HeASe: An AI-powered Framework to Promote Healthy and Sustainable Eating

Alessandro Petruzzelli, Cataldo Musto*, Michele Ciro Di Carlo, Giovanni Tempesta and Giovanni Semeraro

University of Bari Aldo Moro, via Orabona 4, Bari, 70125, Italy

Abstract

This paper introduces Healthy And Sustainable eating (HeASe), a comprehensive framework designed to promote healthy and sustainable eating by leveraging *large language models* and *food retrieval techniques*. As global concerns about nutrition and environmental sustainability escalate, the need for effective solutions that allow people to better nourish and improve their knowledge and self-awareness about food becomes imperative. To this end, given an input recipe, our framework first identifies a set of substitute meals by exploiting a retrieval strategy based on macro-nutrients, then relies on large language models to re-rank candidate recipes based on their healthiness and sustainability. As shown in our experiments, the methodology has the ability to expose individuals to better dietary choices, potentially contributing to overall well-being and reducing the ecological footprint of food consumption.

Keywords

Food Recommendation, Large Language Models, Health-aware Recommender Systems, Sustainability

1. Introduction

Today, the food industry is efficient and offers a variety of fresh and processed options. However, every step of the agricultural and food chain raises environmental concerns. Land use, water consumption, and air emissions all have an impact on the environment. While technological advancements create new markets and opportunities, they must also address these environmental challenges. To mitigate the environmental footprint of the food chain, a fundamental shift in consumer behavior is essential. Indeed, we must transition towards a dietary paradigm that prioritizes both individual health and environmental sustainability [1]. This necessitates a move away from conventional consumption patterns and towards a more mindful approach to food choices. All these principles are in lines with several Sustainable Development Goals (SDGs), in particular SDG3 (*Good Health and Well-being*) and SDG12 (*Responsible Consumption and Production*).

In recent years, *food recommendation systems (RSs)* [2] have emerged as a promising avenue to guide consumers toward healthier and more sustainable dietary choices. These systems can be categorized into two primary types: health-aware and sustainable-aware RSs [3]. Health-aware food RSs [4] aim to assist users in defining daily diets that align with their nutritional needs and health goals. These systems typically achieve this by balancing user preferences with various health-related factors. Previous methods have tried to incorporate healthiness by replacing ingredients with healthier alternatives [5, 6] or incorporating nutritional facts as function constraints [7, 8]. In [9], a post-filtering method has been proposed to score recipes based on health criteria.While these approaches have shown promise in promoting healthier eating habits, they often face limitations. Notably, methods that directly substitute ingredients or impose hard constraints on healthiness can significantly alter the recipe's original

STAI'24: International Workshop on Sustainable Transition with AI (Collocated with the 33rd International Joint Conference on Artificial Intelligence 2024), August 05, 2024, Jeju, Republic of Korea.

*Corresponding author.

alessandro.petruzzelli@uniba.it (A. Petruzzelli); cataldo.musto@uniba.it (C. Musto); m.dicarlo6@studenti.uniba.it (M. C. Di Carlo); g.tempesta16@studenti.uniba.it (G. Tempesta); giovanni.semeraro@uniba.it (G. Semeraro)

0009-0008-2880-6715 (A. Petruzzelli); 0000-0001-6089-928X (C. Musto); 0009-0001-0461-8276 (M. C. Di Carlo); 0009-0000-6211-7173 (G. Tempesta); 0000-0001-6883-1853 (G. Semeraro)

characteristics, potentially compromising user satisfaction. Additionally, post-filtering approaches may discard potentially healthy recipes that fall below an arbitrary threshold, limiting user choice.

On the other hand, *sustainability-aware food RSs* solely consider the environmental impact related to food consumption. For instance, in [3], the authors introduce a system that exploits the information about water footprint. In particular, it promotes recipes with ingredients whose production needs a lower quantity of water. While being of interest and certainly novel, this approach fails to capture the complete picture of a recipe's impact ignoring other sustainability aspects such as *carbon emissions* [10], that play a key role in assessing the sustainability of a recipe. To sum up, the analysis of the state of the art showed that there is a scarcity of systems that jointly tackle the problem of providing food suggestions that are healthy and sustainable at the same time.

Accordingly, we propose a novel framework that aims to fill in this gap by exploiting large language models (LLMs) and a recipe similarity formula based on macro-nutrients. In particular, given an input (not sustainable) recipe, we first use macro-nutrients to identify suitable alternative, then we rank them based on our sustainability score and we finally exploit large language models (*i.e.,* GPT 3.5 Turbo [11]) to select an alternative recipe that is both healthy and sustainable. Up to our knowledge, the use of LLMs to identify sustainable food alternative is a completely novel research direction.

