface of the plate and blur the boundaries between different food groups. While more challenging to annotate than 'Organised' images, both organised and semi-organised images are presented in the dataset and treated equally within our pipeline.

3.3 Model Selection and Experimental Design

3.3.1 Selection of Multimodal LLM Models

In this study, two families of multimodal large language models, namely the GPT families by OpenAI and the Gemini families by Google DeepMind, are studied as the two most relevant, commercially available general-purpose multimodal AI systems for visual understanding and interaction. Both models have been shown to be very powerful for several language-based benchmarks, such as instruction following [19], [32].

The study includes two model generations per family, and in the initial part of the experiment, GPT-4.1 mini and Gemini 2.5 Pro were tested on zero-shot tasks without prior context. The first assessment tested two model families because their base performance needed evaluation. OpenAI introduced GPT-4.1 mini to the market as their latest product on April 2025. Google DeepMind launched Gemini 2.5 Pro as their public product in March 2025 [19], [32]. After the initial comparisons were performed, successor models more suitable for the main experiments were selected. The two models used for the three conditions were GPT-5.4, an OpenAI model available through the OpenAI API as part of the GPT-5 frontier model series, and Gemini 3.1 Pro Preview, released by Google DeepMind in February 2026 [20], [21]. These more advanced language models have improved multimodal capabilities and better performance on the instruction following task required in this study. Importantly, Gemini 3.1 Pro Preview has a context window of up to one million tokens, which is practically relevant when incorporating large structured datasets, such as the Fineli nutritional database, into the system prompt.

Both models, OpenAI and Gemini, were accessed through their official APIs. API keys were stored securely as environment variables and loaded at runtime using the Python dotenv library.

3.3.2 Experimental Conditions and Contextual Scenarios

The study consists of four analyses. The first part of the study applies the older models (GPT-4.1 mini and Gemini 2.5 Pro) under zero-shot conditions. The latter three analyses were carried out by applying the more advanced models (GPT-5.4 and Gemini 3.1 Pro Preview), and comprise three different experimental conditions, namely, the zero-shot task performed without external context and two tasks in which external context was provided, namely, the Sodexo menu and the Fineli nutritional database. Table 3.1 outlines the four analyses conducted in the study.

Table 3.1: Summary of analyses, model configurations, and contextual conditions

Analysis	**Model name and conditions**
Analysis 1	GPT-4.1 mini and Gemini 2.5 Pro, no contextual information
Analysis 2	GPT-5.4 and Gemini 3.1 Pro Preview, no contextual information
Analysis 3	GPT-5.4 and Gemini 3.1 Pro Preview, Sodexo daily menu as context
Analysis 4	GPT-5.4 and Gemini 3.1 Pro Preview, Fineli nutritional database as context

In all four analyses, we utilized the same 192 image dataset, and both OpenAI and Gemini models were run on the images within the same pipeline. The results from each model were then processed separately and compared. This enabled us to conduct three different types of comparisons: between model families, model generations, and contextual conditions.

3.4 Contextual Data Preparation

3.4.1 Sodexo Menu Preparation and Cleaning

In Analysis 3, models were provided with contextual information, the daily Sodexo menu for the different lunch lines at Flavoria. Sodexo Oy, which manages Flavoria's catering, produces a daily menu that lists the dishes that will be available on each lunch line [8]. We obtained access to the raw menu data in a JSON format raw_context_menu.json. Importantly, raw file contains very complex nested struc-

tures and fields, and does not directly serve as context in a prompt. The python script, clean_menu.py, was used to clean the raw_context_menu.json file by extracting nutritional values for energy, protein, carbohydrates, and fat from each one. Missing calorie values were removed, and the result was saved to a clean_menu.json file. This cleaned file was then loaded at runtime and serialised as a JSON string for injecting into the system prompt.

3.4.2 Fineli Database Preparation and Cleaning

For Analysis 4, the following context information was injected to the models: the Finnish national food composition database, Fineli. Fineli is published by the Finnish Institute for Health and Welfare and it contains information on nutritional content of over 4 000 food items which have been analysed. Information covers energy, protein, carbohydrates, fat, and other nutrients per 100 g serving [10]. The raw data is available in the unstructured format in the Excel file Unstructured_fineli_database.xlsx.

To clean the Unstructured_fineli_database.xlsx file, the script clean_fineli.py is used. The script first renames the columns to more standardised names, then it removes the less-than symbols from the values, multiplies the energy 'kJ' values by 4.184 to convert them to 'kcal', changes the food names to lowercase and removes any duplicate entries. It then removes any rows with missing values, and finally it rounds all the values to two decimal places. The resulting cleaned dataset has been saved as fineli_cleaned_analysis4.csv.

During inference, the cleaned Fineli dataset was loaded and converted into a readable text format. For each food entry, a structured block was generated containing the dish name, energy per 100 grams, protein per 100 grams, carbohydrates per 100 grams, and fat per 100 grams. This full context string was then injected into the system prompt before inference.

3.5 Prompt Design and Context Injection Strategy

3.5.1 Baseline Prompt Design

The prompting strategy used within this thesis is based on two parts of prompting, the system prompt and the user prompt. This type of prompting is standard when instructing large language models via their API interfaces [33]. Therefore, the used prompting strategy was consistently applied to the four analyzed scenarios.

The user prompt was maintained constant for all four analyses within the thesis. It instructed the model to identify all visible food items on the plate, estimate the portion size of each item in grams, and calculate the total nutritional values, including calories, protein, carbohydrates, and fat. The prompt contains several rules including estimation on the safe side, not guessing of hidden ingredients, only one JSON string of text returned to the client etc. The JSON output required to be returned by the model is specified to contain an array of items, where each item of food is defined by its name and weight in grams, together with a total object containing the sum of calories and macronutrients for all the food on the plate.

For the baseline condition, the system prompt was kept minimal and simply instructed the model to return only valid JSON with no extra text. This approach does not provide the model with any information concerning the task and therefore, the model's ability to recognize foods and estimate quantities in the absence of any external information can be evaluated in a true zero-shot scenario.

3.5.2 System Prompt Context Injection

The external nutritional information was injected into the system prompt for the contextual experiments. The information was not included in the user prompt to keep the variable between the baseline and the contextual situation as small as possible. In prompt engineering, the system prompt is used to set the model's operational

context and prior knowledge [33]. Therefore in our experiments the system prompt and the user prompt were kept as separate entities, and the information to be used in the contextual situations was injected into the system prompt.

In analysis 3, the cleaned Sodexo menu was injected into the system prompt, where the model has access to a reference database of dishes and their nutritional values and that it must use this dataset for identifying food items by matching them to the closest dish in the dataset and for calculating all nutritional values. The model was not allowed to invent any food items not included in the reference database and prior knowledge was not allowed to be used for the calculation of the nutritional values. Additionally the model was instructed to scale the nutritional values per 100 grams, according to the estimated weight of the found items. In Analysis 4 the same system prompt structure as in Analysis 3 was used. The only difference was that instead of the Sodexo menu, the full cleaned Fineli database context string was used. The rest of the instructions were identical in terms of word count and structure.

There is a significant technical difference between how the two models are processing the system prompt context. The OpenAI API supports a system message role, and thus the context is passed as a separate system-level input. The Gemini API used in this study, however, does not support a separate system prompt parameter, and thus the system prompt content and the user prompt are concatenated into a single combined prompt string that is then sent to the model. This difference in how the two APIs process prompt information must be taken into account when comparing the behavior of the models in different contextual conditions.

The exact user and system prompts used across all experimental conditions are provided in Appendix A for reproducibility purposes.

3.6 Automated Inference Pipeline

3.6.1 Pipeline Architecture

The experimental pipeline for this study was implemented in Python, processing all images in a single automated run through the entire dataset of 192 images for all four different conditions, with the main part of the pipeline code remaining the same for each condition with a minor modification in each experimental condition. Several python libraries are used in this experiment such as pandas to load, process and store the images, the requests library to download images from URLs, the json module to parse model output, openpyxl to read and write .xlsx files, and the dotenv module to manage API keys as environment variables.

Each image is then processed in a sequential manner. For each image, the corresponding information was loaded from the input Excel files, downloaded the image from the image URL, and then send the image to the OpenAI and Gemini APIs with the corresponding prompts. The cleaned-up model output is then stored. This process is repeated for all 192 images in the dataset. Importantly, both models are processing the same image, and the resulting model output is stored as separate rows in the resulting file for easy comparison between the two models.

3.6.2 Image Retrieval and API Integration

The source data for the pipeline is in an Excel spreadsheet with a list of image URLs. In the pipeline, the list of image URLs is then used to download the images in bytes. The MIME type is automatically determined by the file extension of the image.

OpenAI API images are encoded as base64 in data URLs in the requests. The system prompt is submitted as a “system” role message, followed by the user prompt and the base64 encoded image for that prompt. For Gemini API requests, the image is sent as binary part using the API client’s Part.from_bytes method. The combined

context and user prompt for each image is sent as a single input, rather than as separate "system" and "user" messages.

The pipeline includes error handling. In cases where the system failed to download an image, or an error occurred during an API call, the failure is noted, and the pipeline continues with processing the rest of the images in the dataset.

3.6.3 JSON Output Handling and Storage

The models were asked to only return output in the structured JSON format. The output was returned in structured JSON format because that allows for the models' output to be parsed automatically, the food items to be returned with their corresponding weights, and to be compared against the ground truth values that were returned in the lunchline JSON [34]. Models were also returning the JSON output of the LLM wrapped in markdown code fences, and so a function was created to remove any markdown formatting from the output that was returned by the models. This function checks for the existence of code fence markers and removes them, returning the cleaned JSON as text.

The cleaned output from the models were stored as Excel files for each analysis. Each row of the output file corresponds to a model processing a single image. The columns are: image_id (unique id of the image), image_url (url of the image), tray_id (unique id of the tray on which the images were placed), plate_capture_condition (the condition under which plates were captured in the lunchline), model_name (the name of the model), llm_output (the raw LLM output), error (any error that the model encountered), and ground_truth_lunchline_json (the ground truth lunchline JSON for the images in the tray). A test script, test_parsing.py, was also written to test that all of the output files could be parsed correctly as JSON.

3.7 Food Name Normalisation and Mapping

Evaluating Generative LLMs for food recognition has several challenges, one major challenge is that the models can output free-text food names for the same image, and for the same food item in different images and also in different inferences. For example, the models might output: “meatballs”, “swedish meatballs”, “meat patties”, “small meatballs” for the same dish, and therefore, a simple string comparison against ground truth food names will not be accurate and will also result in a low recognition accuracy. Here A manual food name normalisation was done in this work. A dictionary was created where different models-generated food name variants were all mapped to four different canonical food names, that are meat balls, cooked potato, cream sauce, and green salad. These four food items are the basic food items in the dishes in the Flavoria dataset. The mapping was created by first running the entire pipeline on the data from the Flavoria dataset, collecting all the unique names of food items found by the two models. All these names were then manually reviewed and assigned according to the four canonical categories.

The number of entries required for the cream sauce category was the largest among the models, since the category consisted of many variations describing the same dish, such as peppercorn cream sauce, brown gravy, creamy brown sauce, gravy, and many more. In addition, variants like iceberg lettuce, shredded lettuce, and even cabbage were required for the green salad category. Names for objects and food that did not fall into any of the four categories were marked as unknown.

Additional mapping for the Fineli context condition was added. There were cases in which the models worked perfectly and output exactly what was in the Fineli database. Some of the mappings that were added for this analysis (4) are: The average of industrial products that are classified as meatballs, potato that has been peeled and boiled with or without salt, and cream sauce that was made with a meat bouillon.

