The Journal of Nutrition 156 (2026) 101712

American Society for Nutrition
Excellence in Nutrition Research and Practice

JN THE JOURNAL OF NUTRITION

journal homepage: https://jn.nutrition.org/

Nutritional Methodologies and Mathematical Modeling

Evaluation of Energy and Nutrient Estimates from Large Language Models Using Text-Based Queries

Razi Lawabni [1], Olivia R Solano [1], Leo L Kampen [1], Jonathan Kurasch [1], Nabil Alshurafa [2,3], Jung-Su Chang [4], Richard Evans [5], Dongji Feng [6,†], Annie W Lin [1,*,†]

[1] *The Hormel Institute, University of Minnesota, Austin, MN, United States;* [2] *Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States;* [3] *Department of Computer Science, McCormick School of Engineering, Northwestern University, Evanston, IL, United States;* [4] *School of Nutrition and Health Sciences, College of Nutrition, Taipei Medical University, Taipei, Taiwan, Republic of China;* [5] *Clinical and Translational Science Institute, University of Minnesota, Minneapolis, MN, United States;* [6] *School of Computing and Design, California State University, Monterey Bay, Seaside, CA, United States*

A B S T R A C T

Background: Large language models (LLMs) have emerged as promising tools for estimating energy and nutrient values, yet most existing evaluations focus on image-based queries rather than text-based queries. Few studies compare LLM estimates with reference databases commonly used in nutrition research.

Objectives: This study aimed to examine agreement between LLM estimates and a research food composition database for energy and nutrients and to determine if agreement varies by food group.

Methods: We conducted a cross-sectional analysis of energy and nutrient estimates for frequently consumed food items in the United States. Food items were entered as text prompts into 4 LLMs (ChatGPT 5.2, Claude Opus 4.5, Gemini 3 Pro Preview, and Llama 4 Maverick), which provided energy and nutrient estimates. Corresponding foods were matched to the Nutrition Coordinating Center Food and Nutrient Database, and agreement between LLM and database values was assessed using intraclass correlation coefficients and Bland–Altman analyses. Agreement was also evaluated within the 3 most frequently consumed food groups.

Results: Agreement was high for energy and macronutrients for all LLMs. Variability was observed for several micronutrients, particularly vitamin D, folate, and iron. Claude Opus 4.5 showed consistently high agreement, with no nutrients classified as poor. Other LLMs exhibited poor agreement for ≥1 micronutrient. Certain food categories, including condiments and mixed dishes, contributed disproportionately to variability. However, agreement remained high within the most frequently consumed broader food groups.

Conclusions: LLMs show promise for estimating energy and macronutrients. However, performance for micronutrients requires further improvement and may affect overall dietary assessment.

Keywords: dietary assessment, mobile health, reliability, artificial intelligence, Nutrition Care Process, just-in-time adaptive interventions

Introduction

Noncommunicable diseases such as obesity, diabetes, and cardiovascular disease are partially managed through adherence to high-quality dietary patterns that meet individual energy requirements [1]. Nutrition interventions designed to improve dietary habits are increasingly focused on delivering support at the right time and with an appropriate level of intensity [2,3]. These just-in-time adaptive interventions (JITAIs) are frequently implemented through digital health platforms [4–6]. Growing interest has emerged in the application of artificial intelligence (AI) to deliver JITAIs, reflecting AI systems' capacity to *1*) process user inputs, *2*) identify behavioral patterns, *3*) generate personalized feedback, and *4*) support preference-tailored

Abbreviations: AI, artificial intelligence; ICC, intraclass correlation coefficient; ChatGPT, Chat Generative Pretrained Transformer; JITAI, just-in-time adaptive intervention; LLM, large language model; NCC, Nutrition Coordinating Center; NDSR, Nutrition Data System for Research; NFS, not further specified; WWEIA, What We Eat in America.

This article is part of a special issue entitled: AI, ML, and Precision Nutrition published in The Journal of Nutrition.

* Corresponding author. *E-mail address:* awlin@umn.edu (A.W. Lin).

† DF and AWL contributed equally as senior authors.

https://doi.org/10.1016/j.tjnut.2026.101712

Received 16 February 2026; Received in revised form 26 June 2026; Accepted 1 July 2026; Available online 4 July 2026

conversational interactions [7–9]. The effectiveness of these interventions depends on reliable dietary assessment to determine the appropriate timing and level of support [10].

Dietary assessment methods can identify diet patterns, nutrient gaps, and behavioral habits that signal when an intervention is needed [11]. The Nutrition Care Process provides a structured framework to improve dietary intake and places equal emphasis on assessment and intervention [12]. Evidence from earlier digital health tools demonstrates that self-monitoring features support behavior change, but their impact partly relies on the reliability of the food and nutrient data [13–16]. Before the emergence of AI conversational systems, commercial mobile health apps relied on structured diet tracking features that drew from proprietary and open-source food composition databases [17]. Researchers have reported that nutrient estimates vary across mobile health applications and across food groups compared with a reference database [17, 18]. Further examination determined that this variability contributes to differences in the underlying food composition databases used by these applications. The quality and transparency of these databases vary across applications, and discrepancies in database content contribute to differences in nutrient values. As interest in AI-supported nutrition interventions grows, similar evaluation of nutrient estimates generated by large language models (LLMs) against established reference databases is warranted before implementation.

As a type of AI conversational tool, LLMs can generate timely dietary guidance, as well as energy and nutrient estimates from reported food intake [19–22]. These models are trained on large datasets to capture basic patterns, in addition to reinforcement learning from human feedback to provide accurate, evidence-based guidance [23–25]. Dietary assessment research using AI has primarily focused on image-based methods to estimate dietary intake. Prior studies have designed and trained open-source multimodal LLMs to identify and categorize foods [26–28]. Relatively fewer studies have evaluated the agreement of popular closed-source conversational models with established food composition databases, especially for text-based queries [29–33]. Closed-source models are proprietary systems developed by commercial entities, and details such as training data and model architecture are not publicly accessible. Evaluating closed-source LLMs using text-based queries, or prompts, is crucial as this format closely reflects real-world use in clinical and health settings [34]. This approach allows for more practical performance assessment and helps establish benchmarks for their use in specialized domains, thus guiding future development and validation. Systems such as Chat Generative Pre-trained Transformer (ChatGPT; OpenAI), Gemini (Google), and Claude (Anthropic) are marketed as being user-friendly and accessible with a large user base [35–40]. Yet, there is limited transparency about the quality of the training data and underlying model architecture that informs the dietary output.

An additional consideration when evaluating the performance of these models is whether estimations differ by food group. Prior studies evaluating AI-supported dietary assessment have reported energy and nutrient values for an overall meal, rather than examining agreement at the food-item level [29–33]. Item-level analyses allow the opportunity for more granular investigation into potential sources of variability in LLM estimates. We hypothesize that certain categories of foods may be more susceptible to inaccurate nutrient estimation, given differences in nutrient composition across food groups [41]. The primary objective of the current study was to investigate the agreement between energy and nutrient estimates by LLMs and a research-based food composition database using commonly consumed foods in the United States. A secondary objective was to evaluate how model performance varied across food group categories.

Methods

Study design

This cross-sectional study used data from What We Eat in America (WWEIA), the dietary intake component of the NHANES [42]. Participants were asked to complete ≤two 24-h dietary recalls, with the first one conducted in-person at the Mobile Examination Center and the second one administered by phone on a nonconsecutive day. Collectively, these recalls provide population-level estimates of dietary intake. The WWEIA dietary data include an "Individual Foods" file containing line-by-line records of all foods and beverages consumed in the previous 24 h and a "Total Nutrient Intake" file aggregating daily energy and nutrient intake. Nutrient values are derived using the Food and Nutrient Database for Dietary Studies, which provides the nutrient composition for each reported food code. Because NHANES dietary intake data are publicly available and deidentified, this study did not require institutional review board approval.