Figure 1: A toy example of HeASe framework

In our vision, this approach acknowledges that health-conscious consumers often consider not only the nutritional value of food but also its environmental impact. So, by incorporating a sustainability score for each ingredient, the framework can identify recipes that encompass both individual well-being and environmental responsibility. A toy example showing the behavior of the framework is presented in Figure 1, while the contribution of the paper can be summarized as follows:

- **Sustainability Score:** we introduce a strategy to estimate the sustainability of a recipe based on the information about *water and carbon footprint* of its ingredients.
- **Dataset**: we release a new dataset that extends HUMMUS [12] with *sustainability and healthiness* scores for ingredients. In particular, we provided all the recipes in the dataset with information about environmental aspects. This will encourage and foster research in the area of sustainability-aware food RSs.
- **HeASe Framework**: we propose a framework that provides users with more sustainable and healthier recipes by exploiting: *(a)* recipe similarity based on macro-nutrients; *(b)* sustainability and healthiness scores; *(c)* selection mechanism based on LLMs.
- **Evaluation**: we showed that our sustainability scores allowed to identify similar but more sustainable recipes. Moreover, we also showed the LLMs can be particularly effective in selecting the most suitable alternative given a pool of candidate recipes. Both these directions have been scarcely investigated in the state of the art.

2. Assessing Healthiness and Sustainability

2.1. Calculating Healthiness of Recipes

Determining the "healthiness" of a recipe is a complex issue, heavily influenced by its nutrient composition and individual dietary needs. The concept of healthy food has experienced significant evolution, with past approaches focusing on factors like calories information [4], cholesterol levels [13], or multi-nutrients like protein, sodium, and saturated fats [14].

Today, we have a more comprehensive framework based on guidelines from international health organizations like the World Health Organization (WHO) [15]. The WHO recommends daily intake ranges for 15 macro-nutrients. Based on these intakes, in the HUMMUS dataset [12] the authors created a single score reflecting a recipe's overall healthiness. In particular, the method relies on the *"traffic light"* system proposed by [16]: each macro-nutrient range is assigned a color based on its perceived healthfulness (green for healthy, yellow for moderate, red for unhealthy) , and each color is mapped to a range of scores. The individual scores of the macro-nutrients are then added up and normalized to create a final WHO score ranging from 0 (very healthy) to 14 (very unhealthy) for each recipe. Given a recipe r, from now on the healthiness of the recipes calculated as we just described is indicated as $WHO(r)$. For more details on the formula, we suggest to refer to [12].

2.2. Calculating Sustainability of Recipes

While the task of calculating the healthiness of a recipe has some previous attempts, the assessment of the sustainability is relatively newer and scarcely investigated. Indeed, sustainability is a complex and constantly developing field, with no single universally accepted method. This makes it challenging to objectively compare the environmental impact of different recipes. Only of the first attempts in this direction is represented by the SU-EATABLE Life (SEL) dataset [17], that provides *carbon footprint (WC)* and *water footprint (WF)* data for various food ingredients.

In this work, we tackle the task of assessing the sustainability of the recipes available in the HUMMUS dataset by properly processing the information encoded in SEL dataset. In particular, the process is organized as follows:

1. **Pre-process the SU-EATABLE Life (SEL) dataset**. We remove noise by eliminating items lacking both footprints, removing unnecessary characters from names, and filtering out stopwords and adjectives.
2. **Match ingredients with recipes**: We match ingredients in the SEL dataset with those in each recipe from the HUMMUS dataset.
3. **Handle missing ingredients**: To ensure comprehensive matching, we perform additional steps:
   - Check if the SEL ingredient name is contained within the recipe ingredient name.
   - Check if the recipe ingredient name is contained within the SEL ingredient name.
   - If the above steps find matchings, we utilize transformers[1] to calculate the similarity between missing ingredients and matched ones in SEL, with a threshold of 0.98. We manually reviewed similarities further refined the matches.
4. **Manual intervention for high-occurrence missing ingredients**: We manually addressed 87 missing ingredients with over 1000 occurrences, identifying 19 potential associations.

Based on the previous strategy, given an ingredient K we can obtain its corresponding water and carbon footprints, labeled as $WP_f(K)$ and $CP_f(K)$.

Next, to evaluate the overall environmental impact of an ingredient we designed a new metric named *Ingredient Sustainability Score (ISS)*, calculated as follows:

$$ISS(K) = \alpha \times WF_f(K) + \beta \times CF_f(K) \tag{1}$$

where:

- K represents the specific ingredient.
- $WF_f(K)$ denotes the water footprint of ingredient K.
- $CF_f(K)$ represents the carbon footprint of K.
- α and β are weighting factors, with $\alpha = 0.2$ and $\beta = 0.8$[2]

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] This weighting scheme prioritizes the carbon footprint over the water footprint, reflecting the generally greater environmental impact of greenhouse gas emissions compared to water use. Of course, different weighting schemes may be adopted as well.

