Healthy and Sustainable Meals Recommendation Exploiting Food Retrieval and Large Language Models

Alessandro Petruzzelli
alessandro.petruzzelli@uniba.it
University of Bari Aldo Moro
Bari, Italy

Cataldo Musto
cataldo.musto@uniba.it
University of Bari Aldo Moro
Bari, Italy

Michele Ciro Di Carlo
m.dicarlo6@studenti.uniba.it
University of Bari Aldo Moro
Bari, Italy

Giovanni Tempesta
g.tempesta16@studenti.uniba.it
University of Bari Aldo Moro
Bari, Italy

Giovanni Semeraro
giovanni.semeraro@uniba.it
University of Bari Aldo Moro
Bari, Italy

ABSTRACT

Given the rising global concerns about healthy nutrition and environmental sustainability, individuals need more and more support in making good choices concerning their daily meals. To this end, in this paper we introduce HeaSE, a framework for Healthy And Sustainable Eating. Given an input recipe, HeaSE identifies healthier and more sustainable meals by exploiting retrieval techniques and large language models. The framework works in two steps. First, it uses food retrieval strategies based on macro-nutrient information to identify candidate alternative meals. This ensures that the substitutions maintain a similar nutritional profile. Next, HeaSE employs large language models to re-rank these potential replacements while considering factors beyond just nutrition, such as the recipe's environmental impact. In the experimental evaluation, we showed the capabilities of LLMs in identifying more sustainable and healthier alternatives within a set of candidate options. This highlights the potential of these models to guide users towards food choices that are both nutritious and environmentally responsible.

KEYWORDS

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

ACM Reference Format:

Alessandro Petruzzelli, Cataldo Musto, Michele Ciro Di Carlo, Giovanni Tempesta, and Giovanni Semeraro. 2024. Healthy and Sustainable Meals Recommendation Exploiting Food Retrieval and Large Language Models. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3640457.3688193

1 INTRODUCTION

Today, the food industry is efficient and offers diverse fresh and processed options. However, each step in the agricultural and food chain raises environmental concerns, impacting land use, water consumption, and air emissions [12]. In this context, to reduce the food chain's environmental footprint, consumers must shift towards a diet prioritizing healthiness and sustainability of meals [7].

In recent years, food recommender systems (RSs) [14] showed a good capability in moving away people from conventional food consumption patterns [3]. In this research line, research attempts trying to nudge users towards better food choices fall into two main categories: *health-aware* and *sustainability-aware* food RSs. The first help users create diets aligned with their nutritional needs and health goals [6], by balancing preferences with health-related factors. Previous methods include ingredient substitution [13] and incorporating nutritional facts [4]. However, direct ingredient substitution or strict health constraints may alter recipes, affecting user satisfaction. Post-filtering methods [15] could also limit user choice by excluding potentially healthy recipes below a threshold. Conversely, sustainability-aware RSs focus solely on environmental impact, such as water footprint [5]. While novel, these systems may overlook other crucial sustainability factors like *carbon emissions* [10], necessary for a comprehensive assessment of a recipe's sustainability. **Accordingly, there is a scarcity of systems that jointly tackle the problem of providing food suggestions that are healthy and sustainable at the same time.**

To this end, in this paper we introduce HeaSE, a framework for Healthy And Sustainable Eating based on retrieval techniques and large language models (LLMs). As shown in Figure 1, given an input recipe, our framework first identifies suitable alternatives based on their macro-nutrient. Then, it ranks these alternatives using a novel score that combines healthiness and sustainability. Finally, it employs a LLM, *i.e.,* GPT 3.5 Turbo [2], to pick the most suitable alternative given a pool of candidate recipes. Up to our knowledge, the use of LLMs to identify healthy and sustainable meals is a novel and never investigated research direction. In our vision, this approach aims to improve self-knowledge and self-awareness of the users, driving them towards food choices that also take into account the environmental impact of meals. To sum up, the key contributions of this work are as follows:
(1) 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. **(2) Dataset**: we release a new dataset that extends HUMMUS [1] with *sustainability and*

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688193

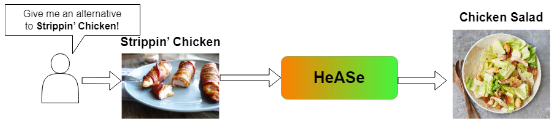

Figure 1: HeaSE framework - High-level Workflow

healthiness scores for ingredients. This will encourage and foster research in the area of sustainability-aware food RSs. **(3) 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. **(4) Evaluation**: we showed that our framework identifies 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 recipes. Both directions have been scarcely investigated so far.