Identification of most frequently consumed foods in the United States

Data from the NHANES 2017 to March 2020 prepandemic cycle were used in the current analyses [43]. Most frequently consumed foods were identified using day 1 dietary recall data to avoid disproportionate weighting of participants with 2 recalls. Each food item was summed by frequency and ranked from most to least frequently consumed. The USDA food codes reported in the WWEIA dataset were mapped to standardized food names using the Food Patterns Equivalents Database food code key. The 200 most frequently consumed items were selected for analysis to capture a proportion of commonly reported foods in the United States diet while maintaining a manageable set across models [44]. To refine the list, the preliminary list of foods was reviewed by the primary author to create a standardized protocol where food items were excluded only if a more specific type of food was already present in the list (e.g., "Not Further Specified, NFS"). The removal of the items was discussed with a registered dietitian before analysis, who agreed with the changes. This qualitative validation served as a sensitivity check to determine that the final selection accurately represented the most common contributors to the United States diet without redundancy. Six items were removed: "Cheese, NFS," "Rice, white, cooked, NS as to fat," "Milk, NFS," "Lemon juice, 100%, NS as to form," "Salsa, NFS," and "Bacon, for use on a sandwich." The item, "Pork bacon, NS as to fresh, smoked or cured, cooked," was retained because it captured multiple preparation types and reflected broader use.

The final analytic dataset included 194 unique foods (Supplemental Table 1), with report frequencies ranging from 135 to

11,397 reports for a single day across the study sample. Notably, bottled and tap water each exceeded 10,000 reports, followed by brewed coffee ($n = 3542$). Each item was assigned to a major food group based on the Nutrition Coordinating Center (NCC) definitions: "Fruits, Vegetables, Grains, Dairy and Nondairy Alternatives, Protein (Meat, Fish, Poultry, Eggs, Nuts, Seeds, and Meat Alternatives), Fats, Sweets, Beverages, and Miscellaneous Foods" [45]. A "Mixed Dishes" group was created to reflect dishes that are typically queried as single dishes (e.g., pizza). All food group assignments underwent a final review by an investigator with expertise in nutrition.

Extraction of nutrient values from the comparator database

The 2025 NCC Food and Nutrient Database served as the comparison database in this study, which was developed by the NCC, University of Minnesota, Minneapolis, MN. The database contains ~19,500 food items, 42% of which are branded products [45]. This database was selected for its use in research and mobile health applications, transparent data sources, and inclusion of single ingredients and mixed dishes [46,47]. Each food item from the final analytical dataset was matched to foods in the NCC database. Exact or close NCC food equivalents were identified by a single team member for standardization and independently verified by a registered dietitian [18]. Exact matches were defined as items with near-identical descriptors in both NHANES and NCC database. Close matches accounted for ~9% of the food items; these occurred when the NCC description was less specific than the NHANES options, particularly regarding fruit juice concentrations or preparation methods for bread (Supplemental Table 2). For these items, systematic rules were applied to prioritize the most general version of the food or beverage. Any discrepancies in food matching were resolved through consensus discussion ($n = 1$). Items unavailable as standalone entries in the NCC database were constructed as recipes by a registered dietitian using the Nutrition Data System for Research (NDSR), version 2024. Energy and nutrient values per 100 g for each food item were extracted and used for analysis. Nutrients examined included those mandated on the nutrition facts label and had public health relevance such as total carbohydrate, total protein, total fat, dietary fiber, total sugars, added sugars, saturated fat, monounsaturated fat, polyunsaturated fat, cholesterol, sodium, vitamin D, calcium, iron, potassium, and total folate. Trans fat was excluded because artificial sources have been largely removed from the food supply [48].

Extraction of nutrient values from LLMs

Four LLMs that were available at the time of this study, namely, ChatGPT 5.2, Claude Opus 4.5, Gemini 3 Pro Preview, and Llama 4 Maverick, were evaluated. The first 3 are closed-source commercial systems, whereas Llama 4 Maverick is an open-weight model with publicly available weights but limited transparency regarding training data. All models were accessed from 13 January, 2026, to 18 January, 2026, through the OpenRouter platform [35–37,49,50]. The platform's default temperature parameter of 1.0 (on a scale of 0.0 to 2.0) was used to preserve standard model behavior. Each LLM received a comma-separated values file containing the list of food items, along with the following initial prompt:

> "You are a nutrition expert. Estimate the nutritional content for a 100 g serving of each of the top 194 foods in the USA. Your csv should include: First column with the food item name. From column 2-15, they should be titled in the following order, with each row giving the estimated nutrition value for each food item: Calories (kcal); Total carbohydrate (g); Total protein (g); Total fat (g); Dietary fiber (g); Total sugars (g); Added sugars (g); Saturated fat (g); Monounsaturated fat (g); Polyunsaturated fat (g); Trans fat (g); Cholesterol (mg); Sodium (mg); Vitamin D (mcg); Calcium (mg); Iron (mg); Potassium (mg); Total Folate (Mcg DFE). In addition to the above you will be evaluated on the following: Correctness of nutritional content estimate for the 100 g serving. Generate the complete CSV for all 194 foods in a single response, do not stop partway, and do not provide commentary or additional messages, only output the full table with all rows and columns as specified."

This prompt was developed based on the Turn, Expression, Level of Details, and Role (TELeR) taxonomy and reviewed by investigators with expertise in prompt engineering [51]. Energy and nutrient values per 100 g were exported from the LLM outputs as comma-separated value files. The LLM output used the provided food names accurately. Therefore, no manual corrections or adjustments were necessary to match the output to the original food list. After nutrient estimates were generated, the following prompt was submitted to request justification from each LLM about how it derived its estimates:

> "Now, please provide the references or data sources used to generate these nutritional content estimates."

A clarification prompt was submitted to ChatGPT only, as the meaning of the term "hand-estimated" was unclear: "What do you mean by "hand-estimated"? Please provide the references or data sources you actually used while generating these nutrition estimates (if none were consulted at generation time, say so explicitly), and then describe the logic/estimation methodology you applied."

Additional details on each model, including reported data sources and estimation approach, are provided in Table 1.

Statistical analysis

All analyses were conducted in R (version 4.5.2; R Core Team, 2025) using the readxl, dplyr, tidyr, stringr, irr, ggplot2, plotly, and htmlwidgets packages [52–60]. Agreement for each nutrient between LLMs and NCC database was evaluated with a 2-way random effects intraclass correlation coefficients (ICCs) model to determine absolute agreement for single measures. This approach was chosen to maximize the generalizability of the findings by treating the selected LLMs as representative of AI technology, with the food list as a representative sample of common dietary choices. Absolute agreement was also chosen as a more conservative metric than consistency. ICC values were categorized as excellent ($\geq$0.90), good (0.75–$<$ 0.90), moderate (0.50–$<$ 0.75), and poor ($<$0.50) [61]. Overall ICCs assessed agreement across all LLMs and the NCC database to evaluate intermethod reliability, and pairwise ICCs examined agreement between each LLM and the NCC database [62,63]. Bland–Altman plots were used to evaluate systematic bias and limits of agreement for energy, total carbohydrate, protein, and total fat,

TABLE 1
Data source and estimation logic reported by each large language model

		ChatGPT 5.2	Claude Opus 4.5	Gemini 3 Pro Preview	Llama 4 Maverick
Data sources	Similarities	USDA FoodData Central			
	Differences	No real-time database access Values generated from internal training patterns consistent with Nutrition Facts label formats	FNDDS (2017–2020) Branded Foods FDA labeling/standards (self-reported) NCC Food and Nutrient Database Manufacturer nutrition information (selected brands) USDA technical reports/ Handbook 8 (general) Dietary Guidelines 2020–2025 (general) Standard Reference (SR Legacy/SR)	Branded Foods Standard Reference (SR Legacy/SR)	FNDDS (2015–2016) Manufacturer labels Peer-reviewed literature USDA: Added Sugars Database (Release 1, 2016) ORAC[6] Database (Release 2, 2010) Standard Reference (SR Legacy/SR) NHANES (2015–2016)
Estimation assumptions	Similarities	Standardized to per 100 g, assumed a generic version of each food			
	Differences	Assumed a typical United States version, based macros on common food patterns, split natural vs. added sugars by typical recipes, estimated fat types and sodium by food category, assigned cholesterol only to animal foods, estimated key vitamins/minerals from common sources, set trace amounts to 0	Trace amounts listed as 0, adjusted for preparation and cooking method, added sugars estimated from "typical formulations"	Composite averaging for broad food descriptions from a variety of brands, nutrients assumed absent set to 0	Direct extraction from reference databases, imputation from similar foods, and statistical modeling

Abbreviations: FDA, Food and Drug Administration; FNDDS, Food and Nutrient Database for Dietary Studies; NCC, Nutrition Coordinating Center; ORAC, Oxygen Radical Absorbance Capacity; SR, standard reference, legacy release.

as these macronutrients contribute to total energy content [64, 65].