Next, based on the ISS scores for ingredients, we define a scoring function for recipes. To this end, we first rank the ingredients $i_1 \dots i_n$ based on their ISS. Then, we define the Recipe Sustainable Score (RSS) for a recipe R as:

$$RSS(R) = \sum_{k=0}^{|N|-1} ISS(i_k)e^{-i} \tag{2}$$

Where i_k represents the k-th ingredient of the recipe, based on the previous ranking.

The intuition behind this formula is to give a *greater importance* to the ingredients with higher carbon and water footprint (*i.e.,* those that have a greater environmental impact). Differently from a simple average, that gives identical importance to the ingredients, this strategy gives more importance to ingredients that are not sustainable. Indeed, this discounting mechanism ensures that the overall recipe score reflects the dominance of the main ingredient while incorporating the influence of additional ingredients. Finally, the ultimate sustainability score (SuS) of a recipe was computed as:

$$\text{SuS(R)} = 1 - \frac{RSS(R) - MinRSS}{MaxRss - MinRss} \tag{3}$$

Where MinRSS and MaxRSS are the minimum and maximum RSS scores obtained over the dataset of recipes, respectively, and are used as a normalization factor. It is important to note that the Sustainability Score is calculated based on the water and carbon footprint of all the ingredients of the recipe. These have negative environmental impacts, so a higher overall score indicates a more sustainable recipe. A qualitative evaluation of the effectiveness of our formula is provided next.

2.3. Description of the Dataset

As mentioned in the previous steps, one of the contributions of the paper is a new dataset providing information about sustainability of recipes. Our dataset is based on Health-aware User-centered recoMMendation and argUment-enabling data Set (HUMMUS) dataset. This dataset is built on top of the existing FoodKG [18] knowledge graph. The authors have added more data to the graph by collecting additional information for each recipe. They have also included valuable features such as nutritional scores from WHO, FDA, and Nutriscore. This dataset has over $507,000$ recipes, and each recipe contains details about ingredients, macro-nutrients (calories, total fat, etc.), and other relevant information organized into *tags*. The tags provide information about key recipe aspects like main ingredients (meat, pork, fruit) and dish category (main course, dessert, breakfast). The dataset contains a set of 902 unique tag values.

To ensure the dataset's quality, we performed some prep-rocessing steps. We removed duplicate recipes, those missing any tags, and those lacking any listed ingredients. This process helped to refine the dataset and improve its overall usability, reducing the number of recipes to $214,800$.

Next, we applied the pipeline described in section 2.2 to calculate the SuS score for each recipe. However, during this process, we noticed that not all ingredients could be matched, even after manual checking. To maintain the overall quality of the dataset, we decided to remove recipes where more than 30% of ingredients could not be matched in the SEL dataset. This additional filtering reduced the number of recipes to 100,870.

Finally, we categorized recipes with three sustainability labels based on their sustainability scores:

- High ($score \geq 0.9$): Representing highly sustainable recipes (16,433 recipes).
- Medium ($0.5 < score < 0.9$): Representing moderately sustainable recipes (79,157 recipes).
- Low ($score \leq 0.5$): Indicating recipes with low sustainability (5,280 recipes).

Some examples of the recipes that were classified in each category will be provided next. Moreover, the dataset together with the labels we calculated was used in our experiment to assess the effectiveness of the strategy and was released as a contribution of the work.

3. Description of the Framework

This section introduces the HeASe framework. As previously stated (see Figure 1), the goal of the framework is to automatically suggest a similar-but-healthier and more sustainable alternative of an input recipe given a by user. For better understanding the framework, we break down the process into four steps, each corresponding to a component in Figure 2.

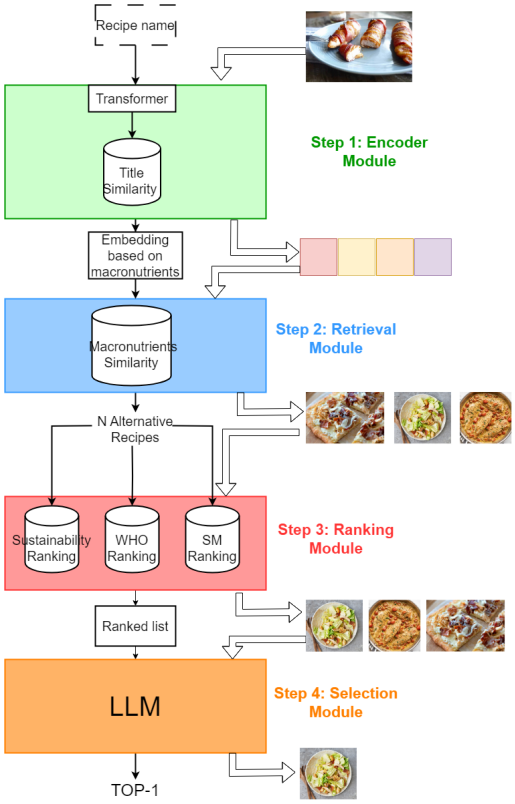

Figure 2: A schematic diagram of HeASe. The framework takes a recipe name as input and outputs a more sustainable and healthiness alternative. The framework consists of four modules: (1) Encoding (2) Retrieval (3) Ranking, and (4) Selection.

