FoodPrint: Emissions and Nutrition Analysis of Food Products with LLaMA3

Abstract

Food and agriculture account for 26% of total global greenhouse gas (GHG) emissions, according to research from Project Drawdown. Consequently, our everyday food choices can significantly impact the fight against climate change. This study aims to empower consumers with the information needed to make environmentally responsible and nutritionally informed food choices. Due to the lack of transparency in proprietary recipe information on food products, assessing their emission impact is challenging. To address this, we employ an open-source model, LLaMA3, to approximate the recipes of various food products. We then create a comprehensive emissions database for these products. Additionally, we calculate a nutrition score for each product using the updated NutriScore system. By combining GHG emissions data with nutritional scores, we provide a holistic view that helps consumers make more sustainable and health-conscious decisions.

1 Introduction

Food Life Cycle Assessment (LCA) is an emerging area of research that involves evaluating the sustainability of the supply chain for food commodities. From a study conducted by Poore and Nemecek (2018), the total emissions generated by our food supply chains can measure up to ~13.7 billion metric tons of carbon dioxide equivalents (CO2eq) contributing to 26% of total anthropogenic emissions. There has been an increased consumer awareness regarding the environmental impact of their food choices which has led to consumers preferring local or regional foods (Nemecek et al. (2016)) and looking for ecolabels, to guide their decision-making. However, Potter et al. (2021) argues that food selection based solely on nutrient profiling is insufficient; it is essential to consider the emissions associated with consumer-driven choices as well.

Food profiling is an interesting problem but comes with its own limitations, especially for complex food items (complex foods are food items that contain more than one ingredient). Clark et al. (2022) studies some of these challenges such as 1) complex supply chain - each ingredient could be sourced from a different part of the world, and without a transparent supply chain, it's difficult to calculate the emission of the ingredient accurately, 2) proprietary information - companies do not share the proprietary information like ingredients and their proportions which makes it difficult to calculate the total emission of each product.

To overcome the challenges of proprietary recipes, our approach leverages open-source language models like LLaMA3 (Touvron et al. (2023)) to estimate recipe of various food products As depicted in Figure 1, we utilize "USDA Global Branded Food Products Database" (The U.S. Department of Agriculture (2019)) which contains detailed information on approximately 400,000 store bought food products, along with their ingredients and nutrition labels. This data is integrated with the total emissions data and nutritional score for each product. The nutritional score is calculated by using the updated NutriScore metric (Sarda et al.). Although there are many alternate metrics as discussed by Drewnowski (2005), most of them are complex to calculate and require detailed prior knowledge, such as daily value for each nutrient. The resulting data, including each product's NutriScore and emission score, is stored in a query-able vector database, so when a user scans any product, related products with lower impact but more nutritional value would be retrieved from the database.

Our main contribution through this paper is to 1) showcase how LLMs can help bridge the gap

