Multi-Task Visual Food Recognition by Integrating an Ontology Supported with LLM

Daniel Ponte[a,*], Eduardo Aguilar[a,b], Mireia Ribera[a], Petia Radeva[a,c]

[a] *Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona, 08007, Spain*
[b] *Dept. de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Angamos 0610, Antofagasta, 1270398, Chile*
[c] *Computer Vision Center, Campus UAB, Edifici O, Cerdanyola, 08193, Barcelona, Spain*

Abstract

Food image analysis is a crucial task with far-reaching implications across various domains, including the culinary arts, nutrition, and food technology. This paper presents a novel approach to multi-task visual food analysis, using large language models to obtain recipes and support the creation of a comprehensive food ontology. The approach integrates the food ontology into an end-to-end model, with prior knowledge on the relationships of food concepts at different semantic levels, within a multi-task deep learning visual food analysis approach, in order to generate better and more consistent class predictions. Evaluated on two benchmark datasets, MAFood-121 and VireoFood-172, this method demonstrates its effectiveness in single-label food recognition and multi-label food group classification. The ontology enhances accuracy, consistency, and generalization by effectively transferring knowledge to the learning model. This method addresses the challenges posed by the diverse appearances of food and enriches the model's understanding of food concepts and their intricate relationships. This study underscores the potential of ontology-based methods to address food image classification complexities, with implications for a wide range of applications, including automated recipe generation and nutritional assessment.

Keywords: Food Ontology, Food Image Analysis, Multi-task Learning, Large Language Models

*Corresponding author
Email address: `dponteva163@alumnes.ub.edu` (Daniel Ponte)

1. Introduction

Food image recognition is a crucial research area with broad applications in culinary arts, nutrition, and food technology [1]. Accurate and efficient food recognition systems are essential for advancements in automated dietary assessment, recipe generation, and nutritional analysis, significantly impacting dietary monitoring, public health initiatives, and personalized nutrition plans [2]. In the culinary field, these systems assist in creating innovative dishes by analyzing existing recipes and ingredient combinations [3], and play a vital role in food safety and quality control by enabling automated identification of defective or mislabeled products [4]. The technology also extends to smart kitchen appliances and applications that provide real-time feedback and recommendations based on ingredient and dish identification, helping users make healthier dietary choices and reduce food waste by suggesting recipes with available ingredients [5]. Moreover, accurate food recognition supports individuals with dietary restrictions or allergies by helping them identify safe options [6], and its integration into mobile applications and wearable devices offers potential for continuous dietary monitoring and improving overall health [7].

Despite significant advances, food image recognition still faces challenges due to high variability in food appearance, preparation styles, and presentation [8]. Different cooking methods, ingredient combinations, and cultural differences result in Variations in color, shape, and texture result [2]. Traditional methods struggle to accurately classify complex dishes with multiple ingredients or from diverse cuisines [9]. These challenges are compounded by the need for large volumes of labeled data, which are costly and time-consuming to produce, leading to inconsistencies and errors in labeling [10]. The dynamic nature of culinary practices adds complexity to maintaining up-to-date datasets. Existing machine learning approaches may not effectively capture intricate relationships between food components, resulting in misclassifications that affect dietary assessments and nutritional analysis [11]. These limitations highlight the need for advanced techniques that can better model semantic relationships between food items and their components, providing more robust and reliable recognition capabilities [12].

To address these challenges, previous research has explored methodologies like convolutional neural networks (CNNs) for feature extraction and classification [13]. While CNNs are promising in extracting relevant features from food images, they often fall short in capturing the contextual information necessary for accurate classification [14]. These methods typically ignore the rich semantic relationships between foods and their ingredients, resulting in insufficient accuracy when recognizing complex dishes with overlapping ingredients [12]. For example, a CNN might correctly identify a tomato in a salad, but struggle to differentiate between similar dishes like minestrone soup and vegetable stew, where context and ingredient combinations are crucial. This limitation highlights the need for more sophisticated approaches that integrate contextual knowledge to enhance the robustness of food image recognition systems. Methods combining multimodal data, such as visual information with textual descriptions of recipes and ingredient lists, have shown potential for improving recognition accuracy [15]. Additionally, transfer learning techniques, where models pre-trained on large, diverse datasets are fine-tuned for specific food recognition tasks, can help address the scarcity of labeled data. However, these techniques still require a comprehensive understanding of the relationships between food items, as many dishes share similar visual features but differ in key ingredients or preparation methods. For example, dishes like "paella" and "seafood risotto" may appear similar but can be differentiated by recognizing specific ingredients like saffron or distinct types of seafood. Integrating ontologies [16], which capture semantic relationships between food groups and ingredients, and large language models (LLMs) [17], which assist in constructing these ontologies, can offer significant benefits. These ontology-based methods allow systems to distinguish visually similar dishes by leveraging ingredient lists and broader contextual information.

The proposed approach is notable for introducing a novel method for multi-task visual food analysis, using LLMs to obtain recipes and support the creation of comprehensive food ontologies [18]. LLMs, such as GPT-4, have transformed the field of natural language processing with their ability to comprehend and produce text that closely resembles human language [19]. These models enable the creation of detailed and accurate representations of food recipes and concepts. By integrating these ontologies which encapsulate prior knowledge about the relationships between food concepts at various semantic levels, the deep learning-based approach improves the accuracy and consistency of predictions [20]. This integration enables the system not only to identify individual foods, but also to understand the broader context in which they appear, ultimately improving the overall performance of the recognition process. The main contributions are:

- A specialized food ontology was meticulously constructed using LLMs, which provided accurate and reliable recipes.
- Better consistency in multi-task results as a consequence of integrating prior knowledge extracted from the ontology into the food image classification model, ensuring a more refined and accurate classification process.
- The proposed ontology-based method demonstrated significant performance improvements in both the individual dish and food groups levels compared to the baseline approach. This advancement highlights the effectiveness of combining ontological knowledge with image recognition techniques.

A very preliminary resulñts of this study was presented in [21]. In contrast to it, here we introduce LLMs to obtain recipes for all dishes and create detailed ontologies. Unlike prior work that only used part of the MAFood-121 dataset [22] focused on Mexican cuisine, we utilized the entire MAFood-121 dataset, allowing for a more comprehensive ontology. Automating the recipe collection process enabled us to gather more recipes and cover a broader range of cuisines, enhancing the ontology's ability to capture nuanced relationships between food concepts. Furthermore, we evaluated the proposed method on an additional dataset, VireoFood-172 [23], a Chinese food dataset notable for its diversity. Incorporating VireoFood-172 allowed us to assess the model's generalization capabilities across a more diverse array of dishes beyond MAFood-121, demonstrating the versatility and applicability of our method to more complex food recognition tasks. Two distinct ontologies were created to model the unique cuisines from each dataset, enhancing the precision of the food recognition model within each cultural context. In addition to previous processes, we used an LLM to obtain recipes and integrated a Linear Programming (LP) method to select main tokens for each ingredient, which was not present in our prior work. Our approach benefits from implementing new network architectures DenseNet-121 [24] and EfficientNet-B4 [25] alongside the previously used ResNet-50 [26], ensuring robust performance across different network designs. The results demonstrate that our proposal, we call LLM-guided Ontology Multi-task Food Recognition (LLMO-MFR), improves upon previous ontology-based food recognition research and significantly speeds up ontology building. This advancement highlights the effectiveness and versatility of our methodology in recognizing food dishes and food groups, confirming its superiority over baselines.

2. Related Work

Visual food analysis has advanced significantly in dish recognition due to deep learning methodologies, particularly CNNs. Initial methods relying on handcrafted features faced limitations due to complex variability in food presentations. Modern approaches leveraging CNNs and techniques like the Hough transform have shown notable improvements [27]. Studies have demonstrated the efficacy of deep learning in classifying a wide range of food items and estimating nutritional information [28]. The robustness of CNNs in handling diverse cuisines has been validated in research focusing on Indian food [29, 30] and traditional Indonesian dishes [31]. The integration of neural network classifiers and image segmentation techniques has further enhanced food recognition accuracy [32]. Machine learning-based approaches for nutritional estimation highlight the potential of combining food recognition with dietary assessments [33, 34]. The superiority of CNNs over traditional methods has been confirmed, showcasing their effectiveness in processing complex visual data [35]. Research has increasingly focused on food group recognition, providing significant advancements in dietary analysis and health-focused applications. Unlike individual dish recognition, food group recognition involves collections of dishes, posing unique challenges due to high intra-class variability and inter-class similarity [36]. Deep learning methodologies, particularly CNNs, have propelled these advancements. Modern techniques using CNNs and advanced image processing algorithms have shown notable improvements in classifying diverse food groups and estimating their nutritional content [11, 37]. The robustness of CNNs in handling various cuisines has been validated by research on different cultural food groups, showing enhanced accuracy and reliability [38]. Additionally, integrating neural network classifiers and image segmentation techniques has further improved recognition accuracy [39]. These advancements underscore the transformative impact of deep learning in dish and food group recognition, offering robust solutions for food identification and nutritional analysis across various culinary traditions. The continuous refinement of these technologies promises significant applications in dietary monitoring, health management, and culinary analysis.