3.1. Step 1: Encoding Module

The workflow starts with the Encoding Module. In a nutshell, this module takes as input the *input recipe* and returns a vector encoding the *characteristics of the recipe* in terms of macro-nutrients. This is a mandatory step, since we want to identify recipes that are healthier and more sustainable, but also *similar* to the input. Accordingly, it is necessary to understand nutritional values and characteristics of a recipe.

To this end, we exploited a pre-trained transformer fine-tuned on the recipe domain[3] to encode the input recipe based on the *name* of the recipe. Next, we calculate the similarity between the *input recipe* and the names of the other recipes available in the dataset. If a match with a similarity score exceeding 0.99 is found, we obtain a *precise match*. It means that a recipe with (almost) the same name exists in the dataset. Otherwise, the k most similar recipes are returned. In this way, the framework is able to manage *both* exact and non-exact matching.

In case of exact match, the output of the module is a vector encoding the values of the *macro-nutrients* of the matched recipe, together with the descriptive *tags* available in the dataset. Conversely, in case of non-exact matching, the macro-nutrients of the input recipe are obtained as the centroid vector of the macro-nutrients of the k similar recipes previously identified by the transformer.

[3]https://huggingface.co/davanstrien/autotrain-recipes-2451975973

3.2. Step 2: Retrieval Module

As mentioned in the previous step, the Encoding module generates a representation of the input recipe based on its macro-nutrients. Such a representation is then used to search for similar recipes. To address this task, we calculated the similarity in terms of macro-nutrients between the input recipe (as returned by the Encoding module) and all the recipes in the dataset, based on the *cosine similarity*. This allowed us to retrieve recipes that closely matched the input recipe in terms of their nutritional composition.

Moreover, we also used the *tags* that are available for each recipe as a further element to improve the quality of the retrieved recipes. In particular, we only return recipes that are similar *and* share *at least* one tag (i.e., pasta, breakfast, japanese, etc.) with the input recipe provided by the user. In this way, we avoid that very different recipes could be included in the output of the Retrieval module.

3.3. Step 3: Ranking Module

Once similar recipes are obtained, it is necessary to rank them in order to identify an alternative that is more sustainable and healthier. This role is played by the Ranking module, whose goal is to take as input the recipes previously returned by the Retrieval module and identify the better alternatives for the user. To rank the recipes, we defined a new function called **HeaSe Score** (HS), defined as follows:

$$\text{HS(R)} = \alpha \cdot \text{Sustainability}(R) + \beta \cdot \text{WHO}(R) \tag{4}$$

- Where R represents a recipe.
- SuS(R) is a function that returns the sustainability score of R, as described in Section 2.2
- WHO(R) is a function that returns the WHO score of a given recipe.
- α and β hyperparameters that allow you to weight the importance of each factor.

At the end of this step, a list of ranked alternative recipes is obtained. It is worth emphasizing that the workflow can also stop after this step, by returning to the user the top-1 recipe retrieved by the systems based on the HeaSe score. However, we also implemented a Selection module based on LLMs to assess whether the knowledge encoded in large language models can be exploited to better handle this task.

3.4. Step 4: Selection Module

Finally, in the Selection module, the output previously obtained from the Ranking module is processed by using LLMs, specifically GPT-3.5 turbo, in order to select the *most suitable alternative* of the recipe provided as input by the user. To carry out this step we specifically designed a strategy inspired by Retrieval-Augmented Generation (RAG) [19] which takes as input the list of candidate recipes and asks the LLM to select the most suitable one. This is done through a *zero-shot prompt* that is used to query the LLM, leaving it the task to identify the most suitable candidate recipe based on the knowledge encoded in the language model. An example of such a prompt is provided below. As shown in the example, we populate the prompt with the recipes previously identified and we let GPT pick the more sustainable alternative recipe. To mitigate potential biases like positional bias [20], the retrieved recipes are shuffled and inserted into the prompt without any additional information.

```
Using your knowledge, please rank (if
necessary) the following recipes from
most to least recommended based on a
balance of sustainability and
healthiness:
1. Recipe: Healthy Salad
2. Recipe: Quinoa Bowl
3. Recipe: Veggie Stir-Fry
Which one should I choose?
Return just the name.
```

It is crucial to note that the lack of information about the input recipe is intentional and derives from the experiment's ultimate objective. We aim to assess the LLM's ability to accurately identify the recipe with higher values of sustainability and healthiness without relying on specific recipe details.

Of course, one of the goals of the experiment will be to assess the effectiveness of LLMs in the task of automatically identifying healthy and sustainable recipes.