In order to ensure that the normalisation is done in a correct manner, the mapping script is run as part of the analysis, not during inference. This ensures that the models are able to behave naturally, and that the pipeline is not affected by any bias that the normalisation could introduce. The script goes through all the model output, it extracts all the predicted food items, it goes through the mapping dictionary and it applies the mapping to the extracted food items. It also stores the original food items, the mapped food items and the estimated weights in a CSV file for further analysis.

3.8 Evaluation Metrics and Accuracy Measurement

To assess the performance of the models proposed in this thesis, the evaluation of the respective models can be performed on four different levels: total meal level, macronutrient level, dish level, and error distribution level. To assess the nutritional models on these four levels, a set of error measures are used to provide a complete assessment of the respective models in terms of accuracy, bias and consistency. For the initial comparison of the two models in Analysis 1, the primary metric was the Mean Absolute Percentage Error (MAPE) of total calories for all images in the test set. The MAPE metric is a simple average of the absolute percentage errors for all images, expressed as a percentage of the ground truth, or actual value, for that image. MAPE is a very easy to understand and used metric, because it can be reported for comparisons at different scales [12]. The absolute percentage error for an image is the absolute difference between the predicted number of calories and the ground truth number of calories for that image, divided by the ground truth number of calories for that image, multiplied by 100.

When investigating the performance of the models in more depth in Analyses 2, 3 and 4 the Mean Absolute Error (MAE) was the chosen metric for the nutritional content as well as the total weight of the dishes. The MAE for a set of predictions

is the average of the absolute values of the differences between the predictions and the corresponding ground truth values. The MAE is expressed in the same units of measurement as the data, so for example the MAE for the protein content of a set of images would be expressed in grams. The MAE for the energy of a set of images would be expressed in kilocalories. MAE is chosen as the primary error metric for these predictions as it is a simple to interpret metric, and it is robust to the occurrence of the odd extremely large error [4].

Signed error analysis for all the nutritional variables under study was also performed. The signed error for each nutrient is calculated by subtracting the ground truth value from the predicted value. Thus, it retains the sign of the error, enabling the model to be assessed for instances of constant overestimation or underestimation. Such instances of systematic errors can provide valuable insights into a model's performance under specific contexts. Error distributions were visualized by plotting histogram graphs including kernel density estimation curves for the error of the various nutrients. In these histogram graphs a red dashed vertical line at zero was added. This line of zero indicates the value where the model would be perfectly correct. The histogram graphs were plotted side by side to easily compare the error distributions of the OpenAI and Gemini models for the various nutrients.

For the evaluation of the models on a dish level, the predicted weights of the canonical food items for each image have been aggregated and compared to the ground truth dish weights for each lunch that have been extracted from the lunchline JSON. This way the performance of the models on individual food items can be determined, which provides a more detailed view than the evaluation on total meal level and helps to identify the food items for which the models perform best and worst [4], [12].

3.9 Ethical Considerations and Data Privacy

There are several ethical considerations that are taken into account during this study: These meal images were collected in a controlled environment for the research purposes. Thus, the meal images do not contain any personal information of individual customers. The way in which data is collected and processed in the research environment is in line with the research ethics that are required by the university [8].

Also when transferring image data of meal tray images through commercial AI APIs of OpenAI and Google DeepMind, questions arise how the data is handled and stored by the third-party service providers. The data transferred, consisting of images of food, does not contain any sensitive personal data, such as faces or identification information. However, it is good to be aware of the data retention and processing of the service provider [35].

The nutritional information generated by the nutritional estimates of the meal images in this study are experimental information and are not intended to be used in any clinical nutritional purposes or as a substitute for a professional diet assessment. The information and results in this study are merely to evaluate the capabilities of general-purpose multimodal LLMs in experimental settings.

3.10 Summary

In this chapter, the methodology of the conducted research of this thesis is presented. The research is of quantitative nature and is based on an experimental-comparative design with partially exploratory in nature. The study tested the use of multimodal LLMs in the task of food recognition and estimation of nutritional values in a real Finnish setting of a cafeteria. The dataset of the study consisted of 192 images that were taken in the Flavoria research restaurant.

The experimental design is presented in four analyses using two generations of models from two families, namely the OpenAI GPT family and the Google Gemini family. The oldest models from the families were evaluated in a zero-shot experiment. The latest models from the families were evaluated in three experiments, namely zero-shot, Sodexo menu context, and Fineli database context. The context was injected to the models in two different ways. For the OpenAI GPT family the system prompt was modified to accommodate the context. For the Google Gemini family the combined prompt (containing both the user and the system prompt) was modified to accommodate the context.

All of the 192 images of meal plates were processed by both models for each of the analyses and accordingly, all of the experimental analyses were implemented as fully automated Python code. The food name normalization used during the analysis of the results was also performed using a manually constructed mapping dictionary. The results of the experiments with the two LLMs for the 192 images of meal images were evaluated using the mean absolute percentage error (MAPE), mean absolute error (MAE), a signed error analysis, a dish level comparison of the results for the two models for each of the images, and plots of error distributions. In addition, the implications of the results with respect to the ethical issues of privacy, the transmission of data by APIs, and responsible use of AI-generated information for assessing nutritional value were discussed.

4 Results And Analysis

4.1 Overview of Experimental Results

This chapter presents the results of the four experiments that were set up in this thesis. Two different multimodal LLMs are compared, the GPT-series from OpenAI and the Gemini-series from Google. The goal of the experiments is to assess how different contextual information affects model behaviour and compare the performance of the two LLMs for food item identification and for nutritional estimation from images of meals that were captured in the Flavoria research restaurant. The chapter is organized in a sequential manner, the results of the first experiment that was set up with older models is presented first, followed by the results of the experiments with the updated models under three different contextual conditions. A separate section is dedicated to the discussion of the food name inconsistency problem that was one of the hardest issues to be handled in the course of the experiments. The chapter concludes with a cross-analysis summary of the findings that were obtained in the four experiments.

The four analyses of the thesis that were performed are summarized below. In Analysis 1, the experiment with the older models GPT-4.1 mini and Gemini 2.5 Pro were evaluated without any contextual information. The experiment calculated the calorie estimation accuracy by means of MAPE (Mean Absolute Percentage Error). In Analysis 2, the updated models GPT-5.4 and Gemini 3.1 Pro Preview

were evaluated without any contextual information. The scope of this experiment was extended compared to the first analysis. The full macronutrient MAE (Mean Absolute Error) as well as the dish-level accuracy were calculated and the error distribution was analyzed in detail. In Analysis 3 and 4, the updated models were provided with contextual information. In Analysis 3, the Sodexo daily menu was used as context, while in Analysis 4, the Fineli nutritional database was used as context.

4.2 Analysis 1: Initial Baseline: Old Models Without Context

The first analysis was conducted by comparing GPT-4.1 mini and Gemini 2.5 Pro (the baseline models) against the ground truth data. The analysis was performed by feeding the models a meal image and a prompt asking them to identify food items, estimate portion sizes, and calculate total nutritional values. The analysis was confined to the prediction of the calories for each meal, and used the metrics of MAPE and MAE to compare the performance of the models against the ground truth for each meal. The purpose of the analysis was to compare the performance of the two model families (the GPT models and the Gemini models) and to establish which model family performs best.

Table 4.1: Comparison of model performance based on MAE, MAPE, and standard deviation

Model	**MAE**	**MAPE (%)**	**STD**
Gemini	145.55	53.50	30.98
OpenAI	182.56	72.23	49.45

The results showed that Gemini 2.5 Pro scored a MAPE of 53.50% compared to the GPT-4.1 mini's MAPE of 72.23%. When expressed in terms of absolute error, Gemini scored an average of 145.55 kcal per meal compared to 182.56 kcal per meal

for the GPT-4.1 mini. As was observed in previous research, general-purpose LLMs lack sufficient contextual grounding and thus fail to provide reasonably accurate calorie estimates for meals [12]. While both Gemini 2.5 Pro and GPT-4.1 mini fell short of this benchmark, the performance of the two models could be contrasted in terms of their consistency. The standard deviation of the difference for GPT-4.1 mini (49.45) was greater than that of Gemini 2.5 Pro (30.98), indicating that the latter model provided more consistent estimates of the calorie content of meals. A strong baseline performance was demonstrated by Gemini in the estimation of calorie values without any context given. As a result of these findings, we chose to continue our analysis by utilizing the updated models in Analysis 2, which are from the same model families but more advanced in terms of their reasoning, and ability to process multimodal information.

4.3 Analysis 2: Updated Models Without Context

Analysis 2 is also a no-context experiment with the updated models: GPT-5.4 and Gemini 3.1 Pro Preview. However, this analysis included more than just the calorie, it also included the full macronutrient MAE for all the dishes as well as dish-level percentage errors and error distribution analysis for each of the 4 nutritional components.

4.3.1 Calorie and Macronutrient Estimation

Table 4.2 shows the calorie-level results for Analysis 2, and Table 4.3 shows the full macronutrient MAE including protein, carbohydrates, fat, calories, and total weight estimation.

Analysis 1 and Analysis 2 both showed very different results for Gemini 3.1 Pro Preview and GPT-5.4. Gemini 3.1 Pro Preview improved the calorie estimation

Table 4.2: Comparison of model performance based on MAPE values in Analysis 2

Model	**MAPE (%)**
Gemini	45.46
OpenAI	80.36

without context for the family of Gemini 2.5 Pro from a MAPE of 53.50% to 45.46%. This was an improvement for Gemini 3.1 Pro Preview over Gemini 2.5 Pro. On the other hand, the MAPE for GPT-5.4 without context was 80.36% as opposed to 72.23% for GPT-4.1 mini. This was an unexpected result. It appears that the GPT-5.4 is being more conservative in its nutritional estimations of this type of cafeteria food without context.

We see in Table 4.3 that Gemini 3.1 Pro Preview was much better than GPT-5.4 at the protein, fat, calorie, and total weight MAE for these 192 samples of cafeteria food without context. It is only the carbohydrate that GPT-5.4 estimated better, with an MAE of 5.66g versus 10.42g by Gemini. We see in later analysis that OpenAI estimates carbohydrates much better than it estimates protein and fat, and Gemini estimates protein and fat much better than it estimates carbohydrates. This same result is found in several subsequent analyses.

Table 4.3: Comparison of mean absolute error (MAE) values across nutritional attributes

Model	**Protein(g)**	**Carbs(g)**	**Fat(g)**	**Calories(kcal)**	**Weight(g)**
Gemini	5.35	10.42	9.16	127.68	66.45
OpenAI	11.96	5.66	19.19	219.54	139.68

The weight estimation MAE reveals an important underlying issue. The weight MAE for GPT-5.4 reached 139.68g per meal while Gemini showed a lower weight MAE of 66.45g. The nutritional values depend on estimated weights which means that any weight errors will directly lead to higher errors in both calorie and macronutrient

calculations [18]. This explains a significant part of why GPT-5.4 shows much higher calorie MAE than Gemini.

4.3.2 Dish-Level Performance Analysis

This analysis compares the performance of the models on an individual dish basis after the name normalization has been applied using the name normalization mapping. The average percentage error for each dish is shown in Table 4.4 for both Gemini and OpenAI. The percentage error here refers to the error in estimated dish weight compared to the ground truth weight.