Multi-task learning based on ontologies has shown significant potential in food recognition and analysis by utilizing deep neural networks to jointly learn multiple related tasks, enhancing overall performance and efficiency. A novel multi-task learning (MTL) architecture integrating multi-scale and label dependency learning effectively improved food and ingredient recognition by capturing fine-grained details and label dependencies [40]. Flexible and compact architectures for MTL in food recognition have been explored, highlighting the importance of architecture search in optimizing models without compromising accuracy [41]. Ontologies play a crucial role in enhancing MTL models by providing structured knowledge for more accurate predictions, as evidenced by their use in aspect-level sentiment analysis in the food domain [20]. Specific applications include MVANet, a multi-task guided multi-view attention network for Chinese food, demonstrating the effectiveness of attention mechanisms in handling diverse food images [42]. Optimizing parameter sharing in MTL models enhances efficiency and applicability in food classification tasks [43]. The integration of ontologies in MTL extends beyond recognition tasks to include sentiment analysis, enhancing interpretability and contextual relevance [44]. Combining deep learning with MTL and ontological knowledge has practical benefits in predicting food categories and nutrients [45], and supports various applications in dietary assessment and analysis [46]. These advancements underscore the transformative impact of MTL and ontologies in food recognition and analysis, offering robust solutions for food identification and nutritional analysis across various culinary traditions.

Ontology Advancements via LLMs have proven that LLMs are powerful tools for constructing various applications, including robust food ontologies. These tools allow for precise ontological data extraction, as proposed in [47] and knowledge graph construction, as explored in [48]. LLMs also enhance the accuracy of question-answering systems through the use of ontologies, as detailed in [49]. Additionally, [50] illustrates how LLMs can be used to create personalized food recommendation systems, leveraging ontologies that integrate appropriate semantic levels and accurate data. These applications demonstrate the potential of LLMs to efficiently and effectively build and manage food ontologies, facilitating the creation of more precise and useful information systems; however, none of these works explored how to leverage an end-to-end model by such ontologies.

3. Methodology

This section details the methodology used for the construction of the foods ontologies, as well as its subsequent application in the task of classifying food groups and specific dishes (see Fig. 1). The construction of this ontology was based on two semantic levels of food concepts: dish names (including examples such as "tacos", "tostadas", "Braised Pork" or "Noodles with Wonton") and food groups (such as "meat", "bread", "FishAndSeafood" or "VegetablesAndLegumes"). In the following subsections, each phase of the proposed method is described in detail. Each of these phases is essential to build and apply the food ontology.

3.1. Recipe collection by LLMs

The first stage involves compiling recipes related to a specific dish from various world cuisines, such as Mexican, Thai, or Chinese, using well-known LLMs, such as ChatGPT or Gemini, to identify the ingredients commonly used in their preparation. In the previous approach [21], web scraping techniques with Python's Beautiful Soup were used to retrieve these recipes from popular food sites like Yummly [51] and AllRecipes [52]. However, a challenge was encountered in that many recipes omitted the main ingredient; for example, some "biryani" recipes did not include "rice". Additionally, many dishes were not available on these popular websites, making it necessary

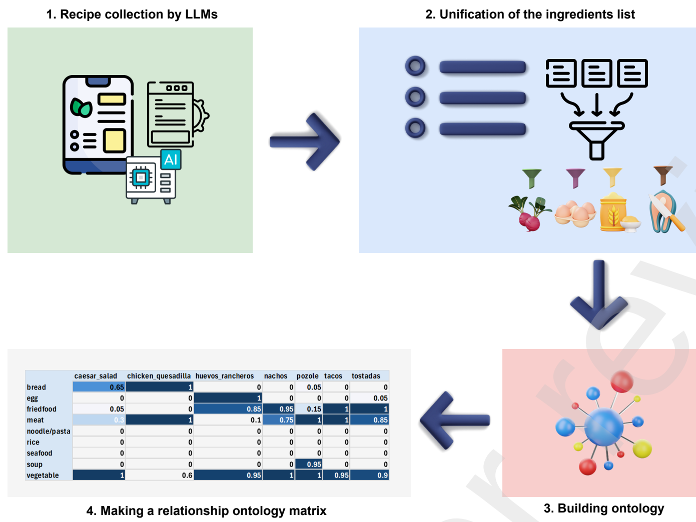

Figure 1: Diagram of the creation of the relationship ontology matrix.

to search different sites to cover all recipes in their entirety, which required constantly adjusting the code configurations for each website, thus consuming more time. Additionally, common web scraping challenges were encountered, such as IP blocking by target sites and the implementation of anti-scraping mechanisms that made data extraction difficult. Currently, ChatGPT is used as an LLM to obtain recipes, significantly reducing the time required to gather them, regardless of their country of origin, because ChatGPT searches several existing public food databases and, based on this information, provides the recipes. ChatGPT retrieves the ingredients of recipes ensuring the key components necessary to be analyzed. For example, given a traditional chicken quesadilla recipe, ChatGPT recovers the essential ingredients such as chicken, cheese, and tortillas.

In Fig. 2, the prompt used for the process of obtaining several chicken quesadilla recipes is illustrated, while Fig. 3 illustrates how the recipe from the website aligns with the one produced by ChatGPT. However, the link of the website provied by ChatGPT is not always accessible or directly connected to the recipe. Despite this, we have manually verified that the ingredients provided are consistent with the requested recipe.

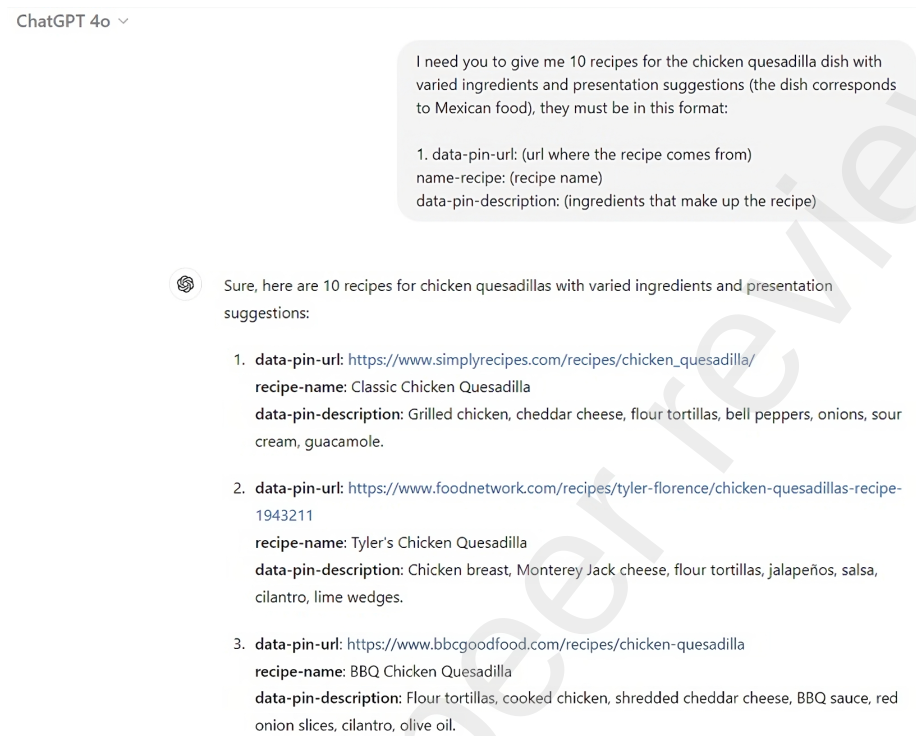

Figure 2: Prompt used to obtain recipes with ChatGPT.

3.2. *Unification of the ingredients list*

After associating each recipe with the corresponding dishes from the target dataset (see section 5.1), the next step involves extracting a list of ingredients using the information provided in the recipes. For this purpose, we employ natural language processing methods using the NLTK library, which includes the elimination of stopwords, such as “the”, “is”, and “in”, which add little semantic value, the removal of any text within parentheses, the exclusion of words containing numbers, the conversion of all words to lowercase to ensure consistency and avoid duplicates like “Pepper” and “pepper,” the lemmatization of words to transform plurals into their singular forms, such as converting “anchovies” to “anchovy”, the removal of non-alphabetic characters, the cleaning of unnecessary whitespace, and finally, the filtering out of empty entries that may result from these transformations.