2 DATA PRE-PROCESSING

In this section, we describe the data pre-processing phase. It is a fundamental step to provide HeaSE with the necessary information to identify healthier and more sustainable alternatives. To this end, we first describe the dataset we considered in this work, then we introduce our strategy to assess the sustainability of an input recipe.

2.1 Data Sources

In this work, we leverage the HUMMUS (Health-aware User-centered recoMMendation and argUment-enabling) dataset [1] for health-aware food recommendation. HUMMUS boasts over 507, 000 recipes with detailed ingredients, nutrients, and recipe tags. While it incorporates health information based on World Health Organization (WHO) guidelines [9], it lacks data on recipe sustainability. To address this gap, we propose a novel contribution: a sustainability score for each recipe and ingredients, that is used to extend the HUMMUS dataset. The scoring strategy will be detailed next.

Preliminary, to get information about the environmental impact of ingredients, we relied on the data available in SU-EATABLE LIFE [11] database. This database provides carbon footprint (CF) and water footprint (WF) values for various food commodities. By focusing on CF and WF, SU-EATABLE LIFE offers important indicators for evaluating the environmental impact of food production, processing, and distribution. The SEL dataset encompasses 3349 CF values and 937 WF values. The dataset adheres to a multilevel categorization structure, encompassing group (e.g., agricultural processed products or fishing), topology (including specific examples like peas or lentils), sub-topology (such as shellfish), and individual items like tomatoes.

2.2 Calculating Sustainability of Recipes

Determining the sustainability of a recipe is a complex task, since a universal strategy to calculate a sustainability score is currently lacking. In this work, to evaluate this aspect of the recipes included in the HUMMUS dataset, we leverage information available in SEL dataset through the following process: **(1) Clean SEL data:** We remove inconsistencies, and irrelevant words from ingredient names in SEL. **(2) Match ingredients**: We link items available in SEL to ingredients mentioned in the HUMMUS dataset. To address partial matching, we exploited transformers[1] and we mapped ingredients having a similarity higher than 98%. **(3) Manual intervention:** We manually reviewed and potentially associated high-occurrence missing ingredients in HUMMUS recipes with matches in SEL.

Based on the previous strategy, given an ingredient K in HUMMUS, 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 introduced the *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. $WF_f(K)$ and $CF_f(K)$ denote the water and footprint of ingredient K, respectively. α and β are weighting factors, with $\alpha = 0.2$ and $\beta = 0.8$[2].

Based on the ISS scores for ingredients, we then define a scoring function for recipes. To ensure data quality, we filtered out some recipes in HUMMUS. This includes removing duplicates, and recipes having less than 70% of their ingredient matching with the SEL dataset. This process resulted in a dataset of 100, 870 recipes, suitable for sustainability analysis. To calculate the sustainability score of a recipe R, we first rank the ingredients $\{i_1 \ldots i_n\}$ based on their ISS (*i.e.,* ingredients with a higher impact are put on top of the list). Then, we define the Recipe Sustainable Score (RSS) 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). With respect to the simple average of the scores, this discounting mechanism ensures that the overall recipe score reflects the dominance of the main ingredient while incorporating the influence of the others. Finally, the ultimate sustainability score (SuS) 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. Some examples of the scores returned by exploiting our strategy will be provided next. To guarantee the reproducibility, we released the entire pipeline to process HUMMUS and SEL dataset and to calculate the score itself.

3 DESCRIPTION OF THE FRAMEWORK

This section introduces the HeaSE framework. As shown in Figure 2, we split the process into four steps, which are described next.

[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. Of course, different weighting schemes will be considered as future work.

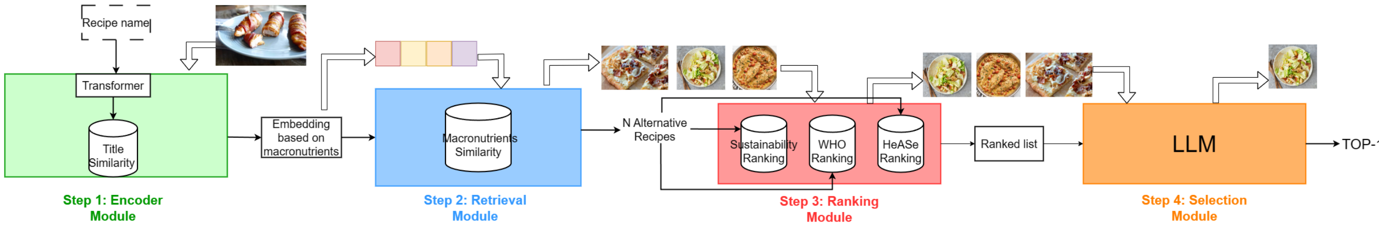

Figure 2: Worflow of HeaSE. The framework takes a recipe name as input and outputs a more sustainable and healthiness alternative. The framework consists of four modules: (1) Encoding (2) Retrieval (3) Ranking, and (4) Selection.