4. Experimental Evaluation

This section explores the effectiveness of the proposed metrics and framework through experiments addressing the following Research Questions (RQs):
RQ1 - Scoring Effectiveness: Can SuS and HeASe scores actually rank recipes based on sustainability and healthiness?
RQ2 - Retrieval Effectiveness: Is the framework able to successfully identify suitable food alternatives?
RQ3 - LLM-based Selection Effectiveness: Can LLMs be leveraged to automatically select sustainable alternatives?

4.1. Experimental Setting

Dataset and Evaluation Protocol All the experiments rely on the dataset previously described in Section 2.3, that is also available online on our repository[4]. Based on this dataset, we evaluated the performance of the framework by providing an input recipe and by checking whether the alternative identified by the framework is healthier and/or more sustainable. To guarantee the soundness of the protocol, we evaluated the performance of HeaSe system across diverse scenarios:

1. **Low Sustainability:** based on 100 randomly selected recipes labeled as "Low" in sustainability.
2. **Medium Sustainability:** based on 100 randomly selected recipes labeled as "Medium" in sustainability.
3. **High Health:** based on 100 randomly selected recipes with a WHO score above average.
4. **Unknown Recipes:** based on 30 Recipes not present in the recipe dataset.

These scenarios allow us to assess the framework's efficacy in different contexts. For instance, for the "Low Sustainability" scenario we expect significant improvements in the output recipe's sustainability and healthiness compared to the input. However, we also evaluate the framework's performance in more challenging settings (*i.e.,* high health, based on recipes that are already healthy, or unknown, in order to also assess the effectiveness of non-exact matching in the retrieval phase).

Implementation Details and Model Parameters The model uses a pre-trained transformer encoder with a hidden dimensionality of 768. This allows the model to efficiently find similarities between the input text and recipe titles, even when the input doesn't perfectly match the recipe title. As for the RETRIEVAL module, the number of alternative recipes based on macro-nutrient similarity which is returned is set to 100. The recipe representation is based on its macro-nutrients, which include: `Calories [cal], Total Fat [g], Saturated Fat [g], Cholesterol [mg], Sodium [mg], Dietary Fiber [g], Sugars [g], and Protein [g]`. As regards the scoring function in the RANKER module, the best configuration for the model was achieved by setting the alpha and beta values in the formula 4 to 0.7 and 0.3, respectively.

Evaluation Metric We evaluate the performance of the HeASe system by calculating the mean percentage increment of each metric for each scenario. Given an input recipe (R) and a list of N possible alternatives (A) returned by the system, we compute the following:

$$\text{WHO\_incr} = \frac{\frac{1}{N}\sum_{i=0}^{N} WHO(A_i) - WHO(R)}{WHO(R)} \times 100 \tag{5}$$

[4] https://github.com/swapUniba/HeASe

$$\text{SuS\_incr} = \frac{\frac{1}{N}\sum_{i=0}^{N} SuS(A_i) - SuS(R)}{SuS(R)} \times 100 \tag{6}$$

$$\text{HeASe\_incr} = \frac{\frac{1}{N}\sum_{i=0}^{N} HeASe(A_i) - HeASe(R)}{HeASe(R)} \times 100 \tag{7}$$

Intuitively, these metrics calculate the increase (if any) in terms of healthiness and sustainability of the recipe retrieved by the framework compared to the input one.

Sensitivity Analysis. Finally, to investigate the performance of the system on varying of different parameters, we also carried out a sensitivity analysis based on the following key factors:

- *Tags matching:* This option controls how strictly the recipe tags need to match between the input recipe and the retrieved items. By setting it to true, the framework only outputs recipes that share all the same tags with the input recipe.
- *Retrieved items:* This parameter determines the number of alternative recipes retrieved as recommendations.

4.2. Discussion of the Results

RQ1 - Scoring Function Effectiveness: To answer RQ1, we present the top-5 and worst-5 recipes based on SuS and HeASe scores.

- **Top-5 Recipes** (Tables 1 and 3): as shown in the tables, this includes recipes like "Homemade Oatmeal," "Quinoa-Toasted," and "Seasoned Rice", which excel in both sustainability and healthiness, achieving high SuS and HeASe scores. These options likely prioritize plant-based ingredients and simple preparation methods, reducing environmental impact and promoting nutritional value. Generally speaking, we can state that the list of the more sustainable and healthy recipes confirms the effectiveness of the scoring function we designed.
- **Worst-5 recipes** (Tables 2 and 4): Conversely, recipes like "Rich Lamb Curry," "Five Meat Chili," and "Middle Eastern Stew" score poorly in both categories. These dishes likely contain significant amounts of meat, which can contribute to a higher environmental footprint and potentially lower overall health benefits. Also, in this case, we can state that the poorly sustainable recipes are correctly identified through our scoring function.