Additionally, after the initial text processing, we apply a LP approach to further

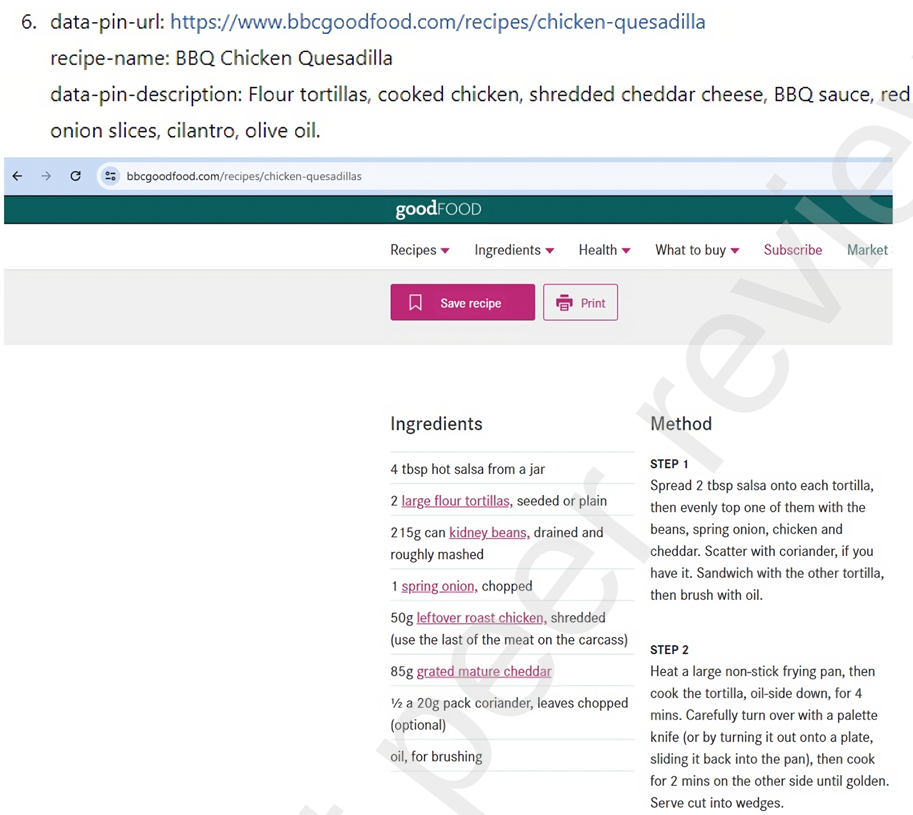

Figure 3: Alignment between the website recipe and the one obtained through ChatGPT.

refine the selection of relevant tokens from a list of ingredients. In this context, tokens refer to the individual words or terms extracted from a sentence, while sub-ingredients represent the key components or parts of an ingredient list that need to be identified. To efficiently select the most relevant tokens, we formalize this task using LP. The objective of this method is to select a subset of tokens x that represent the main sub-ingredients. The function to minimize is the cost associated with selecting these tokens, represented as $c^T x$, where c is a vector assigning penalties to each token. Primarily, tokens corresponding to verbs (VB) or adjectives (JJ), as well as those appearing early in the text, are penalized. These penalties can be based on factors such as token frequency, relevance in the context of the ingredients, or other metrics (e.g., whether the token is a common word or a key part of the ingredient). Nouns (NN), which typically represent key entities such as ingredients, are generally favored in the selection as they play a crucial role in identifying the sub-ingredients.

Consider that the matrix A is a binary matrix where each row represents an ingredient and each column a token, indicating whether the token is present in the ingredient, and the term Ax representing the number of tokens selected for each ingredient. Then, the token selection is subject to the following constraints:

- $-Ax \geq -2$: Up to two tokens are allowed to be selected as main sub-ingredients. This is useful in cases of complex ingredients where a single token is not sufficient. For example in compound ingredients such as "tomato paste", where both "paste" and "tomato" may be needed to fully represent the ingredient.

- $Ax \geq 1$: Each ingredient must have at least one token selected as its representative.

- $x \geq 0$: This constraint ensures that the decision variables x are non-negative, since they represent the presence or absence of a token in the selection.

The cleaning process (see Fig. 4) is used to normalize the ingredients obtained from both ChatGPT and the HELIS ontology [53]. HELIS organizes foods hierarchically into different levels of categories and subcategories, consisting of 16 food categories such as Cereals and Grain Products, and Meat, among others. This ontology provides a standardized basis for aligning ingredients across datasets. The cleaning process includes the removal of irrelevant ingredients, such as salt, pepper and vinegar, to ensure consistency and relevance. The LP method then selects key components from both sets of ingredients, which facilitates the matching process between the ingredients provided by ChatGPT and those in HELIS. By applying the same cleaning process to both data sources, subsequent matching to retrieve the food group related to each ingredient is streamlined, allowing for a more accurate and efficient alignment.

3.3. Building ontology

After processing the datasets and extracting the relevant information, the next step involves linking the recipes and their corresponding ingredients to each dish. This

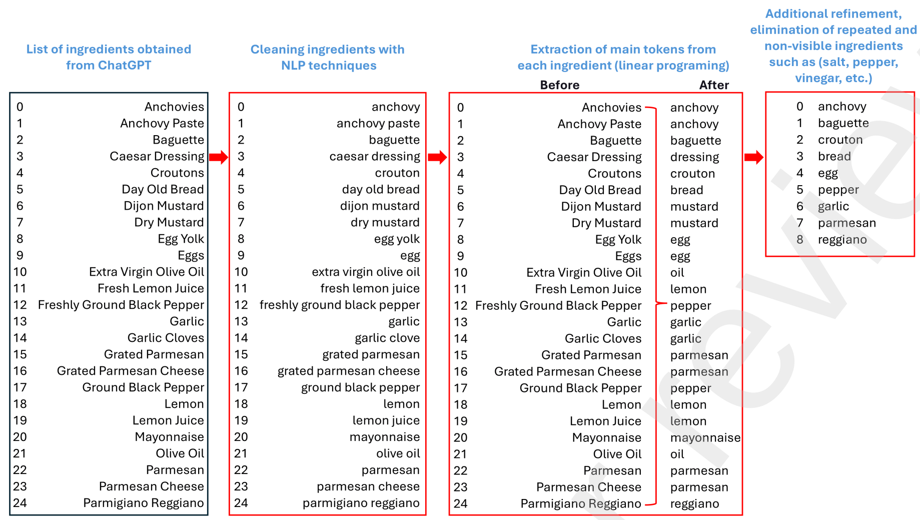

Figure 4: Cleaning and selection of tokens.

ensures that all components are properly associated with their respective dishes, laying the groundwork for constructing the ontology.

For the MAFood-121 dataset, the ingredients are grouped into high-level food concepts (food groups), such as “meats”, “vegetables”, “fruits”, “seafood”, “dairy products”, “spices”, among others. Similarly, for the VireoFood-172 dataset, the ingredients are grouped into broader food categories such as “CerealsAndGrainProducts”, “Eggs”, “FishAndSeafood”, “Meat”, “NutSeedAndOliveProducts”, “VegetablesAndLegumes”, among others, based on the HELIS ontology. These groupings are identified through SPARQL queries on the HELIS ontology (see Fig. 5). The GraphDB tool [54] is used as a visualization and query interface for the HELIS ontology, allowing for efficient extraction of the necessary data through SPARQL queries. It is important to note that these food groups may differ from those available in the target dataset. Therefore, an additional step is necessary to align the extracted food groups with the available annotations. For example, in the case of MAFood-121, food groups like “beans” were categorized as vegetables, and “crabs and fish” were classified as seafood to maintain consistency with the original annotations. In the case of VireoFood-172, ingredients like “Black rice”, “Bread”, and “Cold steamed rice noodles” were grouped under “Ce-

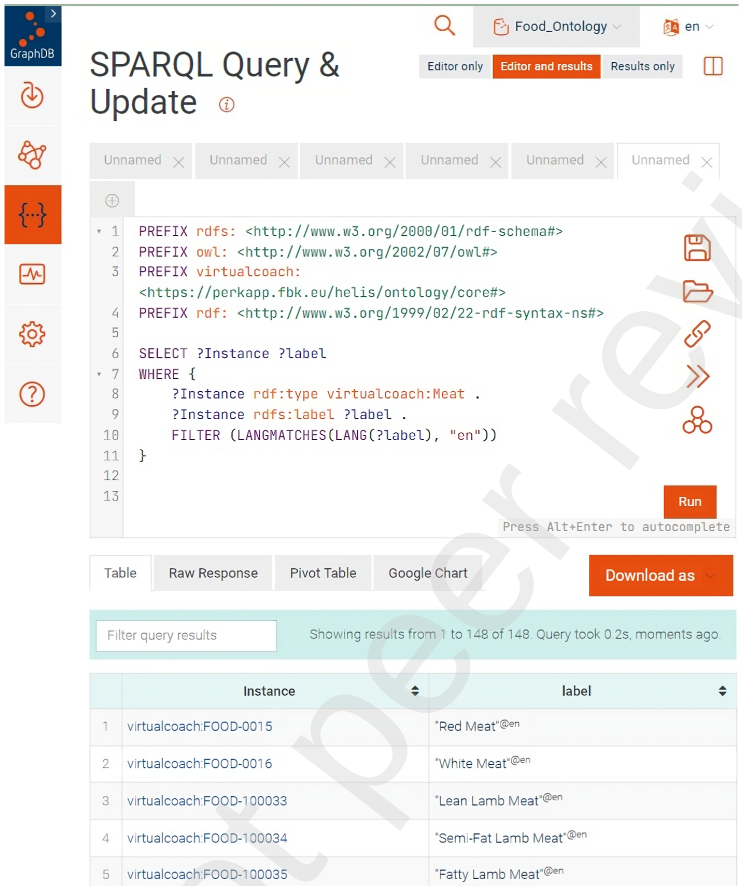

Figure 5: SPARQL queries.

realsAndGrainProducts", following the categories from the HELIS ontology.

3.4. Making a relationship ontology matrix

Once the ontology is constructed and the food groups are defined, the next step is to develop a relationship ontology matrix. This matrix, also referred to as a coexistence matrix, captures the interactions between food groups and their corresponding dishes. By mapping these relationships, we can analyze the frequency and patterns of ingredient usage across various dishes. It is important to highlight that this matrix is scalable to additional semantic levels, allowing for a more comprehensive and detailed representation of the ontology.