Table 4.4: Dish-level percentage error comparison between Gemini and OpenAI models

Model	**Dish**	**Percentage Error (%)**
Gemini	Cooked potato	39.95
Gemini	Cream sauce	49.65
Gemini	Green salad	35.08
Gemini	Meat balls	20.89
OpenAI	Cooked potato	48.51
OpenAI	Cream sauce	83.11
OpenAI	Green salad	69.54
OpenAI	Meat balls	58.12

Gemini performed better than GPT-5.4 across all four items. However, for both of the models, the highest error was for the cream sauce dishes, with 49.65% average error for Gemini, and 83.11% for GPT-5.4. The findings of this study support the literature that amorphous food, such as sauces and soups, are the hardest food to estimate visually, as they do not have clear boundaries and their volume cannot be easily estimated from a 2D image [4]. The lowest error for Gemini was for the

meatballs, with an average error of 20.89%, as meatballs are distinct objects, generally of a fixed size, and therefore are fairly easy to size up as meatballs. For GPT-5.4, the lowest error was for the cooked potato, with an average error of 48.51%. The high errors of GPT-5.4 on green salad (69.54%) and cream sauce (83.11%) indicate that GPT-5.4 struggles with low-density foods and amorphous foods where the visual borders are not clearly given and the actual volume cannot be estimated from the 2D image. In contrast, Gemini estimated all dish types with a high level of accuracy in this condition, showing little variation only in the amount of error for each individual dish type.

4.3.3 Error Distribution and Bias Analysis

For an in-depth assessment of both models, error distribution plots for Analysis 2 have been created. These graphs show the signed error (predicted value minus ground truth value) of the models for all dishes, where a positive value indicates overestimation and a negative value indicates underestimation. The red dashed line at zero indicates the perfect prediction. The error distribution plots of Analysis 2 reveal the error's sign and consistency, or in other words, whether the error was an overestimation or an underestimation and to what extend this overestimation or underestimation has occurred. The values of the error's sign are the differences between the predicted values and the ground truth values, they are positive if there was an overestimation and negative if there was an underestimation. The error distribution plots are shown in the following figures.

First, the error distribution plot of the predicted energy is shown. As can be seen in Figure 4.1, both models obviously overestimated the calorie content, and the values of the error sign are mostly positive. While the errors of GPT-5.4 have spread out widely from about 50 kcal to about 450 kcal, the errors of the energy prediction of Gemini are more or less contained and peak at about 200 kcal. The

error distribution plot clearly indicates that both models overestimated the calorie content. The overestimation of the energy of GPT-5.4 is very inconsistent, whereas the errors of Gemini are more consistent and mostly between 50 and 300 kcal.

Figure 4.1 illustrates the distribution of prediction errors for energy estimation produced by OpenAI and Gemini models.

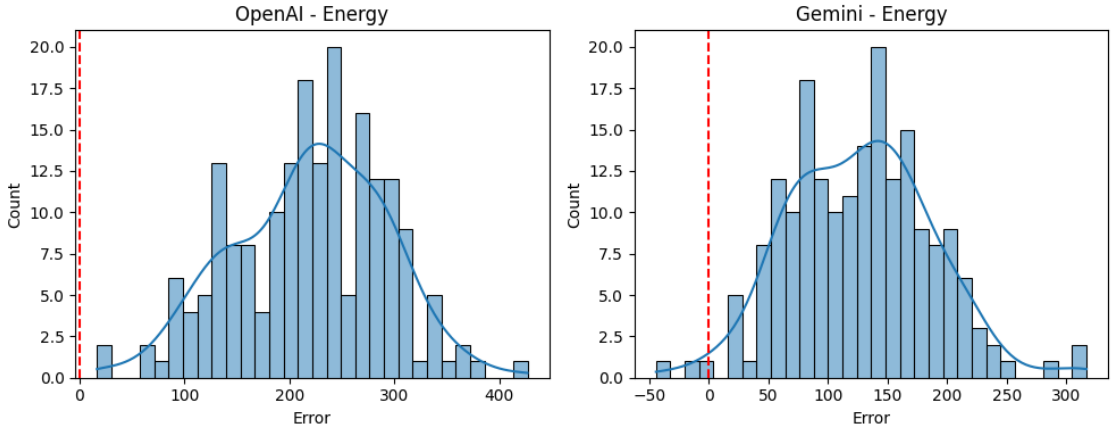

Figure 4.1: Error distribution comparison of energy predictions between OpenAI and Gemini models.

The error distribution plot for fat in Analysis 2 is shown in Figure 4.2. The errors for fat prediction are for the most part above ground truth values, with GPT-5.4 errors concentrated in the 15 to 25g overestimation range. The errors for fat prediction by Gemini fall into a 5g to 15g overestimation range with most of the errors being close to 5g overestimation. As with the energy errors, both models for fat prediction overestimated the nutrient for all dishes and did not make any underestimations.

Figure 4.2 shows the prediction error distribution for fat content estimation using OpenAI and Gemini models.

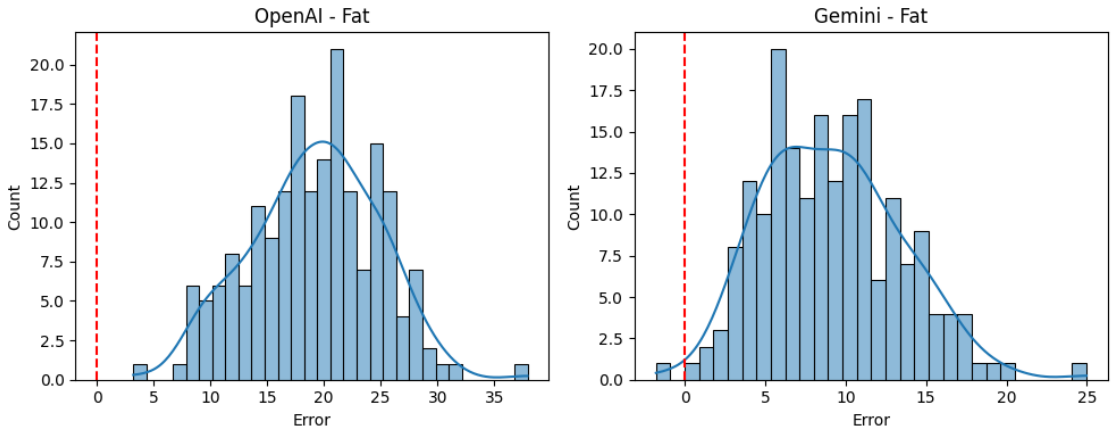

Figure 4.2: Error distribution comparison of fat predictions between OpenAI and Gemini models.

The signed error for protein (Fig. 4.3) shows both models have a positive bias with GPT-5.4 errors peaked around 10–15g of overestimation and Gemini errors peaked around 4–6g of overestimation. Both models overestimated protein in all cases. The overestimation of protein, combined with that of fat, is most likely due to the models' overestimation of portion sizes of high-protein and fat foods, for example, the meatballs in this analysis.

Figure 4.3 presents the distribution of protein prediction errors for OpenAI and Gemini models.

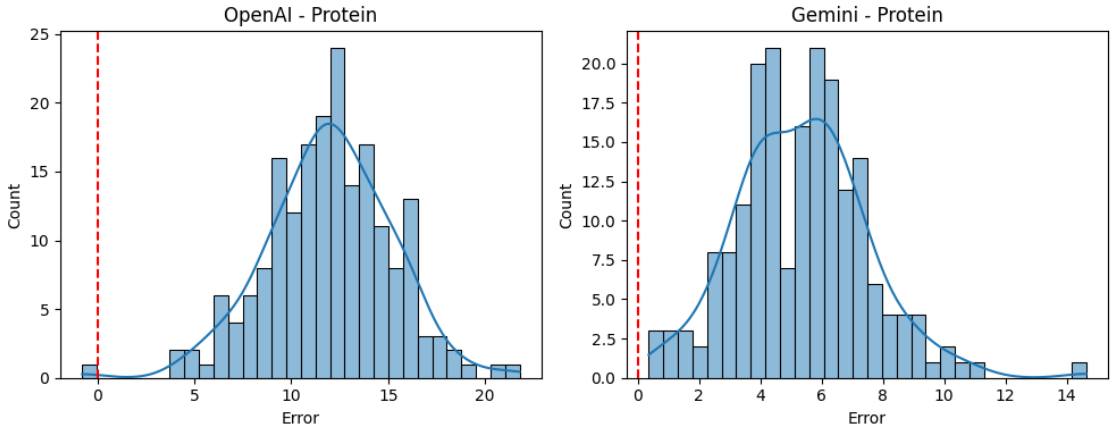

Figure 4.3: Error distribution comparison of protein predictions between OpenAI and Gemini models.

The error distribution plots for the analysis of carbohydrates show a completely different behavior than before. The errors of GPT-5.4 are centered around zero, thereby showing a slight positive bias. The errors of Gemini, however, are centered around a positive value as well and peak around 5 to 10g of overestimation, which is the only nutrient where GPT-5.4 is more accurate in terms of bias than Gemini, confirming the results of the MAE (see Table 4.3) as well.

Figure 4.4 presents the error distribution for carbohydrate predictions generated by OpenAI and Gemini models.

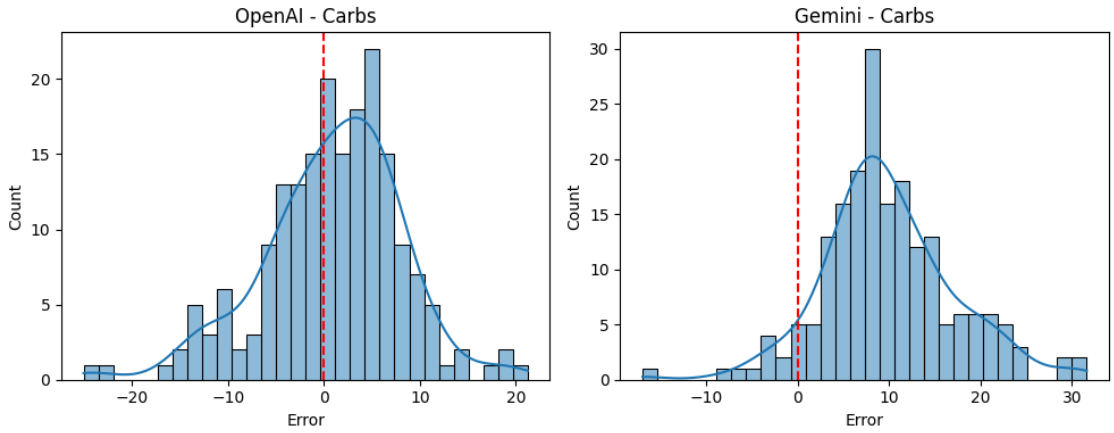

Figure 4.4: Error distribution comparison of carbohydrate predictions between OpenAI and Gemini models.

4.4 Analysis 3: Updated Models With Sodexo Menu Context

For Analysis 3, the Sodexo daily menu was provided as contextual information to understand the potential performance of the models when the full set of daily menu items is provided. The cleaned menu data containing dish names and their nutritional values per 100 grams was injected into the system prompt before inference for OpenAI as well as for Gemini as a system-level message containing the context for the foods on that day. For Gemini, the system prompt and the corresponding user

prompt were concatenated to form a single combined prompt that was presented to Gemini. The user prompt, with corresponding instruction to the models, remained identical to Analysis 2. The goal was to evaluate whether providing the models with information about what food was being served on that day would improve their food identification and nutritional estimation accuracy.

4.4.1 Impact of Menu Context on Nutritional Estimation

Table 4.5 presents the Mean Absolute Error (MAE) comparison between Gemini and OpenAI models using contextual information.

Table 4.5: MAE comparison between Gemini and OpenAI models using contextual information.

Model	**Protein(g)**	**Carbs(g)**	**Fat(g)**	**Calories(kcal)**	**Weight(g)**
Gemini	2.70	8.26	3.38	75.23	93.75
OpenAI	2.11	6.88	2.92	66.86	98.53

Analysis 3 with Sodexo Menu Context produced very consistent improvements for both Gemini and GPT-5.4, as measured by MAE for calories, protein, fat, and carbohydrates, as well as by average weight. For all of these metrics, performance in the No Context condition of Analysis 2 was significantly improved by the Sodexo menu context in Analysis 3.