The disparity between metrics: Interestingly, the top and bottom scorers for SuS do not entirely overlap with those for HeASe. "Boiled Radishes" and "Granita" for example, rank highly in SuS but not in HeASe. This suggests that some sustainable practices might not always translate directly to health benefits, and vice versa, highlighting the need for a balanced metric like HeASe.

To sum up, we can answer RQ1 by stating that the qualitative analysis we provided generally confirmed the effectiveness of the scoring function we introduced in this paper.

Table 1
Top-5 Recipes ordered for HeASe Score

Recipe Title	SuS	WHO	HeASe
Homemade Oatmeal	0.983	0.461	0.827
Quinoa-Toasted	0.975	0.444	0.816
Seasoned Rice	0.979	0.423	0.812
Fat Free Whole Wheat Tortillas	0.975	0.418	0.808
Plain Rice	0.977	0.383	0.801

Table 2
Worst-5 Recipes ordered for HeASe Score

Recipe Title	SuS	WHO	HeASe
Rich Lamb Curry	0.039	0.153	0.074
Five Meat Chili	0.028	0.198	0.079
Middle Eastern Stew	0.031	0.206	0.084
Roast Leg of Lamb	0.049	0.213	0.098
Curried Lamb on Rice	0.049	0.224	0.101

Table 3
Top-5 Recipes ordered for SuS metric

Recipe Title	SuS	WHO	HeASe
Boiled Radishes	0.997	0.293	0.786
Horseradish Applesauce	0.997	0.314	0.792
Granita	0.996	0.236	0.768
Rehydrated Onions	0.995	0.268	0.777
Pot Onion Chops	0.995	0.260	0.775

Table 4
Worst-5 Recipes ordered for SuS metric

Recipe Title	SuS	WHO	HeASe
Five Meat Chili	0.029	0.198	0.079
Middle Eastern Stew	0.032	0.206	0.084
Rich Lamb Curry	0.040	0.153	0.074
Curried Lamb on Rice	0.049	0.224	0.101
Roast Leg of Lamb	0.049	0.213	0.098

RQ2 - Retrieval Effectiveness To answer RQ2, we conducted several tests to evaluate the effectiveness of the framework, that is to say, to assess whether the alternative recipes retrieved through our pipeline are healthier and more sustainable w.r.t. the input recipe. In particular, for each of the 100 recipes in each scenario (see Section 4.1) we retrieved the 100 most similar recipes based on macronutrients, we ranked them based on our HeaSe score, and we calculated the average increase in terms of healthiness and sustainability for all the recipes. The results are reported in Table 5.

As shown in Table 5, the results confirmed the effectiveness of the approach, since the proposed alternative recipes are healthier and more sustainable, on average, in all the experimental scenarios we considered. It is worth emphasizing that the results are consistent across all the different scenarios, even if the gaps of course reflect the complexity of the task. Indeed, when poorly sustainable recipes are used as input of the framework, a huge average increase emerges from all the alternatives. Even though this was expected, it is important to see that the increase we obtained is really huge, on average. It is also important to note that an average increase in terms of sustainability is obtained when recipes that are already healthy are used as input. Next, the results of the sensitivity analysis are shown in Figures 3 and 4. Due to space constraints, we only reported the plot for two scenarios, *i.e.*, the "Low Sustainability" scenario and the "High Health" scenario. The other scenarios follow a similar trend. Plots clearly show that the framework achieves better performance as the number N of alternative recipes increases, and it confirmed our choice of choice of retrieving and ranking 100 similar recipes. In particular, as shown in Figure 4a, this is a necessary choice for the "high health" scenario, since by considering the top-1 and top-10 recipes retrieved we have an average decrease in sustainability. Conversely, by increasing the number of recipes, the overall healthiness and sustainability are higher. While this suggests that alternative strategies for retrieval and ranking need to be investigated in the future, proper tuning of the parameters still guarantees good performance.

Finally, Figures 3b and 4b show the results on varying of the tag matching strategy. The results reveal slight differences, with configurations that don't require matching all tags generally producing better results. This means that when the retrieved recipes need to match *all* the tags of the input recipe, non-relevant recipes may be generally returned. To sum up, all the results of the sensitivity analysis

showed that the platform generally performs well, but a proper choice of parameter may lead to more effective results.

Table 5
Performance of the HEaSe framework in the retrieval task

Scenario	WHO_incr	SuS_incr	HeASe_incr
Low Sustainability	+12.70%	+139.03%	+112.89%
Medium Sustainability	+69.27%	+22.70%	+21.38%
High Health	+5.51%	+20.19%	+17.67%
Unknown Recipes	+16.43%	+17.87%	+17.51%

To conclude the analysis, in Table 6 we report some qualitative examples showing the real behavior of the HEaSe framework. In particular, for each experimental scenario, we present the output generated by the platform based on different input recipes. As shown in the table, in all the reported settings the alternative recipe is healthier more sustainable, and sufficiently similar to the input one. This definitely confirmed the effectiveness of the design choices. More tests can be carried out by running our online demo[5].