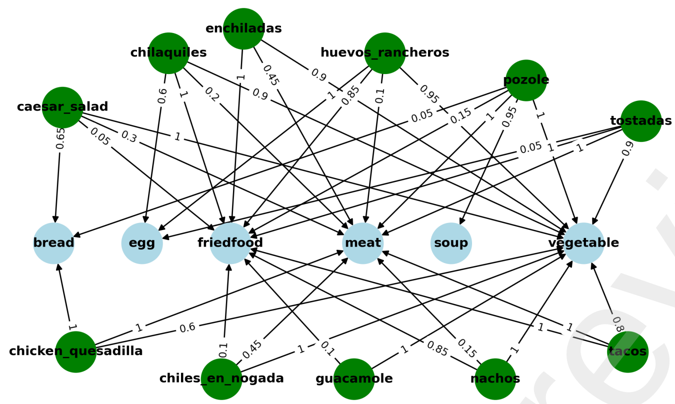

Figure 6: Ontological representation of Mexican cuisine: This diagram illustrates the relationships between various Mexican dishes and their corresponding food groups. The numerical values indicate the strength of the relation between each dish and food groups.

The primary goal of this matrix is to measure the incidence of food groups in the dishes, as illustrated in Fig. 6. Each cell in the matrix indicates the frequency with which a specific food group appears in a particular dish. This provides a quantitative analysis of the food composition of each dish.

The relationship among semantic levels is mirrored in the structure of the relationship matrix (RM) formally defined as follows:

$$\mathrm{RM} = \begin{bmatrix} \frac{\sum_{r=1}^{|R^1|} |FG_1 \in R_r^1|}{|R^1|} & \dots & \frac{\sum_{r=1}^{|R^D|} |FG_1 \in R_r^D|}{|R^D|} \\ \dots & \frac{\sum_{r=1}^{|R^d|} |FG_g \in R_r^d|}{|R^d|} & \dots \\ \frac{\sum_{r=1}^{|R^1|} |FG_G \in R_r^1|}{|R^1|} & \dots & \frac{\sum_{r=1}^{|R^D|} |FG_G \in R_r^D|}{|R^D|} \end{bmatrix}$$

where $|R^d|$ corresponds to the number of recipes linked to the d-th dish, R_r^d - the list of food groups for the r-th recipe linked to the d-th dish and FGg - the g-th food group.

Note that the columns denote the names of dishes, while the rows pertain to food groups. The values in the matrix cells denote the strength of the relationship between food groups and dishes. This not only reflects how often a food group appears in a dish, but also the diversity of food groups present within it. The matrix further aids in understanding the relationship between the semantic levels of food groups and their distribution across dishes.

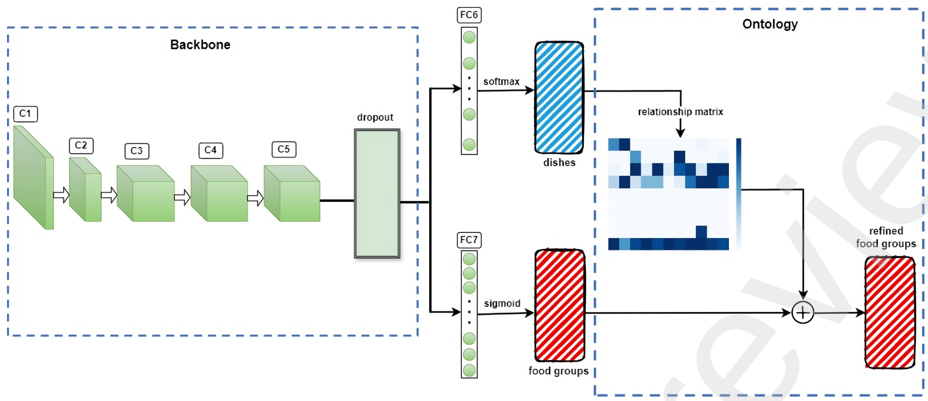

Figure 7: The framework of the proposed method.

4. Multi-Task Visual Food Recognition Supported with LLM

The proposed method, which uses LLM as a support tool for obtaining recipes that serve as a basis for building the ontology for multi-task food recognition, is illustrated in Fig. 7. In deep learning, a multi-task approach can range from a generic network, where all parameters are shared to extract features, to a specific network, where independent networks are used for each task [55]. For multi-task food recognition, a generic network is considered due to the similarity of the tasks involved (food recognition and food group recognition), both of which can benefit from general features extracted from the backbone (e.g., ResNet-50 [26] , DenseNet-121 [24] or EfficientNet-B4 [56]). Atop the backbone, a dropout layer is implemented to prevent overfitting, followed by a fully connected layer dedicated to each task.

For food recognition, a softmax activation is applied at the logits layer to provide a probability of the most likely dish. For food group recognition, similar to any multi-label task, a sigmoid activation is used at the logits layer to provide an independent probability for each group. Subsequently, the probability that each food concept belongs to each task is determined. A critical component of this network is the integration of a food ontology, specifically the RM, which serves as an additional layer reflecting the hierarchical and semantic relationships between food groups and dishes. The output from this layer is combined with the output from the food groups to provide a refined probability of the same. This ontology offers contextual information to the network, enabling a better understanding of the composition of dishes and more accurate classification. The interaction between the ontology and the network is facilitated by a custom layer, which weights the model's predictions based on the relationships established in the ontology.

The probability of the refined food groups is formally defined as follows:

$$p(y_g|\mathrm{W}, \mathrm{RM}) = \lambda \cdot p(y_g|\mathrm{W}) + (1 - \lambda) \cdot \sum_{d=1}^{D} p(y_d = d|\mathrm{W}) \cdot p(y_g|y_d = d)$$

$$p(y_g|\mathrm{W}) = \frac{1}{1 + e^{f_g^W(x)}}, \quad p(y_d|\mathrm{W}) = \frac{e^{f_d^W(x)_d}}{\sum_{k=1}^{K} e^{f_d^W(x)_k}}, \quad p(y_g|y_d) = \mathrm{RM}[g, d],$$

where $p(y_g|\mathbf{W})$ represents the conditional probability that a specific ingredient y_g is present in the image; $\mathbf{W}$ - the model weights; $f_g^W(x)$ - the logits outputs for the food groups; $p(y_d|\mathbf{W})$ - the probability that a specific dish y_d is the correct class; $f_d^W(x)_d$ - the d-th logits output for the dishes; $p(y_g|y_d)$ - the probability, extracted from the RM of having a food group y_g given the dish y_d; K and D - the number of dishes; x - the input image; and λ - a hyperparameter to weight the contribution of both terms.

During the training of the model, we use two loss functions, each given equal weight: Cross-Entropy Loss (CELoss) for the dish recognition task and Binary Cross-Entropy Loss (BCELoss) for categorizing food groups. The equation representing it for a single input image is expressed as follows:

$$\mathrm{CELoss} = -\sum_{d=1}^{D} \hat{y}_d \cdot log(p(y_d|W)),$$

where $\hat{y}_d$ is the Ground Truth (GT) label in one-hot encoding and $p(y_d|W)$ is the probability given by the model for the d-th dish. The CEloss is calculated individually for each image and then averaged.

Regarding the BCELoss, we formally define it as follows:

$$\mathrm{BCELoss} = \sum_{g=1}^{G} \hat{y}_g \cdot log(p(y_g|W, \mathrm{RM})) + \sum_{g=1}^{G} (1 - \hat{y}_g) \cdot log(1 - p(y_g|W, \mathrm{RM})),$$

where G represents the number of food groups; $\hat{y}_g$ is the GT in one-hot encoding for g-th food group, indicating whether food group g is present or not in the sample; and

$p(y_g|W, \mathrm{RM})$ is the refined probability given for g-th food group. The BCEloss is calculated individually for each image and then averaged over the total images and food groups.

It should be highlighted that while the proposed method directly enhances the accuracy of food group predictions, the use of dish probabilities in conjunction with the RM also indirectly improves the predictions for the dishes themselves.

5. Validation

This section outlines the two datasets employed, the experimental setup, and the various evaluation metrics utilized.

5.1. Food Datasets

Two datasets specifically for culinary analysis are introduced. These datasets are designed to support multiple tasks, allowing for a broad range of applications in food-related studies and innovations.

5.1.1. MAFood-121

The MAFood-121 dataset, detailed in [22], consists of 21,175 images representing traditional dishes from eleven of the world's most renowned cuisines, showcasing the culinary diversity and richness of these regions. The dataset is organized into ten distinct food groups ("bread", "egg", "friedfoods", "meat", "pasta/noodles", "rice", "seafood", "soups", "dumplings", and "vegetables") and includes a total of 121 different dishes. Each image is carefully annotated, providing a detailed and context-rich resource for global culinary analysis.

This dataset supports a classification approach that covers both well-known dishes, such as "tacos" and "pad thai", as well as lesser-known but equally important dishes, ensuring a diverse examination of global cuisines. The dataset is divided into 72.5% for training, 12.5% for validation, and 15% for testing, maintaining the original proportions and allowing for a robust evaluation of the classification models used. Furthermore, it emphasizes the relationship between food groups and dishes, providing a valuable resource for MTL approaches, where food recognition extends beyond individual dishes to their broader categories.