However, for GPT-5.4, the improvement was still dramatic. The calorie MAE decreased from 219.54 kcal to 66.86 kcal, and the protein and fat MAE's decreased from 11.96g to 2.11g and from 19.19g to 2.92g respectively. There is little difference between the results of Gemini and GPT-5.4 in Analysis 3. In terms of calorie MAE, Gemini's results show the greatest decrease in MAE from the no-context condition, with a MAE of 75.23 kcal, while GPT-5.4 results show a decrease to a MAE of 66.86 kcal, the lowest MAE for calorie estimation of all conditions. The results for

protein MAE and fat MAE also show great similarity between Gemini and GPT-5.4 in Analysis 3, with both models estimating protein and fat in a very accurate manner. The greatest difference in the results of the two models in this analysis is in terms of carbohydrate MAE, with GPT-5.4 results showing a much lower MAE than those of Gemini, 6.88g compared to 8.26g respectively.

Although average absolute error of weight for Gemini increased from 66.45g to 93.75g in the Sodexo context, this was somewhat unexpected. On the other hand, the performance of GPT-5.4 at estimating weight of portions of foods improved in the Sodexo context. Its average error decreased from 139.68g to 98.53g. Therefore, although Gemini's estimates of individual portion sizes may have been altered in some way by the Sodexo context, they were changed in a way that caused Gemini's overall weight estimates to match the serving sizes that were listed in the menu more closely than the visual inspection of the image of the menu had suggested. In contrast, GPT-5.4's estimates of portion sizes were, on average, more accurate in the Sodexo context than they had been in the context in which the image of the menu was visually inspected.

4.4.2 Dish-Level Performance Analysis

Table 4.6 presents the dish-level percentage error comparison between Gemini and OpenAI models using contextual information for Sodexo meal components. The percentage error here refers to the error in estimated dish weight compared to the ground truth weight.

Table 4.6: Dish-level percentage error comparison between Gemini and OpenAI models using contextual information for Sodexo meal components.

Model	**Dish**	**Percentage Error (%)**
Gemini	Cooked Potato	40.49
Gemini	Cream Sauce	65.22
Gemini	Green Salad	34.73
Gemini	Meat Balls	37.86
OpenAI	Cooked Potato	36.58
OpenAI	Cream Sauce	67.89
OpenAI	Green Salad	43.82
OpenAI	Meat Balls	46.97

Cream sauce remains the hardest dish to estimate for both models even with Sodexo context, with errors of 65.22% for Gemini and 67.89% for GPT-5.4. This is consistent with the pattern seen in Analysis 2 and confirms that the difficulty with cream sauce is primarily a visual problem — knowing from the menu that cream sauce is on the plate does not help the model estimate how much sauce is present, because the sauce spreads across the plate and its boundaries remain visually ambiguous regardless of contextual knowledge [4].

For Gemini, green salad showed the best dish-level result at 34.73%, which is an improvement over the 35.08% seen in Analysis 2. Cooked potato improved slightly from 39.95% to 40.49%. However, meat balls worsened from 20.89% to 37.86% is an unexpected increase. This could be because the Sodexo menu context constrained Gemini to match items to specific menu entries, and the menu's meatball entry may have a different nutritional profile than what Gemini estimated visually in Analysis 2. For GPT-5.4, cooked potato showed the best result at 36.58%, and cream sauce remained the worst at 67.89%. The overall pattern at dish level for GPT-5.4 is similar to Analysis 2, suggesting that the main benefit of the Sodexo context for

GPT-5.4 came through improved total nutritional calculation rather than improved dish-level weight estimation.

4.4.3 Error Distribution and Bias Analysis

The error distribution plots for Analysis 3 show a striking improvement for both models compared to Analysis 2, and reveal that the two models now behave much more similarly to each other when they both properly receive the Sodexo context.

For energy, both models now show distributions centered very close to zero (see Figure 4.5). GPT-5.4's distribution peaks around 0 to 50 kcal above ground truth, a major improvement from the 150 to 300 kcal overestimation seen in Analysis 2. Gemini's energy distribution is also now closely centered around zero, with a shape that closely mirrors GPT-5.4's. Both distributions are much narrower and more symmetric than in Analysis 2, indicating that the Sodexo context helped both models calibrate their energy estimates effectively.

For fat, both models show near-perfect centering around zero (see Figure 4.6). The distributions for both GPT-5.4 and Gemini are tightly clustered between approximately -2g and +5g, with the peak very close to the zero line. This is a dramatic improvement from Analysis 2, where GPT-5.4 showed 15 to 25g overestimation and Gemini showed 5 to 15g overestimation. The Sodexo context essentially eliminated the fat overestimation bias for both models.

For protein, both models are again centered very close to zero (see Figure 4.7). GPT-5.4's distribution is centered just below zero, meaning it very slightly underestimates protein. Gemini's distribution is centered just above zero with a slight positive shift. Both distributions are narrow and well-behaved compared to Analysis 2.

The models show different patterns of carbohydrate results according to the carbohydrate testing results in Figure 4.8. The distribution of GPT-5.4 shows a

negative bias because its results shift leftward from the zero point which causes the system to underestimate carbohydrate content. Gemini exhibits a small positive bias because its distribution moves rightward from the actual value peaking between 5 and 10 grams above the true value. This difference in carbohydrate bias between the two models is the most notable remaining distinction in their behaviour under the Sodexo context condition and is consistent with the MAE results in Table 4.5.

Figure 4.5 illustrates the distribution of energy prediction errors for Sodexo context with OpenAI and Gemini models.

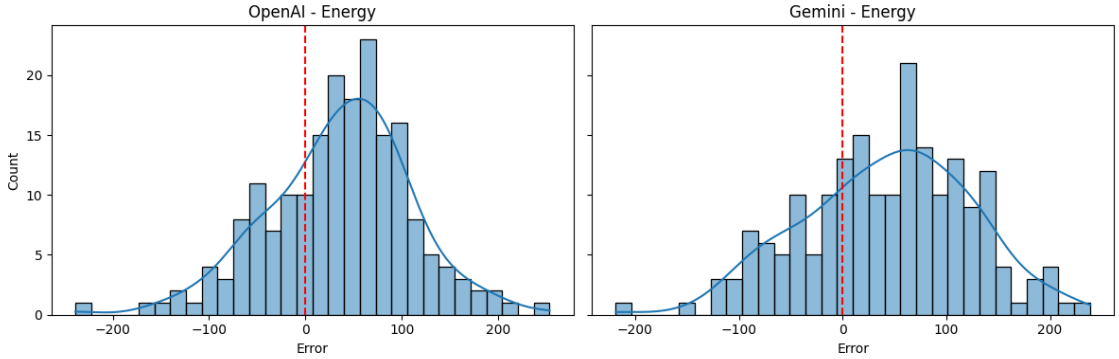

Figure 4.5: Error distribution comparison of energy predictions for context-aware OpenAI and Gemini models.

Figure 4.6 shows the prediction error distribution for fat estimation using Sodexo context with OpenAI and Gemini models.

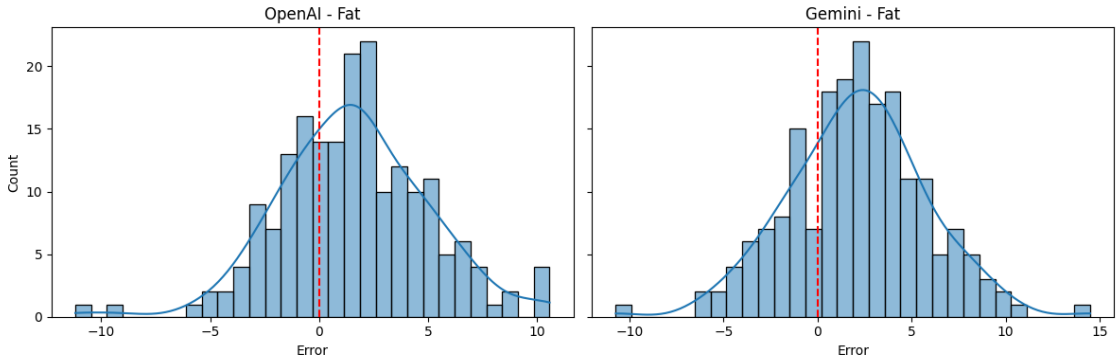

Figure 4.6: Error distribution comparison of fat predictions for context-aware OpenAI and Gemini models.

Figure 4.7 presents the distribution of protein prediction errors for context-aware OpenAI and Gemini models.

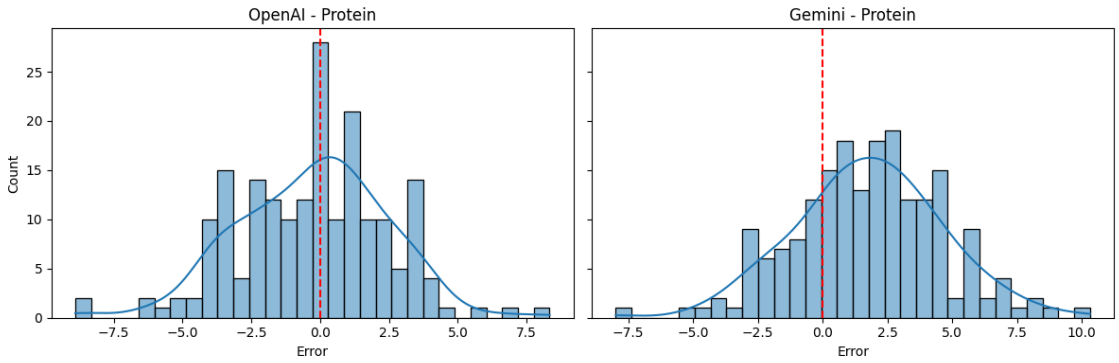

Figure 4.7: Error distribution comparison of protein predictions for context-aware OpenAI and Gemini models.

Figure 4.8 presents the carbohydrate prediction error distribution for context-aware OpenAI and Gemini models.

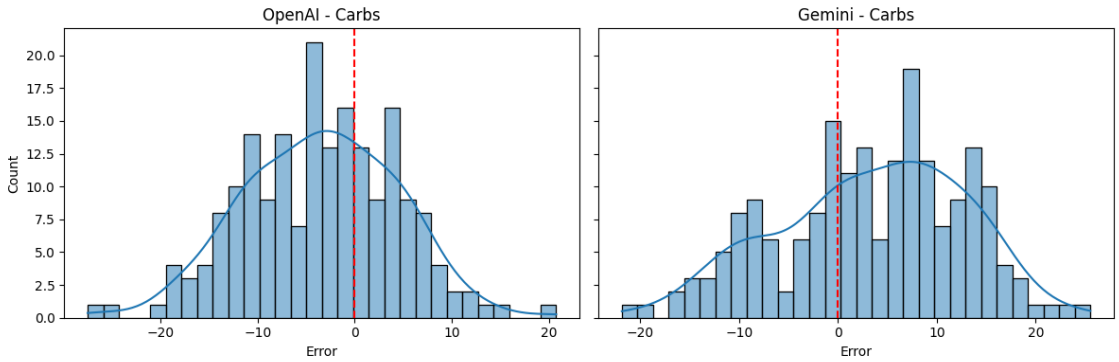

Figure 4.8: Error distribution comparison of carbohydrate predictions for context-aware OpenAI and Gemini models.

4.5 Analysis 4: Updated Models With Fineli Context

Analysis 4 is the most important experiment in this thesis. Instead of the focused Sodexo menu from Analysis 3, the full Fineli national food composition database was

used as context in this experiment. Fineli contains nutritional information for over 4,000 Finnish foods. The information given for each food includes energy, protein, carbohydrates, and fat all expressed per 100 grams of food [28]. The models were instructed to match as many of the foods found in the images to foods in the Fineli database and to use the nutritional information for the closest matching item to calculate the nutritional information for that food in the results. This analysis compares models grounded in the most comprehensive and authoritative Finnish nutritional database for Finnish foods to the results from the no-context and the Sodexo-context analyses.