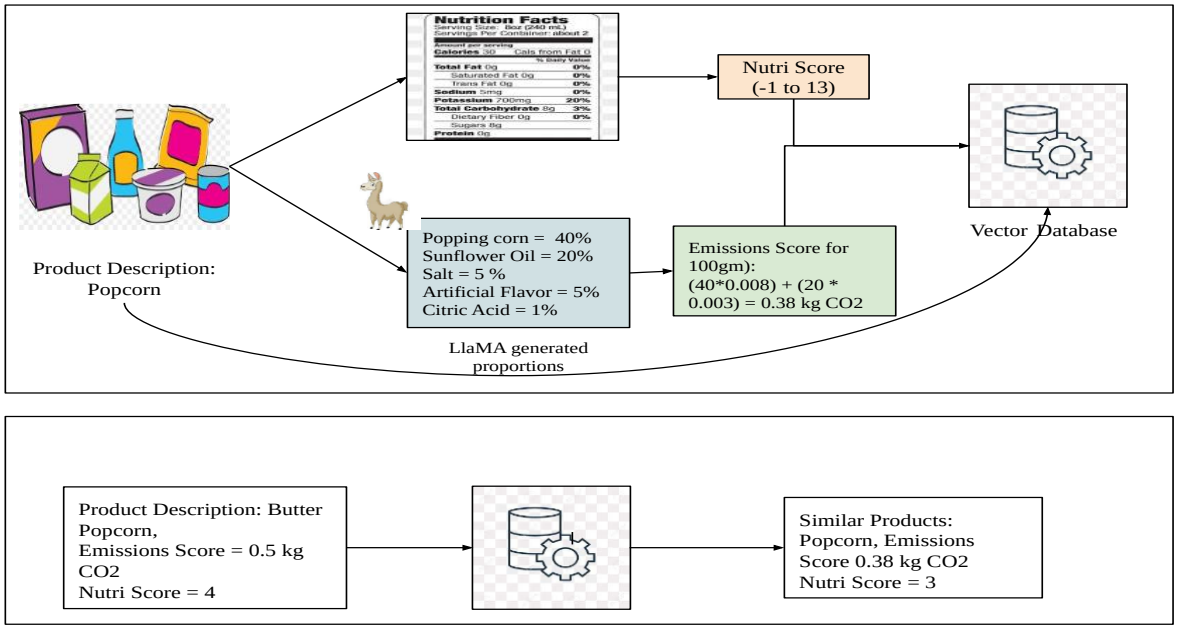

Figure 1: Data collected from USDA is used to generate the NutriScore and Emission score for each food product. This is stored in a query-able vector database.

between the known and the unknown factors of the food system and 2) to develop an application that can help consumers make healthy and sustainable food choices.

2 Literature Review

Stylianou et al. (2021) discusses how consumers need to make dietary changes to have sustainable food systems. Most of the consumers make their food choices either based on the product's price or its nutritional value since those are the only two pieces of information available on the packaging. Muzzioli et al. (2023) argues that a single perspective approach is not sustainable, and we need to consider the environmental impact of various food products. Having said that, calculating a life cycle assessment of a composite food is an arduous process as there is a lack of transparency in the food supply chain.

Most of the previous studies focus on getting the impact score for basic food commodities like fruits. This is not very helpful as most of the products that a consumer buys are composite food items like canned foods or frozen foods, which are a combination of more than one item. Clark et al. (2022) developed an algorithm to populate the environmental impacts of food products by using publicly available information. But their methodology requires some datapoints with their respective food composition, to estimate proportions for rest of the products. In the US, this data is not required nor available and hence is a huge roadblock when it comes to calculating impact score.

Our method uses LLaMA3 70b instead of GPT4, since it is currently the best open-source model, meticulously trained on large amount of internet data. This extensive training enables LLaMA3 to generate accurate food proportions by utilizing its comprehensive prior knowledge. We also use these models to source the greenhouse gas emission (in kg CO2eq) of various ingredients, rather than fetching the data manually.

3 Method

This section outlines the data collection process to calculate emission and nutritional scores. To standardize the process, both metrics are calculated on 100gm of any product, which makes it easier to compare products.

3.1 Emissions Data

To collect the emission data, the first step is to collect the proportion of each ingredient. This is done by prompting LLaMA3 to generate an approximate percentage of each ingredient present in a product. Figure 2 illustrates a sample prompt and the corresponding response generated by LLaMA3. To validate these results, we used ChatGPT (Brown et al. (2020)), which has the

capability to reference real-time data, and used the same prompt to generate a response. The Root Mean Square Error (RMSE) score between the ingredient proportions generated by LLaMA3 and ChatGPT is shown in Table 1 under *LLaMA generated proportions*. We only take into account the ingredients which had proportion > 5% to calculate the total Emission Score of the product. Using this method, recipes for 30,000 products were collected, with ~30,000 different ingredients. The biggest challenge in this process is bucketing similar ingredients together, since one ingredient can be written in several ways. To overcome this, ingredients were lemmatized and then bucketed together based on Levenstein distance. The outcome of the bucketed list was still not comprehensive as ingredients like *apricot* and *dried apricot* were not part of the same group. A list of 4000 ingredients (after bucketing) was passed through LLaMA3, to get a final list of 200 unique ingredients.