5.1.2. VireoFood-172

The Vireo Food-172 [23] dataset contains 110,241 images categorizing 172 Chinese food dishes, including 353 distinct ingredient labels. The dataset is split into 60% for training, 10% for validation, and the remaining 30% for testing. While this dataset does not include predefined food groups, the ingredients, such as broccoli, cauliflower, and bread, are categorized into food groups using the HELIS ontology. For example, ingredients like broccoli and cauliflower are classified under Vegetables and legumes, while bread is classified under Cereals and grains. In total, the original list of 353 ingredients was reduced to 18 food groups. Its diversity and complexity make it particularly suitable for computational tasks like ingredient recognition and MTL for recipe retrieval.

This structured categorization facilitates more detailed studies into the relationships between different dishes and their ingredients, contributing significantly to advancements in ingredient recognition algorithms and food categorization. The methodical approach to training, validation, and testing ensures that the dataset's full potential is utilized, enhancing the accuracy and effectiveness of the applied computational models.

5.2. Experimental setup

To implement the proposed method, three primary architectures were chosen: ResNet-50, DenseNet-121, and EfficientNet-B4. These networks are also used for comparison purposes. Initially, pre-trained on ImageNet, both the reference and proposed methods are retrained for a period of 20 epochs using empirically selected hyperparameters. A learning rate (LR) of 0.001 is set, and the batch size is fixed at 64 to balance computational efficiency and training stability. For regularization, a Dropout layer with a rate of 0.1 is added after the last convolutional layer in all three methods to prevent overfitting. This inclusion effectively regulates and enhances each of the model's ability to generalize to unseen data. The Adam optimizer is used to minimize the loss function, and simple data pre-processing includes resizing images to 224×224

pixels and normalizing them with a mean and standard deviation of 0.5. Additionally, the number of recipes per dish is set at $|R^d| = 20$, noting that increasing this number leads to the retrieval of recipes that do not accurately represent the consulted dish. Lastly, λ is adjusted from 0.5 to 0.9 in increments of 0.1.

5.3. Evaluation metrics

This section defines the metrics used for food recognition, food group recognition, and evaluation of the multi-tasking problem as a whole.

5.3.1. Food Recognition

Food recognition is a single-label problem that aims to classify the overall context of images with the most likely food. For single-label tasks, the Accuracy (A) is the most popular metric used. This metric is defined as the proportion of correct predictions concerning the total number of images:

$$\mathrm{A} = \frac{1}{N}\sum_{n=1}^{N} |y_i \cap \hat{y}_i|,$$

where N is the total number of test examples, $\hat{y}_i$ denotes a set with the GT label and y_i a set with the label predicted by the model for the i-th example. Cardinality is denoted by the operator $|.|$. It is important to note that, in a single-label classification problem, unlike multi-label ones, the sets have only one value.

5.3.2. Food Groups Recognition

Food group recognition is a multi-label classification problem that involves categorizing food images with food groups representing each ingredient contained in the food. For multi-label classification, it is essential to define metrics to evaluate how the model handles simultaneous prediction of multiple labels. Four key metrics stand out in this context: Precision (P), Recall (R), F1 Score (F_1) and Jaccard Index (JI).

P calculates the proportion of correctly predicted instances relative to the total instances predicted as that label. The formal definition is as follows:

$$\mathrm{P} = \frac{\sum_{i=1}^{N} |y_i \cap \hat{y}_i|}{\sum_{i=1}^{N} |y_i|}$$

R is crucial to evaluate the ability of the model to capture all GT labels. For each label, R calculates the proportion of correctly predicted instances relative to actual instances of that label. It measures how effectively the model recovers true labels from a wide range of possibilities. The formal definition is as follows:

$$\mathrm{R} = \frac{\sum_{i=1}^{N} |y_i \cap \hat{y}_i|}{\sum_{i=1}^{N} |\hat{y}_i|}$$

F_1 combines P and R using a harmonic mean into a single metric. This balance is especially important when the model must predict multiple labels in the context of food classification. It considers both false positives and false negatives, providing a comprehensive assessment of model performance. The metric is defined as follows:

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Finally, the JI , also known as Intersection over Union (IoU), measures the overlap between the true labels $\hat{y}$ and the predicted labels y. In multi-label food recognition, the JI evaluates how well the predicted labels align with the actual labels and quantifies the degree of similarity in the set of labels. It takes into account the similarities and differences between the predicted and true labels, providing information on how well the model predicts the multi-label nature of food images. The metric is defined as follows:

$$\mathrm{JI} = \frac{\sum_{i=1}^{N} |y_i \cap \hat{y}_i|}{\sum_{i=1}^{N} |y_i \cup \hat{y}_i|}$$

5.3.3. Multi-task Food Recognition

In the process of evaluating the model, a key metric known as MTA [22] was applied to quantify the consistency in the predictions of the model in relation to multiple classification tasks. The MTA is defined by the following expression:

$$\mathrm{MTA} = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \frac{|y_i^t \cap \hat{y}_i^t|}{|y_i^t \cup \hat{y}_i^t|}$$

where T denotes the number of classification tasks being evaluated, corresponding to various labels related to ingredients and food dishes. The term $|y_i^t \cap \hat{y}_i^t|$ reflects the size of the intersection between the model's predictions and the GT labels for task t in sample i, while $|y_i^t \cup \hat{y}_i^t|$ relates to the size of the union between the model's predictions and the true labels for task t in sample i.

The MTA metric becomes essential, as it allows for evaluating the performance of the model on multiple classification tasks in a consistent manner. This consistency is crucial to ensure that the model predictions are accurate and relevant, especially in a context where the classification of food groups and dishes coexist. The use of the MTA metric in the model responds to the need to evaluate its ability to coherently classify both food groups and dishes, a central aspect of the research.

6. Results

To evaluate the performance of multi-task classification methods applied to the challenge of identifying dishes and their food groups from images, we consider the following models: ResNet-50, DenseNet-121, and EfficientNet-B4, together with their improved versions using ontologies built from recipes provided by LLMO-MFR. Two distinct ontologies were used: one for the MAFood-121 dataset and another for the VireoFood-172 dataset. These evaluations utilized both datasets. For the variants LLMO-MFR-ResNet-50, LLMO-MFR-DenseNet-121, and LLMO-MFR-EfficientNet-B4, the λ parameter was determined after analyzing the results in the validation set. Five experiments were conducted with λ values ranging from 0.5 to 0.9. The results obtained for the training and validation sets are presented in Fig. 8 and Fig. 9.

Fig. 8 shows that in the MAFood-121 dataset, both training and validation, a low λ tends to offer more consistent and superior performance than a high λ. For example, in the training set, a λ value of 0.6 showed superior performance in terms of key metrics (F_1, JI, A, and MTA) compared to the baseline for models ResNet-50, DenseNet-121, and EfficientNet-B4, which correspond to a regularization parameter λ of 1.0. Although λ values of 0.8 and 0.9 also demonstrated good performance, it was the λ of 0.6 that consistently outperformed the baseline in important metrics for single-label, multi-label, and multi-task tasks across all evaluated models. In the validation set, a

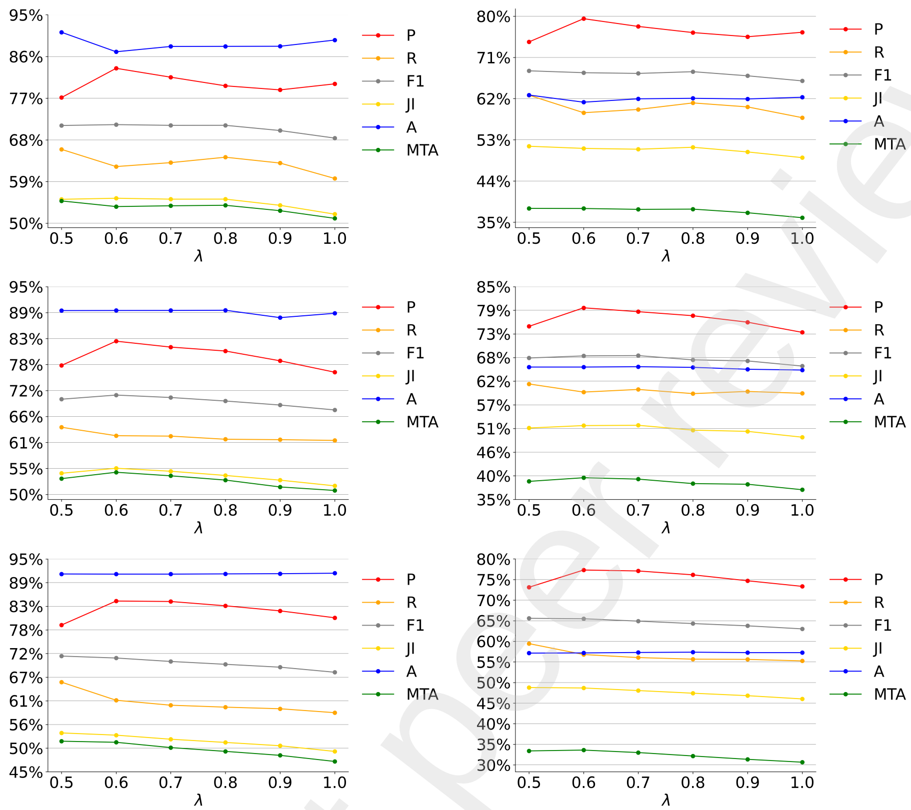

Figure 8: The performance of LLMO-MFR models, including ResNet-50 (first row), DenseNet-121 (second row), and EfficientNet-B1 (third row), was evaluated on the MAFood-121 dataset on both the training (first column) and validation (second column) sets, using a range of λ values from 0.5 to 0.9.