4.5.1 Impact of Fineli Context on Nutritional Estimation

Table 4.7 presents the Mean Absolute Error (MAE) comparison between Gemini and OpenAI models using contextual information with the Fineli database.

Table 4.7: MAE comparison between Gemini and OpenAI models using contextual information of Fineli database.

Model	**Protein(g)**	**Carbs(g)**	**Fat(g)**	**Calories(kcal)**	**Weight(g)**
Gemini	5.08	4.89	6.20	87.38	67.35
OpenAI	9.12	6.39	12.95	153.17	127.06

Gemini's performance in the Analysis 4, compared to the Sodexo context is mixed. For carbohydrates Gemini achieved its best result of all analyses with an MAE of 4.89g (compared to 8.26g in the Sodexo condition). The estimation of weight also improved from 93.75g to 67.35g. However, as opposed to the Sodexo condition, the MAE for calories, protein and fat all increased rather than decreased. Thus, whereas the Fineli context improved the Gemini results for carbohydrates and weight, the results for the other nutrients were all worse than the Sodexo condition. This is somewhat unexpected, as Gemini had already shown improvement with the Sodexo

menu context in Analysis 3, and a more comprehensive database like Fineli was expected to improve its performance further. As mentioned before, the large amount of foods in the Fineli database makes it more difficult for Gemini to match the identified foods to the appropriate database entries.

GPT-5.4 however performed worse in Fineli context than in Sodexo menu analysis. The calorie MAE increased from 66.86 kcal to 153.17 kcal (more than double), protein MAE increased from 2.11g to 9.12g, fat MAE increased from 2.92g to 12.95g. The only slight improvement was in the carbohydrates MAE, which decreased from 6.88g to 6.39g. The weight MAE also increased from 98.53g to 127.06g. These results indicate that the GPT-5.4 has severe problems in processing a large nutritional database like Fineli in trying to match the visually identified foods with the correct database entries. The small Sodexo menu was in this case much easier for the model to process, as it had a small search space and was well defined.

Notably, while both models were able to use the large nutritional database Fineli as context to generate a variety of nutritionally annotated outputs of the foods shown in examples, it was clear that the Sodexo-focused menu context was far superior for both models at nutritionally annotating the examples. The large database contained many foods and many variations on foods, and it appears that the number of relevant items in the database, as well as their organization and relevance, are as or more important than the size of the database in this task.

4.5.2 Dish-Level Performance Analysis

Table 4.8 presents the dish-level percentage error comparison between Gemini and OpenAI models using contextual information for the Fineli dataset. The percentage error here refers to the error in estimated dish weight compared to the ground truth weight.

Table 4.8: Dish-level percentage error comparison between Gemini and OpenAI models using contextual information for the Fineli dataset.

Model	**Dish**	**Percentage Error (%)**
Gemini	Cooked Potato	31.54
Gemini	Cream Sauce	55.26
Gemini	Green Salad	37.59
Gemini	Meat Balls	22.26
OpenAI	Cooked Potato	42.48
OpenAI	Cream Sauce	88.21
OpenAI	Green Salad	63.71
OpenAI	Meat Balls	54.80

At the dish level, Gemini showed some interesting improvements in Analysis 4. Meatballs achieved its best result across all analyses at 22.26%, and cooked potato also improved to 31.54% — the best cooked potato result for Gemini across all four analyses. Green salad was at 37.59%, which is similar to previous conditions. Cream sauce remained difficult at 55.26%, though this is slightly better than the 65.22% seen in Analysis 3. For GPT-5.4, the dish-level results worsened across almost all categories compared to Analysis 3. Cream sauce reached its highest error across all analyses at 88.21%. Meatballs increased to 54.80% and green salad to 63.71%. Only cooked potato showed slight improvement at 42.48%. This is consistent with the macronutrient results — the Fineli context did not help GPT-5.4 estimate dish-level weights more accurately.

Across all four analyses, cream sauce consistently had the highest error for both models. Cream sauce is an amorphous liquid that spreads across the plate, making it inherently difficult to estimate visually regardless of how much contextual information is provided [4]. This is a fundamental limitation of 2D image-based food estimation that contextual grounding alone cannot fully overcome.

4.5.3 Error Distribution and Bias Analysis

The error distribution plots for Analysis 4 show that compared to Analysis 3, the distributions have shifted back towards a positive bias for both models — meaning both models tend to overestimate more with the Fineli context than with the Sodexo context.

For energy, both models show a clear positive shift in Analysis 4 (see Figure 4.9). GPT-5.4's energy errors are spread widely from approximately 0 to 400 kcal above ground truth, peaking around 150 kcal overestimation. Gemini's energy errors are more concentrated, peaking around 50 to 100 kcal above ground truth — still a positive bias but more contained than GPT-5.4. Both distributions are noticeably shifted further from zero compared to Analysis 3, where both models were centered very close to zero.

For fat, GPT-5.4 shows a clear positive bias with most errors between 10 and 20g of overestimation (see Figure 4.10). Gemini shows a smaller but still positive fat bias, with most errors clustered between 2 and 7g above ground truth. Both models overestimate fat more in Analysis 4 than in Analysis 3, where fat distributions were almost perfectly centered at zero.

For protein, GPT-5.4 errors are spread between 0 and 20g of overestimation, peaking around 6 to 9g (see Figure 4.11). Gemini's protein errors are smaller and more tightly clustered, peaking around 3 to 5g above ground truth. Again, both models show more overestimation in Analysis 4 compared to Analysis 3 where protein was very well centered.

For carbohydrates, the picture is more balanced (see Figure 4.12). Both models show distributions that are closer to zero compared to their fat and energy distributions. GPT-5.4 has a slight positive bias peaking around 3 to 5g above ground truth. Gemini's carbohydrate distribution is the most well-centered of all nutrients in Analysis 4, with errors spread almost symmetrically around zero. This confirms

the carbohydrate MAE improvement seen in Table 4.7 and suggests that the Fineli database provided Gemini with particularly accurate carbohydrate reference values for the foods in this dataset.

Figure 4.9 illustrates the distribution of energy prediction errors for context-aware OpenAI and Gemini models using the Fineli dataset.

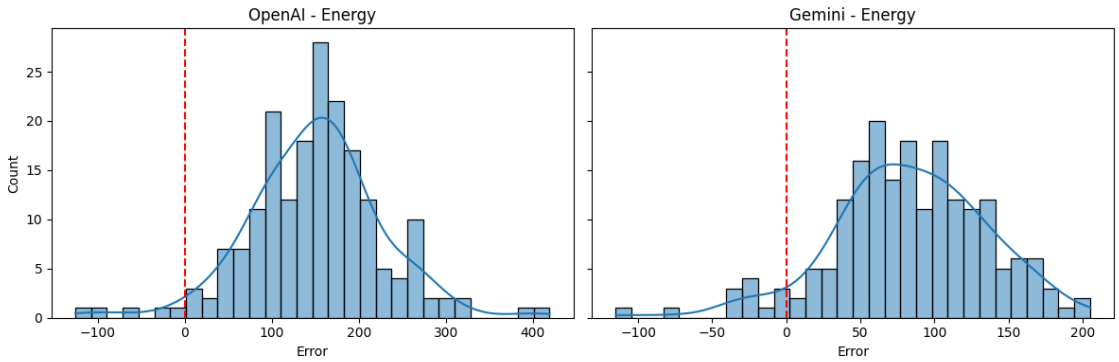

Figure 4.9: Error distribution comparison of energy predictions for context-aware OpenAI and Gemini models using the Fineli dataset.

Figure 4.10 shows the prediction error distribution for fat estimation using context-aware OpenAI and Gemini models on the Fineli dataset.

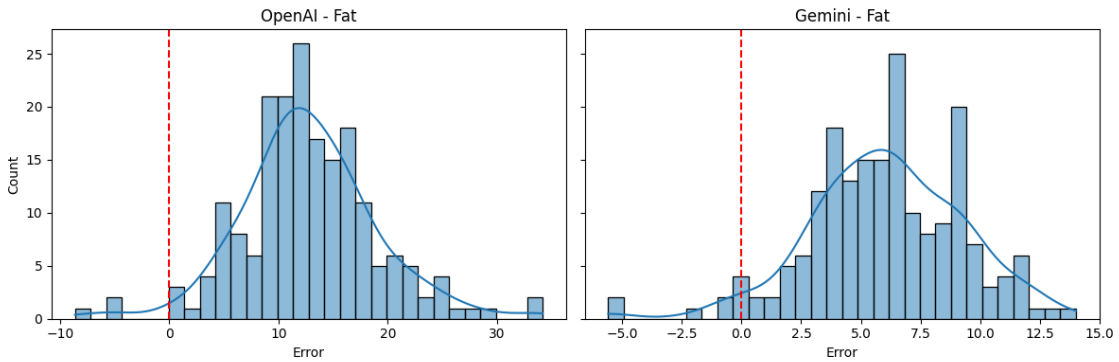

Figure 4.10: Error distribution comparison of fat predictions for context-aware OpenAI and Gemini models using the Fineli dataset.

Figure 4.11 presents the distribution of protein prediction errors for context-aware OpenAI and Gemini models using the Fineli dataset.

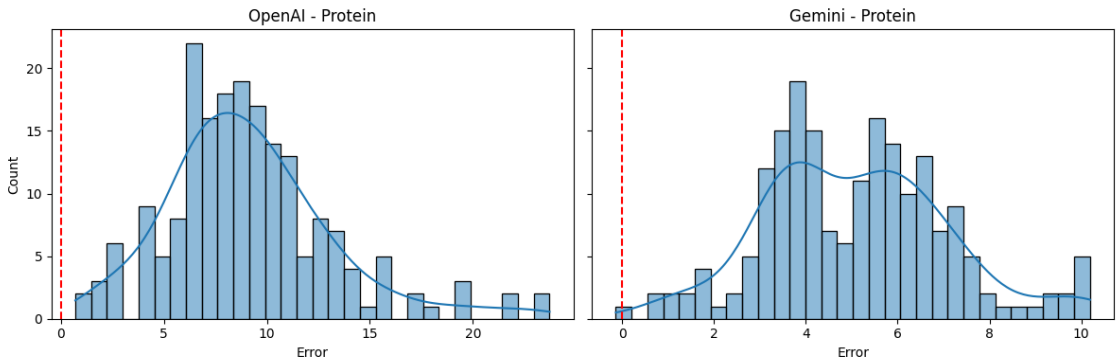

Figure 4.11: Error distribution comparison of protein predictions for context-aware OpenAI and Gemini models using the Fineli dataset.

Figure 4.12 presents the carbohydrate prediction error distribution for context-aware OpenAI and Gemini models using the Fineli dataset.

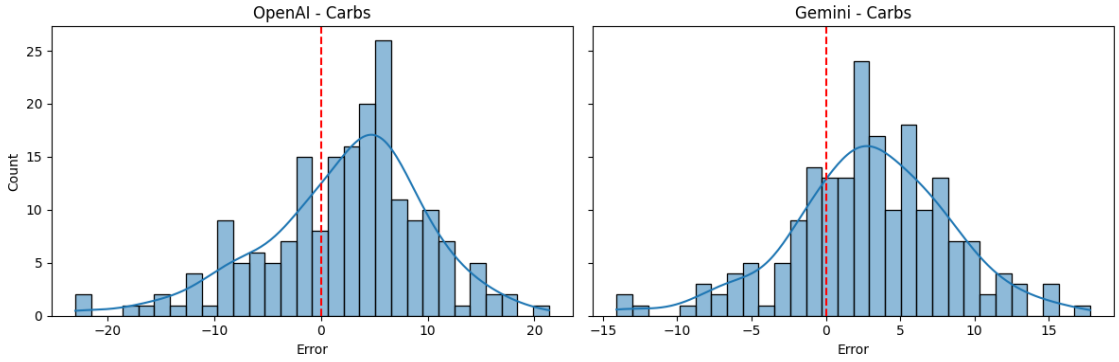

Figure 4.12: Error distribution comparison of carbohydrate predictions for context-aware OpenAI and Gemini models using the Fineli dataset.