Food variability was examined for its impact on agreement between the LLMs and the NCC database. In the Bland–Altman analysis, food items that fell outside the limits of agreement were categorized by predefined food groups to identify potential contributors to variability. ICCs were also calculated for the most frequently consumed food groups to determine whether agreement differed by food group for energy. Macronutrients relevant to each food group were selected based on the predominant nutrient profile of that group. Supplemental Table 1 lists the food items within each food group to provide context for the selected nutrients and support these classifications.

Results

The greatest proportion of items was classified in the Grains group (22%), followed by Beverages (18%) and Fruits (11%). The smallest proportions were observed in the Miscellaneous Foods, Mixed Dishes, and Sweets food groups (5% for each group). Supplemental Table 1 provides the distribution of items across food groups and the list of food items included in the analysis. No energy or nutrient data were missing across LLMs. Four foods not found in the NCC Food and Nutrient Database were prepared as per recipes in the NDSR software to obtain nutrient information: *1*) iced, brewed black tea, unsweetened; *2*) iced, bottled green tea; *3*) iced, brewed black tea, presweetened with sugar; and *4*) white rice, cooked, prepared with oil. The following sections will describe these results: *1*) comparison of model outputs against the NCC database for baseline accuracy; *2*) measurement of agreement and systematic biases in nutrition estimation; and *3*) comparison of model agreement within the 3 most frequently consumed food groups.

When evaluating overall reliability pooled across all 5 sources (4 LLMs and NCC database), several nutrients demonstrated excellent agreement: energy; macronutrients including total carbohydrates, protein, and total fat; fiber; SFAs; and micronutrients including cholesterol, sodium, calcium, and potassium (overall ICC range: 0.90–0.99). Total and added sugars, MUFAs and PUFAs, and iron showed good agreement (overall ICC range: 0.77–0.88). Folate and vitamin D had moderate agreement pooled across all analyzed models and the NCC database (overall ICC range: 0.50–0.69).

Further analyses evaluated nutrient estimates generated by each LLM against the NCC database as a clinically relevant reference standard. Overall, agreement was excellent or good for energy and most nutrients across models, with poorer agreement primarily observed for vitamin D, total folate, and iron. Gemini 3 Pro Preview demonstrated excellent agreement for energy and most nutrients, with good agreement for MUFAs and moderate agreement for PUFAs and vitamin D (Table 2); however, total folate showed poor agreement (ICC = 0.48; 95% CI: 0.36, 0.58). Claude Opus 4.5 exhibited excellent agreement for most nutrients and did not demonstrate poor agreement for any nutrient (Table 2), although agreement ranged from good to moderate for total and added sugars, MUFAs and PUFAs, vitamin D, and folate (ICC range: 0.52–0.83). ChatGPT 5.2 showed excellent or good agreement for most nutrients, with

TABLE 2
Intraclass correlation coefficients between each large language model and the nutrient coordinating center database[1]

Nutrient	ChatGPT 5.2	Claude Opus 4.5	Gemini 3 Pro Preview	Llama 4 Maverick
Calories (kcal)	0.95 (0.93–0.96)	0.96 (0.95–0.97)	0.95 (0.93–0.96)	0.93 (0.91–0.95)
Total carbohydrate (g)	0.89 (0.86–0.92)	0.96 (0.94–0.97)	0.99 (0.99–0.99)	0.88 (0.85–0.91)
Total protein (g)	0.98 (0.98–0.99)	0.98 (0.98–0.99)	0.99 (0.98–0.99)	0.97 (0.97–0.98)
Total fat (g)	0.92 (0.89–0.94)	0.92 (0.89–0.94)	0.90 (0.87–0.93)	0.90 (0.87–0.92)
Dietary fiber (g)	0.97 (0.96–0.98)	0.97 (0.96–0.98)	0.97 (0.95–0.97)	0.92 (0.89–0.94)
Total sugars (g)	0.84 (0.79–0.88)	0.83 (0.79–0.87)	0.97 (0.96–0.98)	0.81 (0.75–0.85)
Added sugars (g)	0.82 (0.77–0.86)	0.74 (0.67–0.80)	0.90 (0.87–0.93)	0.79 (0.72–0.83)
Saturated fat (g)	0.90 (0.87–0.92)	0.95 (0.94–0.96)	0.94 (0.92–0.96)	0.92 (0.90–0.94)
Monounsaturated fat (g)	0.78 (0.71–0.83)	0.84 (0.79–0.88)	0.87 (0.83–0.90)	0.86 (0.82–0.89)
Polyunsaturated fat (g)	0.70 (0.62–0.76)	0.75 (0.68–0.81)	0.72 (0.65–0.78)	0.75 (0.68–0.81)
Cholesterol (mg)	0.95 (0.94–0.97)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.98 (0.97–0.98)
Sodium (mg)	0.95 (0.93–0.96)	0.97 (0.96–0.98)	0.97 (0.97–0.98)	0.93 (0.91–0.95)
Vitamin D (μg)	0.23 (0.09–0.35)	0.59 (0.49–0.67)	0.66 (0.57–0.73)	0.33 (0.20–0.45)
Calcium (mg)	0.86 (0.82–0.89)	0.92 (0.89–0.94)	0.90 (0.87–0.93)	0.85 (0.80–0.88)
Iron (mg)	0.97 (0.95–0.97)	0.96 (0.94–0.97)	0.96 (0.95–0.97)	0.34 (0.21–0.46)
Potassium (mg)	0.95 (0.94–0.96)	0.96 (0.95–0.97)	0.96 (0.94–0.97)	0.88 (0.84–0.91)
Total folate (μg DFE)	0.44 (0.32–0.55)	0.52 (0.41–0.61)	0.48 (0.36–0.58)	0.48 (0.36–0.58)

Abbreviation: DFE, dietary folate equivalent.
[1] Values reported are intraclass correlation coefficients (95% limits of agreement). $n = 194$ food items included in the analysis.

moderate agreement for PUFAs (ICC = 0.70; 95% CI: 0.62, 0.76), but poor agreement for total folate (ICC = 0.44; 95% CI: 0.32, 0.55) and vitamin D (ICC = 0.23; 95% CI: 0.09, 0.35). Llama 4 Maverick demonstrated excellent or good agreement for energy and most nutrients (ICC range: 0.75–0.98) but had poor agreement for vitamin D (ICC = 0.33; 95% CI: 0.20, 0.45), total folate (ICC = 0.48; 95% CI: 0.36, 0.58), and iron (ICC = 0.34; 95% CI: 0.21, 0.46).

