Datasheet for NutriBench

Andong Hua* **Mehak Preet Dhaliwal*** **Yao Qin**
University of California, Santa Barbara
{dongx1997,mdhaliwal,yaoqin}@ucsb.edu

This document is based on a publicly available template [1] for *Datasheets for Datasets* by Gebru *et al.* [3].

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
This dataset was created to evaluate Large Language Models (LLMs) on the task of nutrition estimation from natural language meal descriptions.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Andong Hua, Mehak Preet Dhaliwal, and Yao Qin at the University of California, Santa Barbara.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
N/A.

Any other comments?
None.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance in the dataset represents a natural language meal description containing up to three food items. Each food item may be present in a single serving (e.g., "1 cup of tea") or multiple servings (e.g., "2 cups of tea"). The corresponding serving units may be natural (e.g., "1 cup") or metric (e.g., "20g"). The data is further divided into instances that may or may not be directly retrievable from an external knowledge database. An example of a single serving, single food item, natural serving unit and directly retrievable instance is "Enjoying a delicious cooked and broiled bison ribeye steak for lunch.". The data is arranged in separate files identifying the subset named in the following format: *serving_unit_retrieval_type_number_of_food_number_of_serving.csv*.

Each instance is labeled with the corresponding macronutrient breakdown for the meal, including protein, carbohydrates, fats, and calories.

How many instances are there in total (of each type, if appropriate)?
There are 5,000 total instances divided into 15 subsets varying in complexity based on the number of food items, serving amounts, serving units, and type of retrieval. The following table documents the number of instances in each subset.

	Number of Food Item	Single		Double		Triple
	Number of Serving	Single	Multiple	Single	Multiple	Single
Natural Serving	Direct Retrieval	500	500	500	500	500
	Indirect Retrieval	200	200	200	200	200
Metric Serving	Indirect Retrieval	300	300	300	300	300

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The food items in the dataset are sampled from the April 2024 version of the FoodData Central (FDC) [1]- the food composition information center of the US Department of Agriculture (USDA) [2]. We include both common and uncommon food items in the data instances to ensure wide coverage. The commonness score is measured using an embedding-based cosine similarity matrix constructed from all food items in the database using OpenAI's "text-embedding-3-large" model. Further details of our commonness-based sampling approach are included in our paper.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of a natural language meal description

[1] https://github.com/AudreyBeard/Datasheets-for-Datasets-Template

along with macro-nutrient labels for carbohydrates, proteins, fats, and calories. An example of an instance is:

Meal Description	For dinner, I am enjoying a refreshing orange slice and a cup of great northern beans.
Carbohydrate	115.08
Protein	40.05
Fat	2.09
Calories	624.05

Is there a label or target associated with each instance? If so, please provide a description.
Each instance is labeled with the corresponding meal description's carbohydrate, protein, fat, and calorie amount.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
N/A.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
The individual instances are not related.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
The dataset is created for testing.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
N/A.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
N/A.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
N/A.

Any other comments?
N/A.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The food items and nutrition labels associated with the meal description in each instance were obtained from the April 2024 version of the FoodData Central [1] (Data download page).

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The data was collected from the most recent version at the

time of writing (April 2024).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The data was directly downloaded from the official FoodData Central (Data download page).

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[4] for approaches in this area.)
Generating queries from the food items obtained through FoodData Central [1] involved accessing GPT-3.5 via the OpenAI API for inference. Overall, the cost of generating NUTRIBENCH queries was approximately $5.93.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The final NUTRIBENCH dataset is composed of a sample of the food items in the FoodData Central [1]. We include both common and uncommon food items in the data instances to ensure wide coverage. The commonness score is measured using an embedding-based cosine similarity matrix constructed from all food items in the database using OpenAI's "text-embedding-3-large" model. Further details of our commonness-based sampling approach are included in our paper.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process was done by the authors.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
N/A.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
N/A.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
N/A.

Any other comments?
None.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Data collected from the FoodData Central [1] was cleaned, filtered, sampled, and converted to natural language meal descriptions. The end-to-end data creation process is documented in our paper. All code for constructing the final instances from the raw data will be publicly released.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The raw data used for creating the final data instances was collected from the April 2024 version of the FoodData Central [1] (link).

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
The process used to preprocess/clean/label the instances is

documented in our paper. All code for constructing the final instances from the raw data will be publicly released.

Any other comments?
None.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
The dataset has been used to evaluate several LLMs on the task of carbohydrate estimation from natural language meal descriptions in our paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
The papers or systems using the dataset can be tracked via the NutriBench paper citations.

What (other) tasks could the dataset be used for?
The dataset can be used to evaluate LLMs on estimating other macro-nutrients including proteins, fats, and calories from the natural language meal descriptions. The estimation can be done separately or simultaneously for all macro-nutrients.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
N/A.

Are there tasks for which the dataset should not be used? If so, please provide a description.
N/A.

Any other comments?
None.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
N/A.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is publicly available and hosted on Github.

When will the dataset be distributed?
The dataset is publicly available as of June 8, 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is distributed under the Creative Commons Attribution Non Commercial Share Alike 4.0 license

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
N/A.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
N/A.

Any other comments?
None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is hosted at `https://github.com/DongXzz/NutriBench/`and maintained by Andong Hua, Mehak Dhaliwal, and Yao Qin.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Please contact Andong Hua or Mehak Dhaliwal at `{dongx1997,mdhaliwal}@ucsb.edu` or qinlab01@gmail.com

Is there an erratum? If so, please provide a link or other access point.
None.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
All updates will be communicated at the hosting site on GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
All updates will be communicated at the hosting site on GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Contributions to the dataset can be made via pull requests to the data Github repository `https://github.com/DongXzz/NutriBench/`

Any other comments?
None.

References

[1] Fooddata central. `https://fdc.nal.usda.gov/`. Accessed: 2024-06-05.

[2] Naomi K Fukagawa, Kyle McKillop, Pamela R Pehrsson, Alanna Moshfegh, James Harnly, and John Finley. Usda's fooddata central: what is it and why is it needed today? *The American journal of clinical nutrition*, 115(3):619–624, 2022.

[3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.