These plots support our prior analysis. In addition to producing better-calibrated and less-biased nutritional quantities for the Sodexo menu context as compared to the Fineli database context, the Sodexo menu context appears to have far less overestimation for both models, particularly for GPT-5.4. The one case where the Fineli database context appears to be better for Gemini is in its carbohydrate estimation for the set of analyzed foods. This is likely because the Fineli database contains such a large amount of information on nutrients for many foods.

4.6 Food Name Inconsistency and Mapping Challenge

One practical challenge observed during the evaluation was the inconsistency in how the models named food items. Unlike a CNN classifier that outputs from a fixed vocabulary, a generative LLM produces free-text food names that can vary significantly across different images even for the same dish. This required a dedicated name normalization step before any evaluation could be performed.

The variation analysis in Analysis 2 showed a clear difference between the two models. GPT-5.4 generated approximately 769 total food item names with 171 unique variants, while Gemini 3.1 Pro Preview generated approximately 766 total names with only 77 unique variants. This means GPT-5.4 used more than twice as many different names for the same four food categories. Cream sauce showed the highest variability, over 60 different name variants were recorded across both models, including peppercorn cream sauce, brown gravy, creamy brown sauce, and pepper sauce among many others. Green salad was also variable, with models sometimes labeling it as iceberg lettuce, shredded lettuce, romaine lettuce, or even cabbage.

To handle this, a manual mapping dictionary was built to standardize all name variants to four canonical categories: meat balls, cooked potato, cream sauce, and green salad. For Analysis 4, additional Fineli specific names were added since models sometimes adopted exact Fineli terminology in their outputs.

4.7 Summary of Key Findings

This chapter presented the results of four analyses evaluating two multimodal LLMs for food recognition and nutritional estimation in a real Finnish cafeteria environment. Based on the results, the following key findings are identified.

First, both models can identify food items from cafeteria meal images and produce

nutritional estimates, but the errors are high across all conditions. Neither model comes close to the precision needed for clinical nutritional monitoring or individual dietary tracking.

Second, Gemini 3.1 Pro Preview consistently outperformed GPT-5.4 in the no-context conditions across most nutritional metrics. It produced lower calorie MAE, lower protein and fat errors, and more consistent dish-level estimates. GPT-5.4 was only better in carbohydrate estimation without context.

Third, when the Sodexo menu context was provided to both models, both showed significant improvements in nutritional estimation compared to the no-context condition. The two models also achieved their closest performance gap in this condition. Gemini with a calorie MAE of 75.23 kcal and GPT-5.4 with 66.86 kcal. This shows that a focused, relevant menu context is beneficial for both model families.

Fourth, the Fineli database context produced different effects on the two models. Gemini benefited mainly in carbohydrate estimation, where its MAE dropped from 8.26g to 4.89g, its best carbohydrate result across all analyses. However, for most other nutrients, Gemini performed slightly worse with Fineli than with the Sodexo menu. GPT-5.4 performed considerably worse with Fineli context across almost all metrics compared to the Sodexo condition. This suggests that a large comprehensive database does not automatically produce better results than a smaller focused menu, the relevance and size of the contextual information matters significantly.

Fifth, cream sauce was the most difficult food item to estimate across all models and all conditions without exception. Its amorphous and visually ambiguous nature makes it fundamentally hard to estimate from a 2D image regardless of what contextual information is provided. Meatballs was the most consistently and accurately estimated food item for Gemini across the analyses.

Sixth, food name inconsistency, particularly in GPT-5.4's outputs is a significant

practical challenge. GPT-5.4 generated 171 unique food name variants compared to Gemini's 77 for the same set of four dishes, requiring careful post-processing through a name normalisation mapping before evaluation could be performed.

5 Discussion

5.1 Overview

This chapter presents an analysis and discussion of the results obtained in the experiments performed in this thesis. The aim of this chapter is to provide a deeper understanding of the results of the experiments. Why do the models behave in a particular way? What do the results obtained in the experiments of this thesis mean for the application of multimodal LLMs for food recognition in a real-life cafeteria scenario, and how do the results of the experiments of this thesis compare to the results of other studies conducted on the same research question? Furthermore, this chapter describes the problems encountered during the implementation of the experiments and the limitations of this study.

The results of the experiments give deep insights to both the strengths and the weaknesses of general-purpose multimodal LLMs for the tasks of food recognition and for estimating nutritional values.

5.2 Interpretation of Model Performance Without Context

For the no-context condition, both models correctly classified the foods into several categories. However, their estimates were a systematic overestimation. Given that

the models have no knowledge of the specific foods in the sample and are relying solely on their pre-learned knowledge of portion sizes, their best case scenario is to provide positive bias (overestimation) of calories and macronutrients, similar to prior studies that used general-purpose LLMs for dietary assessment [12]. Therefore, no-context estimates from general-purpose LLMs are unlikely to be accurate in specific food environments, and models will require some form of contextualization to produce reliable results.

One reason that Gemini performed better in the no-context scenario may be due to its native multimodal architecture. The model processes visual information and textual information within the same model, simultaneously [19]. Therefore, Gemini is better able to integrate nutritional information with the visual observation of food, than GPT-5.4 which encodes visual information into an encoding that then is processed by the model for nutritional information. The internal workings of these models are not publicly documented in full detail, therefore it is purely an inference based on the results.

Interestingly, despite generally superior performance in previous studies, GPT-5.4 performed worse than the earlier GPT-4.1 mini model in no-context calorie estimation. This finding challenges the notion of newer generations of language models generally performing better on all tasks, and suggests that while GPT-5.4 is certainly a more capable general reasoning model, its application to specialized tasks such as this can in fact produce poorer results. The model's more deliberate processing of visual content in the absence of a context for a task such as nutrition estimation for example, results in different assumptions about portion size. This finding supports previous work [7] that notes that model capability in general does not equal performance in specialized applications.

5.3 Effect of Contextual Information on Model Behaviour

5.3.1 Sodexo Menu Context

The Sodexo menu context improved both models because it simplified the task. Instead of estimating what the food is and its nutritional values from scratch, the model only needs to estimate how much of each dish is on the plate. This anchoring effect of contextual information has been observed in existing literature as well [22]. Performance between models converged when both were using the correct, focused context. This implies that the difference between models, given the correct context, does not matter in this scenario. When using a daily menu that is always available for context in a real cafeteria setting, perhaps the way one structures their context will end up mattering more than the model itself.

For the difficult case of cream sauce, performance was again very poor with the menu context. As previously concluded, it appears that there is a strong visual component to this object, and knowing it is a sauce on a plate does not assist in estimating the amount of sauce when it has spread irregularly across the plate, potentially with unclear boundaries [4]. Indeed, no amount of additional contextual information will resolve this fundamental limit for 2D image-based estimation of 3D quantities.

It's worth noting the very different carbohydrate performance of the two models: GPT-5.4 tended to slightly underestimate while Gemini tended to slightly overestimate carbohydrates. This appears to result from the very different ways in which each model integrates provided contextual information into its prior. The way GPT-5.4 integrates such information appears to be to 'anchor' to provided values, while Gemini allows provided context to be influenced more by its prior for carbohydrates, such as from knowledge about typical portion sizes [6].

5.3.2 Fineli Database Context

The key insight from the Fineli experiment is that a larger database is not automatically better than a focused menu. GPT-5.4 performed worse than Sodexo while Gemini performed better on the carbohydrate content, but not for other nutrients. We believe that the size of the database to match against, i.e. thousands of entries for Fineli, makes it harder for the model to pick the correct match for a given observed food while the Sodexo menu for the experiment had only the dishes for that day, thus it was easier to match. This is in line with the work [7] where a retrieval-based pipeline was used to first select the most relevant database entries for a given input as opposed to injecting the full database into the model's context.

An important observation with Gemini's performance with respect to Fineli is that Gemini handled large context better than GPT-5.4, this is consistent with Gemini being designed for very long context processing [21]. Looking at context injection strategy in addition to data quality is an important research direction for improving performance. For example, using Sodexo's daily menu as input and then retrieving and injecting relevant portions of the thousands of entries of the Fineli database as relevant context for each menu item, would provide a model with the specificity it needs to perform well for daily served menus, and the nutritional accuracy of Fineli.

5.4 Food Name Inconsistency as a Methodological Challenge

The food name inconsistency problem deserves a dedicated discussion because it is not just a technical inconvenience, it reflects something fundamental about how generative LLMs produce outputs compared to traditional classifiers. A CNN trained for food recognition outputs probabilities over a fixed set of class labels. A generative

LLM describes what it sees in free text, which means the same dish can be described in many different ways depending on how it appears visually, how the prompt is phrased, and what terminology the model prefers.

The difference between GPT-5.4 generating 171 unique name variants and Gemini generating only 77 for the same four dishes is striking. GPT-5.4's higher verbosity and descriptive specificity which is generally considered a strength, actually becomes a weakness in an automated evaluation pipeline where consistency is more important than descriptive richness. Gemini's more standardised outputs were easier to work with and required less post-processing. This is an important practical consideration for anyone building food recognition systems using generative LLMs.

The case-insensitive normalisation using Python's .lower() and .strip() methods was a simple but essential fix. Without it, semantically identical labels like Meatballs and meatballs would be treated as different items, inflating the apparent inconsistency. The broader mapping dictionary then handled the remaining semantic variability mapping over 60 cream sauce variants and more than 20 green salad variants to their canonical categories. It is worth noting that the mapping dictionary was deliberately kept conservative. Only names that were clearly identifiable as belonging to one of the four categories were mapped. Names that were too ambiguous were left as unknown and excluded from dish-level evaluation. This conservative approach was chosen to avoid introducing bias if a model consistently produces names that could plausibly refer to different dishes, forcing them all into one category would artificially improve the apparent accuracy. The trade-off is that some valid predictions may have been excluded, potentially slightly underestimating true dish-level accuracy.

In Analysis 4, when the Fineli database was provided as context, both models sometimes adopted exact Fineli database terminology in their outputs, for example describing a dish meatball as meatballs average of industrial products. This is an interesting side effect of context injection that was not anticipated by providing

database terminology as context, the model starts using that terminology in its outputs. A small number of these Fineli-specific names were added to the mapping dictionary to handle this, but again conservatively. This shows that the relationship between context and output terminology is not one-directional. The context shapes not only the nutritional estimates but also the vocabulary the model uses to describe what it sees.

5.5 Technical Challenges Encountered

Several practical challenges were encountered during implementation that are worth discussing, as they reflect the real-world complexity of working with commercial LLM APIs for research. One of the technical challenges encountered was the difference in how the two APIs handle system prompts. The OpenAI API has a dedicated system message role that is clearly separate from the user message, making it straightforward to inject contextual information at the system level. The Gemini API handles prompting differently, it does not have an equivalent separate system message channel in the same way, which means contextual information needs to be incorporated directly into the combined prompt sent to the model. This architectural difference is not immediately obvious when starting to work with both APIs and highlights an important lesson — before implementing a multi-model pipeline, it is necessary to understand how each API structures its inputs and handles context. Treating different APIs as interchangeable without studying their individual architectures can lead to inconsistent experimental setups where models receive different information even when the intention is to give them the same input [33].