Bland–Altman analyses for energy showed patterns consistent with the ICC results. Mean differences ranged from –0.09 to 3.47 kcal across models (Figure 1). Despite small mean differences, the 95% limits of agreement were wide and exceeded ±100 kcal. Such limits indicate large variability in individual food estimates relative to reference values, despite minimal overall bias. Foods contributing most to this variability were condiments or discretionary additions, including sauces, dressings, fats, and sugar substitutes (Supplemental Table 3). For total carbohydrates, mean differences were minimal across all LLMs, with the largest difference equal to 1 g (Figure 2). The widest 95% limits of agreement were observed for ChatGPT,

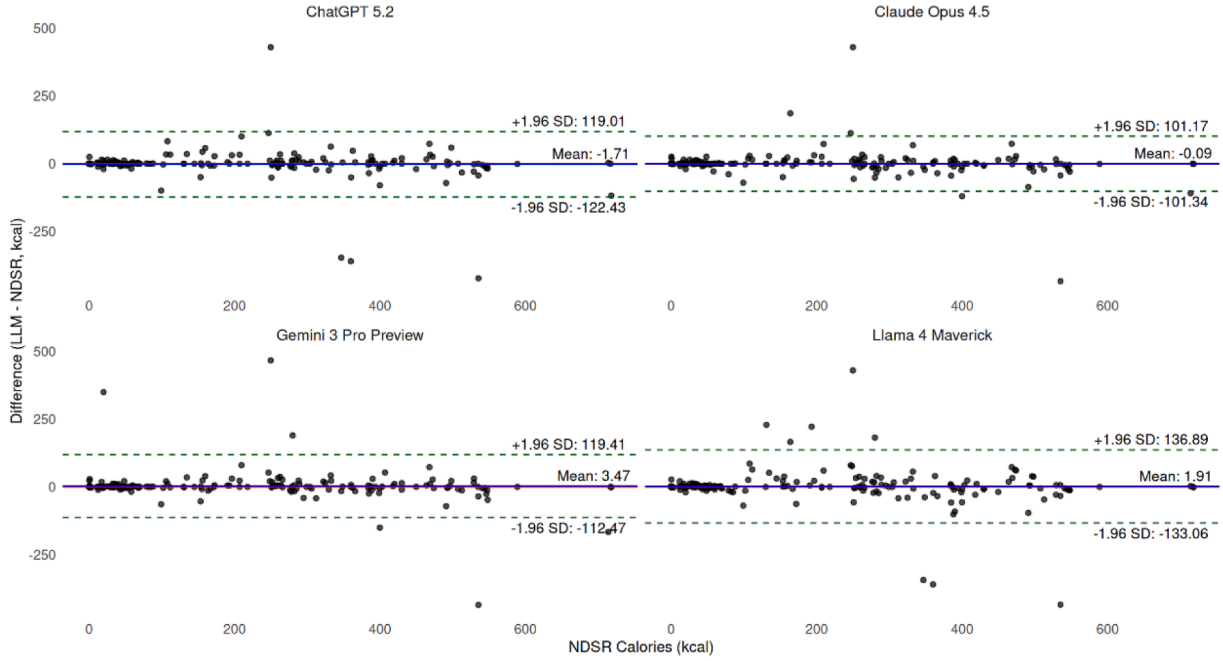

FIGURE 1. Bland–Altman agreement between Large Language Model (LLM) estimated calories and the Nutrition Coordinating Center Food and Nutrient Database ($n = 194$ foods). Each point represents a single food item. Models shown in clockwise order are ChatGPT 5.2, Claude Opus 4.5, Llama 4 Maverick, and Gemini 3 Pro Preview. The blue line indicates the reference line (0 difference), the red line indicates the mean difference, and the dashed green lines indicate the 95% limits of agreement for each model. ChatGPT, Chat Generative Pretrained Transformer; NDSR, Nutrition Data System for Research.

Claude, and Llama, with half-widths ranging from 16 to 26 g. Several foods contributing to variability in total carbohydrate estimates overlapped with those identified in calorie discrepancies, particularly Miscellaneous Foods and Mixed Dishes. Additional variability was driven by foods from the Grains and Sweets groups. For total fat, mean differences between LLMs and the NCC database were small and remained below 1 g (Figure 3). The 95% limits of agreement were ±12 to 14 g. Foods contributing to fat variability mostly overlapped with those contributing to calorie variability, with the exclusion of artificial sweeteners, macaroni and cheese, and chewing gum. For total protein, mean differences were also <1 g (Figure 4). The 95% limits of agreement ranged from ±2 to 3 g. Unlike the calorie outliers, foods contributing to protein variability were primarily items within the Meat, Fish, and Poultry and Mixed Dishes groups.

We also explored agreement by model for energy, carbohydrate, protein, and fat estimations within the 3 most frequently consumed food groups. Fruits and Grains food groups had excellent reliability with the NCC database for energy across all LLMs (ICC range: 0.96–1.00). Unlike the other food groups, Beverages group showed lower ICCs for energy and macronutrients for all LLMs, although the agreement remained good (ICC range: 0.76–0.87). For total carbohydrates, agreement was excellent for all models except for Llama 4 Maverick (ICC = 0.88; 95% CI: 0.79, 0.93). Total protein and fat demonstrated excellent agreement for all models in the Grains group. For this study, protein and fat results were presented for the Grain group only, as these nutrients are present in meaningful amounts in grain-based foods. In Fruits and Beverages group, protein and fat values clustered near zero, which may restrict variability and influence ICC estimates.

Discussion

To our knowledge, this study is among the first to investigate calorie and nutrient agreement between 3 closed-source (ChatGPT 5.2, Claude Opus 4.5, and Gemini 3 Pro Preview) and 1 open-weight LLMs (Llama 4 Maverick) with the reference NCC food and nutrition database using text-based queries. The food group was further examined as a potential source of variability between each LLM and the reference database. Model-specific differences were evident across dietary components. Claude Opus 4.5 showed no poor agreement across any dietary components, whereas the other LLMs had poor agreement for vitamin D, total folate, and/or iron. Bland–Altman analyses demonstrated that although mean differences were generally small, limits of agreement were wide for energy and macronutrients. Variability was driven largely by condiments, other discretionary additions, and mixed dishes. Agreement between each LLM and the NCC database for energy and macronutrients was similar in each of the 3 most commonly consumed food groups: Grains, Beverages, and Fruits.

All LLMs investigated in this study had excellent agreement in energy relative to the NCC database. Although prior studies largely focused on image-based estimation of meals, our results suggest similar performance using text-based food queries [29–33]. The findings for Claude Opus 4.5 agreed with a previous study reporting that Claude 3.5 Sonnet achieved a strong correlation with a reference dataset ($r = 0.78$; $P < 0.05$) [30]. A separate study documented a 17% overestimation in calorie estimation relative to reference values from nutrition labels, but our Bland–Altman analyses showed no uniform overestimation for the Claude model [31]. Our ChatGPT 5.2 findings were consistent with reports of strong associations with reference