λ value of 0.5 showed superior performance in the metrics of (F_1), JI, A, and MTA compared to the baseline for ResNet-50, effectively covering the tasks of single-label, multi-label, and multi-task. For DenseNet-121 and EfficientNet-B4, a λ of 0.6 stood out by providing superior performance in key metrics (F1, JI, A, and MTA) compared to the baseline, surpassing it in all tasks.

Regarding VireoFood-172 (see Fig. 9), for the three models evaluated in the training set (ResNet-50, DenseNet-121, and EfficientNet-B4), a λ value of 0.6 provides superior performance in key metrics ($\mathbf{F}_1$, JI, A, and MTA) compared to the baseline. Although v values of 0.7 and 0.8 also show good performance, it is the v of 0.6 that

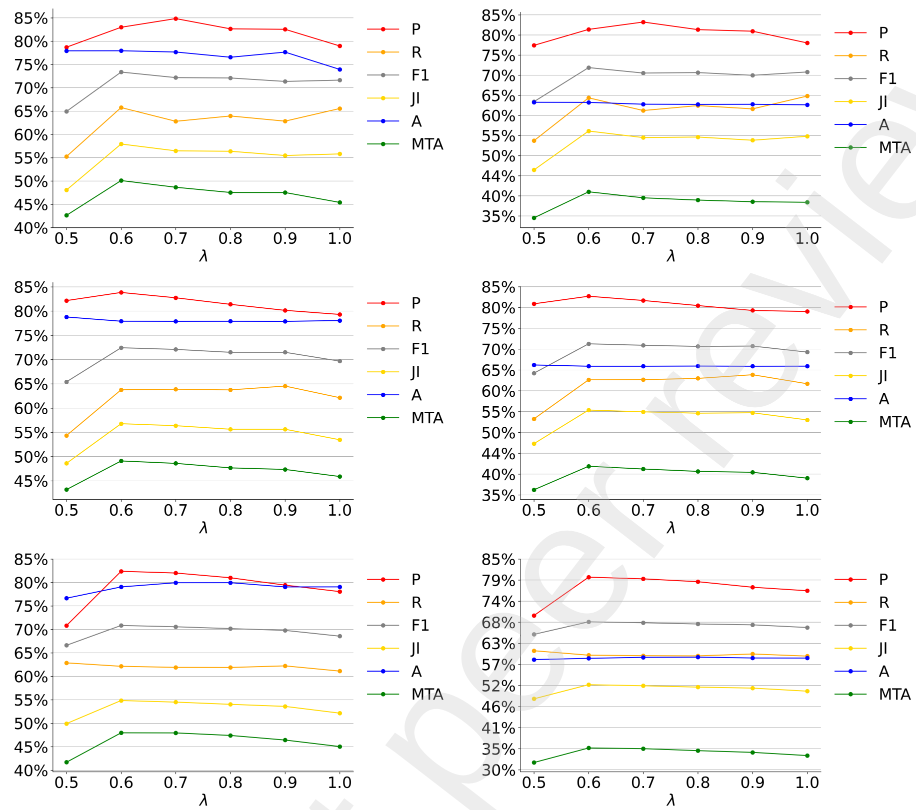

Figure 9: The performance of LLMO-MFR models, including ResNet-50 (first row), DenseNet-121 (second row), and EfficientNet-B1 (third row), was evaluated on the VireoFood-172 dataset across both the training (first column) and validation (second column) sets, using a range of λ values from 0.5 to 0.9.

consistently surpasses the baseline in important metrics for the three tasks (single label, multi-label, and multi-task). In the validation set, a λ value of 0.6 also provides superior performance in the metrics of $\mathbf{F}_1$, JI, A, and MTA for the three models evaluated compared to the baseline, effectively covering the all tasks.

Therefore, a λ of 0.6 was used to evaluate the model performance in the test set. This analysis is based on the consistency of the performance of λ of 0.6, which surpassed the baseline in all important metrics for single-label, multi-label and multi-task tasks.

Method	λ	P	R	F_1	JI	A	MTA
ResNet-50	-	0.8313	**0.8478**	0.8395	0.7234	0.7125	0.5539
OD-ResNet-50	0.8	0.8462	0.8371	0.8416	0.7266	0.7220	**0.5777**
LLMO-MFR-ResNet-50	0.8	**0.8473**	0.8447	**0.8460**	**0.7332**	**0.7252**	0.5763

Table 1: Performance comparison on Mexican Food datasets using Webscraping and LLM methods.

6.1. *Comparative Performance Based of Web Scraping and LLM Methods for obtaining recipe datasets*

Comparative analysis of the results in Table 1 shows a clear advantage of the LLMO-MFR-ResNet-50 method over the previous web scraping-based approach (OD-ResNet-50) and the baseline ResNet-50 on Mexican food datasets. In addition to the performance improvements, a crucial factor to consider is the ontology-building time. With the web scraping method, collecting data (recipes) for a single dish class can take approximately one week, since the recipes are scattered across multiple food websites. In contrast, leveraging LLMs reduces the building time significantly to just one or two days per dish. If a large number of dishes are involved, such as MAFood-121 or VireoFood-172, web scraping could extend the process over many months, whereas the LLM-based approach could complete it in just two to three months. Analyzing the metrics in the table, the LLMO-MFR-ResNet-50 method outperforms both the baseline and the previous approach in nearly all aspects. Using the same regularization parameter, λ of 0.8, LLMO-MFR-ResNet-50 achieves the highest values in P 0.8473, F1 0.8460, JI 0.7332, and A 0.7252, surpassing both OD-ResNet-50 and the baseline ResNet-50 in these key performance metrics. Although the OD-ResNet-50 method slightly outperforms the new approach in MTA, with 0.5777 compared to 0.5763 for LLMO-MFR-ResNet-50, this difference is minimal. Moreover, the overall performance improvements, particularly in JI, F1, and A, highlight the effectiveness of the LLMO-MFR method. This parity in MTA, combined with superior results in other critical metrics, strengthens the case for the LLM-based approach. Notably, these improvements were achieved using the same regularization parameter configuration, λ of 0.8, as employed by the baseline model, emphasizing the efficiency of the new method without requiring changes to the regularization setup. Additionally, process optimization not only accel-

Method	λ	P	R	F_1	JI	A	MTA
ResNet-50 - MAFood-121							
ResNet-50	-	0.7631	0.5722	0.6540	0.4859	**0.6396**	0.3662
LLMO-MFR-ResNet-50	0.6	**0.7897**	**0.5911**	**0.6761**	**0.5107**	0.6342	**0.3986**
DenseNet-121 - MAFood-121							
DenseNet-121	-	0.7407	**0.6002**	0.6631	0.4960	0.6604	0.3776
LLMO-MFR-DenseNet-121	0.6	**0.7923**	0.5971	**0.6810**	**0.5163**	**0.6670**	**0.4099**
EfficientNet-B4 - MAFood-121							
EfficientNet-B4	-	0.7378	0.5500	0.6302	0.4601	0.5855	0.3218
LLMO-MFR-EfficientNet-B4	0.6	**0.7679**	**0.5615**	**0.6487**	**0.4800**	**0.5864**	**0.3507**

Table 2: Comparative performance of the LLMO-MFR method using ResNet-50, DenseNet-121, and EfficientNet-B4 as backbones on the MAFood-121 dataset, with regularization parameter λ of 0.6.

erates ontology construction but also enables faster iterations and better adaptability as project needs evolve.

6.2. Comparative Performance Analysis of ResNet, DenseNet, and EfficientNet on the MAFood-121 Dataset

Analysis of the results obtained in Table 2 reveals key findings when comparing the baseline models (ResNet-50, DenseNet-121, and EfficientNet-B4) with their improved versions using the LLMO-MFR method with a regularization parameter λ of 0.6 on the MAFood-121 dataset. For ResNet-50, the LLMO-MFR method improves the performance metrics considerably. The P increases from 0.7631 to 0.7897, while the R rises from 0.5722 to 0.5911, suggesting a better balance between A and sensitivity. The F_1 improves from 0.6540 to 0.6761, and the JI increases from 0.4859 to 0.5107. Although the A decreases slightly from 0.6396 to 0.6342, the increase in MTA from 0.3662 to 0.3986 suggests better multi-task handling capability. On DenseNet-121, the P increases from 0.7407 to 0.7923, which is the largest improvement among the evaluated models. Although the (R) decreases from 0.6002 to 0.5971, the F_1 and JI show significant improvements, increasing from 0.6631 to 0.6810 and from 0.4960 to 0.5163, respectively. The A also improves slightly, from 0.6604 to 0.6670, and the MTA rises from 0.3776 to 0.4099, reaffirming the model's ability to handle multiple

Method	λ	P	R	F_1	JI	A	MTA
ResNet-50 - VireoFood-172							
ResNet-50	-	0.7796	**0.6475**	0.7075	0.5473	0.6238	0.3840
LLMO-MFR-ResNet-50	0.6	**0.8112**	0.6412	**0.7162**	**0.5579**	**0.6268**	**0.4040**
DenseNet-121 - VireoFood-172							
DenseNet-121	-	0.7864	0.6150	0.6902	0.5270	**0.6627**	0.3916
LLMO-MFR-DenseNet-121	0.6	**0.8244**	**0.6244**	**0.7106**	**0.5511**	0.6567	**0.4152**
EfficientNet-B4 - VireoFood-172							
EfficientNet-B4	-	0.7657	0.5977	0.6713	0.5052	0.5960	0.3426
LLMO-MFR-EfficientNet-B4	0.6	**0.8016**	**0.6009**	**0.6869**	**0.5231**	**0.5963**	**0.3636**

Table 3: Comparative performance of the LLMO-MFR method using ResNet-50, DenseNet-121, and EfficientNet-B4 as backbones on the VireoFood-172 dataset, with a regularization parameter λ of 0.6.

tasks. On EfficientNet-B4, the P and R improve moderately, from 0.7378 to 0.7679 and from 0.5500 to 0.5615, respectively. The F_1 and JI also improve, although less sharply, while the A barely rises, from 0.5855 to 0.5864. However, the MTA increases from 0.3218 to 0.3507, highlighting a modest improvement. In summary, the improved version of DenseNet-121 using LLMO-MFR is the most significant in terms of overall improvements in P, F_1, JI and MTA, standing out as the most balanced and effective model among the compared versions.