The next step is to collect emissions data for each of these ingredients. The main source for the ingredient level emissions data was "Our World in Data" (Ritchie et al. (2022)), which provides emissions data for 250 food items. To estimate the emissions for all the other ingredients that are not covered in the above dataset, we prompt LLaMA3

Prompt:
What is the quantity in % of each of these ingredients in GREEN & BLACK'S ORGANIC CHOCOLATE BAR SEA SALT. Ingredients: Ingredients: Cocoa mass, raw cane sugar, cocoa butter, Anglesey sea salt (0.4%), emulsifier (soya lecithin), vanilla extract.
Output:

```
{
  "cocoa mass": (30, 40),
  "raw cane sugar": (30, 40),
  "cocoa butter": (20, 30),
  "Anglesey sea salt": (0.2, 0.6),
  "emulsifier (soya lecithin)": (1, 5),
  "vanilla extract": (1, 5)
}
```

Figure 2: Example of a prompt given as input to LLaMA model.

to generate approximate emission (which includes transportation, packaging, and production). For accurate assessment of the emission, one needs to consider the LCA for each ingredient, but that data is not easily available. We test the accuracy of the model's generated emission score with data collected from the internet, and the RMSE score is reported in Table 1 under *GHG Score from LLaMa3*.

To calculate the final emission score for 100gm of a product, a weighted sum is calculated for each ingredient and its emission score, as shown in eq 1. Here p_j is the proportion, and e_j is the emission (of 1gm of the ingredient) for ingredient j where $j \in Z$.

$$Emission\ Score = \sum_j^n p_j * e_j \qquad (1)$$

3.2 Nutritional Value

The nutritional value of a food product can be determined by using the nutrition label on the packaging. The USDA Food database provides nutrition label information for each of the products. This database contains 477 unique nutrients such as Protein, Fiber, Total Sugar, Fats, various Vitamins and Minerals etc.

For this study, we employ the NutriScore metric system, which assigns a numerical score between -15 to 40 to profile foods based on their nutritional quality. These scores can be categorized into 5 buckets: A (<= -1), B (0 to 2), C (3 to 10), D (11 to 18), and E (19 to 40). NutriScore is calculated based on the quantity of various nutrients present in a product. It gives negatives score for nutrients that should be consumed in less quantity such as fat, saturated fat, sugar, and sodium and awards positive scores for beneficial components like protein, fibers and fruits and vegetables. Although the score has its limitations, like it only considers limited nutrients, or it does not factor in if the food is ultra processed, it is still overall a simple way to calculate the health benefit for a food product.

Given that the nutritional value of each product is already stored in the database, we can directly apply the NutriScore algorithm to compute the final nutrition score for each product.

3.3 Vector Embedding

Vector embedding is a way to represent data in latent space, such that similar data points are closer to each other. The vector database can query and retrieve similar products quickly and allows the flexibility of filtering on metadata. We generate the vector embeddings of the food descriptions using a basic sentence-transformer model (*all-MiniLM-L6-v2*), introduced by Reimers and Gurevych (2019).

These embeddings are then stored in a vector database. As explained by Han et al. (2023), a vector database is used to store and retrieve high-dimensional vector embedding of unstructured data quickly and accurately. ChromaDB, along with their respective NutriScore and Emission Scores. The final data that goes in the vector data base is,

$\sum_{i=n}\{s_i, metadata: \{n_i,\ e_i\}\}$; where s is the description, e is the Emission Score, and n is the NutriScore for a food item.