The decision to keep the user prompt fixed across all experimental conditions was deliberate and methodologically important. Changing the user prompt alongside the system prompt would have introduced an additional variable, making it impossible to isolate the effect of contextual information alone. While it is possible that a different

user prompt might have produced better results in some conditions, maintaining consistency was the right choice for the validity of the comparison.

Injecting the full Fineli database into the system prompt raised concerns about API cost and context window limits. API cost is affected by the number of input tokens, output tokens, the model used, and the total number of API calls. The Fineli database significantly increased the input token count per request. Running a small-scale test before the full dataset run helped estimate the cost and confirmed that the database fitted within the models' context windows. Both GPT-5.4 and Gemini 3.1 Pro Preview support very large context windows, which made this technically feasible [21].

The Gemini API also imposed rate limits that caused the pipeline to stop mid-run when requests were sent too rapidly. This is a known practical issue with commercial APIs and was resolved by introducing a time delay between requests. This kind of issue is not unusual when running large-scale inference pipelines and highlights the importance of building robust error handling into automated pipelines.

5.6 Limitations of the Study

Some of the limitations of the study reported in this thesis are addressed in this section. The number of images of foods of different categories used in the study was relatively small, 192 images representing four food categories of meals in a single cafeteria environment. The research site, where the study was conducted, provided the necessary environment for the author to carry out the specialized research. However, the findings of the study reported in this paper would not apply to foods of different categories in different cafeteria environments; to various different foods; and to complex meals.

The mapping dictionary that was created for this study relied on manual curation by the author. Therefore, the mapping of the model's output to the names of the

dishes created by the author in this study were subjective in nature. It is also worth noting that the author in this study deliberately left 10% of the predicted food names provided by the model for a dish unmapped in order to avoid over-mapping uncertain predictions made by the model for the dishes in the study. The unmapped predicted food names included some of the correct names of the foods in the study. Future research must be carried out to develop methods for the construction of the dictionary that would be used for the assessment of the accuracy of the predictions made by the model for the identification of foods and for their nutritional estimates, in order to create a more objective dictionary for the assessment of the model's accuracy.

5.7 Summary

In this chapter, the results of the research have been presented from the interpretive perspective. In the study, the results have been confirmed that the multimodal LLMs are able to recognize the foods and to estimate the nutrients from the images of the cafeteria meals. However, the results also revealed that the errors of the models are too large for using the models for precise nutritional monitoring. In addition, the results of the study revealed that the use of the contextual information, i.e., a focused daily menu, improved the results of both models significantly. In addition, the results also showed that a large comprehensive database, i.e., Fineli, produces model-dependent results. In the study, the problem of the food name inconsistency has been discussed from the perspective of the generative LLMs and traditional classifiers. In the study, several technical challenges were faced during the implementation. The challenges have been caused by the differences in the APIs, by the design decisions of the prompts, and by the design decisions of the context injection strategy. The results of the study are consistent with the existing literature, but they add new insights into how the different models are able to handle the

different types of contextual grounding in the real Finnish cafeteria environment.

6 Conclusion And Future Work

6.1 Summary of the Study

This thesis investigated the use of multimodal large language models for food item recognition and nutritional estimation from meal images captured in the Flavoria research restaurant at the University of Turku, Finland. The study was motivated by prior CNN-based work in the same environment and aimed to explore whether general-purpose multimodal LLMs without any domain-specific training could perform comparable food recognition tasks, and how the provision of structured contextual information affects their outputs.

A dataset of 192 meal plate images was used across four analyses. The first analysis evaluated older model versions GPT-4.1 mini and Gemini 2.5 Pro under zero-shot conditions to establish an initial baseline. The remaining three analyses used updated models GPT-5.4 and Gemini 3.1 Pro Preview across three experimental conditions: no context, Sodexo daily menu as context, and Fineli national food composition database as context. Both models were evaluated against objectively measured ground truth nutritional data from the Flavoria lunch line sensor system. Performance was measured using MAPE, MAE, signed error distribution analysis, and dish-level percentage errors.

6.2 Answers to Research Questions

RQ1 asked how accurately multimodal LLMs can identify food items and estimate nutritional values without external contextual information. The results showed that both models could identify the general food categories present on the plate, but nutritional estimation errors were high. Gemini 3.1 Pro Preview achieved a calorie MAPE of 45.46% and GPT-5.4 achieved 80.36% in the no-context condition. These models performed systematically poorly for the calorie and macronutrient estimation task because they relied on their built-in knowledge about standard portions rather than on the information in the images of the food in front of them. The error rates presented here for such general-purpose multimodal LLMs lacking contextual grounding to apply their knowledge to a situation as in nutritional monitoring are too high to be trusted.

RQ2 asked how providing a focused daily restaurant menu as context affects model performance. The Sodexo menu context produced clear improvements for both models. GPT-5.4's calorie MAE dropped from 219.54 kcal to 66.86 kcal, and Gemini's dropped from 127.68 kcal to 75.23 kcal. Both models also showed significant improvements in protein and fat estimation. This condition produced the smallest performance gap between the two models across all analyses, suggesting that a focused and relevant contextual input can reduce the influence of model-specific architectural differences. The improvement demonstrates that menu-based context is a practically viable approach for improving LLM-based food recognition in institutional food service settings.

RQ3 asked how providing a comprehensive national nutritional database affects model performance compared to a focused menu context. The Fineli database context produced different outcomes for the two models. Gemini improved notably in carbohydrate estimation, its best estimated carbohydrate across all analyses, but showed slight regression in calorie and fat estimation compared to the Sodexo

condition. GPT-5.4 performed considerably worse with Fineli than with the Sodexo menu across most metrics. These results indicate that the size and relevance of the contextual information matter significantly. A large database introduces ambiguity in food-to-entry matching, and this appears to affect GPT-5.4 more than Gemini. The Fineli experiment also revealed that Gemini handles large context more effectively, which is consistent with its architectural design.

RQ4 asked how different LLM families and generations compare across experimental conditions. Gemini consistently outperformed GPT-5.4 in the no-context conditions and responded more positively to the Fineli database context. GPT-5.4 responded more effectively to the focused Sodexo menu context and showed the most dramatic improvement in that condition. The comparison between old and new model generations showed that newer models do not automatically produce better results in all conditions. GPT-5.4 performed worse than GPT-4.1 mini in the no-context calorie estimation, which suggests that model capability in general does not directly translate to better performance in specialised dietary assessment tasks without appropriate grounding.

6.3 Conclusion

The findings of this thesis demonstrate the potential of multimodal LLMs for food recognition in a real cafeteria environment. Although the accuracy of the best results is still not on a level to allow for individual dietary monitoring or precise clinical nutrition. Here, large errors of even the best results, as seen in this thesis, must not be taken as a failure of the model, because the model was not designed for this type of task at hand. Instead, it is a characteristic of the task itself, because analyzing weight and nutritional information of food from 2D images without depth information is a very challenging task, regardless of the model that is used for the analysis.

Both models' accuracy improved after using context information. Both LLMs benefit significantly from Sodexo daily menu information injected into the prompt in a simple and realistic way. This approach requires no model retraining, no additional data collection, and no specialised infrastructure.

Also, the experiment with the larger and more comprehensive food database Fineli was used as context information. However, the accuracy of model outputs did not improve much when using Fineli database, and in some cases even decreased. This leads to the conclusion that in context injection approach, more information is not always better. The quality, relevance, and scope of the information has to be taken into account, and possibly different approaches have to be taken for different models.

Unlike CNN-based systems [4] for specific classification tasks, the models examined in this thesis are general-purpose, and their use in food recognition leads to low absolute accuracy but extreme flexibility. In other words, there is a great variety of objects that a free-running LLM can describe using a food name, or even a list of food names. Therefore, one can use them to recognize any food items and respond to contextual instructions in natural language. This means that the flexibility of the model for a specific purpose needs to be weighed against the desired accuracy of the results.

One issue that was discovered during the test runs of the free-text-based LLMs was the fact that the systems were unable to consistently and correctly name the food items in the test runs. While this does not pose a significant problem in practice, it does highlight the challenges that free-text-based systems present to automated evaluation, particularly in contrast to fixed-vocabulary classifiers. It is likely that any working implementation of food recognition using a generative LLM will require to implement some form of post-processing as well as possibly some constraints on the allowed structure of the output.

This thesis explores the capabilities and limitations of multimodal LLMs in food recognition task within a Finnish cafeteria setting. The results are grounded in reality, and point to specific areas that are worth exploring in more detail.

6.4 Future Work

Multiple directions of future work can be shown in this thesis, but the most promising one can be, instead of injecting the contextual information statically into the system prompt for every image, the information could be retrieved dynamically for each image using retrieval-augmented generation (RAG). This would be more powerful because the information that is most relevant for a given image could be retrieved for that image in particular, rather than information from the full database or even the focused menu being injected into the prompt for every image. This would allow the best information to be used for a given image, in particular for the nutritional calculations. For example, after the model has identified the food categories visible in an image, the most relevant entries from the Fineli database for the dishes on the day's Sodexo menu could be retrieved for that image, and those targeted entries could then be used for the nutritional calculations. This would combine the specificity of the focused menu used for the GPT-5.4 results reported in [29] with the nutritional accuracy of the Fineli database used in this thesis, and would be better than either the focused menu or the full database.

Apart from the above direction, future work on multimodal LLMs for dietary monitoring should include the evaluation of larger datasets, containing more images of different food categories in various cafeteria environments, under different lighting conditions, and with a variety of cuisines and ways of serving food. Such an evaluation could help to assess the strengths and weaknesses of LLMs for automated dietary monitoring and give a more complete picture of their potential use.

Further avenues for this work could include an investigation into automated

food name normalisation as an alternative to the manually-mapped dictionary used within this thesis. Utilising an embedding-based semantic matching approach where predicted food names are matched against canonical food categories using vector similarity, rather than exact string match, would address many of the current issues surrounding the handling of unseen food name variants, whilst also removing much of the inherent subjectivity found within the current manually-mapped dictionary. Additionally, an evaluation of the performance of domain-specific fine-tuned LLMs on the aforementioned data, within the identical experimental framework, could provide further insight into the extent to which the current accuracy gap between general-purpose LLMs and those that have been task-specifically trained could be addressed via the process of domain adaptation.

References

[1] A. Doulah, M. A. Mccrory, J. A. Higgins, and E. Sazonov, "A systematic review of technology-driven methodologies for estimation of energy intake", *IEEE Access*, vol. 7, pp. 49 653–49 668, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2910308. Accessed: May 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8686084.

[2] P. Chotwanvirat, A. Prachansuwan, P. Sridonpai, and W. Kriengsinyos, "Advancements in using AI for dietary assessment based on food images: Scoping review", *Journal of Medical Internet Research*, vol. 26, no. 1, e51432, Nov. 15, 2024. DOI: 10.2196/51432. Accessed: May 3, 2026. [Online]. Available: https://www.jmir.org/2024/1/e51432.

[3] L. M. Amugongo, A. Kriebitz, A. Boch, and C. Lütge, "Mobile computer vision-based applications for food recognition and volume and calorific estimation: A systematic review", *Healthcare*, vol. 11, no. 1, p. 59, Jan. 2023, ISSN: 2227-9032. DOI: 10.3390/healthcare11010059. Accessed: May 3, 2026. [Online]. Available: https://www.mdpi.com/2227-9032/11/1/59.

[4] T. Sarapisto, L. Koivunen, T. Mäkilä, A. Klami, and P. Ojansivu, "Camera-based meal type and weight estimation in self-service lunch line restaurants", in *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, Jun. 2022, pp. 1–7. DOI: 10.1109/ICPRS54038.2022.9854056. Accessed:

May 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9854056.