6.3. Comparative Performance Analysis of ResNet, DenseNet, and EfficientNet on the VireoFood-172 Dataset

The analysis of the results in Table 3 shows the impact of the LLMO-MFR method on the ResNet-50, DenseNet-121 and EfficientNet-B4 models, evaluated on the VireoFood-172 dataset, comparing their baseline versions with the improved ones using a regularization parameter λ of 0.6. For ResNet-50, the improvements are notable, with an increase in P from 0.7796 to 0.8112, reflecting a better ability of the model to make correct predictions.

Although the R decreases slightly from 0.6475 to 0.6412, the F_1 rises from 0.7075 to 0.7162, indicating a better balance between A and sensitivity. The JI also improves from 0.5473 to 0.5579, suggesting improved performance on class segmentation. The

A improves from 0.6238 to 0.6268, and the MTA experiences a significant increase from 0.3840 to 0.4040, reinforcing its generalization capability across multiple tasks. For DenseNet-121, the improvements are also considerable, with an increase in the P from 0.7864 to 0.8244, along with an improvement in the R from 0.6150 to 0.6244. The F_1 rises from 0.6902 to 0.7106, and the JI from 0.5270 to 0.5511, reinforcing its ability to handle class segmentation well. Although A decreases slightly from 0.6627 to 0.6567, MTA increases from 0.3916 to 0.4152, highlighting its ability to handle multiple tasks. In EfficientNet-B4, the improvements are more moderate. P increases from 0.7657 to 0.8016, while R improves marginally from 0.5977 to 0.6009. F_1 and JI also improve slightly, although the increase in A is minimal. MTA, however, increases from 0.3426 to 0.3636, showing a relative improvement. Overall, the improved version of DenseNet-121 with the LLMO-MFR method is the most significant in terms of overall improvement in P, R, F_1 and JI, suggesting that it is the model most benefited by the $\lambda = 0.6$ regularization on this dataset. In particular, it excels in complex tasks, represented by the MTA, and overall, it consolidates as the model with the most significant improvements, making it the most favored by the LLMO-MFR method.

6.4. *Qualitative Comparison of LLMO-MFR-DenseNet-121 and DenseNet-121 on the MAFood-121 and VireoFood Datasets*

Qualitative results, presented in Fig. 10, are shown for the LLMO-MFR-DenseNet-121 and DenseNet-121 (without ontology) models, using the MAFood-121 and VireoFood datasets, each accompanied by a specific ontology generated with the help of an LLM. A threshold of 0.5 was applied for classifications, so values below this threshold are interpreted as unclassified by the model. In both datasets, the LLMO-MFR-DenseNet-121 model shows significant improvements in the classification of dishes and food groups. For example, in MAFood-121, the dish “kebap” achieved a confidence score of 0.57 compared to 0.37 for the baseline model, and in VireoFood, “Pork Ribs and White Gourd Soup” reached 0.85 versus 0.80.

In terms of food groups, the proposed model outperforms the baseline in most categories, such as “vegetable” in MAFood-121 (0.93 vs 0.89) and “Eggs” in VireoFood (0.73 vs 0.44). These results demonstrate the positive influence of the ontology and

Figure 10: Qualitative results showcasing model performance on both the MAFood-121 and VireoFood-172 datasets.

the optimal weighting of 0.6 in the classification process, improving the A in identifying relevant food groups. However, some failure cases were identified, such as "prime rib" in MAFood-121 and "Noodles with Wonton" in VireoFood, where the baseline outperformed the proposed model in some food group predictions. This suggests that, while the ontology provides valuable prior knowledge, the model must be particularly precise in applying this knowledge to avoid missing correct predictions. In these cases, it is crucial that the model balances the additional information provided by the ontology to maximize its effectiveness, avoiding potential interferences that could negatively impact certain predictions.

In summary, the integration of a data-rich ontology, supported by LLMs in its construction, into the LLMO-MFR-DenseNet-121 model provides substantial improvements in the classification of dishes and food groups across both datasets. The increase in available data within the ontology enhances the model's ability to recognize and categorize a broader range of dishes and ingredients, leading to more accurate and consistent results. This highlights the relevance of using ontologies with extensive data for future applications in computer vision and food classification.

7. Conclusions

This study has demonstrated how the incorporation of LLM-supported ontologies significantly improves accuracy in food image classification. By supporting the process of ontology construction, LLMs enable the extraction of more representative recipes and accelerate the grouping of ingredients into food categories. This results in a faster and more efficient ontology creation process, which is crucial when working with large-scale datasets. Furthermore, LLMs allow the retrieval of recipes from a wide variety of global cuisines, ensuring that the ontology is both comprehensive and adaptable to diverse culinary contexts.

Moreover, the use of ontologies in conjunction with image data has significantly enriched the model's learning process. This integration allows the model to extract and process information more precisely and efficiently, leading to improved accuracy in food classification. The ability to adjust parameters, such as different lambdas, further enhances the robustness of the model. LLMs, by dynamically generating relevant and contextual content, offer a scalable solution that surpasses traditional methods of data compilation. This adaptability paves the way for future research where the system can continuously and automatically evaluate new dishes and regional variations.

For future research, the integration of nutritional data and allergen detection through ontologies remains a promising area, which could not only improve accuracy in food classification but also have significant implications for food safety and diet personalization. Furthermore, expanding this approach to different cultures and regions of the world could provide valuable insights into global culinary diversity and its impact on food recognition technologies. Additionally, the exploration of Graph Neural Networks GNN to further enhance the understanding of complex relationships in food ontologies promises deeper insights and even more accurate classifications.

Therefore, the support of LLMs in ontology creation and their integration into deep learning models represents a significant advancement in the field of food image analysis. This approach opens new avenues for research and development in the realm of artificial intelligence applied to nutrition and global gastronomy.

Acknowledgements

This work has been partially supported by the Spanish project PID2022-136436NB-I00 (AEI-MICINN), Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), Beatriu de Pinós Programme and the Ministry of Research and Universities of the Government of Catalonia (2022 BP 00257), and Agencia Nacional de Investigación y Desarrollo de Chile (ANID) (Grant No. FONDECYT INICIACIÓN 11230262). D. Ponte acknowledges the support of Secretaría Nacional de Ciencia, Tecnología e Innovación Senacyt Panamá (Scholarship No. 270-2022-125).

References

[1] G. Chhabra, S. Kumar, S. Gupta, and P. Nagpal, “Food recognition and nutrition estimation using deep learning,” in *Artificial Intelligence to Analyze Psychophysical and Human Lifestyle*. Springer, 2023, pp. 77–91.

[2] M. Mansouri, S. Benabdellah Chaouni, S. Jai Andaloussi, and O. Ouchetto, “Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: A systematic review,” *SN Computer Science*, vol. 4, no. 5, p. 513, 2023.

[3] L. Boyd, N. Nnamoko, and R. Lopes, “Fine-grained food image recognition: A study on optimising convolutional neural networks for improved performance,” *Journal of Imaging*, vol. 10, no. 6, p. 126, 2024.

[4] K. B. Chhetri, “Applications of artificial intelligence and machine learning in food quality control and safety assessment,” *Food Engineering Reviews*, vol. 16, no. 1, pp. 1–21, 2024.

[5] N. Nath, R. P. Salunke, and W. Mishra, “Smart refrigerator system: Enabling healthy cooking,” in *International Conference on Research into Design*. Springer, 2023, pp. 1081–1091.

[6] M. Mishra, T. Sarkar, T. Choudhury, N. Bansal, S. Smaoui, M. Rebezov, M. A. Shariati, and J. M. Lorenzo, “Allergen30: detecting food items with possible allergens using deep learning-based computer vision,” *Food Analytical Methods*, vol. 15, no. 11, pp. 3045–3078, 2022.

[7] J. Ojha, Shikha, A. Nayak, V. Goyal, K. Sharma, and S. Reddy, “An analysis of smart kitchen gadgets and a smartphone application for the smart cooker,” in *International Conference on Smart Trends for Information Technology and Computer Communications*. Springer, 2023, pp. 823–837.

[8] G. A. Tahir and C. K. Loo, “A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment,” in *Healthcare*, vol. 9, no. 12. MDPI, 2021, p. 1676.