(1) Encoding Module. The workflow begins with the Encoding Module. It takes the input recipe and returns a vector-space representation of the recipe based on its macro-nutrient. This is done by using a pre-trained transformer fine-tuned on recipes[3].

To obtain such a representation based on macro-nutrients, the module first searches for an exact match (*i.e.,* a recipe in the dataset having the exact name of the input one). To this end, we compare the name of the recipe to all the others in the dataset. If a match with a similarity score exceeding 99% is found, the macro-nutrient vector and the descriptive tags associated with that recipe are gathered. Conversely, if no exact match exists, the module first identifies the k most similar recipes based on the transformer's output. Next, vector space representation of the recipe is obtained as the centroid vector of the macro-nutrient of the k most similar recipes. This centroid vector becomes the encoded representation of the input recipe's characteristics. Of course, k is a tunable hyper-parameter.

(2) Retrieval Module. Next, the framework aims to retrieve a pool of similar recipes. This set represents the set of *candidate* alternative meals, that will be ranked next in order to identify a healthier and more sustainable alternative. To address this task, we calculated the *cosine similarity* in terms of macro-nutrients between the input recipe (as returned by the Encoding module) and all the recipes in the dataset. This allowed us to retrieve recipes that closely matched the input recipe in terms of their nutritional composition.

To further refine results, we incorporate recipe *tags.* 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) 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. To rank the k candidate recipes, we defined a new function called **HeaSE Score** (HS), defined as follows:

$$\mathrm{HS(R)} = \delta \cdot \mathrm{Sustainability}(R) + \gamma \cdot \mathrm{WHO}(R) \tag{4}$$

Where R represents a recipe, $\mathrm{SuS}(R)$ is a function that returns the sustainability score of R, as described in Section 2.2, $\mathrm{WHO}(R)$ returns the WHO score of a given recipe, δ and γ hyperparameters that allow to weight the importance of each factor.

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

(4) Selection Module. Finally, in the Selection module, the output from the Ranking module is processed using LLMs, specifically GPT-3.5 turbo, to select the *most suitable alternative* of the recipe provided by the user. We developed a strategy inspired by Retrieval-Augmented Generation (RAG) [8] for this step. This involves inputting the list of candidate recipes and using a *zero-shot prompt* to query the LLM, which identifies the most suitable candidate based on its encoded knowledge. To mitigate biases like positional bias [16], the recipes are shuffled before being inserted into the prompt. The prompt has two parts: the Fixed Part which remains constant and provides the context and specific request, and the Dynamic Part which is populated at runtime with the retrieved recipes.

```
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.
```

The output of the prompt returned by the LLM is finally returned to the user as alternative meal recommendation.

4 EXPERIMENTAL EVALUATION

Our experiments aimed to answer the following Research Questions (RQs):

- **RQ1 - Scoring Effectiveness:** Do SuS and HeaSE scores reflect the principles of healthiness and sustainability when ranking recipes?
- **RQ2 - Retrieval Effectiveness:** Can the framework actually recommend recipes that are both healthier and more sustainable?
- **RQ3 - LLM-based Selection Effectiveness:** Can we leverage LLMs to automatically recommend sustainable food alternatives?

4.1 Experimental Setting

Dataset and Evaluation Protocol. The experiments leverage the dataset described in Section 2.1, available online in our repository[4], together with the source code of the framework.

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

As regards the protocol, we evaluate healthiness and sustainability of the recipes returned by HeaSE. To ensure more robustness, we evaluated the performance across various scenarios:

(1) ***Low Sustainability:*** 100 randomly selected recipes labeled "Low" in sustainability (SuS score ≤ 0.5) are used as input.
(2) ***Medium Sustainability:*** 100 based on randomly selected recipes labeled as "Medium" in sustainability (SuS score between 0.5 and 0.9).
(3) ***High Health:*** 100 random recipes with a WHO score above 0.9 are considered.
(4) ***Unknown Recipes:*** 30 recipes not present in the dataset to assess the performance with unknown inputs.

The rationale behind the choice of these scenarios lies in the idea of evaluating the framework through inputs that are increasingly complex. Indeed, while the identification of more sustainable alternatives can be easy for the *low* or *medium* sustainability scenarios, this becomes more challenging when a recipe that is already healthy is provided as input. Results will be discussed next.