Table 6
Input-Output examples per scenario

Scenario	Input	Output	HeASe
Medium Sustainability	Rockin Cheddar Ranch Turkey Burgers! Strippin' Chicken! (Bacon Strip Chicken)	Ginger, Lemon and Garlic Swordfish Steak. Super Simple Chicken Salad	**+24.70%** **+24.12%**
Low Sustainability	Beef Stir-Fry Turkey-beef Kebabs	Tofu Hot wings Slow-Cooker Swiss Steak	**+104.80%** **+92.17%**
High Health	Chili Dog Casserole Wedding Cakes	No-fuss Burgers Spice Cookies	**+119.23%** **+10.84%**

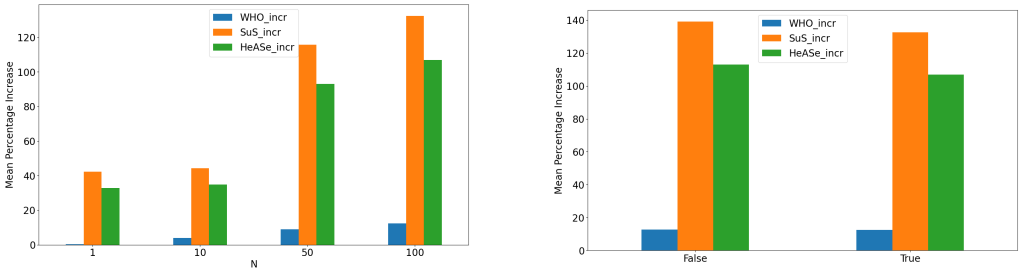

Figure 3: Mean percentage increments on the three metrics on Low Sustainability Scenario on different configuration

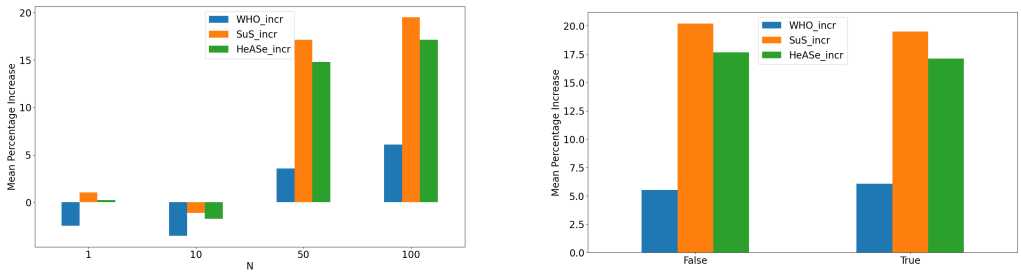

Figure 4: Mean percentage increments on the three metrics on High Health Scenario on different configuration

[5] https://github.com/GiovTemp/SustainaMeal_Case_Study

Table 7
Experiments on the Selection based on LLMs

WHO_incr	SuS_incr	HeASe_incr	gpt_rerank
+3.26%	**+71.33%**	**+56.07%**	True
+2.77%	+68.41%	+54.27%	False

RQ3 - LLM-based Selection Effectiveness: Finally, to answer RQ3, we evaluated the ability of GPT-3.5 Turbo to automatically pick the more sustainable alternative in a pool of candidate recipes retrieved by the system. The process follows the step described in the Selection module of the framework. Due to limitations in prompt length, we experimented with a smaller set of alternatives (*i.e.,* 10 candidate recipes). The analysis with a longer prompt is left as future work. In Table 7, we compare the healthiness and sustainability of the recipe with the highest score calculated by the Ranker to the recipe identified by GPT among the top-10 returned by the Ranker as well. As shown in the table, the results show that the LLM showed an unexpected and surprising ability to exploit its own knowledge about responsible food consumption to automatically select the best recipe in a pool of 10 candidates. Indeed, when compared with the top-1 recipes previously picked, the average sustainability and healthiness of the recipes is generally higher. These findings suggest that LLMs can effectively leverage the strengths of both retrieval and generation techniques to identify recipes that are both sustainable and healthy. This is an important finding of this work, showing the effectiveness of LLMs in a novel and scarcely investigated research direction.

5. Discussion and Future Works

The framework described in this paper aligns with SDG3 and SDG12. In particular, we foresee the following impact:

- **SDG 3 - Good Health and Well-being: Promoting Healthier Diets.** The framework focuses on encouraging individuals to adopt healthier eating habits. By leveraging our system users can explore and choose recipes that contribute to a balanced and nutritious diet. This directly contributes to the goal of ensuring good health and well-being by promoting better nutrition and reducing the risk of diet-related diseases.
- **SDG12 - Responsible Consumption and Production: Ingredient Substitution:** The framework contributes to responsible consumption by helping users identify more sustainable substitute ingredients in recipes. This aligns with SDG 12's focus on ensuring sustainable consumption by promoting eco-friendly and ethically sourced ingredients.