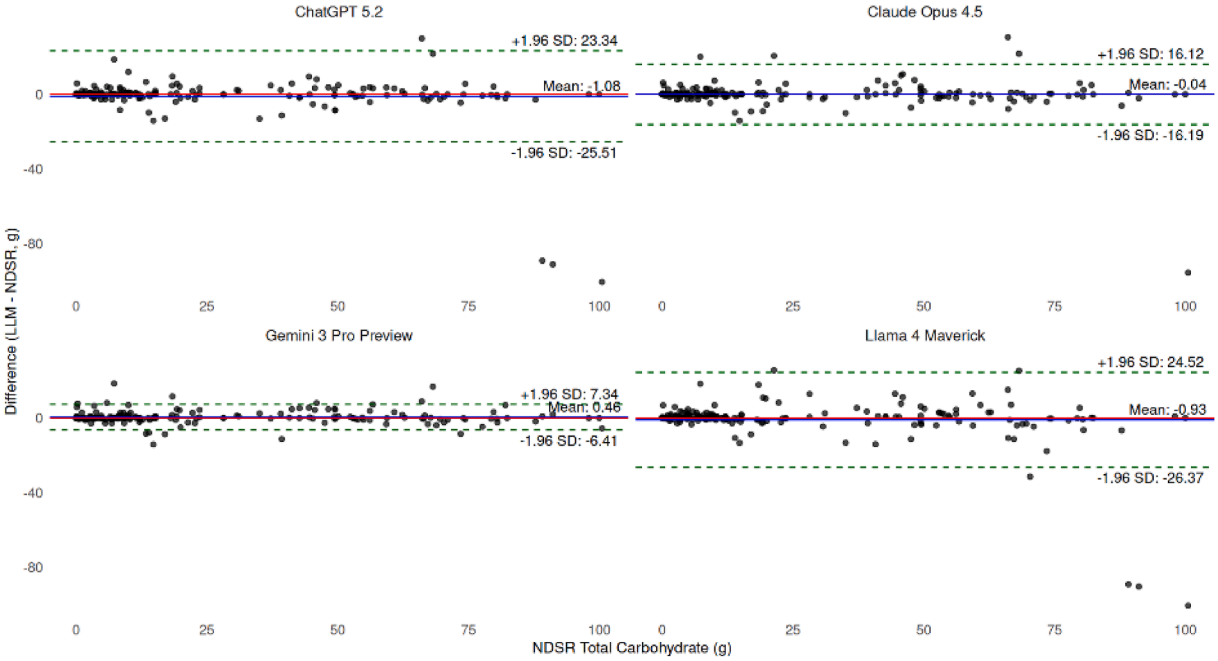

FIGURE 2. Bland–Altman Agreement between Large Language Model (LLM) estimated total carbohydrates (grams) and the Nutrition Coordinating Center Food and Nutrient Database (n = 194 foods). Each point represents a single food item. Models shown in clockwise order are ChatGPT 5.2, Claude Opus 4.5, Llama 4 Maverick, and Gemini 3 Pro Preview. The blue line indicates the reference line (0 difference), the red line indicates the mean difference, and the dashed green lines indicate the 95% limits of agreement for each model. ChatGPT, Chat Generative Pretrained Transformer; NDSR, Nutrition Data System for Research.

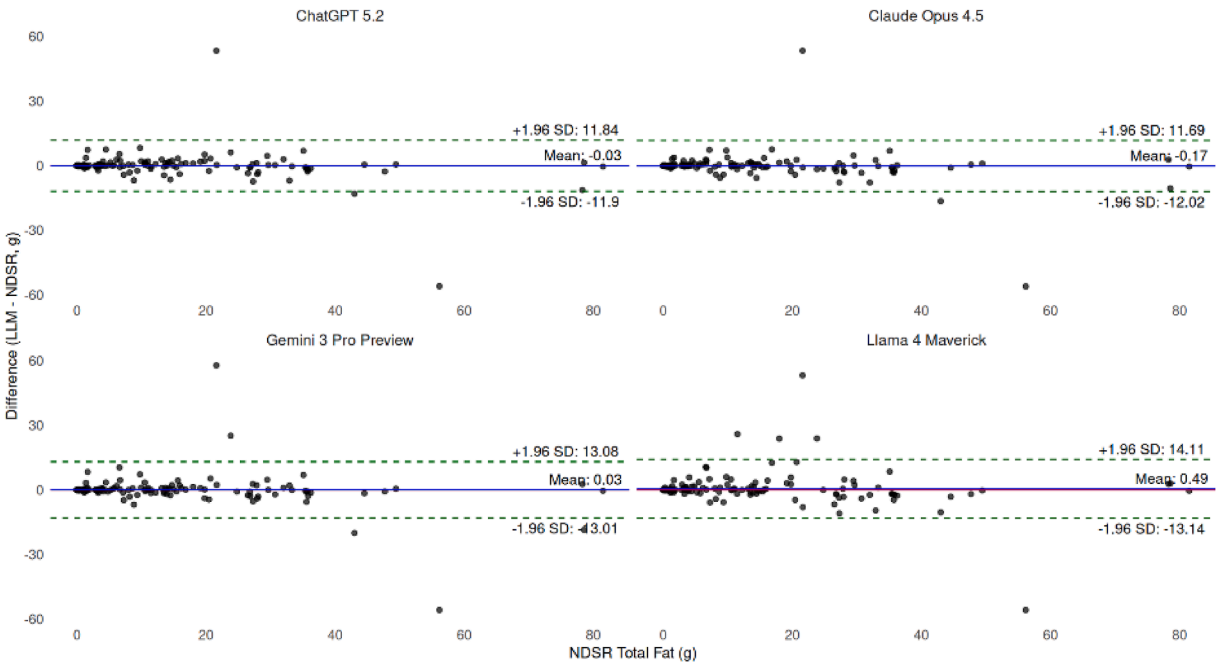

FIGURE 3. Bland–Altman Agreement between Large Language Model (LLM) estimated total fat (grams) and the Nutrition Coordinating Center Food and Nutrient Database ($n = 194$ foods). Each point represents a single food item. Models shown in clockwise order are ChatGPT 5.2, Claude Opus 4.5, Llama 4 Maverick, and Gemini 3 Pro Preview. The blue line indicates the reference line (0 difference), the red line indicates the mean difference, and the dashed green lines indicate the 95% limits of agreement for each model. ChatGPT, Chat Generative Pretrained Transformer; NDSR, Nutrition Data System for Research.

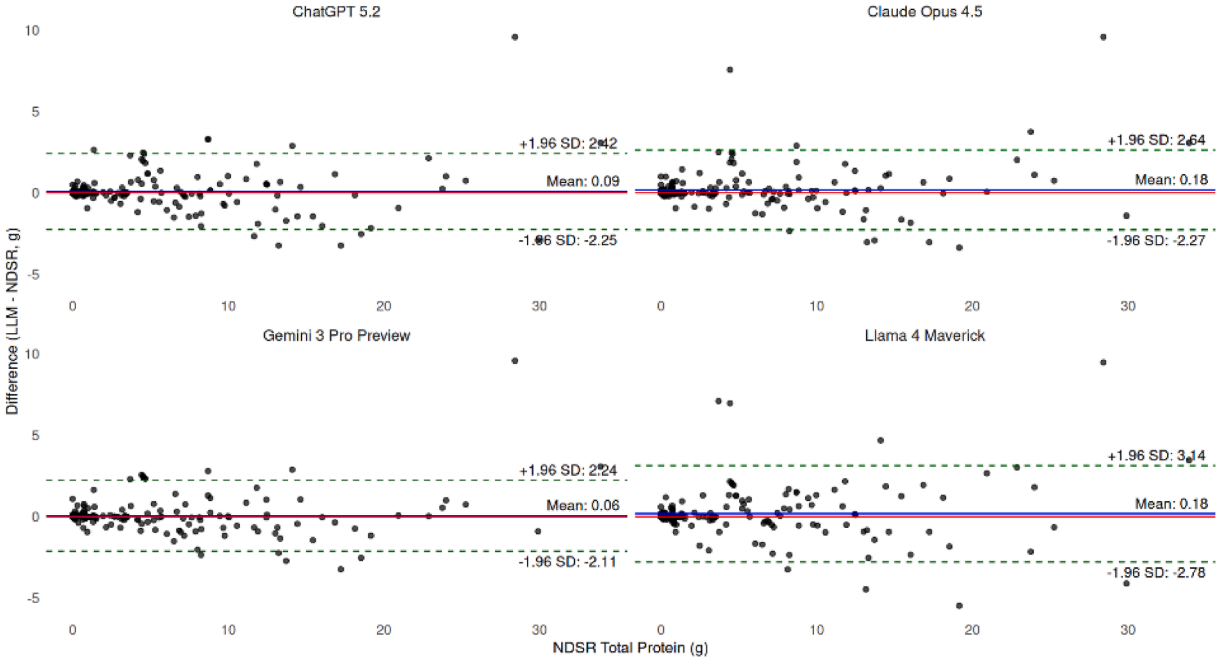

FIGURE 4. Bland–Altman Agreement between Large Language Model (LLM) estimated total protein (grams) and the Nutrition Coordinating Center Food and Nutrient Database ($n = 194$ foods). Each point represents a single food item. Models shown in clockwise order are ChatGPT 5.2, Claude Opus 4.5, Llama 4 Maverick, and Gemini 3 Pro Preview. The blue line indicates the reference line (0 difference), the red line indicates the mean difference, and the dashed green lines indicate the 95% limits of agreement for each model. ChatGPT, Chat Generative Pretrained Transformer; NDSR, Nutrition Data System for Research.