Implementation Details and Model Parameters. The model utilizes a pre-trained transformer with size=768. The Retrieval module retrieves 100 alternative recipes based on macro-nutrient similarity. Recipe representation is based on the following macronutrients: `Calories, Total Fat, Saturated Fat, Cholesterol, Sodium, Dietary Fiber, Sugars, and Protein`. As for the Ranker module, after a rough tuning δ and γ values in Equation 4 were set to 0.7 and 0.3, respectively[5]. To answer RQ1 and RQ2 we run the pipeline by *excluding* the LLM, whose performance are assessed in RQ3.

Evaluation Metric: The performance of the HeaSE system is evaluated by calculating the mean percentage increment in terms of healthiness and sustainability, calculated through the WHO score and the SUS score, respectively, between the input recipe (R) and the list of alternatives (A) returned by the system.

	HeaSE Score		SuS metric	
	Recipe Title	*HeaSE*	Recipe Title	*SuS*
Top-5	Homemade Oatmeal	0.827	Boiled Radishes	0.997
	Quinoa-Toasted	0.816	Horseradish Applesauce	0.997
	Seasoned Rice	0.812	Granita	0.996
	Fat Free Whole Wheat Tortillas	0.808	Rehydrated Onions	0.995
	Plain Rice	0.801	Pot Onion Chops	0.995
Worst-5	Rich Lamb Curry	0.074	Five Meat Chili	0.029
	Five Meat Chili	0.079	Middle Eastern Stew	0.032
	Middle Eastern Stew	0.084	Rich Lamb Curry	0.040
	Roast Leg of Lamb	0.098	Curried Lamb on Rice	0.049
	Curried Lamb on Rice	0.101	Roast Leg of Lamb	0.049

Table 1: Top and Bottom 5 Recipes Ranked by Health and Sustainability Scores

4.2 Discussion of the Results

RQ1 - Scoring Function Effectiveness: To answer RQ1, we analyzed the top and bottom five recipes based on both HeaSE and SuS scores presented in Table 1.

Top-5 Recipes: the table shows that recipes like "Homemade Oatmeal," "Quinoa-Toasted," and "Seasoned Rice" achieved high HeaSE scores, indicating a focus on healthy and sustainable ingredients. These likely prioritize plant-based ingredients and simple preparation methods, aligning with HeaSE's design principles.

Worst-5 recipes: Conversely, the table also reveals the five recipes with the lowest HeaSE scores, including "Rich Lamb Curry," "Five Meat Chili," and "Middle Eastern Stew". These dishes likely contain significant amounts of meat, contributing to a higher environmental impact and potentially lower health benefits. The scoring function effectively identified these less sustainable options by penalizing the use of meat-heavy ingredients.

The disparity between metrics: Interestingly, the top and bottom ranked recipes for sustainability (SuS) did not entirely overlap with those ranked by HeaSE. For example, "Boiled Radishes" and "Granita" scored highly in SuS but not in HeaSE. This highlights the strength of a balanced metric like HeaSE. While SuS might focus primarily on environmental factors, HeaSE considers both health and sustainability, ensuring that a recipe with a low environmental impact but limited health benefits doesn't rank exceptionally high.

RQ2 - Retrieval Effectiveness Next, to answer RQ2, we assessed the gain in healthiness and sustainability of the recipes retrieved by the system for all the scenarios mentioned in Section 4.1. Results are calculated as the average WHO and SUS scores of the 100 most similar options ranked by HeaSE. As previously stated, the impact of LLMs is not considered here.

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%

Table 2: Performance of HeaSE in the retrieval task

Table 2 illustrates the effectiveness of our approach, showing that alternative recipes result as healthier and more sustainable, on average. The findings are valid across all the different scenarios, even the more challenging ones. These results confirm the robustness and promise of our methodology.

To conclude the analysis, in Table 3 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 confirmed the effectiveness of the design choices. More tests can be carried out by running our online demo[6].

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

Table 3: Input-Output examples per scenario

RQ3 - LLM-based Selection Effectiveness: As for RQ3, we assessed the ability of a LLM (*i.e.,* GPT-3.5 Turbo) to automatically pick the most suitable alternative from a pool of candidates. Results are presented in Table 4, comparing the average gain in WHO, SUS and

[5]Due to space reasons, we just report the results with these parameters. Tuning and further experiments are left as future work.

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

HeaSE score by considering the top-1 recipe (based on the original ranking of the framework) w.r.t. the average gain when the top-1 recipe is selected by the LLM from a pool of 10 candidates[7].