[9] Z. Wang, W. Min, Z. Li, L. Kang, X. Wei, X. Wei, and S. Jiang, “Ingredient-guided region discovery and relationship modeling for food category-ingredient prediction,” *IEEE Transactions on Image Processing*, vol. 31, pp. 5214–5226, 2022.

[10] T. Fredriksson, D. I. Mattos, J. Bosch, and H. H. Olsson, "Data labeling: An empirical investigation into industrial challenges and mitigation strategies," in *International Conference on Product-Focused Software Process Improvement*. Springer, 2020, pp. 202–216.

[11] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "A study of multi-task and region-wise deep learning for food ingredient recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 1514–1526, 2020.

[12] H. Zhao, K.-H. Yap, and A. C. Kot, "Fusion learning using semantics and graph convolutional network for visual food recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1711–1720.

[13] A. A. Barbhuiya, R. K. Karsh, and R. Jain, "Cnn based feature extraction and classification for sign language," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 3051–3069, 2021.

[14] Y. Liu, H. Pu, and D.-W. Sun, "Efficient extraction of deep image features using convolutional neural network (cnn) for applications in detecting and analysing complex food matrices," *Trends in Food Science & Technology*, vol. 113, pp. 193–204, 2021.

[15] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, 2023.

[16] P. Castellano-Escuder, R. González-Domínguez, D. S. Wishart, C. Andrés-Lacueva, and A. Sánchez-Pla, "Fobi: an ontology to represent food intake data and associate it with metabolomic data," *Database*, vol. 2020, p. baaa033, 2020.

[17] J. M. Rodríguez-de Vera, P. Villacorta, I. G. Estepa, M. Bolaños, I. Sarasúa, B. Nagarajan, and P. Radeva, "Dining on details: Llm-guided expert networks for fine-grained food recognition," in *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, 2023, pp. 43–52.

[18] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] M. Belguith, C. Aloulou, and B. Gargouri, "Aspect level sentiment analysis based on deep learning and ontologies," *SN Computer Science*, vol. 5, no. 1, p. 58, 2023.

[21] D. Ponte, E. Aguilar, M. Ribera, and P. Radeva, "Ontology-driven deep learning model for multitask visual food analysis," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2024, Volume 2: VISAPP, Rome, Italy, February 27-29, 2024*. SCITEPRESS, 2024, pp. 624–631.

[22] E. Aguilar, M. Bolaños, and P. Radeva, "Regularized uncertainty-based multi-task learning model for food analysis," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 360–370, 2019.

[23] C.-w. N. Jing-jing Chen, "Deep-based ingredient recognition for cooking recipe retrival," *ACM Multimedia*, 2016.

[24] S. Nandhini and K. Ashokkumar, "An automatic plant leaf disease identification using densenet-121 architecture with a mutation-based henry gas solubility optimization algorithm," *Neural Computing and Applications*, vol. 34, no. 7, pp. 5513–5534, 2022.

[25] C. Li, Z. Qiao, K. Wang, and J. Hongxing, "Improved efficientnet-b4 for melanoma detection," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. IEEE, 2021, pp. 127–130.

[26] L. Zhang, Y. Bian, P. Jiang, and F. Zhang, "A transfer residual neural network based on resnet-50 for detection of steel surface defects," *Applied Sciences*, vol. 13, no. 9, p. 5260, 2023.

[27] H. Shao, J. Mu, R. Tang, X. Chen, and M. Liu, "Research on automatic dish recognition algorithm based on deep learning," in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE, 2020, pp. 566–570.

[28] R. D. Kumar, E. G. Julie, Y. H. Robinson, S. Vimal, and S. Seo, "Recognition of food type and calorie estimation using neural network," *The Journal of Supercomputing*, pp. 1–22, 2021.

[29] S. Sathish, S. Ashwin, M. A. Quadir, and L. Pavithra, “Analysis of convolutional neural networks on indian food detection and estimation of calories,” *Materials Today: Proceedings*, vol. 62, pp. 4665–4670, 2022.

[30] Y.-C. Liu, D. D. Onthoni, S. Mohapatra, D. Irianti, and P. K. Sahoo, “Deep-learning-assisted multi-dish food recognition application for dietary intake reporting,” *Electronics*, vol. 11, no. 10, p. 1626, 2022.

[31] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, P. Mursanto, and W. Jatmiko, “Traditional food knowledge of indonesia: a new high-quality food dataset and automatic recognition system,” *Journal of Big Data*, vol. 7, pp. 1–19, 2020.

[32] S. J. Minija and W. S. Emmanuel, “Food recognition using neural network classifier and multiple hypotheses image segmentation,” *The Imaging Science Journal*, vol. 68, no. 2, pp. 100–113, 2020.

[33] Z. Shen, A. Shehzad, S. Chen, H. Sun, and J. Liu, “Machine learning based approach on food recognition and nutrition estimation,” *Procedia Computer Science*, vol. 174, pp. 448–453, 2020.

[34] S. M. Wasif, S. Thakery, A. Nagauri, and S. I. Pereira, “Food calorie estimation using machine learning and image processing,” *International Journal of Advanced Research, Ideas and Innovations in Technology*, vol. 5, no. 2, pp. 1627–1630, 2019.

[35] Y. Lu, “Food image recognition by using convolutional neural networks (cnns),” *arXiv preprint arXiv:1612.00983*, 2016.

[36] A. Venkataramanan, M. Laviale, C. Figus, P. Usseglio-Polatera, and C. Pradalier, “Tackling inter-class similarity and intra-class variance for microscopic image-based classification,” in *International conference on computer vision systems*. Springer, 2021, pp. 93–103.

[37] V. Hemalatha Reddy, S. Kumari, V. Muralidharan, K. Gigoo, and B. S. Thakare, “Literature survey—food recognition and calorie measurement using image processing and machine learning techniques,” in *ICCCE 2019: Proceedings of the 2nd International Conference on Communications and Cyber Physical Engineering*. Springer, 2020, pp. 23–37.

[38] Y. Zhang, L. Deng, H. Zhu, W. Wang, Z. Ren, Q. Zhou, S. Lu, S. Sun, Z. Zhu, J. M. Gorriz *et al.*, “Deep learning in food category recognition,” *Information Fusion*, vol. 98, p. 101859, 2023.

[39] O. S. Azeez, H. Z. Shafri, A. H. Alias, and N. A. Haron, "Integration of object-based image analysis and convolutional neural network for the classification of high-resolution satellite image: A comparative assessment," *Applied Sciences*, vol. 12, no. 21, p. 10890, 2022.

[40] F. Shuang, Z. Lu, Y. Li, C. Han, X. Gu, and S. Wei, "Foodnet: Multi-view and label dependency learning based multi-task network for food and ingredient recognition," *Available at SSRN 4248593*, 2022.

[41] J. Zhao, W. Lv, B. Du, J. Ye, L. Sun, and G. Xiong, "Deep multi-task learning with flexible and compact architecture search," *International Journal of Data Science and Analytics*, pp. 1–13, 2023.

[42] H. Liang, G. Wen, Y. Hu, M. Luo, P. Yang, and Y. Xu, "Mvanet: Multi-task guided multi-view attention network for chinese food recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3551–3561, 2020.

[43] X. Sun, R. Panda, R. Feris, and K. Saenko, "Adashare: Learning what to share for efficient deep multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8728–8740, 2020.

[44] J. Zhang, K. Yan, and Y. Mo, "Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms," *Information*, vol. 12, no. 5, p. 207, 2021.

[45] P. Ma, Z. Zhang, Y. Li, N. Yu, J. Sheng, H. K. McGinty, Q. Wang, and J. K. Ahuja, "Deep learning accurately predicts food categories and nutrients based on ingredient statements," *Food Chemistry*, vol. 391, p. 133243, 2022.

[46] J. He, Z. Shao, J. Wright, D. Kerr, C. Boushey, and F. Zhu, "Multi-task image-based dietary assessment for food recognition and portion size estimation," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 49–54.

[47] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol: Large language models for ontology learning," in *International Semantic Web Conference*. Springer, 2023, pp. 408–427.

[48] V. K. Kommineni, B. König-Ries, and S. Samuel, "From human experts to machines: An llm supported approach to ontology and knowledge graph construction," *arXiv preprint arXiv:2403.08345*, 2024.

[49] D. Allemang and J. Sequeda, "Increasing the llm accuracy for question answering: Ontologies to the rescue!" *arXiv preprint arXiv:2405.11706*, 2024.

[50] Z. Yang, E. Khatibi, N. Nagesh, M. Abbasian, I. Azimi, R. Jain, and A. M. Rahmani, "Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an llm-augmented framework," *Smart Health*, vol. 32, p. 100465, 2024.

[51] I. Yummly, "Yummly," 2024. [Online]. Available: https://www.yummly.com/

[52] I. Allrecipes, "Allrecipes," 2024. [Online]. Available: https://www.allrecipes.com/

[53] I. Donadello and M. Dragoni, "Ontology-driven food category classification in images," in *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*. Springer, 2019, pp. 607–617.

[54] R. H. Güting, "Graphdb: Modeling and querying graphs in databases," in *VLDB*, vol. 94. Citeseer, 1994, pp. 12–15.

[55] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[56] D. Singh, V. Kumar, and M. Kaur, "Densely connected convolutional networks-based covid-19 screening model," *Applied Intelligence*, vol. 51, pp. 3044–3051, 2021.