standards but not with reports of nonsignificant associations or caloric underestimation [29–31,33]. Earlier evaluations of Gemini models reported overestimation and ranked them as the weakest model relative to Claude and ChatGPT models [30,31]. Yet, we observed excellent agreement in caloric estimates. To our knowledge, no prior studies have evaluated Llama for this purpose, thus precluding comparisons with our results. Methodological differences may account for these conflicting results,

including our evaluation of more recent model versions, use of text-based queries rather than images, focus on single ingredients and mixed dishes, and comparison with the NCC database.

Agreement between LLMs and the NCC database ranged from good to excellent for macronutrients but had more variable results for micronutrients. Vitamin D, iron, and total folate had poor ICC agreement in ≥1 of the models: ChatGPT for vitamin D and total folate, Gemini for total folate only, and Llama for all 3 nutrients. Similar to a prior image-based evaluation of ChatGPT-4, we found that ChatGPT 5.2 had lower agreement in estimating folate [29]. Same study also reported a high correlation for vitamin D estimation, which contrasts with our finding of poor agreement for ChatGPT. We hypothesize that these discrepancies may partially stem from differences in fortification practices [66–68]. Vitamin D concentrations vary by brand and formulation, whereas iron and folic acid differ according to enrichment status and/or additional fortification. Closed-source LLMs are trained to generate text based on patterns learned from large-scale data; they do not provide transparent evidence of direct lookup from standardized food composition databases. Therefore, nutrient estimates generated through text prompts cannot be assumed to reflect deterministic retrieval of defined database values. In contrast, the NCC database assigns values to defined food items that often reflect generic food profiles [47]. When nutrient concentrations differ widely across products, pattern-based estimates may not match the standardized value assigned to that food. Gemini and Llama cited the use of different nutrient references (e.g., USDA FoodData Central, branded food databases, manufacturer nutrition labels). These references can contain heterogeneous nutrient values for fortified foods, which could influence pattern-based estimates and contribute to discrepancies with standardized NCC values. Claude showed no poor agreement for any nutrient, which may be related to its report that it referenced the NCC database as one of its data sources (Table 1). Reported data sources should be interpreted with caution, as LLMs may generate inaccurate or nonverifiable source attributions [69].

Bland–Altman analyses identified overlapping foods that contributed to discrepancies in calorie and macronutrient estimates between LLMs and the NCC database. These items consisted mainly of condiments and discretionary additions, most often within the Fats and Miscellaneous groups, as well as Mixed Dishes. Higher-fat condiments are energy dense; even small formulation differences per 100 g result in large differences in calorie and macronutrient values. Mixed dishes similarly vary in ingredient composition and relative proportions, further contributing to discrepancies. The mean bias was small in our models, suggesting that current LLMs may be helpful for population-level surveillance of energy and macronutrient intake where individual errors aggregate toward the mean. However, the 95% limits of agreement were wide and should be interpreted with caution for individualized assessment. A single food item consistently consumed by a patient can significantly impact the evaluation of one's diet. This situation is particularly relevant for condiments and discretionary additions as they increase energy and fat intake by adding kilocalories to increase palatability [70]. Another study reported that 85% of participants consumed ≥1 such item per day, contributing a median of 4% of total energy and 5% of fat intake, with notable contributions to vitamin E and sodium [71]. Misestimation of these foods may affect overall dietary assessment and lead to inaccurate nutrition advice when LLMs are used to support food selection [72].

A strength of our study includes the evaluation of text-based nutrient estimation. This study has focused on image-based estimation, yet many conversational AI agents also rely on text input from the user to generate feedback. Evaluating text input tested a separate capability of current closed-source LLMs and provides greater understanding of how these systems may perform in real-world coaching applications. To that end, we selected multiple currently available closed-source LLMs to offer a comparative assessment of available tools that consumers and practitioners may encounter. We also focused on commonly consumed foods in the United States to strengthen the relevance and practical applicability of our findings. Lastly, we examined patterns in specific food types that may explain discrepancies in energy and nutrients between each LLM and the NCC database.

A limitation of this study is the rapid evolution of LLM technology may not reflect future model performance with text-based calorie and nutrient estimation. However, our study emphasizes the importance of understanding how model architecture and underlying nutrition data sources influence estimation accuracy. This understanding requires human expertise and oversight to appropriately interpret and contextualize model-generated estimates. Our evaluation was conducted at the food-item level using standardized food codes rather than complete 24-h dietary recalls or food diaries. This design allowed for controlled comparison between LLM nutrient estimates and a research-based food composition database, but it does not capture additional variability introduced by mixed meals or portion reporting error. Further, we did not conduct repeated testing in this study to evaluate intramodel variability across multiple iterations of the same prompt. Previous work has shown that LLM-generated values can change across iterations, which may limit reproducibility [73]. Our evaluation reflects a single-response scenario per query, consistent with common user interaction. We recommend that further studies account for the intramodel variability and determine if repeated prompting influences LLM agreement with standardized food composition databases. Lastly, this study evaluated several nutrients through multiple models. It is important to acknowledge that the large number of statistical tests performed may increase the potential for type I errors.

As AI becomes integrated into JITAI interventions delivered through mobile health applications, understanding the accuracy of underlying models is critical. This study advances digital nutrition intervention research by evaluating text-based energy and nutrient estimation across multiple closed-source LLMs with a standardized nutrition database. Agreement was strong for energy and macronutrients across models, although variability emerged for certain micronutrients, particularly vitamin D, folate, and iron. These findings suggest that the practical application of LLMs should be tailored to the specific clinical objective. For energy and macronutrient estimation in self-monitoring apps or specific JITAI contexts, LLMs may be acceptable provided that users and practitioners remain aware of the current limitations regarding micronutrients. For applications assessing deficiency risks or evaluating fortification policies, current LLM performance may require manual

verification by practitioners or integration of retrieval-augmented generation with verified food databases.

Claude Opus 4.5 demonstrated consistently high agreement without any nutrients classified as poor, whereas other models showed poor agreement for ≥1 micronutrient. Differences in estimation for all models appeared to be partly driven by specific nutrients and food types rather than systematic discrepancies in energy estimation. We also identified food categories that contributed to variability, although agreement remained strong within the most frequently consumed broader food groups. Our findings demonstrated the importance of understanding underlying data sources and structural assumptions, as these factors can influence estimation accuracy in LLM outputs. Because LLM technology is rapidly advancing, these findings should be viewed as a snapshot of model behavior as of January 2026. Subsequent model iterations may yield different performance characteristics. Yet, a key constant remains: rigorous prompt engineering and query structuring are essential for high-quality outputs when using closed-source systems lacking transparent training data. Future research should examine how prompt engineering and repeated queries influence nutrient estimation accuracy and agreement in real-world contexts.

Acknowledgments

We thank the Community Outreach and Engagement Department at the Hormel Institute for supporting this work and providing this summer research opportunity. We also thank the Nutrition Coordinating Center for permitting the use of the Food and Nutrient Database.