As shown in the Table, GPT-3.5 Turbo exhibited an unexpected proficiency in leveraging its knowledge about responsible food consumption to identify the optimal recipe. Indeed, the average SUS and WHO scores were generally higher for the LLM-selected recipes compared to the top selections from the Ranker. This suggests that LLMs can effectively combine retrieval and generation techniques to identify recipes that are both sustainable and healthy. These findings highlight the effectiveness of LLMs in a novel and under-explored area, and finally confirm the effectiveness of our design.

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

Table 4: Experiments on the Selection based on LLMs

5 DISCUSSION AND FUTURE WORKS

Our findings demonstrate the effectiveness of the HeaSE framework in promoting healthy and sustainable food choices. By leveraging HeaSE scoring system, the framework retrieves alternative recipes that are demonstrably better for both the user's well-being and the environment. HeaSE achieves this by identifying recipes with healthier and more sustainable ingredients.

These functionalities contribute to positive user outcomes. Users can adopt healthier eating habits and potentially reduce the risk of diet-related diseases. Furthermore, HeaSE promotes environmentally friendly practices by encouraging responsible consumption patterns. Looking ahead, we plan to explore several avenues for future work. One area of focus will be the evaluation of different ranking and selection strategies for alternative recipes. This could involve investigating new methods for prioritizing options based on user preferences, dietary restrictions, and taste profiles. Finally, while this work focused on a limited set of LLM-generated candidate recipes, a promising area for future exploration lies in a more comprehensive analysis of LLM integration. Investigating how LLMs can be leveraged to not only retrieve but also generate novel recipes that adhere to HeaSE's health and sustainability principles holds significant potential for future advancements.

ACKNOWLEDGMENTS

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) 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

REFERENCES

[1] Felix Bölz, Diana Nurbakova, Sylvie Calabretto, Armin Gerl, Lionel Brunie, and Harald Kosch. 2023. 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* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3604915.3609491

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Ayoub El Majjodi, Alain D Starke, and Christoph Trattner. 2022. Nudging towards health? examining the merits of nutrition labels and personalization in a recipe recommender system. In *Proceedings of the 30th ACM conference on user modeling, adaptation and personalization*. 48–56.

[4] David Elsweiler, Morgan Harvey, Bernd Ludwig, and Alan Said. 2015. Bringing the "healthy" into Food Recommenders. In *International Workshop on Decision Making and Recommender Systems*. https://api.semanticscholar.org/CorpusID:1838398

[5] Ignazio Gallo, Nicola Landro, Riccardo La Grassa, and Andrea Turconi. 2022. Food Recommendations for Reducing Water Footprint. *Sustainability* 14, 7 (2022). https://doi.org/10.3390/su14073833

[6] Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware Food Recommender System. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) *(RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 333–334. https://doi.org/10.1145/2792838.2796554

[7] Christina Hartmann, Gianna Lazzarini, Angela Funk, and Michael Siegrist. 2021. Measuring consumers' knowledge of the environmental impact of foods. *Appetite* 167 (2021), 105622.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[9] World Health Organization. 2020. Healthy diet. https://www.who.int/news-room/fact-sheets/detail/healthy-diet.

[10] Divya Pandey, Madhoolika Agrawal, and Jai Shanker Pandey. 2011. Carbon footprint: current methods of estimation. *Environmental monitoring and assessment* 178 (2011), 135–160.

[11] Tashina Petersson, Luca Secondi, Andrea Magnani, Marta Antonelli, Katarzyna Dembska, Riccardo Valentini, Alessandra Varotto, and Simona Castaldi. 2021. A multilevel carbon and water footprint dataset of food commodities. *Scientific data* 8, 1 (2021), 127.

[12] Irtiqa Shabir, Kshirod Kumar Dash, Aamir Hussain Dar, Vinay Kumar Pandey, Ufaq Fayaz, Shivangi Srivastava, and Nisha R. 2023. Carbon footprints evaluation for sustainable food processing system development: A comprehensive review. *Future Foods* 7 (2023), 100215. https://doi.org/10.1016/j.fufo.2023.100215

[13] Chun-Yuen Teng, Yu-Ru Lin, and Lada A Adamic. 2012. Recipe recommendation using ingredient networks. In *Proceedings of the 4th annual ACM web science conference*. 298–307.

[14] Christoph Trattner and David Elsweiler. 2017. Food recommender systems: important contributions, challenges and future research directions. *arXiv preprint arXiv:1711.02760* (2017).

[15] Christoph Trattner and David Elsweiler. 2017. 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*. 489–498.

[16] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL]

[7] The number of candidates was set due to the length limitation of GPT prompts. Longer prompts with more candidates will be evaluated as future work.