4 Results

To the best of our knowledge, no comprehensive dataset exists that store both the emissions impact and nutritional score of composite foods. Therefore, to evaluate our methodology, we picked simple examples where comparative emissions of various food products would be common knowledge and manually labelled them as correct or incorrect. For instance, a pepperoni pizza will have a higher emission score than a veggie pizza, since vegetables have a smaller environmental impact than pepperoni.

Figure 3 shows the 10 most commonly occurring ingredients for products with low emission score (10th percentile) and high nutrition score (10th percentile, low NutriScore is better). On the other hand, Figure 4 shows 10 commonly occurring ingredients for products with high emission score (90th percentile) and low nutrition score (90th percentile, high NutriScore is not good). From these figures one can understand that plant-based food like vegetables and fruits usually have low environmental impact and are high in nutritional value, whereas meat based or dairy products have very high environmental impact.

Task	Metric	Score
GHG Score from LLaMa3	RMSE	5.44
LLaMA3 generated proportions	RMSE	6.38

Table 1: Metrics

5 Conclusion

This study presents a methodology to estimate the environmental and nutritional impacts of food products using open-source large language models, due to the absence or lack of availability of comprehensive datasets in this domain. Recognizing the significant role of the food and agriculture sector in global greenhouse gas emissions, our approach equips consumers with the necessary information to make sustainable and health-conscious choices.

This study focuses on the GHG emissions of a product, although there are other aspects that need to be considered. In the future we will focus on augmenting the data with more environmental factors such as water usage, land usage, effect on biodiversity etc.

This paper is particularly focused on driving decisions based on the impacts and nutrition, but most of us choose food products based on their price. Incorporating the price point with the impact scores can make a true impact when it comes to consumers' choices.

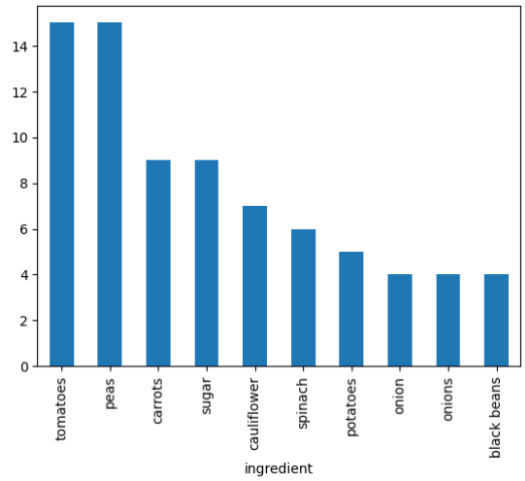

Figure 3: most commonly occurring ingredients for products with low emission score and low NutriScore

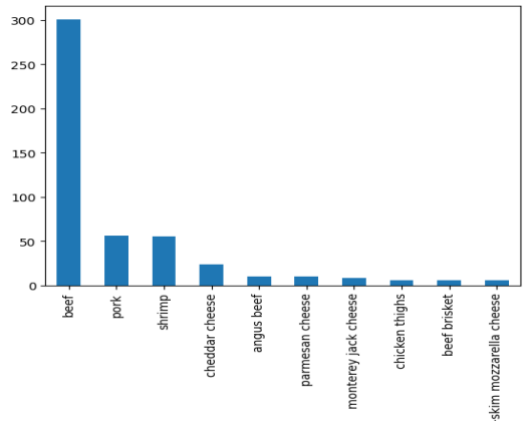

Figure 4: most commonly occurring ingredients for products with high emission score and high NutriScore

References

Ahn, Chiyoung, and Chung Gun Lee. "*Effect of NUTRI-SCORE Labeling on Sales of Food Items in Stores at Sports and Non-Sports Facilities.*" Preventive Medicine Reports, vol. 29, 21 July 2022, p. 101919, pubmed.ncbi.nlm.nih.gov/35911572/, https://doi.org/10.1016/j.pmedr.2022.101919.

Brown, Tom B., et al. "*Language Models Are Few-Shot Learners*." Arxiv.org, vol. 4, 28 May 2020, arxiv.org/abs/2005.14165.