Author contributions

The authors' responsibilities were as follows – RL, ORS, LLK, DF, AWL: conceptualized and designed the study; RL: collected, managed, and analyzed the data, with informational support from RE; RL, ORS, AWL: drafted the initial manuscript; and all authors contributed to data interpretation, critically revised the manuscript, and approved the final version.

Conflict of interest

The authors report no conflicts of interest.

Funding

Funding was provided by the SURE Intern Program at The Hormel Institute. The funding source had no role in the study and imposed no restrictions on publication.

Data availability

The data generated and analyzed in this study are available from the corresponding author on reasonable request. Nutrient values obtained from the Nutrition Coordinating Center Food and Nutrient Database are subject to proprietary licensing restrictions and cannot be publicly shared.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (version 5.2) in order to assist with grammar and language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tjnut.2026.101712.

References

[1] Center for Science in Public Interest. 2025-2030 Uncompromised DGA [Internet]. Available from: https://www.cspi.org/UncompromisedDGA.
[2] I. Nahum-Shani, S.N. Smith, B.J. Spring, L.M. Collins, K. Witkiewitz, A. Tewari, et al., Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support, Ann. Behav. Med. 52 (2018) 446–462.
[3] S.H. Adams, J.C. Anthony, R. Carvajal, L. Chae, C.S.H. Khoo, M. E. Latulippe, et al., Perspective: guiding principles for the implementation of personalized nutrition approaches that benefit health and function, Adv. Nutr. 11 (2020) 25–34.
[4] B.T. Nezami, C.G. Valle, H.M. Wasser, L. Hurley, K.E. Hatley, D.F. Tate, Optimizing a mobile just-in-time adaptive intervention (JITAI) for weight loss in young adults: rationale and design of the AGILE factorial randomized trial, Contemp. Clin. Trials. 150 (2025) 107808.
[5] J.R. Golbus, V.S.E. Jeganathan, R. Stevens, W. Ekechukwu, Z. Farhan, R. Contreras, et al., A physical activity and diet just-in-time adaptive intervention to reduce blood pressure: the myBPmyLife study rationale and design, J. Am. Heart Assoc. 13 (2024) e031234.
[6] S.P. Goldstein, F. Zhang, P. Klasnja, A. Hoover, R.R. Wing, J.G. Thomas, Optimizing a just-in-time adaptive intervention to improve dietary adherence in behavioral obesity treatment: protocol for a microrandomized trial, JMIR Res. Protoc. 10 (2021) e33568.
[7] Y.J. Oh, J. Zhang, M.-L. Fang, Y. Fukuoka, A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss, Int. J. Behav. Nutr. Phys. Act. 18 (2021) 160.
[8] M. Menictas, M. Rabbi, P. Klasnja, S. Murphy, Artificial intelligence decision-making in mobile health, Biochem (Lond). 41 (2019) 20–24.
[9] H.S.J. Chew, The use of artificial intelligence–based conversational agents (chatbots) for weight loss: scoping review and practical recommendations, JMIR Med, Inform 10 (2022) e32578.
[10] R.L. Bailey, Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies, Curr. Opin. Biotechnol. 70 (2021) 91–96.
[11] B. Olendzki, D. Seres, Dietary Assessment in Adults - UpToDate, Wolters Kluwer, Waltham, MA, United States, 2025.
[12] Academy of Nutrition and Dietetics, The Nutrition Care Process (NCP). Academy of Nutrition and Dietetics, 2026. Available from: https://www.ncpro.org/nutrition-care-process.
[13] I. Cavero-Redondo, V. Martinez-Vizcaino, R. Fernandez-Rodriguez, A. Saz-Lara, C. Pascual-Morena, C. Álvarez-Bueno, Effect of behavioral weight management interventions using lifestyle mHealth self-monitoring on weight loss: a systematic review and meta-analysis, Nutrients 12 (2020) 1977.
[14] N. Teasdale, A. Elhussein, F. Butcher, C. Piernas, G. Cowburn, J. Hartmann-Boyce, et al., Systematic review and meta-analysis of remotely delivered interventions using self-monitoring or tailored feedback to change dietary behavior, Am. J. Clin. Nutr. 107 (2018) 247–256.
[15] S.F. Schakel, Y.A. Sievert, I.M. Buzzard, Sources of data for developing and maintaining a nutrient database, J. Am. Diet. Assoc. 88 (1988) 1268–1271.
[16] D.K.N. Ho, W.-C. Chiu, J.-W. Kao, H.-T. Tseng, C.-Y. Lin, P.-H. Huang, et al., Reliability issues of mobile nutrition apps for cardiovascular disease prevention: comparative study, JMIR Mhealth Uhealth 12 (2024) e54509.
[17] A.W. Lin, N. Morgan, D. Ward, C. Tangney, N. Alshurafa, L. Van Horn, et al., Comparative validity of mostly unprocessed and minimally processed food items differs among popular commercial nutrition apps compared with a research food database, J. Acad. Nutr. Diet. 122 (2022) 825–832.e1.
[18] C. Griffiths, L. Harnack, M.A. Pereira, Assessment of the accuracy of nutrient calculations of five popular nutrition tracking applications, Public Health Nutr 21 (2018) 1495–1502.