In summary, the HeaSe framework contributes to SDG 3 by promoting healthier diets and better well-being and to SDG 12 by encouraging responsible consumption and production practices. By combining technology-driven solutions with user engagement and education, the project seeks to address the interconnected challenges of health and sustainability in the context of food choices. In future work, we will evaluate different strategies for the selection of alternative recipes, and we evaluate the effectiveness with real users.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI under the NRRP MUR program funded by the NextGenerationEU and project PHaSE (CUP H53D23003530006) - Promoting Healthy and Sustainable Eating through Interactive and Explainable AI Methods, funded by MUR under the PRIN program. Additionally, we acknowledge the CINECA award under the ISCRA initiative (class C project: IscrC_LLM_REC), for the availability of high-performance computing resources and support

References

[1] C. Hartmann, G. Lazzarini, A. Funk, M. Siegrist, Measuring consumers' knowledge of the environmental impact of foods, Appetite 167 (2021) 105622.

[2] C. Trattner, D. Elsweiler, Food recommender systems: important contributions, challenges and future research directions, arXiv preprint arXiv:1711.02760 (2017).

[3] I. Gallo, N. Landro, R. La Grassa, A. Turconi, Food recommendations for reducing water footprint, Sustainability 14 (2022). URL: https://www.mdpi.com/2071-1050/14/7/3833. doi:10.3390/su14073833.

[4] M. Ge, F. Ricci, D. Massimo, Health-aware food recommender system, in: Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 333–334. URL: https://doi.org/10.1145/2792838.2796554. doi:10.1145/2792838.2796554.

[5] C.-Y. Teng, Y.-R. Lin, L. A. Adamic, Recipe recommendation using ingredient networks, in: Proceedings of the 4th annual ACM web science conference, 2012, pp. 298–307.

[6] D. Elsweiler, C. Trattner, M. Harvey, Exploiting food choice biases for healthier recipe recommendation, in: Proceedings of the 40th international acm sigir conference on research and development in information retrieval, 2017, pp. 575–584.

[7] D. Elsweiler, M. Harvey, B. Ludwig, A. Said, Bringing the "healthy" into food recommenders, in: International Workshop on Decision Making and Recommender Systems, 2015. URL: https://api.semanticscholar.org/CorpusID:1838398.

[8] Y.-K. Ng, M. Jin, Personalized recipe recommendations for toddlers based on nutrient intake and food preferences, in: Proceedings of the 9th international conference on management of digital ecosystems, 2017, pp. 243–250.

[9] C. Trattner, D. Elsweiler, Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 489–498.

[10] D. Pandey, M. Agrawal, J. S. Pandey, Carbon footprint: current methods of estimation, Environmental monitoring and assessment 178 (2011) 135–160.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] F. Bölz, D. Nurbakova, S. Calabretto, A. Gerl, L. Brunie, H. Kosch, Hummus: A linked, healthiness-aware, user-centered and argument-enabling recipe data set for recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1–11. URL: https://doi.org/10.1145/3604915.3609491. doi:10.1145/3604915.3609491.

[13] A. Starke, C. Trattner, H. Bakken, M. Johannessen, V. Solberg, The cholesterol factor: Balancing accuracy and health in recipe recommendation through a nutrient-specific metric, in: Proceedings of the 1st Workshop on Multi-Objective Recommender Systems (MORS 2021), 2021.

[14] R. Yera Toledo, A. A. Alzahrani, L. Martínez, A food recommender system considering nutritional information and user preferences, IEEE Access 7 (2019) 96695–96711. doi:10.1109/ACCESS.2019.2929413.

[15] W. H. Organization, Healthy diet, https://www.who.int/news-room/fact-sheets/detail/healthy-diet, 2020.

[16] G. Sacks, M. Rayner, B. Swinburn, Impact of front-of-pack 'traffic-light'nutrition labelling on consumer food purchases in the uk, Health promotion international 24 (2009) 344–352.

[17] T. Petersson, L. Secondi, A. Magnani, M. Antonelli, K. Dembska, R. Valentini, A. Varotto, S. Castaldi, A multilevel carbon and water footprint dataset of food commodities, Scientific data 8 (2021) 127.

[18] S. Haussmann, O. Seneviratne, Y. Chen, Y. Ne'eman, J. Codella, C.-H. Chen, D. L. McGuinness, M. J. Zaki, Foodkg: A semantics-driven knowledge graph for food recommendation, in: C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon (Eds.),

The Semantic Web – ISWC 2019, Springer International Publishing, Cham, 2019, pp. 146–162.
[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.
[20] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, 2023. arXiv:2305.17926.