Clark, Michael, et al. "*Estimating the Environmental Impacts of 57,000 Food Products.*" Proceedings of the National Academy of Sciences, vol. 119, no. 33, 8 Aug. 2022, www.pnas.org/doi/10.1073/pnas.2120584119, https://doi.org/10.1073/pnas.2120584119.

Clark, Michael A, et al. "*Multiple Health and Environmental Impacts of Foods.*" Proceedings of the National Academy of Sciences, vol. 116, no. 46, 28 Oct. 2019, p. 201906908, https://doi.org/10.1073/pnas.1906908116.

Drewnowski, Adam. "*Concept of a Nutritious Food: Toward a Nutrient Density Score.*" The American Journal of Clinical Nutrition, vol. 82, no. 4, 1 Oct. 2005, pp. 721–732, https://doi.org/10.1093/ajcn/82.4.721. Accessed 24 May 2020.

Han, Yikun, et al. "*A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge*." ArXiv.org, 18 Oct. 2023, arxiv.org/abs/2310.11703.

Muzzioli, Luca, et al. "*How Much Do Front-Of-Pack Labels Correlate with Food Environmental Impacts?*" Nutrients, vol. 15, no. 5, 26 Feb. 2023, p. 1176, https://doi.org/10.3390/nu15051176.

Potter, Christina, et al. "*The Effects of Environmental Sustainability Labels on Selection, Purchase, and Consumption of Food and Drink Products: A Systematic Review.*" Environment and Behavior, vol. 53, no. 8, 20 Feb. 2021, p. 001391652199547, journals.sagepub.com/doi/full/10.1177/0013916521995473, https://doi.org/10.1177/0013916521995473.

Reimers, Nils, and Iryna Gurevych. "*Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.*" ArXiv:1908.10084 [Cs], 27 Aug. 2019, arxiv.org/abs/1908.10084.

Ritchie, Hannah, et al. *"Environmental Impacts of Food Production.*" Our World in Data, 2022, ourworldindata.org/environmental-impacts-of-food.

Stylianou, Katerina S., et al. "*Small Targeted Dietary Changes Can Yield Substantial Gains for Human Health and the Environment*." Nature Food, vol. 2, no. 8, 1 Aug. 2021, pp. 616–627, www.nature.com/articles/s43016-021-00343-4, https://doi.org/10.1038/s43016-021-00343-4.

Touvron, Hugo, et al. "*LLaMA: Open and Efficient Foundation Language Models.*" ArXiv:2302.13971 [Cs], 27 Feb. 2023, arxiv.org/abs/2302.13971.

U.S. Department of Agriculture. "*FoodData Central.*" Usda.gov, 2019, fdc.nal.usda.gov/.

Vidergar, Petra, et al. "*A Survey of the Life Cycle Assessment of Food Supply Chains.*" Journal of Cleaner Production, vol. 286, no. 0959-6526, 1 Mar. 2021, p. 125506, https://doi.org/10.1016/j.jclepro.2020.125506.

Poore, Joseph, and Thomas Nemecek. *"Reducing Food's Environmental Impacts through Producers and Consumers."* Science, vol. 360, no. 6392, 1 June 2018, pp. 987–992, https://doi.org/10.1126/science.aaq0216.

Nemecek, Thomas, et al. *"Environmental Impacts of Food Consumption and Nutrition: Where Are We and What Is Next?"* The International Journal of Life Cycle Assessment, vol. 21, no. 5, 2 Mar. 2016, pp. 607–620, link.springer.com/article/10.1007/s11367-016-1071-3, https://doi.org/10.1007/s11367-016-1071-3.

Sarda, Barthélemy et al. *"Complementarity between the Updated Version of the Front-of-Pack Nutrition Label Nutri-Score and the Food-Processing NOVA Classification."* Public Health Nutrition 27.1 (2024): e63. Web.