[19] Z. Su, M.C. Figueiredo, J. Jo, K. Zheng, Y. Chen, Analyzing description, user understanding and expectations of AI in mobile health applications, AMIA Annu. Symp. Proc. 2020 (2021) 1170–1179.
[20] A. Deniz-Garcia, H. Fabelo, A.J. Rodriguez-Almeida, G. Zamora-Zamorano, M. Castro-Fernandez, M.D.P. Alberiche Ruano, et al., Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges, J. Med. Internet Res. 25 (2023) e44030.
[21] G.G. Panayotova, Artificial intelligence in nutrition and dietetics: a comprehensive review of current research, Healthcare (Basel) 13 (2025) 2579.
[22] C. O'Hara, G. Kent, C.L. Leydon, N.M. Walsh, E.R. Gibney, D. Skoutas, et al., Language models in nutrition and dietetics: a scoping review, Am. J. Clin. Nutr. 123 (2025) 101127.
[23] G.A. Tahir, C.K. Loo, A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment, Healthcare (Basel) 9 (2021) 1676.
[24] A. Jankelow, A. Godneva, M. Rein, D. Samocha-Bonet, D. Weissglas-Volkov, S. Zohar, et al., NutriMatch: harmonizing food composition databases with large language models for enhanced nutritional prediction, NPJ Digit, Public Health 1 (2026) 1.
[25] N. Lambert, Reinforcement learning from human feedback [Internet]. arXiv, Available from: http://arxiv.org/abs/2504.12501.
[26] R. Yan, H. Luo, J. Lu, D. Liu, H. Posluszny, M.P. Dhaliwal, et al., DietAI24 as a framework for comprehensive nutrition estimation using multimodal large language models, Commun. Med. (Lond.) 5 (2025) 458.
[27] Y. Lu, T. Stathopoulou, M.F. Vasiloglou, L.F. Pinault, C. Kiley, E. K. Spanakis, et al., goFOODTM: An artificial intelligence system for dietary assessment, Sensors (Basel) 20 (2020) 4283.
[28] H. Zhou, L. Chow, L. Harnack, S. Panda, E.N.C. Manoogian, M. Li, et al., NutriRAG: unleashing the power of large language models for food identification and classification through retrieval methods, J. Am. Med. Inform. Assoc. 33 (2026) 802–811.
[29] C. O'Hara, G. Kent, A.C. Flynn, E.R. Gibney, C.M. Timon, An evaluation of ChatGPT for nutrient content estimation from meal photographs, Nutrients 17 (2025) 607.
[30] J. Fridolfsson, E. Sjöberg, M. Thiwång, S. Pettersson, Performance evaluation of 3 large language models for nutritional content estimation from food images, Curr. Dev. Nutr. 9 (2025) 107556.
[31] C.-F. Hsuan, Y.-J. Lee, H.-C. Hsu, C.-M. Ouyang, W.-C. Yeh, W.-H. Tang, Comparison of accuracy in the evaluation of nutritional labels on commercial ready-to-eat meal boxes between professional nutritionists and chatbots, Nutrients 17 (2025) 3044.
[32] M. Haman, M. Školník, M. Lošťák, AI dietician: unveiling the accuracy of ChatGPT's nutritional estimations, Nutrition 119 (2024) 112325.
[33] Y.N. Hoang, Y.-L. Chen, D.K.N. Ho, W.-C. Chiu, K.-J. Cheah, N. R. Mayasari, et al., Consistency and accuracy of artificial intelligence for providing nutritional information, JAMA Netw. Open 6 (2023) e2350367.
[34] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nat. Med. 29 (2023) 1930–1940.
[35] OpenAI, Introducing GPT-5.2, 2025. Available from: https://openai.com/index/introducing-gpt-5-2/.
[36] DeepMind. Gemini 3 Pro [Internet], Google DeepMind, Alphabet/DeepMind, 2026. Available from: https://deepmind.google/models/gemini/pro/.
[37] Anthropic, Introducing Claude Opus 4.5, Anthropic PBC, 2025. Available from: https://www.anthropic.com/news/claude-opus-4-5.
[38] A. Chatterji, T. Cunningham, D.J. Deming, Z. Hitzig, C. Ong, C.Y. Shan, et al., How people use ChatGPT, National Bureau of Economic Research, 2025. Available from: https://www.nber.org/papers/w34255.
[39] S. Pichai, Q4 earnings call: remarks from our CEO, Google, 2026. Available from: https://blog.google/company-news/inside-google/message-ceo/alphabet-earnings-q4-2025/.
[40] R. Appel, M. Massenkoff, P. McCrory, M. McCain, R. Heller, T. Neylon, et al., Anthropic economic index report: economic primitives, 2026. Available from: https://www.anthropic.com/research/anthropic-economic-index-january-2026-report.
[41] A.K. Kant, A. Schatzkin, G. Block, R.G. Ziegler, M. Nestle, Food group intake patterns and associated nutrient profiles of the US population, J. Am. Diet. Assoc. 91 (1991) 1532–1537.
[42] U.S. Department of Agriculture, Agricultural Research Service. WWEIA/NHANES Overview : USDA ARS, 2025. Available from: https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/wweianhanes-overview/.
[43] Centers for Disease Control and Prevention, 2017-March 2020 Pre-Pandemic Dietary Data - Continuous NHANES, 2026. Available from: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary&Cycle=2017-2020.
[44] U.S. Department of Agriculture, Agricultural Research Service, FPED databases: USDA ARS, Food Surveys Research Group, 2026. Available from: https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fped-databases/.
[45] Nutrition Coordinating Center. Foods, nutrients and food groups, 2025. Available from: https://www.ncc.umn.edu/about-ncc/foods-nutrients-and-food-groups/.
[46] Cronometer, Cronometer Software Inc [Internet]. 2011. Available from: https://apps.apple.com/us/app/cronometer-calorie-counter/id1145935738.
[47] Nutrition Coordinating Center, NCC nutrient information sources, 2026. Available from: https://www.ncc.umn.edu/products/nutrient-information-sources/.
[48] U.S. Food and Drug Administration, Trans fat, FDA, 2024. Available from: https://www.fda.gov/food/food-additives-petitions/trans-fat.
[49] OpenRouter, OpenRouter [Internet]. 2023. Available from: https://openrouter.ai.
[50] AI Meta, The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Meta AI, Meta Platforms, Inc., 2025. Available from: https://ai.meta.com/blog/llama-4-multimodal-intelligence/.
[51] S.K.K. Santu, D. Feng, TELeR: a general taxonomy of LLM prompts for benchmarking complex tasks, arXiv (2023). Available from: http://arxiv.org/abs/2305.11430.
[52] R Project Developers. R: the R project for statistical computing [Internet]. R Foundation for Statistical Computing; 2025. Available from: https://www.r-project.org/.
[53] H. Wickham, J. Bryan, Read excel files, Posit (2026). Available from: https://readxl.tidyverse.org/.
[54] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, dplyr: a grammar of data manipulation, Posit, 2026. Available from: https://dplyr.tidyverse.org/.
[55] H. Wickham, D. Vaughan, M. Girlich, Tidy messy data, Posit (2026). Available from: https://tidyr.tidyverse.org/.
[56] H. Wickham, Simple, consistent wrappers for common string operations [Internet]. Available from, https://stringr.tidyverse.org/.
[57] M. Gamer, J. Lemon, I. Fellows, P. Singh, irr: Various coefficients of interrater reliability and agreement, 2005, 0.84.1. Available from: https://CRAN.R-project.org/package=irr.
[58] H. Wickham, W. Chang, L. Henry, T.L. Pedersen, K. Takahashi, C. Wilkie, et al., Create elegant data visualisations using the grammar of graphics, Posit (2025). Available from: https://ggplot2.tidyverse.org/.
[59] C. Sievert, Interactive web-based data visualization with R, plotly, and shiny [Internet]. Available from, https://plotly-r.com/.
[60] R. Vaidyanathan, ramnathv/htmlwidgets, 2026. Available from: https://github.com/ramnathv/htmlwidgets.
[61] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2016) 155–163.
[62] P.E. Shrout, J.L. Fleiss, Intraclass correlations: Uses in assessing rater reliability, Psychol Bull 86 (1979) 420–428.
[63] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients, Psychological Methods US, 1, American Psychological Association, 1996, pp. 30–46.
[64] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310.
[65] J.M. Bland, D.G. Altman, Measuring agreement in method comparison studies, Stat. Methods Med. Res. 8 (1999) 135–160.
[66] R.F. Hurrell, Iron fortification practices and implications for iron addition to salt, J. Nutr. 151 (Supplement 1) (2021) 3S–14S.
[67] M.S. Calvo, S.J. Whiting, C.N. Barton, Vitamin D fortification in the United States and Canada: current status and data needs, Am. J. Clin. Nutr. 80 (Supplement) (2004) 1710S–1716S.
[68] S.F. Choumenkovitch, J. Selhub, P.W.F. Wilson, J.I. Rader, I. H. Rosenberg, P.F. Jacques, Folic acid intake from fortification in United States exceeds predictions, J. Nutr. 132 (2002) 2792–2798.

[69] A. Chatelan, A. Clerc, P.-A. Fonta, ChatGPT and future artificial intelligence chatbots: what may be the influence on credentialed nutrition and dietetics practitioners? J. Acad. Nutr. Diet. 123 (2023) 1525–1531.
[70] R.L. Best, K.M. Appleton, Comparable increases in energy, protein and fat intakes following the addition of seasonings and sauces to an older person's meal, Appetite 56 (2011) 179–182.
[71] M. Whatnall, E.D. Clarke, T. Schumacher, M.E. Rollo, T. Bucher, L. M. Ashton, et al., Do sauces, condiments and seasonings contribute important amounts of nutrients to Australian dietary intakes? J. Hum. Nutr. Diet. 36 (2023) 1101–1110.
[72] Y.J. Chen, C.-C. Chang, Y.N. Hoang, A.W. Lin, W.-L. Lin, C.-Y. Lin, et al., Customized multimodal Diabot-GPT-4o enhances accuracy of image-based dietary assessments in dietetic trainees in Taiwan: validation against weighed food records, Am. J. Clin. Nutr. 122 (2025) 1836–1849.
[73] V. Ponzo, R. Rosato, M.C. Scigliano, M. Onida, S. Cossai, M. De Vecchi, et al., Comparison of the accuracy, completeness, reproducibility, and consistency of different AI chatbots in providing nutritional advice: an exploratory study, J. Clin. Med. 13 (2024) 7810.