[5] G. A. Tahir and C. K. Loo, "A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment", *Healthcare*, vol. 9, no. 12, p. 1676, Dec. 2021, ISSN: 2227-9032. DOI: 10.3390/healthcare9121676. Accessed: May 3, 2026. [Online]. Available: https://www.mdpi.com/2227-9032/9/12/1676.

[6] D. Caffagni et al., "The revolution of multimodal large language models: A survey", in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13 590–13 618. DOI: 10.18653/v1/2024.findings-acl.807. Accessed: May 3, 2026. [Online]. Available: https://aclanthology.org/2024.findings-acl.807/.

[7] S. Yin et al., "A survey on multimodal large language models", *National Science Review*, vol. 11, no. 12, nwae403, Dec. 1, 2024, ISSN: 2095-5138. DOI: 10.1093/nsr/nwae403. Accessed: May 3, 2026. [Online]. Available: https://doi.org/10.1093/nsr/nwae403.

[8] L. Koivunen et al., "Increasing customer awareness on food waste at university cafeteria with a sensor-based intelligent self-serve lunch line", in *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, Jun. 2020, pp. 1–9. DOI: 10.1109/ICE/ITMC49519.2020.9198571. Accessed: May 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9198571.

[9] "Flavoria", Flavoria, Accessed: May 3, 2026. [Online]. Available: https://www.flavoria.fi/en/.

[10] "Frontpage - fineli", Accessed: May 3, 2026. [Online]. Available: https://fineli.fi/fineli/en/index.

[11] S. Romero-Tapiador et al., "Are vision-language models ready for dietary assessment? exploring the next frontier in ai-powered food image recognition", in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2025, pp. 430–439. DOI: 10.1109/cvprw67362.2025.00047. [Online]. Available: http://dx.doi.org/10.1109/CVPRW67362.2025.00047.

[12] J. Fridolfsson, E. Sjöberg, M. Thiwång, and S. Pettersson, "Performance evaluation of 3 large language models for nutritional content estimation from food images", *Current Developments in Nutrition*, vol. 9, no. 10, p. 107 556, Oct. 1, 2025, ISSN: 2475-2991. DOI: 10.1016/j.cdnut.2025.107556. Accessed: May 3, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2475299125030185.

[13] Y. Liu, "Automatic food recognition based on efficientnet and ResNet", *Journal of Physics: Conference Series*, vol. 2646, no. 1, p. 012 037, Dec. 2023, ISSN: 1742-6596. DOI: 10.1088/1742-6596/2646/1/012037. Accessed: May 3, 2026. [Online]. Available: https://doi.org/10.1088/1742-6596/2646/1/012037.

[14] D. Liu, E. Zuo, D. Wang, L. He, L. Dong, and X. Lu, "Deep learning in food image recognition: A comprehensive review", *Applied Sciences*, vol. 15, no. 14, p. 7626, Jan. 2025, ISSN: 2076-3417. DOI: 10.3390/app15147626. Accessed: May 3, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/15/14/7626.

[15] S. Rokhva and B. Teimourpour, *A novel method for accurate & real-time food classification: The synergistic integration of EfficientNetB7, CBAM, transfer learning, and data augmentation*, Oct. 3, 2024. DOI: 10.48550/arXiv.2410.

02304. arXiv: 2410.02304[cs]. Accessed: May 3, 2026. [Online]. Available: http://arxiv.org/abs/2410.02304.

[16] K. A. Nfor, T. P. Theodore Armand, K. P. Ismaylovna, M.-I. Joo, and H.-C. Kim, "An explainable CNN and vision transformer-based approach for real-time food recognition", *Nutrients*, vol. 17, no. 2, p. 362, Jan. 2025, ISSN: 2072-6643. DOI: 10.3390/nu17020362. Accessed: May 3, 2026. [Online]. Available: https://www.mdpi.com/2072-6643/17/2/362.

[17] F. S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "A review of image-based food recognition and volume estimation artificial intelligence systems", *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 136–152, 2024, ISSN: 1941-1189. DOI: 10.1109/RBME.2023.3283149. Accessed: May 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10144465.

[18] Y. Zhao, P. Zhu, Y. Jiang, and K. Xia, "Visual nutrition analysis: Leveraging segmentation and regression for food nutrient estimation", *Frontiers in Nutrition*, vol. 11, Dec. 17, 2024, ISSN: 2296-861X. DOI: 10.3389/fnut.2024.1469878. Accessed: May 3, 2026. [Online]. Available: https://www.frontiersin.org/journals/nutrition/articles/10.3389/fnut.2024.1469878/full.

[19] "Gemini 2.5: Our most intelligent AI model", Google, Accessed: May 3, 2026. [Online]. Available: https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/.

[20] "All models | OpenAI API", Accessed: May 6, 2026. [Online]. Available: https://developers.openai.com/api/docs/models/all.

[21] "Gemini 3.1 pro - model card", Google DeepMind, Accessed: May 6, 2026. [Online]. Available: https://deepmind.google/models/model-cards/gemini-3-1-pro/.

[22] F. P.-W. Lo et al., "Dietary assessment with multimodal ChatGPT: A systematic analysis", *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 12, pp. 7577–7587, Dec. 2024, ISSN: 2168-2208. DOI: 10.1109/JBHI.2024.3417280. Accessed: May 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10565946.

[23] R. K. Anam, *Evaluating gemini LLM in food image-based recipe and nutrition description with EfficientNet-b4 visual backbone*, Nov. 11, 2025. DOI: 10.5281/zenodo.17581078. arXiv: 2511.08215[cs]. Accessed: May 3, 2026. [Online]. Available: http://arxiv.org/abs/2511.08215.

[24] P. Zhou et al., "FoodSky: A food-oriented large language model that can pass the chef and dietetic examinations", *Patterns*, vol. 6, no. 5, May 9, 2025, ISSN: 2666-3899. DOI: 10.1016/j.patter.2025.101234. Accessed: May 3, 2026. [Online]. Available: https://www.cell.com/patterns/abstract/S2666-3899(25)00082-0.

[25] A. Gjorgjevikj et al., "Large language models in food and nutrition science: Opportunities, challenges, and the case of FoodyLLM", *Current Research in Food Science*, vol. 12, p. 101 351, Jan. 1, 2026, ISSN: 2665-9271. DOI: 10.1016/j.crfs.2026.101351. Accessed: May 3, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665927126000511.

[26] M. J. Mirza et al., "Meta-prompting for automating zero-shot visual recognition with LLMs", in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 370–387, ISBN: 978-3-031-72627-9. DOI: 10.1007/978-3-031-72627-9_21.

[27] H. Tanabe and K. Yanai, "Reasoning-driven food energy estimation via multimodal large language models", *Nutrients*, vol. 17, no. 7, p. 1128, Jan. 2025,

ISSN: 2072-6643. DOI: 10.3390/nu17071128. Accessed: May 3, 2026. [Online]. Available: https://www.mdpi.com/2072-6643/17/7/1128.

[28] N. Kaartinen et al., "The finnish national dietary survey in adults and elderly (FinDiet 2017)", *EFSA Supporting Publications*, vol. 17, no. 8, 1914E, 2020, _eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2020.EN-1914, ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2020.EN-1914. Accessed: May 3, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2020.EN-1914.

[29] R. Yan et al., "DietAI24 as a framework for comprehensive nutrition estimation using multimodal large language models", *Communications Medicine*, vol. 5, no. 1, p. 458, Nov. 5, 2025, ISSN: 2730-664X. DOI: 10.1038/s43856-025-01159-0. Accessed: May 3, 2026. [Online]. Available: https://www.nature.com/articles/s43856-025-01159-0.

[30] Flavoria Research Platform, *My Flavoria® App*, Accessed: 2026-06-02, 2026. [Online]. Available: https://www.flavoria.fi/en/for-customers/my-flavoria/.

[31] Y. Chang et al., "A survey on evaluation of large language models", *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, 39:1–39:45, Mar. 29, 2024, ISSN: 2157-6904. DOI: 10.1145/3641289. Accessed: May 11, 2026. [Online]. Available: https://dl.acm.org/doi/10.1145/3641289.

[32] "Introducing GPT-4.1 in the API", OpenAI, Accessed: May 3, 2026. [Online]. Available: https://openai.com/index/gpt-4-1/.

[33] S. Schulhoff et al., *The prompt report: A systematic survey of prompt engineering techniques*, Feb. 26, 2025. DOI: 10.48550/arXiv.2406.06608. arXiv: 2406.06608[cs]. Accessed: May 11, 2026. [Online]. Available: http://arxiv.org/abs/2406.06608.

[34] Z. R. Tam, C.-K. Wu, Y.-L. Tsai, C.-Y. Lin, H.-y. Lee, and Y.-N. Chen, *Let me speak freely? a study on the impact of format restrictions on performance of large language models*, Oct. 14, 2024. DOI: 10.48550/arXiv.2408.02442. arXiv: 2408.02442[cs]. Accessed: May 11, 2026. [Online]. Available: http://arxiv.org/abs/2408.02442.

[35] P. Detopoulou et al., "Artificial intelligence, nutrition, and ethical issues: A mini-review", *Clinical Nutrition Open Science*, vol. 50, pp. 46–56, Aug. 2023, ISSN: 26672685. DOI: 10.1016/j.nutos.2023.07.001. Accessed: May 11, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2667268523000311.

Appendix A Prompts Used in the Experiments

Listing A.1: Baseline User Prompt (Analyses 1–4)

```
Identify all visible food items on the plate.
For each item, estimate the portion size in grams and state
   your confidence level.
If identification is uncertain, list plausible alternatives
   and explain why.
Use conservative estimates and avoid guessing hidden
   ingredients.

Return ONLY JSON with this structure:
{
  "items": [
    {"name": "...", "estimated_weight_g": number}
  ],
  "total": {
    "calories_kcal": number,
    "protein_g": number,
    "carbs_g": number,
    "fat_g": number
```

```
  }
}

Rules:
- If unsure, make your best estimate.
- Do not include explanations outside JSON.
"""
```

Listing A.2: System Prompt: No Context Condition

```
SYSTEM_PROMPT = "Return ONLY valid JSON. No extra text."
```

Listing A.3: System Prompt: Sodexo Menu Context Condition

```
SYSTEM_PROMPT = f"""
You are a nutrition expert. You have access to a reference
   database of dishes and their nutritional values:

{menu_context_str}

You MUST use this dataset for:

- Identifying food items by matching them to the closest dish
   in the dataset
- Calculating all nutritional values

Rules:

- Do NOT invent food items that are not present in the dataset
- Do NOT use external or prior knowledge for nutritional
   values
```

```
- If unsure, select the closest matching item from the dataset
- All nutritional values must be calculated from the dataset
   which contains nutritional information per 100g and scaled
   according to the estimated weight
- Ensure internal consistency: kcal should approximately match
    4xprotein + 4xcarbs + 9xfat (+/-10%)

Return ONLY valid JSON. No extra text.
"""
```

Listing A.4: System Prompt: Fineli Database Context Condition

```
SYSTEM_PROMPT = f"""
You are a nutrition expert. You have access to a reference
   database of dishes and their nutritional values:

{menu_context_str}

You MUST use this dataset for:

- Identifying food items by matching them to the closest dish
   in the dataset
- Calculating all nutritional values

Rules:

- Do NOT invent food items that are not present in the dataset
- Do NOT use external or prior knowledge for nutritional
   values
- If unsure, select the closest matching item from the dataset
```

```
- All nutritional values must be calculated from the dataset
   which contains nutritional information per 100g and scaled
   according to the estimated weight
- Ensure internal consistency: kcal should approximately match
    4xprotein + 4xcarbs + 9xfat (+/-10%)

Return ONLY valid JSON. No extra text.
"""
```